New Statistical Methods for Complex Survival Data with High-dimensional Covariates

by

Tao Sun

BS, Cornell University, 2009

DVM, China Agricultural University, China, 2010

MS, University of Pittsburgh, 2013

MS, Columbia University, 2015

Submitted to the Graduate Faculty of
the Graduate School of Public Health in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2020

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Tao Sun

It was defended on

April 7th 2020

and approved by

Ying Ding, PhD, Associate Professor of Biostatistics,

Graduate School of Public Health, University of Pittsburgh

Wei Chen, PhD, Associate Professor of Pediatrics,

School of Medicine, University of Pittsburgh

Jong H. Jeong, PhD, Professor and Vice Chair of Biostatistics,

Graduate School of Public Health, University of Pittsburgh

Yu Cheng, PhD, Associate Professor of Statistics,

School of Arts and Sciences, University of Pittsburgh

Daniel E. Weeks, PhD, Professor of Human Genetics,

Graduate School of Public Health, University of Pittsburgh

Dissertation Director: Ying Ding, PhD, Associate Professor of Biostatistics,

Graduate School of Public Health, University of Pittsburgh

New Statistical Methods for Complex Survival Data with High-dimensional Covariates

Tao Sun, PhD

University of Pittsburgh, 2020

Abstract

Complex survival outcomes, such as multivariate and interval-censored endpoints, are becoming more commonly used in clinical trials. The revolutionary development of genetics technologies allows the generation of large-scale genetic data. This dissertation proposes new statistical methods for complex survival outcomes with high-dimensional covariates.

In the first part, to deal with bivariate interval-censored data, we propose a flexible two-parameter copula-based model with semiparametric transformation margins. We estimate the model parameters by the sieve likelihood approach and establish the asymptotic properties of the sieve estimators. We demonstrate satisfactory estimation and inference performance in simulation studies. Lastly, we apply our method to the Age-Related Macular Degeneration Study (AREDS) data and successfully identify novel genetic variants associated with the progression of Age-related Macular Degeneration (AMD). An R package *CopulaCenR* is published for analyzing bivariate censored data in a regression setting.

In the second part, we develop a novel information-ratio-based test statistic to evaluate the goodness-of-fit of copula survival models. We establish the asymptotic properties of our test statistic. The simulation studies demonstrate that our method performs well under interval and right censoring. Lastly, we evaluate our results in multiple real data sets. To the best of our knowledge, our method is the first approach that can test any parametric copula model under both interval and right censoring.

In the third part, motivated by recent demanding needs for developing accurate survival prediction models utilizing rich genetic data, we develop a novel framework for constructing and evaluating a deep neural network (DNN) based survival model. Our simulation results clearly demonstrate the high predictive power of the DNN survival model, especially in the presence of complex data structures. We also build an accurate and interpretable DNN survival prediction model for AMD progression using AREDS data.

Public health significance: This dissertation provides a comprehensive set of novel statistical and computational tools for analyzing bivariate survival outcomes with large-scale genetic data, which have the potential to fundamentally improve the current practice in analyzing such clinical studies, and thus to enhance the understanding of disease progression and to increase the success of individualized risk management and precision medicine.

Table of Contents

List of Tables

List of Figures

Preface

First, I would like to thank my dissertation advisor, Dr. Ying Ding. It is my great honor and pleasure to be her student for the past five years. She always takes every effort to make students succeed. She patiently guided me through all these wonderful dissertation projects and opened the door that leads me to an exciting academic career. She is a fantastic advisor. I would also like to thank Dr. Wei Chen, who is my GSR advisor. He taught me a lot of brilliant ideas and mentored me on many research projects in areas such as bioinformatics, single-cell RNA-seq, immunology, pulmonary medicine, and cancer biology. He opened up my eyes to numerous novel applications of statistics and motivated me to develop my future research directions. I would also like to thank Dr. Jong Jeong, who was my GSR advisor in my second year. He mentored me on many statistical consulting projects. I also took his advanced-level survival analysis course, which laid a solid foundation for me in survival theory and methodology. I would also like to thank Dr. Yu Cheng. She provided valuable and insightful suggestions to my dissertation, especially the project on the goodness-of-fit test for copula specification. I would also like to thank Dr. Daniel Weeks and Dr. Qi Yan because my dissertation is partly motivated by their earlier works. Dr. Weeks also provided thoughtful and thorough review comments on my dissertation draft, which significantly enhances the quality of my dissertation. I would also like to thank Dr. Robert Krafty for being my GSR advisor in my second year, as well as his continuous support.

I am also indebted to all who have supported and encouraged me. Thanks to the former and current members of Dr. Ying Ding's and Dr. Wei Chen's groups, who provided a supportive and intellectually challenging research environment. Thanks to my research collaborators Dr. Jay Kolls, Dr. Kong Chen, Dr. Fadi Lakkis, Dr. Dario Vignali, and Dr. Juan Celedon, for sharing with me their expertise in various clinical and biomedical areas. Thanks to Dr. Ada Youk for providing me with the opportunity to teach a core course as the primary instructor. Thanks to Dr. Chung-Chou Chang and the American Statistical Association (ASA) Pittsburgh Chapter for allowing me to practice my leadership skills. Thanks to Dr. Abdus Wahed, Dr. George Tseng, Dr. Gong Tang, Dr. Yongseok Park, and Dr.

Chaeryon Kang for delivering excellent and instructive lectures. Thanks to the Department of Biostatistics for hosting research symposiums that honed my presentation skills. Thanks to all who guided, helped, and encouraged me in my job hunting process. Last, I would like to thank Dr. Joshua Mattila for his timely help and continuous support.

Finally, I reserve my deepest gratitude for the endless patience, support, and encouragement of my family and friends throughout my journey in pursuing my Ph.D. degree. I owe a special thanks to my wife, Jun Zhang, for her deepest love, encouragement, and belief in me at all times. This would have never been accomplished without you.

1.0 Introduction

1.1 Overview

Complex survival outcomes such as multivariate and/or interval-censored endpoints are becoming more commonly used in clinical trials, for example, to study bilateral diseases or diseases with multiple comorbidities. The revolutionary development of genetics technologies allows the generation of large-scale genetic data in modern clinical trials. Motivated by two large clinical trials for studying a bilateral eye disease, Age-related Macular Degeneration (AMD), this dissertation proposes new statistical methods for analyzing complex survival outcomes with high-dimensional covariates: (1) to efficiently test and identify risk factors associated with disease progression in a regression setting, (2) to perform rigorous model diagnosis through a novel goodness-of-fit test and (3) to accurately predict disease progression profiles using a deep learning survival prediction model.

In the rest of this Chapter, I will start by introducing the basic concepts, such as survival data and interval-censored data, in Sections 1.2 and 1.5. In the following, I will talk about some popular regression models for survival data, particularly multivariate and/or interval-censored data, in Section 1.3, 1.4, and 1.6. Then, I will discuss about the existing goodness-of-fit tests for copula specification in Section 1.7. Lastly, I am going to introduce several popular survival prediction models in Section 1.8.

1.2 Failure time data

Failure time data usually represent times to specific event of interest, including death, the onset of a disease, outbreak of an epidemic, and malfunction of a machine. The situation is called “failure” when such an event occurs. The term “failure time data” is equivalent to “survival data”, and the random variable of time to failure is also denoted by “survival time”. Such data primarily arise from medical and biological studies and also widely exist in epidemiological, sociological, economic, and financial studies. More concepts and examples can be found in Kalbfleisch and Prentice [2011].

In the analysis of survival data, also referred to as survival analysis, the common interest is to study the survival function, which is the probability that failure time is greater than a certain time point. There are three associated problems: estimation of survival functions, comparison of survival functions, and regression between survival function and covariates.

The survival analysis distinguishes itself from other statistical fields by the existence of censoring in the data. In practice, the exact failure time may not always be observed. Sometimes, we could only know the failure time is greater or smaller than the observation time, corresponding to right- or left-censored survival data, respectively. The censoring mechanism may depend on failure times so that it further complicates the analysis. Truncation is another source of complexity, where subjects are recruited to a study only if they satisfy certain pre-specified conditions.

1.3 Models for failure time data

We define the survival function of failure time, represented by a non-negative random variable T, as the probability that T is greater than a value t, expressed as:

$$S(t) = P(T > t),\ 0 < t < \infty.$$

When T is absolutely continuous, we have a one-to-one relationship between the survival function $S(t)$, density function $f(t)$ and the hazard function $\lambda(t)$ as defined below:

$$S(t) = e^{-\Lambda(t)},$$

where

$$\Lambda(t) = \int_0^t \lambda(s)ds,\ \lambda(t) = f(t)/S(t).$$

In survival analysis, the survival and hazard functions are usually easier to use for modeling the failure time T than the density function.

As mentioned before, regression between the survival function and covariate effect is an important aspect of survival analysis. Here we introduce some common regression models for survival time.

1.3.1 Proportional hazards model

The proportional hazards (PH) or Cox model [Cox, 1972] is built on the hazard function:

$$\lambda(t; Z) = \lambda_0(t)\exp(Z^T\beta),$$

where $\lambda_0(t)$ is an unspecified baseline hazard function, Z is a vector of covariates, and β is a vector of covariate coefficient parameters. The model is interpreted as the ratio of hazard functions is a constant (independent of time t) between two subjects with different Z. In a special case where $Z = 0$ or 1, we have

$$\frac{\lambda(t; Z = 1)}{\lambda(t; Z = 0)} = \exp(\beta).$$

Define $h(t) = \log[\Lambda_0(t)]$, which is a strictly non-decreasing function of t. Then we can re-write the PH model as

$$h(T) = -Z^T\beta + \epsilon,$$

where the random variable ϵ follows a extreme value distribution with distribution function $F(s) = 1 - \exp[-e^s]$.

Under the PH model, the survival and cumulative hazard functions of T are given by:

$$S(t; Z) = e^{-\Lambda_0(t)\exp(Z^T\beta)} = \{S_0(t)\}^{\exp(Z^T\beta)},\ \Lambda(t; Z) = \Lambda_0(t)\exp(Z^T\beta),$$

where

$$\Lambda_0(t) = \int_0^t \lambda_0(s)ds \quad \text{and} \quad S_0(t) = e^{-\Lambda_0(t)}$$

are the baseline cumulative hazard and baseline survival functions.

The PH model is the most popular regression model in survival analysis due to the partial likelihood approach for right-censored failure time data [Cox, 1972]. The approach is simple and efficient because the partial likelihood only involves the finite-dimensional β parameter without the nuisance infinite-dimensional $\lambda_0(t)$. The resulting β estimate is asymptotically equivalent to that obtained from the full likelihood.

1.3.2 Proportional odds model

The proportional odds (PO) model is also commonly used in survival analysis [Bennett, 1983]. It assumes a constant odds ratio in survival functions between two subjects with different covariate effects. The model is given by:

$$\frac{1 - S(t;Z)}{S(t;Z)} = \frac{1 - S_0(t;Z)}{S_0(t;Z)} e^{Z^T\beta},$$

or

$$logit\{S(t;Z)\} = logit\{S_0(t;Z)\} - Z^T\beta,$$

where $S_0(t)$ denotes baseline survival function.

In the special case that $Z = 0$ or 1, we have the ratio of hazard functions as

$$\frac{\lambda(t;Z=1)}{\lambda(t;Z=0)} = \frac{1}{1 + (e^{-\beta} - 1)S_0(t)}.$$

Unlike in the PH model, the ratio of hazard functions under the PO model is a monotonically increasing function of t and reaches a maximum of 1 when $t \to \infty$.

Let $h(t) = -logit\{S_0(t)\}$, which is a strictly non-decreasing function of t. Then the PO model can also be written as

$$h(T) = -Z^T\beta + \epsilon,$$

where the random variable ϵ follows a standard logistic distribution.

1.3.3 Additive hazards model

The additive hazard model [Holford, 1976] is also built upon the hazard function. However, its formula has an additive form:

$$\lambda(t;Z) = \lambda_0(t) + Z^T\beta,$$

where $\lambda_0(t)$ is an unknown baseline hazard function, and the coefficient β is interpreted as the hazard difference. Again, when $Z = 0$ or 1, we get

$$\lambda(t;Z=1) = \lambda(t;Z=0) + \beta.$$

One nice feature of the additive hazards model is its model simplicity and ease of interpretation. Especially, it is the case for the additive frailty model under which the marginal model still follows the additive hazards assumption, and the coefficient β has the same interpretations under both conditional and marginal settings.

1.3.4 Accelerated failure time model

The accelerated failure time (AFT) model [Wei, 1992] directly defines the relationship between survival time T and covariates Z as follows

$$\log T = Z^T\beta + \epsilon,$$

where β is a vector of coefficients, and ϵ follows an unspecific distribution.

It is interesting to notice that the covariate effect is multiplicative in both PH and AFT models, but on hazard function and (log) survival time, respectively. Next, we will show that the covariate effect on hazard function is different between AFT and PH models.

Let $\lambda_{\epsilon^\star}(t)$ be the hazard function of random variable $\epsilon^\star = \exp(\epsilon)$. Then the AFT model can be re-written as $T = \exp(Z^T\beta)\epsilon^\star$, which has the following hazard and survival functions

$$\lambda(t; Z) = \lambda_{\epsilon^\star}(te^{-Z^T\beta})e^{-Z^T\beta}$$

and

$$S(t; Z) = \exp\{-\Lambda_{\epsilon^\star}(te^{-Z^T\beta})\},$$

where $\Lambda_{\epsilon^\star}(t) = \int_0^t \lambda_{\epsilon^\star}(s)ds$.

If $Z \in \{0, 1\}$, we have

$$S(t; Z = 1) = S(\gamma t; Z = 0)$$

and

$$\lambda(t; Z = 1) = \gamma\lambda(\gamma t; Z = 0),$$

where $\gamma = e^{-Z^T\beta}$.

1.3.5 Linear transformation model

Here we introduce a collection of regression models, known as the linear transformation model [Chen et al., 2002, Fine et al., 1998] defined as

$$h(T) = Z^T\beta + \epsilon,$$

where $h(t)$ is an unknown strictly increasing function of t, Z is the covariate vector, and ϵ is a random variable with a known distribution function F. The linear transformation model includes PH and PO models as special cases, where the random variable ϵ follows the extreme value distribution or the standard logistic distribution, respectively.

An equivalent format of linear transformation model is defined as

$$g\{S(t;Z)\} = h(t) - Z^T\beta,$$

where $g^{-1}(s) = 1 - F(s)$. We can see that it is a semiparametric model for $S(t;Z)$, as the nuisance part $h(t)$ is unknown with infinite dimension, and the finite-dimensional β is the primary goal for estimation and inference. A major advantage of the model is its generality and flexibility, as F can be any specific distribution function. In Chapter 2, we will introduce a copula-based semiparametric transformation model for bivariate interval-censored data.

1.4 Models for multivariate failure time data

The previous section introduces some general regression models for univariate survival data, as only one survival time T is defined in each model. When several correlated survival times are modeled together, their dependence structure needs to be properly modeled. Many papers have addressed the analysis of multivariate failure time data, mostly in the presence of right-censored data. Klein and Moeschberger [2006] and Hougaard [2012] are excellent reference books on this topic. The emphasis of this dissertation is bivariate survival data that includes two survival times T_1 and T_2. There are three general categories of methods for bivariate failure time data: marginal method, frailty models, and copula-based approaches.

1.4.1 Marginal models

The marginal models assume each of the correlated survival times follows a marginal distribution and build the likelihood function without considering the correlation between the margins. As a result, the naive variance estimator of regression coefficients β is underestimated. Instead, one needs to obtain a robust variance estimator to account for the correlation. Among the extensive literature, Wei et al. [1989] and Guo and Lin [1994] developed the marginal proportional hazards models for continuous and discrete right-censored data, respectively. Goggins and Finkelstein [2000] and Kim and Xue [2002] applied the same marginal proportional hazards models to continuous and discrete interval-censored data.

1.4.2 Frailty models

The frailty models assume there exists an unobserved frailty random variable that accounts for the dependence structure of correlated survival times [Clayton and Cuzick, 1985]. Given the frailty term, the survival times are conditionally independent. The frailty model has become a popular approach for bivariate survival data [Oakes, 1989]. It is also used to model survival time and informative censoring time together.

Under the frailty model with a proportional hazard assumption, we first define the conditional cumulative hazard function

$$\Lambda_j(t|u) = u\Lambda_{j,m}(t),\ j = 1, 2,$$

where j denotes the jth margin, u is the frailty random variable with the density function $f_\eta(u)$, $\Lambda_{j,m}(t)$ is the cumulative hazard function at time t when $u = 1$. The corresponding marginal survival function can be derived through the Laplace transformation:

$$S_{j,m}(t) = \int S_j(t|u)f_\eta(u)du = \int e^{-u\Lambda_{j,m}(t)}f_\eta(u)du = \mathscr{L}_\eta(\Lambda_{j,m}(t)),$$

where $\mathscr{L}_\eta(.)$ is the Laplace transformation function with respect to the frailty density function. Thus, we can obtain the following formula by the inverse Laplace transformation:

$$\Lambda_{j,m}(t) = \mathscr{L}_\eta^{-1}(S_{j,m}(t)).$$

By assuming conditional independence of two margins given u, we can write the conditional joint survival function as

$$S(t_1, t_2|u) = S_1(t_1|u)S_2(t_2|u).$$

Finally, the marginal joint survival function [Oakes, 1989] is expressed by:

$$\begin{aligned} S_m(t_1, t_2) &= \int S_1(t_1|u)S_2(t_2|u)f_\eta(u)du \\ &= \int e^{-u\{\Lambda_{1,m}(t)+\Lambda_{2,m}(t)\}} f_\eta(u)du \\ &= \mathscr{L}_\eta[\mathscr{L}_\eta^{-1}\{S_{1,m}(t)\} + \mathscr{L}_\eta^{-1}\{S_{2,m}(t)\}]. \end{aligned}$$

In spite of the frailty model's advantage in modeling the correlation between survival times, it has several limitations: (1) the frailty term accounts for correlation, but its interpretation is not straightforward; (2) the coefficient β typically can only be interpreted upon conditioning on the frailty term.

1.4.3 Copula models

In this dissertation, we will use the copula model for multivariate survival data. The copula model is another commonly used method for bivariate survival data. There are several advantages to applying this model. First, the copula has a property of "scale-invariance" nature of the dependence between two survival times. That is, if α and β are almost surely increasing functions of T_1 and T_2 respectively, then the copula of $\alpha(T_1)$ and $\beta(T_2)$ is the same as the copula of T_1 and T_2. Hence it is the copula that captures the "distribution-free" nature of the dependence between T_1 and T_2 [Nelsen, 2006]. Second, the copula's dependence parameter can be expressed by nonparametric dependence measures such as Kendall's τ. This connection is particularly useful in survival models as multivariate survival times do not always follow a normal distribution. Third, there are many types of copula models, and each could account for a distinct tail dependence. This property renders great flexibility for modeling various correlation patterns. Lastly, marginal survival distributions are independent of the choice of the dependence parameter in the copula model. Thus, one can model the margins and dependence parameter separately. This property stems from the Sklar's theorem [Sklar, 1959], which states that any multi-dimensional joint distribution function may be decomposed into marginal distributions and a copula function that completely describes the dependence structure.

Clayton [1978] first applied the copula model to bivariate survival data. Shih and Louis [1995] proposed a two-stage estimation procedure for the dependence parameter in copula models for bivariate right-censored data. Wang and Ding [2000] and Sun et al. [2006] extended the two-stage estimation of dependence parameter to the bivariate case I and II interval-censored data, respectively.

Concept of copula

Let T_1 and T_2 be two continuous random variables with marginal cumulative distribution functions $F_1(t_1) = P(T_1 \leq t_1)$ and $F_2(t_2) = P(T_2 \leq t_2)$. Define random variables $U_1 = F_1(T_1)$ and $U_2 = F_2(T_2)$ and both follow the uniform distribution on $\mathbf{I}$ ($\mathbf{I} = [0, 1]$). Thus the joint cumulative distribution of (U_1, U_2) is:

$$C(u_1, u_2) = P(U_1 \leq u_1, U_2 \leq u_2).$$

Then, the mapping from $(u_1, u_2) \in \mathbf{I^2}$ to $C(u_1, u_2) \in \mathbf{I}$ is a *copula*. Since $F(t)$ is a monotonically increasing function, the above expression could be rewritten to the joint distribution of (T_1, T_2) :

$$C(u_1, u_2) = P(T_1 \leq F_1^{-1}(u_1), T_2 \leq F_2^{-1}(u_2)).$$

An informal definition of a two-dimensional copula

Suppose the previously defined (T_1, T_2) has the joint distribution function:

$$F(t_1, t_2) = P(T_1 \leq t_1, T_2 \leq t_2).$$

Then for every (t_1, t_2) in $[-\infty, +\infty]^2$, we consider the points in $\mathbf{I^3}$ with coordinates $(F_1(t_1), F_2(t_2), C(t_1, t_2))$. Then, the mapping from $\mathbf{I^2}$ to $\mathbf{I}$ is a *copula.*

A formal definition of a two-dimensional copula

A two dimensional copula is a function C: $\mathbf{I^2} \rightarrow \mathbf{I}$ such that

(1) $C(u_1, u_2)$ is grounded, i.e., $C(0, u_2) = C(u_1, 0) = 0$ for every $(u_1, u_2) \in \mathbf{I^2}$.

(2) $C(1, u_2) = u_2, C(u_1, 1) = u_1$ for every $(u_1, u_2) \in \mathbf{I^2}$.

(3) $C(u_1, u_2)$ is two-increasing, i.e., for $u_1, u_1', u_2, u_2' \in \mathbf{I}$ with $u_1 \leq u_1'$ and $u_2 \leq u_2'$,

$$V_C([u_1, u_1'] \times [u_2, u_2']) = C(u_1', u_2') - C(u_1, u_2') - C(u_1', u_2) + C(u_1, u_2) \geq 0,$$

where V_C is called the C-volume of the rectangle $[u_1, u_1'] \times [u_2, u_2']$. Note that $V_C([0, u_1] \times [0, u_2]) = C(u_1, u_2)$.

In brief words, a copula is a function that maps any point in the unit square to a value on $[0, 1]$. From a probabilistic perspective, a copula is a joint cumulative distribution function whose marginal distributions are uniform, i.e, $V_C([0, u_1] \times [0, u_2]) = C(u_1, u_2)$. A special copula is the product copula, defined as $\Pi(u_1, u_2) = u_1 u_2$.

The informal and formal definitions are connected by the following Sklar theorem.

Sklar's Theorem [Sklar, 1959]

Let $F(t_1, t_2)$ be a two-dimensional distribution function with marginal distributions $F_1(t_1)$ and $F_2(t_2)$. Then, there exists a copula C such that

$$F(t_1, t_2) = C\{F_1(t_1), F_2(t_2)\}.$$

Furthermore, when $F_1(t_1)$ and $F_2(t_2)$ are continuous, then C is unique. Conversely, for any distribution functions F_1 and F_2 and any copula C, the $F(\cdot, \cdot)$ function defined above is a two-dimensional distribution function with margins $F_1(t_1)$ and $F_2(t_2)$.

Based on the Sklar's theorem, the joint distribution function of (t_1, t_2) could be parameterized by two marginal distributions and a copula, indexed by a parameter η:

$$F(t_1, t_2; \eta) = C_\eta(F_1(t_1), F_2(t_2)),$$

where η is the dependence parameter of the copula.

Tail dependence

Joe [1997] defined the tail-dependence coefficient (TDC) in copula. Let T_1 and T_2 be two continuous random variables with marginal distribution $S_1(t_1)$ and $S_2(t_2)$ that are coupled with a copula C_η to form a joint survival function. Then the upper and lower tail dependence coefficients of (T_1, T_2) are defined as:

$$\lambda_L = \lim_{v \to 1^-} P(S_2(T_2) \geq v | S_1(T_1) \geq v) = \lim_{v \to 1^-} \frac{C_\eta(1-v, 1-v)}{1-v}$$

and

$$\lambda_U = \lim_{v \to 0^+} P(S_2(T_2) \leq v | S_1(T_1) \leq v) = \lim_{v \to 0^+} \frac{1 - 2v + C_\eta(1-v, 1-v)}{v}$$

provided that λ_U and λ_L exist.

When $\lambda_L(\lambda_U) \in (0,1]$, T_1 and T_2 are asymptotically dependent in the lower (upper) tail. If $\lambda_L(\lambda_U) = 0$, T_1 and T_2 are asymptotically independent in the lower (top) tail.

Archimedean copula family

One popular copula group is the Archimedean copula family, and it is widely used in areas such as finance, insurance, and health. It has explicit expressions [Nelsen, 2006, Schweizer and Sklar, 2011]:

$$C_\eta(u,v) = H_\eta\{H_\eta^{-1}(u) + H_\eta^{-1}(v)\},\ 0 \leq u, v \leq 1,$$

where $H_\eta : [0,\infty) \to [0,1]$ is a generator function. The generator function is a strictly decreasing function, as illustrated in the copula definition. That is, it is a continuous strictly decreasing and convex function from $[0,+\infty]$ to $\mathbf{I}$, with $H_\eta(0) = 1$.

If H_η is a Laplace transformation of some distribution, the Archimedean copula family will reduce to proportional frailty models [Marshall and Olkin, 1988, Oakes, 1989]. For example, when $H_\eta(u) = (1+u)^{-1/\eta}$, which is the Laplace transformation of a Gamma distribution, it becomes the Clayton copula [Clayton, 1978]:

$$C_\eta(u,v) = (u^{-\eta} + v^{-\eta} - 1)^{-1/\eta},\ \eta \in (0,\infty),$$

where T_1 and T_2 are positively associated and become independent when $\eta \to 0$, and $\lambda(t_2|T_1 = t_1)/\lambda(t_2|T_1 \geq t_1) = \eta - 1$. When $H_\eta(u) = \exp(-u^{1/\eta})$, the Laplace transformation of a positive stable distribution, it becomes the Gumbel copula [Gumbel, 1960, Hougaard, 1986]:

$$C_\eta(u,v) = \exp[-\{(-\log u)^\eta + (-\log v)^\eta\}^{1/\eta}],\ \eta \in (1,\infty),$$

where T_1 and T_2 are positively associated and become independent when $\eta \to 1$.

The Clayton copula models lower tail correlation between u and v, with Kendall's $\tau = \frac{\eta}{\eta+2}$, while the Gumbel copula models upper tail dependence, with Kendall's $\tau = 1 - \frac{1}{\eta}$.

Two-parameter Archimedean copula

To model both upper and lower tail dependence, one can apply the two-parameter copula [Joe, 1997], derived from a more sophisticated generator function and inverse generator function:

$$H_{\alpha,\kappa}(s) = (\frac{1}{1+s^\alpha})^\kappa,\ s \in [0,+\infty),$$

and

$$H_{\alpha,\kappa}^{-1}(u) = (u^{-1/\kappa} - 1)^{\frac{1}{\alpha}},\ u \in [0,1),$$

where α models the upper tail dependence and κ models the lower tail dependence. The joint survival function is defined by the two-parameter copula:

$$\begin{aligned} S(t_1, t_2; \alpha, \kappa) &= C_{\alpha,\kappa}(S_1(t_1), S_2(t_2)) \\ &= H_{\alpha,\kappa}[H_{\alpha,\kappa}^{-1}\{S_1(t_1)\} + H_{\alpha,\kappa}^{-1}\{S_2(t_2)\}] \\ &= [1 + \{(u^{-1/\kappa} - 1)^{1/\alpha} + (v^{-1/\kappa} - 1)^{1/\alpha}\}^{\alpha}]^{-\kappa},\ \alpha \in (0,1],\ \kappa \in (0,\infty), \end{aligned}$$

where $u, v \in [0,1]$.

Both Clayton and Gumbel copulas are special cases of the two-parameter copula model, in which Kendall's $\tau = 1 - \frac{2\alpha\kappa}{2\kappa+1}$. When $\alpha \to 1$, the two-parameter copula becomes Clayton copula, with Kendall's $\tau = \frac{\kappa^{-1}}{\kappa^{-1}+2}$. When $\kappa \to \infty$, we get Gumbel copula, with Kendall's $\tau = 1 - \alpha$. In Chapter 2, we will introduce the first two-parameter copula survival model in bivariate interval-censored data.

1.4.4 Relationship between frailty and copula

Both frailty and Archimedean copula models could model the bivariate dependence structure between two survival times. The form of frailty is determined by choice of Laplace transformation function $\mathscr{L}_\eta(.)$, whereas the form of copula depends on the specification of the generator function $H_\eta(.)$. Oakes [1989] suggested that the two models are intimately connected. To establish the relationship between copula and frailty, we define $H_\eta(.)$ in the Copula model to be the Laplace transformation $\mathscr{L}_\eta(.)$ in the frailty model. Then, the joint survival function under the copula model becomes

$$S_c(t_1, t_2) = \mathscr{L}_\eta[\mathscr{L}_\eta^{-1}\{S_{1,c}(t)\} + \mathscr{L}_\eta^{-1}\{S_{2,c}(t)\}].$$

Note that previously we have defined the joint survival function in the frailty model as

$$S_m(t_1, t_2) = \mathscr{L}_\eta[\mathscr{L}_\eta^{-1}\{S_{1,m}(t)\} + \mathscr{L}_\eta^{-1}\{S_{2,m}(t)\}].$$

When the frailty variable u follows a Gamma distribution $Gamma(1/\eta, 1/\eta)$ with unit mean and variance η ($\eta > 0$), the Laplace transformation becomes:

$$\mathscr{L}_\eta(s) = (1+\eta s)^{-1/\eta},\ \mathscr{L}_\eta^{-1}(s) = (s^{-\eta} - 1)/\eta.$$

Then the joint survival functions of the two models are re-written as:

$$S_c(t_1, t_2) = \Big[\{S_{1,c}(t_1)\}^{-\eta} + \{S_{2,c}(t_2)\}^{-\eta} - 1\Big]^{-1/\eta},$$

$$S_m(t_1, t_2) = \Big[\{S_{1,m}(t_1)\}^{-\eta} + \{S_{2,m}(t_2)\}^{-\eta} - 1\Big]^{-1/\eta},$$

where

$$S_{j,m}(t) = \mathscr{L}_\eta\{\Lambda_{j,m}(t)\} = \int e^{-u\Lambda_{j,m}(t)} f_\eta(u)du = \{1 + \eta\Lambda_{j,m}(t)\}^{-1/\eta},\ j = 1, 2.$$

We notice that the copula formula has become the popular Clayton copula, and more interestingly, the Gamma frailty and Clayton models share the same mathematical expression. However, $S_{j,m}(t)$ contains the frailty parameter η, while $S_{j,c}(t)$, which is the marginal function under copula models, is free of η by its definition. Thus, it leads to different joint survival functions [Goethals et al., 2008]. In fact, only when $\eta \to 0$, which indicates T_1 and T_2 are independent, the two models are equivalent with

$$\lim_{\eta \to 0} S_{j,m}(t) = e^{-\Lambda_{j,m}(t)},$$

which is free of η. In this special case, $S_{j,m}(t) = S_{j,c}(t)$ and $S_m(t_1, t_2) = S_c(t_1, t_2)$.

1.5 Interval-censored failure time data

Interval-censored data are common in many medical studies, in which the exact failure time is only known to lie within a time interval. The exact or right/left-censored failure times can be considered as special cases of interval-censoring, when the interval reduces to a single point or the right/left endpoint of the interval approaches infinity. The analysis of interval-censored data is more challenging and less developed than that of the right-censored data. Many popular methods for right-censored data, such as Cox partial likelihood and Kaplan-Meier estimator, do not apply under interval censoring. More discussions about interval-censored data can be found in Sun [2007]. In Chapters 2 and 3, I will introduce novel statistical methods for modeling bivariate interval-censored data and examining the goodness-of-fit of the fitted models.

1.5.1 Case I interval-censoring

The case I interval censoring is a special and simple case of interval-censoring. Each subject is observed only once during the entire study. As a result, the event of interest is only known to occur before or after the observation time. In this case, the case I data only contain left- or right-censored data, and such data are also referred to as current status data. The case I censoring is usually due to cross-sectional or nature of the experiment.

1.5.2 Case II interval-censoring

Case II interval-censored data are also known as general interval-censored data. Any interval-censored data that is not case I is considered as case II. In other words, case II interval-censored data are interval-censored data that include some finite intervals away from zero. For example, each subject is observed twice, where U and V are two random variables satisfying $U \leq V$. For another example, there exists a set of K (K is random) observation time points (case K or mixed case interval-censored data), which includes the first example as a special case and is a natural representation of interval-censored data arising from longitudinal studies with periodic follow-ups.

1.5.3 Panel count data

Previous examples treat the event of interest as an absorbing event and deal with time to the event or between two events. In practice, the event of interest could repeatedly appear over time, known as the recurrent event. If the recurrent process is monitored at discrete observation times, it leads to interval-censored recurrent event data, in which only the numbers of occurrences of the event are known at each observed time. This type of data is also referred to as panel count data. The counting process technique is commonly used for the analysis of panel count data.

1.6 Models for interval-censored data

1.6.1 Case I data

Unlike the right-censored data, there exists no comparable approach with the partial likelihood for the interval-censored case (including case I data). Thus, one needs to deal with a full likelihood that includes both finite-dimensional regression coefficients and infinite-dimensional nuisance parameters (e.g., the baseline cumulative hazard or the survival function). Among the semiparametric models for the case I data, Huang et al. [1996] and Rossini and Tsiatis [1996] proposed maximum likelihood-based approaches under the proportional hazards and proportional odds assumptions, respectively, and obtained inference of β through sieve maximum likelihood methods. Lin et al. [1998] developed an additive hazards model and made inferences based on estimating equations. Sun and Sun [2005] investigated linear transformation models for the case I interval-censored data.

1.6.2 Case II data

The case II interval-censored data includes more than one observation times for each survival time, so it contains more information than the case I data. However, the analysis of case II interval-censored data is more difficult than the case I data in terms of computation and inference. Finkelstein [1986] proposed a proportional hazards model and applied the Newtown-Raphson algorithm to determine the maximum likelihood estimator of coefficients and finite-dimensional baseline hazards together. Huang and Rossini [1997] and Shen [1998] applied sieve-based approaches in a proportional odds model by estimating the baseline log odds function through piecewise linear and monotonic spline functions, respectively. However, it is hard to choose the number of knots. Rabinowitz et al. [2000] developed an approximate conditional likelihood for the proportional odds model without estimating the baseline log odds function, but the method does not perform well in small samples. For the additive hazards model, Zeng et al. [2006], Chen and Sun [2009] and Zhu et al. [2008] investigated the maximum likelihood, multiple imputation, and transformation approaches, respectively. Wang et al. [2010] and Chen et al. [2007] fitted an additive hazards or proportional odds model, respectively, and established asymptotic properties of β by estimating equation methods. Rabinowitz et al. [1995], Li and Pu [2003] and Betensky et al. [2001] developed accelerated failure time models for case II interval-censored data using score statistics and estimating equation methods. Gu et al. [2005], Zhang et al. [2005] and Zhang and Zhao [2013] employed linear transformation models to obtain rank-based estimators, but they are computationally and statistically inefficient. Zeng et al. [2016] proposed a maximum likelihood estimation algorithm for a frailty-induced transformation model of interval-censored data with time-dependent covariates. The resulted coefficient estimates are consistent and asymptotically efficient. Zeng et al. [2017] further extended the semiparametric transformation models with random effect to multivariate interval-censored data.

Most of the papers mentioned above use the observed Fisher information matrix to estimate the variance-covariance matrix of the maximum likelihood estimators of regression coefficients. One alternative method is the profile likelihood approach, proposed by Huang and Wellner [1997], in which the variance-covariance matrix is estimated by the inverse of the curvature of the profile likelihood. It is feasible when the number of regression parameters β is small, and the profile likelihood is a smooth function of β.

1.6.3 Multivariate interval-censored data

Finkelstein et al. [2002], Chen et al. [2007], Tong et al. [2008] and Chen et al. [2013] fitted marginal models for multivariate interval-censored data under proportional hazards, proportional odds, additive hazards and linear transformation models, respectively. Those marginal approaches ignore the correlation structure within the multivariate data and may lose efficiency. They also assumed a common set of examination time points for all subjects. To account for the underlying dependency, Cook et al. [2008] developed a multistate model approach. Chen et al. [2009] proposed a frailty proportional hazards model for case I data and estimated baseline hazard functions by piece-wise constants. Chen et al. [2014] built a frailty proportional hazards model to case II data and applied an EM algorithm for parameter estimation. Wen and Chen [2013] developed a semiparametric maximum likelihood estimation approach for the gamma-frailty proportional hazards model under mixed-case interval censoring. Wang et al. [2015] employed an EM algorithm for bivariate current status data and estimated the baseline function by splines. More recently, Zhou et al. [2017] implemented a gamma frailty-based linear transformation model for bivariate interval-censored data and estimated regression parameters by a sieve maximum likelihood estimation approach. Besides the multistate and random effect models, Wang et al. [2008] took account of the correlation by implementing a copula model with proportional hazards margins for current status data in which the examination time is parameterized by a Cox model. Cook and Tolusso [2009] and Kor et al. [2013] developed copula models with proportional hazards margins and piece-wise baseline functions for the case I and II interval-censored data, respectively. In Chapter 2, I will introduce a novel two-parameter-copula-based semiparametric transformation model for bivariate data under general interval censoring.

1.7 Goodness-of-fit tests for copula models

Copula models are widely used to account for dependency between correlated distributions, and there are many different types of copula models. Therefore, a goodness-of-fit

(GOF) test for copula specification is highly desired. In the next sections, we will introduce the existing GOF tests for copula models.

1.7.1 Complete data

Wang and Wells [2000] proposed a non-parametric selection procedure for checking whether an Archimedean copula model properly models a random sample of bivariate right-censored data based on the L_2 norm of a truncated Kendall process introduced by Genest and Rivest [1993], which measures the distance between the empirical and model-based estimates of Kendall distribution. They also established the asymptotic behavior of the Kendall process. Later, Genest et al. [2006] extended Wang and Wells [2000] by developing a test statistic for complete data based on the probability integral transformation and the asymptotic behavior of a non-truncated Kendall process. Chen and Fan [2005] introduced pseudo-likelihood ratio tests for selecting semiparametric multivariate copula models in which the marginal distributions are unspecified, but the copula function is parameterized and can be misspecified. The tests compare between two or more candidate copula models, which can be either generalized non-nested or generalized nested. Huang and Prokhorov [2014] proposed an in-sample test statistic based on the subtraction of the expected Hessian matrix of log-likelihood and the expected outer product of the corresponding score function. More recently, Zhang et al. [2016] applied a pseudo-in-and-out-of sample (PIOS) likelihood ratio test statistic to check the goodness-of-fit for semiparametric copula models in identically independent distributed data and time series data. The essential idea of PIOS is to measure how sensitive the assumed copula-based likelihood is to the data change through a jackknife procedure. In particular, the asymptotic behavior of the PIOS test statistic is developed based on the Information Ratio (IR) test statistic, which was applied to check the covariance structure of the generalized estimating equation (GEE) model in Zhou et al. [2012]. One big advantage of the PIOS and IR tests is that they can apply to all parametric copula models with explicit functional forms. Moreover, their calculations are simple and straightforward.

1.7.2 Censored data

Shih [1998] first proposed a goodness-of-fit test for the Clayton family fitted in bivariate right-censored data. Specifically, their test compares unweighted and weighted concordance estimators of the dependence parameter η derived under the same class of estimation equations with different weight functions. The difference should be close to zero if the assumed Clayton model is the true copula model. Emura et al. [2010] further extended this idea to the general Archimedean copula family, and Fine and Jiang [2000] extended a similar idea to testing the Clayton copula with AFT margins in the presence of covariates. Overall, this type of method deletes non-orderable pairs of the bivariate event times from the estimating equation, which is difficult to adapt to bivariate interval-censored data where no exact event times are observed. Andersen et al. [2005] developed three types of bootstrap-based goodness-of-fit test statistics applicable to any pre-specified form of the copula in bivariate right-censored data. Its core idea is to compare the parametric estimate of the assumed copula and a non-parametric estimate via the chi-square type statistic, the Kolmogorov-like statistic, and the weighted difference based statistic. Due to the non-parametric estimation procedure in constructing the test statistics, this method does not have the power as the test of Shih [1998]. Lakhal-Chaieb [2010] extended Wang and Wells [2000] to testing Archimedean copulas under right censoring by developing a non-parametric inverse probability of censoring weighted estimator for Kendall's distribution. Chen et al. [2010] extended the pseudo-likelihood ratio tests of Chen and Fan [2005] to multivariate survival data under the general right censorship. Specifically, the event times are allowed to have different censoring mechanisms, for example, one random and the other fixed or one censored and the other uncensored. Its test hypothesis is to examine whether the assumed copula fits data better than a group of other copula models. Wang [2010] proposed a Fisher Z test statistic for bivariate right-censored data using Archimedean copulas based on multiply imputed complete data. Its statistic is derived from the correlation coefficient between two random variables following the Kendall distribution, which are shown to be independent under the correct Archimedean copula specification, as shown by Genest and Rivest [1993]. More recently, Mei [2016] proposed a likelihood-based PIOS test under right censoring based on the PIOS test for complete data from Zhang et al. [2016]. Lin and Wu [2020] developed a smooth test for copula specification in the right-censored data, which usually requires the selection of a suitable set of moment functions in constructing the test statistic. In addition, Yilmaz and Lawless [2011] developed a procedure for testing $\eta = \eta_0$ under the correct copula setting in bivariate right-censored data, which is completely different from testing copula specification as in the rest papers.

To the best of our knowledge, there is no formal statistical test for copula specification under interval censoring. In Chapter 3, I will introduce a novel information ratio (IR)-based goodness-of-fit test for diagnosing copula in multivariate survival data under complete, right- and interval-censored settings.

1.8 Survival prediction models

1.8.1 Survival prediction models for precision medicine in the big data era

Accurate 'time-to-event' data based survival prediction is fundamental to effective clinical management and precision medicine of human diseases [Chin et al., 2011, Compton, 2018]. It relies on a survival model to predict the dynamic risk profile of a future event over time (e.g., disease onset, recurrence, progression, or death) based on the individual's current status, such as clinical characteristics, genetic information, and medical images. Most importantly, such a prediction addresses the patient's key concern regarding the disease progression pattern in the future and shapes the physician's decision making for the treatment or clinical management strategy. It is to be noted that the survival prediction is fundamentally different from typical prediction models that predict a future event (whether occurs or not) by fixing the time of interest through a binary classification [Castro-Rodríguez et al., 2000, Chi et al., 2007]. Despite its essential role in precision medicine, the survival prediction remains a challenging task [Abrams et al., 2014, Barillot et al., 2012, Schumacher et al., 2012], largely due to the complex nature of diseases and the heterogeneity between patients. Therefore, there is an urgent need for developing accurate and personalized survival prediction models with improved capacity in learning the complex structures and interplays among predictors. Recent advances in high-throughput technologies have generated large volumes of molecular profiling data for each patient, which provides unprecedented opportunities in identifying potential biomarkers and further establishing accurate survival prediction models [Chen et al., 2019, Collins and Varmus, 2015, Sarnowski et al., 2018]. In particular, several national-wide large-scale longitudinal studies, such as the Trans-Omics for Precision Medicine (TOPMed) and All of Us, are underway using whole-genome sequencing and other omics technologies, with the ultimate goal of accelerating precision medicine. However, how to effectively utilize the wealthy amount of data is challenging. The first challenge comes from how to connect high-dimensional predictors with the outcome of interest. This problem is particularly difficult in survival prediction because the events of interest are often censored due to either a short study period or loss of follow-up during the study. The second challenge is how to model the complex structure among numerous biomarkers, where the specific structure is largely unknown. The third challenge is that given the heterogeneity of patients, how to interpret the importance of each predictor for each patient and further how to identify patient subgroups to provide personalized prevention or treatment strategy.

The recent advances in multi-layer deep neural network models have made extraordinary achievements in providing new effective risk prediction models from complex and high dimensional biomedical data, such as omics and biomedical imaging [Grassmann et al., 2018, Min et al., 2016, Miotto et al., 2017, Poplin et al., 2018]. However, the application of deep learning in survival prediction is still limited.

1.8.2 Cox proportional hazards model

The Cox proportional hazards model is the most popular regression model for right-censored survival data. It assumes that the hazard function of survival time T takes the form [Klein and Moeschberger, 2006]

$$h(t|Z_i) = h_0(t)\exp(Z_i^T\theta), \tag{1.8.1}$$

where $h_0(t)$ is the unspecified baseline hazard function at time t, and θ is a vector of covariate effects. The term $Z_i^T\theta$ is called the linear predictor or prognostic index.

The estimator $\hat{\theta}$ can be obtained by maximizing the log partial likelihood [Cox, 1972]

$$-\frac{1}{N_D}\sum_{j\in D}\Big\{Z_j^T\theta - \log\sum_{i\in R_j} e^{Z_i^T\theta}\Big\}, \tag{1.8.2}$$

where D is the set of all events with size N_D; $\{t_j\}$ is the set of unique event times; R_j is the risk set satisfying $Y_i \geq t_j$. The standard optimization algorithm, such as Newton-Raphson, can be used to maximize the log partial likelihood.

To obtain the estimated survival probabilities for each subject i, we have

$$\hat{S}(t|Z_i) = \exp\{-\hat{H}_0(t)e^{Z_i^T\hat{\theta}}\}, \tag{1.8.3}$$

where $\hat{H}_0(t) = \int_0^t \hat{h}_0(u)du$ is the estimated baseline cumulative hazard function [Klein and Moeschberger, 2006].

The Cox model is a flexible semiparametric model that does not assume a parametric distribution for the baseline hazard function $h_0(t)$. However, it suffers from the limitation in the dimension of covariates Z_i. It does not work in the presence of high-dimensional predictors, such as in the genome-wide association study (GWAS). Moreover, it assumes a linear relationship in the prognostic index (i.e., $Z_i^T\theta$), which may not hold in practice. Therefore, more sophisticated survival models are needed to handle both high-dimensional predictors and non-linear structure among predictors.

1.8.3 Cox LASSO model

One approach to handle high-dimensional covariates is the Cox LASSO method. Tibshirani [1996] proposed to shrink regression coefficients by L_1 penalization and later extended the method to the regular Cox proportional hazards model [Tibshirani, 1996]. The loss function for LASSO is the negative log partial likelihood function (formula (1.8.2)) plus the L_1 penalty:

$$-\frac{1}{N_D}\sum_{j\in D}\Big\{Z_j^T\theta - \log\sum_{i\in R_j} e^{Z_i^T\theta}\Big\}+\lambda||\theta||_1, \tag{1.8.4}$$

where λ is the L_1 penalty parameter that enables LASSO to deal with high-dimensional covariates. The optimization of this penalized log partial likelihood function is implemented in the R package *glmnet* [Simon et al., 2011]. However, based on the formula (1.8.4), we can see that LASSO also assumes that the prognostic index is a linear combination of covariates. Therefore, it is still not flexible enough to account for non-linear covariate structures (such as non-linear and interaction effects).

1.8.4 Random survival forest (RSF) model

The random survival forest model [Ishwaran et al., 2008] is a tree-ensemble nonparametric method for survival outcomes. It grows every single tree by randomly drawing bootstrap samples from original data and further randomly selecting a subset of predictors as candidates for splitting at each node. At each node, the best split is found among all binary splits defined by the selected predictors according to a splitting rule, such as the log-rank test. Finally, the model aggregates terminal nodes across all survival trees and obtain a survival prediction ensemble. RSF has become a popular survival prediction method, and it does not assume linearity among predictors. We implement the RSF model through the R package *RandomForestSRC* [Ishwaran and Kogalur, 2007].

1.8.5 Deep neural network (DNN) survival model

Unlike regular Cox or LASSO, the deep neural network model is well known for its capacity in learning complex covariate structures (i.e., non-linearity, interactions) [LeCun et al., 2015]. By the Universal Approximation Theorem [Cybenko, 1989, Hornik et al., 1989], for any continuous function $g(Z;\theta)$, there is guaranteed to be a neural network that approximates this function. Moreover, this theorem holds even if we restrict the neural networks to have just one single hidden layer. Therefore, even a very simple neural network architecture can be extremely powerful. The synergy of the powerful DNN and the popular Cox model leads us to build a DNN survival model and evaluate it together with other already discussed machine learning survival models. More details can be found in Chapter 4.

2.0 Copula-based Semiparametric Regression Model in Bivariate Data under General Interval Censoring

2.1 Introduction

Bivariate time-to-event endpoints are frequently used as co-primary outcomes in biomedical and epidemiological fields. For example, two time-to-event endpoints are often seen in clinical trials studying the progression (or recurrence) of bilateral diseases (e.g., eye diseases) or complex diseases (e.g., cancer and psychiatric disorders). The two endpoints are correlated as they come from the same individual. Bivariate interval-censored data arise when both events are not precisely observed due to intermittent assessment times. Therefore, the event times are only known to belong to an interval (i.e., case II interval-censored). A further complication is that the event status can be indeterminate (i.e., right-censored) for individuals who are event-free at their last assessment time. The special case when there exists only one assessment time, leading to the bivariate current status data (events are either left- or right-censored), can also happen for some individuals. Therefore, the bivariate data we are interested in modeling are under general interval censoring, which may include a mixture of left-, right- and interval-censored data.

Our motivating example of such bivariate general interval-censored data came from a large clinical trial [AREDS Group, 1999] studying the progression of a bilateral eye disease, Age-related Macular Degeneration (AMD), of which the two-eyes from the same patient were periodically examined for late-AMD. The study aims to discover genetic variants that are significantly associated with AMD progression, as well as to characterize both the joint and conditional risks of AMD progression. For example, the joint 5-year progression-free probability for both eyes is a clinically significant measure to group patients into different risk categories. Similarly, for patients who have one eye already progressed, the conditional 5-year progression-free probability for the non-progressed eye (given its fellow eye already progressed) is vital to both clinicians and patients. Therefore, a desired statistical method needs to characterize and predict both joint and conditional risk profiles.

There are several approaches to modeling bivariate interval-censored data. For example, Goggins and Finkelstein [2000], Kim and Xue [2002], Chen et al. [2007], Tong et al. [2008] and Chen et al. [2013] fitted various marginal models for multivariate interval-censored data. All these approaches model the marginal distributions based on the working independence assumption, and thus cannot produce joint or conditional distributions. Another popular method is based on frailty models (for example, Oakes, 1982), which are mixed effects models with a latent frailty variable applied to the conditional hazard functions. For example, Chen et al. [2009] and Chen et al. [2014] built frailty proportional hazards (PH) models with piecewise constant baseline hazards for multivariate current status data and interval-censored data, respectively. Wen and Chen [2013] and Wang et al. [2015] developed Gamma-frailty PH models for bivariate interval-censored data through a nonparametric maximum likelihood estimation approach and bivariate current status data through a sieve estimation approach, respectively. Recently, Zhou et al. [2017] and Zeng et al. [2017] proposed frailty-based transformation models for bivariate or multivariate interval-censored data, and obtained parameter estimates through the sieve maximum likelihood estimation and nonparametric maximum likelihood estimation, respectively. For frailty models, the covariate effects are typically interpreted on the conditional level by conditioning on the random frailty term.

The third popular approach is based on copula models [Clayton, 1978, for example]. Unlike the marginal or frailty approaches, the copula-based methods directly connect the two marginal distributions through a copula function to construct the joint distribution, of which the copula parameter determines the dependence. This unique feature makes the modeling of the margins separable from the copula function, which is attractive from both the modeling perspective and the interpretation purpose. Both joint and conditional distributions can be obtained from copula models. Several copula models have been proposed in the literature. Wang et al. [2008] used sieve estimation in a copula model with proportional hazards margins for bivariate current status data. Cook and Tolusso [2009] and Kor et al. [2013] developed estimating equations for copula models with piecewise constant baseline marginal hazards for clustered current status and interval-censored data, respectively. Hu et al. [2017] developed a semiparametric sieve approach for bivariate current status data using copula framework with proportional hazards margins. To date, most copula-based regression models only handle a specific interval censoring type (e.g., case I current status or case II interval censoring) and are often limited to the PH assumption. Also, the most frequently used copula models, such as Clayton, Gumbel, and Frank, all use only one dependence parameter, which can result in a lack of flexibility.

Goethals et al. [2008] and Wienke [2010] have discussed the connection and distinction between copula and frailty models. For example, the Clayton copula has the same mathematical expression as the Gamma frailty model in terms of the joint survival distribution. However, their marginal survival functions are modeled differently. Specifically, the marginal function under the Clayton model only involves the time and covariate effects, whereas the marginal function under the Gamma frailty model includes not only the time and covariate effects but also the frailty parameter. As a result, the joint distribution functions of the Clayton copula and Gamma frailty models are not equivalent, except when the two margins are independent. More details are discussed in Appendix A.3. In this chapter, the objectives of our real study lead us to choose copula-based models, which offer a straightforward interpretation of covariate effects and dependence strength, as well as an easy generation of joint and conditional survival distributions.

We propose a class of copula-based semiparametric transformation model for bivariate data subject to general interval censoring. Specifically, we build a two-parameter copula model framework, which can handle more flexible dependence structures than one-parameter copulas. Our method incorporates a broad class of semiparametric regression models that includes both PH and PO models. We approximate the infinite-dimensional nuisance parameters using sieves with Bernstein polynomials and propose a novel maximum likelihood estimation procedure which is computationally stable and efficient. We establish the asymptotic normality and efficiency for the sieve estimators of finite-dimensional model parameters. Moreover, we develop a generalized score test with numerical approximations of the score function and observed Fisher information for testing covariate effects.

The chapter is organized as follows. Section 2.2 introduces the model and the joint likelihood function. Section 2.3 presents the sieve maximum likelihood estimation procedure, the asymptotic properties, and the generalized score test. Section 2.4 illustrates extensive simulation studies for the estimation and testing performances of our proposed methods. We analyze the Age-related Eye Disease Study (AREDS) data and present the findings in Section 2.5. Finally, we discuss and conclude in Section 2.6. Additional simulation and analysis results, the regularity conditions, proofs and additional technical details are provided in Appendix A.

2.2 Notation and likelihood

2.2.1 Copula model for bivariate censored data

Assume there are n independent subjects in a study. For subject i, we observe $D_i = \{(L_{ij}, R_{ij}, Z_{ij}), j = 1, 2\}$, where $(L_{ij}, R_{ij}]$ is the time interval that the true event time T_{ij} lies in and Z_{ij} is the covariate vector. When $R_{ij} = \infty$, T_{ij} is right-censored, and when $L_{ij} = 0$, T_{ij} is left-censored. We define the marginal survival function for subject i margin j as $S_j(t_{ij}|Z_{ij}) = pr(T_{ij} > t_{ij}|Z_{ij})$ and the joint survival function for subject i as $S(t_{i1}, t_{i2}|Z_{i1}, Z_{i2}) = pr(T_{i1} > t_{i1}, T_{i2} > t_{i2}|Z_{i1}, Z_{i2})$.

By the Sklar's theorem (Sklar, 1959), so long as marginal survival functions S_j are continuous, there exists a unique function C_η that connects two marginal survival functions into the joint survival function: $S(t_1, t_2|Z_1, Z_2) = C_\eta(S_1(t_1|Z_1), S_2(t_2|Z_2)),\ t_1, t_2 \geq 0$. Here, the function C_η is called a copula, which maps $[0, 1]^2$ onto $[0, 1]$ and its parameter η measures the dependence between the two margins. A signature feature of the copula is that it allows the dependence to be modeled separately from the marginal distributions [Nelsen, 2006].

One favorite copula family for bivariate censored data is the Archimedean copula family, which usually has an explicit formula. Two frequently used Archimedean copulas are the Clayton (Clayton, 1978) and Gumbel (Gumbel, 1960) copula models, which account for the lower or upper tail dependence between two margins using a single parameter.

Here, we consider a more flexible two-parameter Archimedean copula model [Joe, 1997], which is formulated as

$$C_{\alpha,\kappa}(u, v) = [1 + \{(u^{-1/\kappa} - 1)^{1/\alpha} + (v^{-1/\kappa} - 1)^{1/\alpha}\}^\alpha]^{-\kappa},\ \alpha \in (0, 1],\ \kappa \in (0, \infty), \quad (2.2.1)$$

where u and v are two uniformly distributed margins. The two dependence parameters (α and κ) account for the correlation between u and v at both upper and lower tails, and they explicitly connect to the Kendall's τ with $\tau = 1 - 2\alpha\kappa/(2\kappa + 1)$. In particular, when $\alpha = 1$, the two-parameter copula (2.2.1) becomes the Clayton copula, and when $\kappa \to \infty$, it becomes the Gumbel copula. Thus, the two-parameter copula model provides more flexibility in characterizing the dependence than the Clayton or Gumbel copula.

2.2.2 Joint likelihood for bivariate data under general interval censoring

The joint likelihood function using the two-parameter copula model can be written as

$$
\begin{aligned}
L_n(S_1, S_2, \alpha, \kappa \mid D) &= \prod_{i=1}^{n} pr(L_{i1} < T_{i1} \leq R_{i1}, L_{i2} < T_{i2} \leq R_{i2} \mid Z_{i1}, Z_{i2}) \\
= \prod_{i=1}^{n} \Big\{ & pr(T_{i1} > L_{i1}, T_{i2} > L_{i2} \mid Z_{i1}, Z_{i2}) - pr(T_{i1} > L_{i1}, T_{i2} > R_{i2} \mid Z_{i1}, Z_{i2}) \\
& - pr(T_{i1} > R_{i1}, T_{i2} > L_{i2} \mid Z_{i1}, Z_{i2}) + pr(T_{i1} > R_{i1}, T_{i2} > R_{i2} \mid Z_{i1}, Z_{i2}) \Big\} \\
= \prod_{i=1}^{n} \Big[& C_{\alpha,\kappa}\{S_1(L_{i1} \mid Z_{i1}), S_2(L_{i2} \mid Z_{i2})\} - C_{\alpha,\kappa}\{S_1(L_{i1} \mid Z_{i1}), S_2(R_{i2} \mid Z_{i2})\} \\
& - C_{\alpha,\kappa}\{S_1(R_{i1} \mid Z_{i1}), S_2(L_{i2} \mid Z_{i2})\} + C_{\alpha,\kappa}\{S_1(R_{i1} \mid Z_{i1}), S_2(R_{i2} \mid Z_{i2})\} \Big]. \quad (2.2.2)
\end{aligned}
$$

For a given subject i, if T_{ij} is right-censored, then any term involving R_{ij} becomes 0 (since R_{ij} is set to be ∞). Then the joint survival function for subject i reduces to either only one term (if both T_{i1} and T_{i2} are right-censored) or two terms (if one T_{ij} is right-censored). The particular case of current status data can also fit into this model frame, where either L_{ij} is 0 (if the event has already occurred before the examination time, which is R_{ij} in this case) or R_{ij} is ∞ (if the event has not happened upon the examination time, which is L_{ij} in this case). Therefore, the likelihood function (2.2.2) can handle the general form of bivariate interval-censored data.

Next, we will estimate both the dependence parameters (α, κ) and two marginal survival functions (S_1, S_2) together.

2.2.3 Semiparametric linear transformation model for marginal functions

We consider the semiparametric transformation models for marginal survival functions:

$$S_j(t \mid Z_j) = \exp[-G_j\{\exp(Z_j^T\beta_j)\Lambda_j(t)\}],\ j = 1, 2, \tag{2.2.3}$$

where $G_j(\cdot)$ is a pre-specified strictly increasing function, β_j is a vector of unknown regression coefficients, and $\Lambda_j(\cdot)$ is an unknown non-decreasing function of t. In model (2.2.3), the transformation function $G_j(\cdot)$, the regression parameter β_j and the infinite-dimensional parameter $\Lambda_j(\cdot)$ are all denoted as margin-specific (indexed by j) for generality. In practice, some or all of them can be the same for the two margins, and in that case, the corresponding index j can be dropped.

This model (2.2.3) contains a class of survival models. For example, when $G(x) = x$, the marginal survival function follows a proportional hazards model. When $G(x) = \log(1 + x)$, the marginal function becomes a proportional odds model. In practice, the transformation function can also be "estimated" by the data. For example, the commonly used Box-Cox transformation $G(x) = \{(1 + x)^r - 1\}/r$, $r > 0$, or the logarithmic transformation $G(x) = \log(1 + rx)/r$, $r > 0$, can be assumed. The proportional hazards and proportional odds models are special cases in both transformation classes. Then the parameter r in $G(\cdot)$ can be estimated together with other parameters in the likelihood, as we will demonstrate in our simulation studies.

2.3 Estimation and inference

2.3.1 Sieve likelihood with Bernstein polynomials

In our likelihood function, we are interested in estimating the unknown parameter $\theta \in \Theta$:

$$\Theta = \{\theta = (\beta_1^T, \beta_2^T, \alpha, \kappa, \Lambda_1, \Lambda_2)^T \in \mathcal{B} \otimes \mathcal{M} \otimes \mathcal{M}\}.$$

Here $\mathcal{B} = \{(\beta = (\beta_1^T, \beta_2^T)^T, \alpha, \kappa) \in R^p \times R^{(0,1]} \times R^+, \|\beta\| + \|\alpha\| + \|\kappa\| \leq M\}$ with p being the dimension of β and M being a positive constant. We denote by $\mathcal{M}$ the collection of all bounded, continuous and nondecreasing, nonnegative functions over $[c, u]$, where $0 \leq c < u < \infty$. In practice, $[c, u]$ can be chosen as the maximum range of all L_{ij} and R_{ij}.

In our log-likelihood function

$$l_n(\theta; D) = \log L_n(\theta; D) = \sum_{i=1}^{n} \log L(\theta; D_i) = \sum_{i=1}^{n} l(\theta; D_i),$$

there are finite-dimensional parameters of interest (β, α, κ) and two infinite-dimensional nuisance parameters (Λ_1, Λ_2), which need to be estimated simultaneously. Unlike the right-censored data, tools like partial likelihood and martingale can not be applied to the interval-censored data due to the absence of exact event times. Instead, following Huang and Rossini [1997], we employ the sieve approach and form a sieve likelihood. Specifically, similar to Zhou et al. [2017], we use Bernstein polynomials to build a sieve space $\Theta_n = \{\theta_n = (\beta^T, \alpha, \kappa, \Lambda_{1n}, \Lambda_{2n})^T \in \mathcal{B} \otimes \mathcal{M}_n \otimes \mathcal{M}_n\}$. Here, $\mathcal{M}_n$ is the space defined by Bernstein polynomials for both $j = 1$ and 2:

$$\mathcal{M}_n = \left\{ \Lambda_{jn}(t) = \sum_{k=0}^{m_n} \phi_{jk} B_k(t, m_n, c, u) : \sum_{k=0}^{m_n} |\phi_{jk}| \leq M_n;\ 0 \leq \phi_{j0} \leq \cdots \leq \phi_{jm_n}; j = 1, 2 \right\},$$

where t denotes time, $B_k(t, m_n, c, u)$ represents the Bernstein basis polynomial defined as:

$$B_k(t, m_n, c, u) = \binom{m_n}{k} (\frac{t - c}{u - c})^k (1 - \frac{t - c}{u - c})^{m_n - k};\ k = 0, ..., m_n, \tag{2.3.1}$$

with degree $m_n = o(n^\nu)$ for some $\nu \in (0, 1)$, ϕ_{jk} are coefficients, and $M_n = O(n^a)$ with a being a positive constant. We assume the basis polynomials $B_k(t, m_n, c, u)$ are the same between the two margins, while the coefficients ϕ_{jk} can be margin-specific. In practice, one may choose m_n based on model AIC values. With a pre-specified m_n, we solve ϕ_{jk} together with other parameters (β, α, κ). One big advantage of Bernstein polynomials is that they can achieve the non-negativity and monotonicity properties of $\Lambda_j(t)$ through re-parameterization [Zhou et al., 2017]. Another advantage of Bernstein polynomials is that they do not require the specification of interior knots, as seen from (2.3.1), making them flexible for use.

With the sieve space defined above, $\Lambda_j(t)$ will be approximated by $\Lambda_{jn}(t) \in \mathcal{M}_n$. In the next section, we propose an estimation procedure to maximize $l_n(\theta; D)$ over the sieve space Θ_n to obtain the sieve maximum likelihood estimators $\hat{\theta}_n = (\hat{\beta}_n^T, \hat{\alpha}_n, \hat{\kappa}_n, \hat{\Lambda}_{1n}, \hat{\Lambda}_{2n})^T$.

2.3.2 Estimation procedure for sieve maximum likelihood estimators $\hat{\theta}_n$

We develop a novel sieve maximum likelihood estimation procedure that is generally applicable to any choice of Archimedean copulas and marginal models. In principle, we can obtain the sieve maximum likelihood estimators by maximizing the joint likelihood function (2.2.2) in one step. Due to the complex structure of the joint likelihood function, we recommend using a separate step to obtain appropriate initial values for all the unknown parameters. In essence, (β_j, Λ_{jn}) are first estimated marginally in step 1(a). Then their estimators are plugged into the joint likelihood to form a pseudo-likelihood. In step 1(b), the dependence parameters (α, κ) are estimated through maximizing the pseudo-likelihood function. Finally, using initial values from step 1(a) and 1(b), we update all the unknown parameters simultaneously under the joint log-likelihood function in step 2. The estimation procedure is described below:

1. Obtain initial estimates of θ_n:
   a. $(\hat{\beta}_{jn}^{(1)}, \hat{\Lambda}_{jn}^{(1)}) = \arg\max_{(\beta_j, \Lambda_{jn})} l_{jn}(\beta_j, \Lambda_{jn})$, where l_{jn} denotes the sieve log-likelihood for the marginal model, $j = 1, 2$;
   b. $(\hat{\alpha}_n^{(1)}, \hat{\kappa}_n^{(1)}) = \arg\max_{(\alpha,\kappa)} l_n(\hat{\beta}_n^{(1)} = (\hat{\beta}_{1n}^{(1)}, \hat{\beta}_{2n}^{(1)}), \alpha, \kappa, \hat{\Lambda}_{1n}^{(1)}, \hat{\Lambda}_{2n}^{(1)})$, where $\hat{\beta}_{jn}^{(1)}$ and $\hat{\Lambda}_{jn}^{(1)}$ are the initial estimates from (a), and l_n is the joint sieve log-likelihood.
2. Simultaneously maximize the joint sieve log-likelihood to get final estimates: $\hat{\theta}_n = (\hat{\beta}_n, \hat{\alpha}_n, \hat{\kappa}_n, \hat{\Lambda}_{1n}, \hat{\Lambda}_{2n}) = \arg\max_{(\beta,\alpha,\kappa,\Lambda_{1n},\Lambda_{2n})} l_n(\beta, \alpha, \kappa, \Lambda_{1n}, \Lambda_{2n})$ with initial values $(\hat{\beta}_n^{(1)}, \hat{\alpha}_n^{(1)}, \hat{\kappa}_n^{(1)}, \hat{\Lambda}_{1n}^{(1)}, \hat{\Lambda}_{2n}^{(1)})$ obtained from step 1(a) and 1(b).

For the variance-covariance of finite-dimensional parameter estimates $(\hat{\beta}_n, \hat{\alpha}_n, \hat{\kappa}_n)$, we invert the observed information matrix of all parameters including the nuisance parameters (ϕ_{jk}) from the last iteration of step 2 and then take the corresponding block. In section 2.3.3, we establish the asymptotic normality and semiparametric efficiency for the finite-dimensional parameters. However, since the asymptotic variance form is intractable, we adopt this heuristic approach, which has been shown to work well in practice [Ding and Nan, 2011].

Some standard optimization algorithms such as the Newton-Raphson algorithm or the conjugate gradient algorithm can be employed to obtain the maximizers and observed information matrix. Due to the complex structure of the joint sieve log-likelihood, instead of analytically deriving the first and second order derivatives, we propose to use the Richardson's extrapolation (Lindfield and Penny, 1989) to approximate the score function and observed information matrix numerically. As shown in our simulations, the proposed procedure guarantees almost 100% convergence and the computing speed is notably improved by using initial values from step 1.

2.3.3 Asymptotic properties of sieve estimators

This section presents asymptotic properties of the sieve maximum likelihood estimators $\hat{\theta}_n$ with regularity conditions and proofs being supplied in Appendix A.4. Denote P as the true probability measure and $\mathbb{P}_n$ as the empirical measure for n independent subjects. Let $|v|$ be the Euclidean norm for a vector v. Define the supremum norm $\|f\|_\infty = sup_t|f(t)|$ for a function $f(t)$. Also define $\|f\|_{L_2(P)} = (\int |f|^2 dP)^{1/2}$ for a function f under the probability measure P. In particular, the $L_2(P)$ norm for Λ_j is defined as $\|\Lambda_j\|_2^2 = \int[\{\Lambda_j(l)\}^2 + \{\Lambda_j(r)\}^2]dF_j(l,r)$, where $F_j(l,r)$ denotes the joint cumulative distribution function of L_{ij} and R_{ij} $(i = 1, ..., n; j = 1, 2)$. Finally, we define the distance between $\theta_1 = (\beta_1^T, \alpha_1, \kappa_1, \Lambda_{11}, \Lambda_{21})^T \in \Theta$ and $\theta_2 = (\beta_2^T, \alpha_2, \kappa_2, \Lambda_{12}, \Lambda_{22})^T \in \Theta$ as

$$d(\theta_1, \theta_2) = (|\beta_1 - \beta_2|^2 + |\alpha_1 - \alpha_2|^2 + |\kappa_1 - \kappa_2|^2 + \|\Lambda_{11} - \Lambda_{12}\|_2^2 + \|\Lambda_{21} - \Lambda_{22}\|_2^2)^{1/2}.$$

Let $\theta_0 = (\beta_0^T, \alpha_0, \kappa_0, \Lambda_{10}, \Lambda_{20})^T$ denote the true value of $\theta \in \Theta$. The following theorems present the convergence rate, asymptotic normality, and efficiency of the sieve estimators.

Theorem 2.3.1. *(Convergence rate) Assume that Conditions 1-2 and 4-5 in Appendix A.4 hold. Let $m_n = o(n^\nu)$, where $\nu \in (0,1)$ and q be the smoothness parameter of Λ_j as defined in Condition 4, then we have*

$$d(\hat{\theta}_n, \theta_0) = O_p\big(n^{-\min\{q\nu/2,(1-\nu)/2\}}\big).$$

Theorem 2.3.1 states that the sieve estimator $\hat{\theta}_n$ has a polynomial convergence rate. Although this overall convergence rate is lower than $n^{-1/2}$, the following normality theorem states that the proposed estimators of the finite-dimensional parameters (β, α, κ) are asymptotically normal and semiparametrically efficient.

Theorem 2.3.2. *(Asymptotic normality and efficiency) Suppose Conditions 1-5 in Appendix A.4 hold. Define* $\hat{b}_n = (\hat{\beta}_n^T, \hat{\alpha}_n, \hat{\kappa}_n)^T$ *and* $b_0 = (\beta_0^T, \alpha_0, \kappa_0)^T$*. If* $1/(2+q) < \nu < 1/2$*, then*

$$n^{1/2}(\hat{b}_n - b_0) = I^{-1}(b_0) n^{1/2} \mathbb{P}_n l^*(b_0, \Lambda_{10}, \Lambda_{20}; D) + o_p(1) \to_d N\{0, I^{-1}(b_0)\},$$

where $I(b_0) = Pl^*(b_0, \Lambda_{10}, \Lambda_{20}; D)^{\otimes 2}$ *and* $l^*(b_0, \Lambda_{10}, \Lambda_{20}; D)$ *is the efficient score function defined in the proof. Therefore,* $\hat{b}_n$ *is asymptotically normal and efficient.*

2.3.4 Generalized score test

We now separate β into two parts: β_g and β_{ng}, where β_g is the parameter set of interest for hypothesis testing and β_{ng} denotes the rest of the regression coefficients. The likelihood-based tests such as the Wald, score, and likelihood-ratio tests can be constructed to test β_g, and they are asymptotically equivalent. In our motivating study, we aim to perform a GWAS on AMD progression, which contains testing millions of SNPs one-by-one. Therefore, computing speed is critical. We propose to use the generalized score test. One big advantage of the score test in a GWAS setting is, one only needs to estimate the unknown parameters once under the null model without any SNP (i.e., $\beta_g = 0$), since the non-genetic risk factors are the same no matter which SNP is being tested. Therefore, the score test is faster as compared to the Wald and likelihood ratio tests. Moreover, the Wald or likelihood ratio test needs to estimate parameters under each alternative hypothesis (a total of 6 millions in our real data application), which may fail when the estimation procedure fails to converge.

With the sieve joint likelihood, we can obtain the restricted sieve maximum likelihood estimators under H_0 ($\beta_g = 0$ and the rest parameters are arbitrary), and then calculate the generalized score test statistic as defined in Cox and Hinkley [1979]. Similar to our estimation procedure, we also propose to use Richardson's extrapolation to numerically approximate the first and second order derivatives when calculating the score test statistic.

2.4 Simulation study

We first evaluated the parameter estimation of our proposed two-parameter copula sieve model for bivariate data under general interval censoring. Then we assessed the type-I error control, and power performance of the proposed generalized score test. We also evaluated the accuracy in estimating the joint survival probability using our proposed method. Finally, we evaluated the computing speed and convergence rate of our proposed method.

2.4.1 Generating bivariate interval-censored times

The data were generated from various Archimedean copula models (i.e., Clayton, Frank, Ali–Mikhail–Hap (AMH) and Joe) with Loglogistic margins. We first generated bivariate true event times T_{ij} using the conditioning approach described in Sun et al. [2019]. To obtain interval-censored data, we followed the censoring procedure in Kiani and Arasan [2012], which fits the study design of AREDS data. Explicitly, we assumed each subject was assessed for K times with the length between two adjacent assessment times following an Exponential distribution. In the end, for each subject i, L_{ij} was defined as the last assessment time before T_{ij} and R_{ij} was the first assessment time after T_{ij}. The overall right-censoring rate is set to be 25%. For the dependence strength between margins, we set Kendall's τ at 0.2 or 0.6, indicating weak or strong dependence. We assumed that the two event times share a common baseline distribution, for example, PO model with Loglogistic baseline hazards function (scale $\lambda = 1$ and shape $k = 2$) or PH model with Weibull baseline hazards function (scale $\lambda = 0.1$ and shape $k = 2$).

We included both genetic and non-genetic covariates in the simulations. Specifically, each SNP, coded as 0 or 1 or 2, was generated from a multinomial distribution with probabilities $\{(1-p)^2, 2p(1-p), p^2\}$, where $p = 40\%$ or 5% is the minor allele frequency (MAF). We also included a margin-specific continuous variable, generated from $N(6, 2^2)$, and a subject-specific binary variable, generated from Bernoulli ($p = 0.5$).

Under all scenarios, the sample size was set as $N = 500$. For simplicity, we assumed the same covariate effects for two margins, denoted as $(\beta_{ng1}, \beta_{ng2}, \beta_g)$, where β_{ng1} and β_{ng2} are marginal- and subject-specific non-genetic effects, respectively, and β_g is the SNP effect. We set $\beta_{ng1} = \beta_{ng2} = 0.1$. For estimation performance evaluation, we let $\beta_g = 0$ and replicated 1,000 times. For type-I error control evaluation of testing $\beta_g = 0$, we replicated 100,000 times and evaluated at various tail levels: 0.05, 0.01, 0.001 and 0.0001, respectively. For power evaluation, we replicated 1,000 times under each SNP effect size, where a range of β_g's were selected to represent weak to strong SNP effects.

2.4.2 Simulation-I: parameter estimation

In this section, we evaluated the estimation performance of our proposed method under both correct and misspecified settings. For the sieve margins, we used the true linear transformation function. We assumed the same Bernstein coefficients $\phi_{1k} = \phi_{2k}$ with degree $m_n = 3$ $(k = 0, 1, 2, 3)$ for Λ_j, $j = 1, 2$. For the event time range $[c, u]$, we chose $c = 0$ and set u as the largest value of all $\{L_{ij}, R_{ij}\}$ plus a constant.

In Table 2.4.1, the true model is Clayton copula with Loglogistic (proportional odds) or Weibull (proportional hazards) margins, with Kendall's $\tau = 0.6$. We fitted three models: the true parametric copula model (i.e., Clayton copula with Loglogistic or Weibull margins), a two-parameter copula model with sieve margins ("Copula2-S") and a marginal sieve model with the robust variance-covariance estimate ("Marginal-S") (a model also used in Zhou et al., 2017). We obtained estimation biases and 95% coverage probabilities for regression coefficients and dependence parameters. Under the two-parameter copula model, the sieve maximum likelihood estimators are all virtually unbiased, and all empirical coverage probabilities are close to the nominal level. Moreover, their standard errors are virtually the same as the standard errors under the true parametric model, indicating our proposed method works well. For the robust marginal sieve model, the regression coefficient estimates are also unbiased with correct coverage probabilities, but their standard errors are apparently larger.

In the real setting, the value of the transformation function parameter r is often unknown. Therefore, we examined our methods in estimating the transformation function parameter r together with the other parameters in our proposed model. The results are presented in the Table A.1.1 of Appendix A.1, which shows satisfactory estimation performance.

Table 2.4.1: Estimation results for bivariate interval-censored data.

	True			Copula2-S			Marginal-S		
Param	Bias	SE	SEE (CP)	Bias	SE	SEE (CP)	Bias	SE	SEE (CP)
				proportional odds					
β_{ng1}	0.0013	0.0171	0.0163 (0.942)	0.0003	0.0176	0.0165 (0.938)	0.0024	0.0293	0.0300 (0.930)
β_{ng2}	0.0120	0.1300	0.1300 (0.945)	0.0006	0.1330	0.1310 (0.939)	0.0110	0.1510	0.1500 (0.944)
β_g	-0.0007	0.0927	0.0942 (0.953)	-0.0110	0.0951	0.0947 (0.950)	0.0012	0.1050	0.1090 (0.955)
τ	-0.0005	0.0210	0.0208 (0.944)	-0.0045	0.0223	0.0221 (0.950)	NA	NA	NA
				proportional hazards					
β_{ng1}	0.0012	0.0097	0.0103 (0.958)	0.0013	0.0099	0.0105 (0.957)	0.0009	0.0182	0.0187 (0.957)
β_{ng2}	-0.0043	0.0780	0.0789 (0.952)	-0.0040	0.0782	0.0788 (0.951)	-0.0043	0.0960	0.0969 (0.957)
β_g	0.0005	0.0606	0.0569 (0.935)	0.0002	0.0608	0.0569 (0.938)	0.0003	0.0722	0.0701 (0.945)
τ	-0.0003	0.0220	0.0219 (0.952)	-0.0012	0.0224	0.0221 (0.951)	NA	NA	NA

Table 2.4.2: Estimation results using the proposed model when copula is misspecification.

	Frank			AMH			Joe		
Param	Bias	SE	SEE (CP)	Bias	SE	SEE (CP)	Bias	SE	SEE (CP)
β_{ng1}	0.0002	0.0177	0.0176 (0.950)	-0.0011	0.0262	0.0267 (0.953)	0.0016	0.0160	0.0166 (0.962)
β_{ng2}	0.0018	0.1480	0.1470 (0.944)	0.0013	0.1250	0.1250 (0.951)	-0.0027	0.1388	0.1438 (0.954)
β_g	0.0001	0.1050	0.1060 (0.952)	-0.0001	0.0885	0.0901 (0.959)	0.0037	0.0984	0.1043 (0.962)
τ	-0.0036	0.0219	0.0198 (0.937)	-0.0056	0.0318	0.0304 (0.934)	0.0168	0.0195	0.0185 (0.830)

We also examined how the proposed method works in the special case of bivariate current status data (by setting $K = 1$), which is shown in the Table A.1.2 of Appendix A.1. Our proposed method works as well as the true model in this setting too. The larger standard errors are due to less information in current status data as compared to the standard interval censoring case.

We further evaluated the estimation performance of the proposed model on bivariate interval-censored data generated from copula models that do not belong to the two-parameter copula family, such as Frank copula with $\tau = 0.6$, AMH copula with $\tau = 0.2$ (τ is always $< 1/3$ for AMH copula) and Joe copula with $\tau = 0.6$. In Table 2.4.2, the regression coefficient estimates from the two-parameter copula are all unbiased with coverage probabilities close to 95%. The biases for the τ estimates are also minimal with good coverage probabilities except in the scenario when data were generated from a Joe copula (coverage probability = 83%). Overall, the two-parameter copula model family demonstrates good robustness to misspecification in copula functions.

2.4.3 Simulation-II, generalized score test performance

We evaluated the type-I error control of our proposed generalized score test under Copula2-S. Specifically, we tested the SNP effect β_g under different dependence strengths

(Kendall's $\tau = 0.6,\ 0.2$) and two different MAFs (40%, 5%). The true model is Clayton copula with Loglogistic margins. We included score tests of two misspecified copula models, one with misspecified margins but correct copula (i.e., Clayton copula with Weibull margins) and the other with misspecified copula but correct margins (i.e., Gumbel copula with Loglogistic margins). We also included the score test under the correct parametric copula model (i.e., Clayton copula with Loglogistic margins), which served as the benchmark model. Besides, we examined Wald tests from the marginal Loglogistic model with variance-covariance either from the independence estimate or the robust sandwich estimate.

Table 2.4.3 shows type I errors under different tail levels. In the top part where Kendall's $\tau = 0.6$, our proposed score test controls type-I errors as well as the correct parametric model at all tail levels and MAFs. However, type-I errors in the two misspecified copula models are inflated at all scenarios, especially when margins are wrong at MAF = 40%. It is not surprising to observe the greatest inflation occurs with the marginal approach under the independence assumption. After applying the robust variance-covariance estimate, the type-I errors seem to be controlled at MAF = 40% but are still slightly inflated at MAF = 5%. When Kendall's $\tau = 0.2$, the proposed two-parameter model still performs as well as the correct parametric model and outperforms the other models, although the type-I error inflations from other models were smaller due to the weaker dependence.

We also compared the power performance between the score test under our Copula2-S model and score tests from two other models: the true parametric copula model and the Marginal-S model. Figure 2.4.1 presents the power curves of these three tests over a range of SNP effect sizes. Our proposed model yields the similar power performance as the true parametric model and is considerably more potent than the robust marginal sieve model.

2.4.4 Simulation-III: joint survival probability estimation performance

In addition, we evaluated the accuracy for estimating joint survival probabilities under our proposed Copula2-S model. We generated data from the Clayton copula with Weibull margins, and fitted the Clayton-Weibull ("Clayton-WB") and Copula2-S models and obtained the average estimated joint survival probabilities $Pr(T_1 > t, T_2 > t|Z_1, Z_2)$ on a

Table 2.4.3: Type-I error for the genetic effect β_g at various tail levels.

MAF	Tail	Indep-LL	Robust-LL	Clayton-W	Gumbel-LL	Copula2-S	Clayton-LL
				Kendall's $\tau = 0.6$			
40%	0.05	0.141	0.051	0.131	0.065	0.052	0.050
	0.01	0.053	0.010	0.041	0.015	0.010	0.010
	0.001	0.0131	0.0012	0.0074	0.0022	0.0013	0.0012
	0.0001	0.0037	0.0002	0.0012	0.0003	0.0001	0.0001
5%	0.05	0.141	0.056	0.059	0.066	0.053	0.051
	0.01	0.053	0.014	0.012	0.016	0.012	0.011
	0.001	0.0136	0.0018	0.0013	0.0020	0.0013	0.0012
	0.0001	0.0034	0.0003	0.0002	0.0003	0.0002	0.0002
				Kendall's $\tau = 0.2$			
40%	0.05	0.083	0.051	0.103	0.061	0.051	0.050
	0.01	0.022	0.010	0.029	0.013	0.010	0.010
	0.001	0.0036	0.0012	0.0045	0.0017	0.0011	0.0010
	0.0001	0.0006	0.0002	0.0006	0.0003	0.0002	0.0002
5%	0.05	0.083	0.056	0.054	0.060	0.053	0.052
	0.01	0.023	0.013	0.011	0.014	0.012	0.011
	0.001	0.0036	0.0017	0.0013	0.0018	0.0014	0.0013
	0.0001	0.0007	0.0003	0.0001	0.0002	0.0002	0.0001

sequence of pre-specified time points given covariate values. The number of replications is $1,000$. Figure 2.4.2 illustrates that Copula2-S produced an almost identical joint survival profile as "Clayton-WB". In addition, we quantified the estimation error between the estimated and true joint survival probabilities by the mean square errors (MSE) averaged over all time points and replications, which are 0.0004 (sd = 0.0012) and 0.0003 (sd = 0.0005) for Copula2-S and Clayton-WB, respectively.

2.4.5 Simulation-IV, convergence and computing speed

We examined the computational advantages of our proposed sieve estimation procedure as compared to the one-step estimation (directly maximizing the joint likelihood with arbitrary initial values). Data were simulated from a Clayton copula with Loglogistic margins. For $1,000$ replications, the one-step procedure took $1,260$ seconds while our proposed procedure took 925 seconds, saving about 27% computing time. For convergence rate, the proposed procedure failed in 0.1% out of $1,000$ replications, whereas the one-step procedure failed in 1.6%, which is 16 times of the proposed procedure.

We also compared the computing speed (using a 2.4GHz Intel Core i5 processor with 4GB memory) of three likelihood-based tests on testing $1,000$ SNPs under three models: the true Clayton model with Loglogistic margins, our proposed Copula2-S model and the Marginal-S model. The 1,000 genetic variants were simulated from MAF = 40%. The results are shown in Table 2.4.4. We found that the score test is about 3-5 times faster than the Wald test or the likelihood ratio test on average. Within the three score tests, although the score test under our Copula2-S model is the slowest due to model complexity, it is still faster than the Wald test under the Marginal-S model. Given its advantages in robustness, type-I error control, and power performance, we recommend the proposed Copula2-S model with the score test for large-scale testings of bivariate data subject to general censoring.

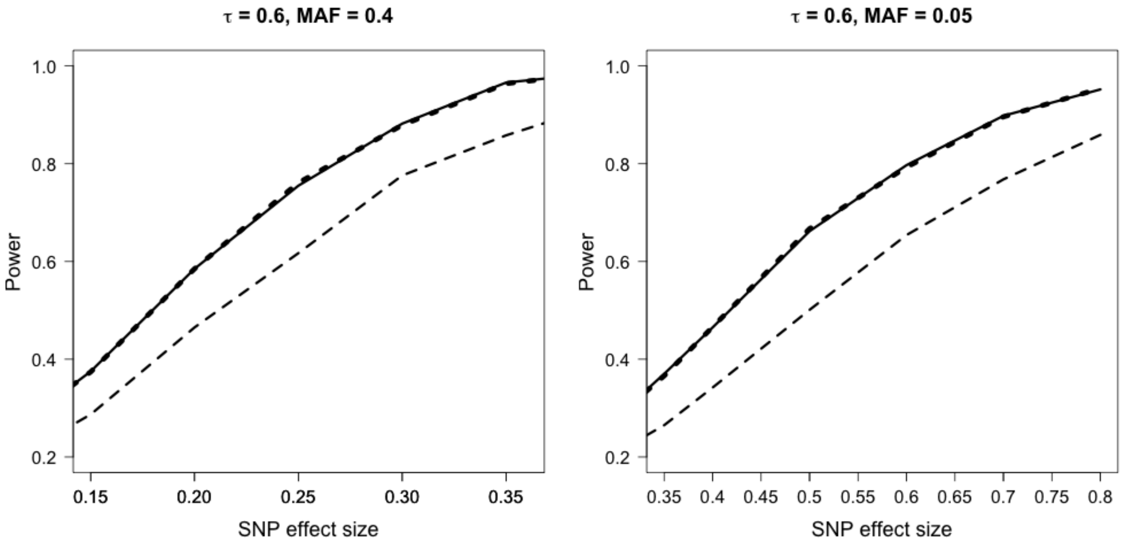

Figure 2.4.1: Simulation results for power performance of the score test.

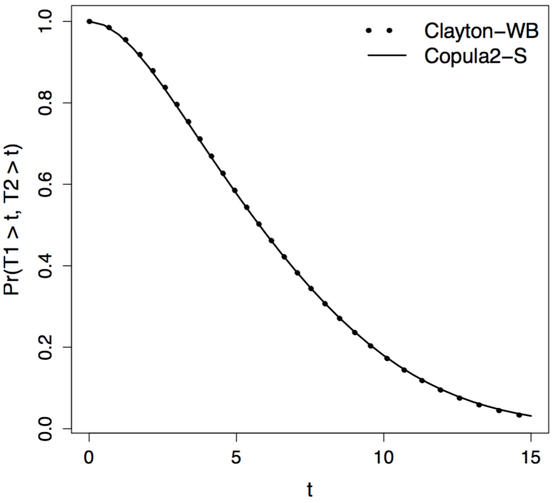

Figure 2.4.2: Estimated joint survival probabilities.

Table 2.4.4: Computing speed for testing 1,000 SNPs by score test.

Time (seconds)	Score	Wald	LRT
Marginal-S	148	693	NA
Clayton-LL	286	1254	1146
Copula2-S	475	1679	1570

2.5 Real data analysis

We implemented our proposed method to analyze the AREDS data. AREDS was designed to assess the clinical course of, and risk factors for the development and progression of AMD. DNA samples were collected from the consenting participants and genotyped by the International AMD Genomics Consortium [Fritsche et al., 2016]. In this study, each participant was examined every six months in the first six years and then once a year after year six. To measure disease progression, a severity score, scaled from one to twelve (with a larger value indicating more severe AMD), was determined for each eye of each participant at every examination. The outcome of interest is the bivariate progression time-to-late-AMD, where late-AMD is defined as the stage with severity score ≥ 9. Both phenotype and genotype data of AREDS are available from the online repository dbGap (accession: phs000001.v3.p1, and phs001039.v1.p1, respectively). By far, all the studies that have analyzed AREDS data for AMD progression treated the outcome as right-censored (e.g., Ding et al. [2017], Yan et al. [2018], and Sun et al. [2019]), and some only used data from the worst eye in each subject (e.g., Seddon et al. [2014]).

We analyzed 2718 Caucasian participants, including 2295 subjects who were free of late-AMD in both eyes at the enrollment, i.e., time 0 (bivariate data indicated as group A), and 423 subjects who had one eye already progressed to late-AMD by enrollment (univariate data indicated as group B). For the jth eye (free of late-AMD at time 0) of subject i, we observe L_{ij}, the last assessment time when the jth eye was still free of late-AMD and R_{ij}, the first assessment time when the jth eye was already diagnosed as late-AMD. For the eye that did not progress to late-AMD by the end of the study follow-up, we assigned a large number to R_{ij}. Since there are two groups of subjects (group A and B), we implemented a two-part model. Specifically, we created a covariate for each eye to indicate whether its fellow eye had already progressed or not at time 0 (i.e., 0 for both eyes of group A subjects and 1 for group B subjects). Then the joint likelihood is the product of the copula sieve model for group A subjects and the marginal sieve model for group B subjects. In addition, we performed a secondary sensitivity analysis using only group A subjects (i.e., subjects who were free of late-AMD in both eyes at time 0) and obtained similar top SNPs as from the two-part model. The secondary analysis results are presented in Table A.2.1 of Appendix A.2.

We included four risk factors as non-genetic covariates, including the baseline age, severity score, smoking status, and fellow-eye progression status. We checked various transformation functions and Bernstein polynomial degrees m_n, and chose the model that produced the smallest AIC, which is the proportional odds model with $m_n = 4$ for both margins.

We performed GWAS on 6 million SNPs (either from exome chip or imputed) with MAF $> 5\%$ across the 22 autosomal chromosomes and plotted their $-\log(p)$ in Figure 2.5.1. As highlighted in the figure, the *PLEKHA1–ARMS2–HTRA1* region on chromosome 10 and the *CFH* region on chromosome 1 have variants reaching the "genome-wide" significance level ($p < 5 \times 10^{-8}$). Previously, these two regions were found being significantly associated with AMD onset from multiple case-control studies [Fritsche et al., 2016]. Moreover, we identified SNPs in another region *ATF7IP2* on chromosome 16, showing moderate to strong association with AMD progression ($5 \times 10^{-8} < p < 1 \times 10^{-5}$). As a comparison, we also fitted the robust marginal sieve model (Marginal-S) and the Gamma frailty sieve model (Frailty-S) [Zhou et al., 2017], and performed the corresponding score tests for each SNP. Overall, their results are consistent with our Copula2-S model, but the p-values are generally larger (as shown in Table 2.5.1). Note that the *CFH* region did not reach the "genome-wide" significance level under the Marginal-S model.

Table 2.5.1 presents the top significant variants of the three regions denoted in Figure 2.5.1. Besides Copula2-S, we also present score test p-values from Frailty-S and Marginal-S. The odds ratio of an SNP was calculated by fitting a Copula2-S model including this SNP and those non-genetic factors. For example, $rs2284665$, a known AMD risk variant from *HTRA1* region, has an estimated odds ratio of 1.66 (95% CI = $[1.46, 1.89]$), which implies its minor allele has a "harmful" effect on AMD progression. Under this model, the estimated dependence parameters are $\hat{\alpha} = 0.90$ and $\hat{\kappa} = 1.00$, corresponding to $\hat{\tau} = 0.40$, which indicates moderate dependence in AMD progression between two eyes.

For variant $rs2284665$, we obtained both estimated joint and conditional survival functions from the fitted Copula2-S model. The left panel of Figure 2.5.2 plots the joint progression-free probability contours for subjects who are smokers with the same age (= 68.6) and AMD severity score (= 3.0 for both eyes) but different genotypes of $rs2284665$. The right panel of Figure 2.5.2 plots the corresponding conditional progression-free probability of remaining years (after year 5) for one eye, given its fellow eye has progressed by year 5. It is clearly seen that in both plots, the three genotype groups look well separated, with the GG group having the largest progression-free probabilities. These estimated progression-free probabilities provide valuable information to characterize or predict the progression profiles for AMD patients with different characteristics.

2.6 Conclusion and discussion

We proposed a flexible copula-based semiparametric transformation model for analyzing and testing bivariate (general) interval-censored data. Unlike the approach proposed by Hu et al. [2017], which approximated the copula function by sieves, our approach kept the copula function in its parametric form but flexibly modeled the margins through semiparametric transformation models. In this way, our method guaranteed to produce consistent estimates for both regression and copula parameters, which then led to reliable joint distribution estimates. On the other hand, Hu et al. [2017] focused on estimating regression parameters only but with possible biased estimates for the copula function. Our proposed method

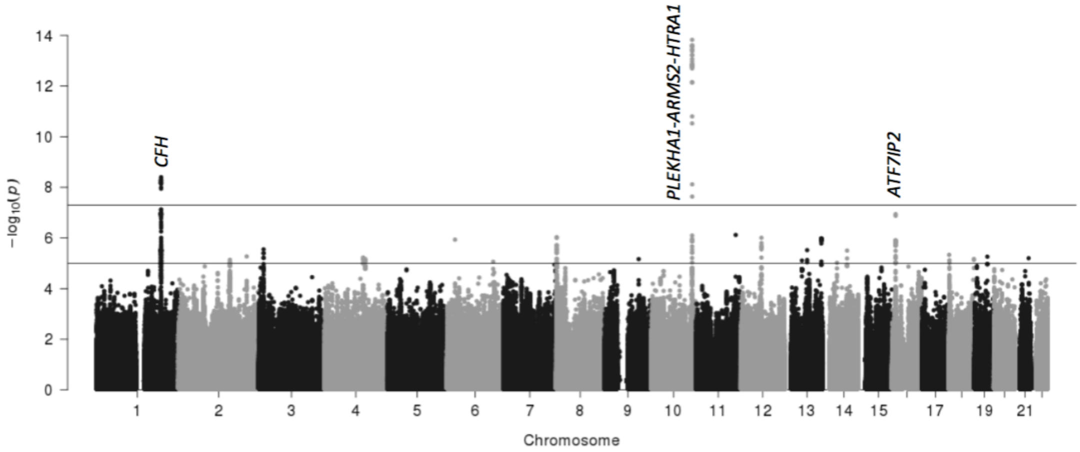

Figure 2.5.1: Manhattan plot for GWAS results of AMD progression.

Table 2.5.1: The top SNPs identified to be associated with AMD progression.

SNP	Chr	Gene	MAF	OR	p (Copula2-S)	p (Frailty-S)	p (Marginal-S)
*rs*2284665	10	*HTRA1*	0.33	1.66	1.5×10^{-14}	2.7×10^{-12}	1.6×10^{-10}
*rs*2293870	10	*ARMS2-HTRA1*	0.33	1.65	2.5×10^{-14}	2.5×10^{-12}	2.4×10^{-10}
*rs*3750846	10	*ARMS2-HTRA1*	0.34	1.62	1.6×10^{-13}	8.5×10^{-12}	8.7×10^{-10}
*rs*58649964	10	*PLEKHA1*	0.24	1.63	3.0×10^{-11}	1.0×10^{-9}	2.0×10^{-8}
*rs*10922109	1	*CFH*	0.28	0.64	4.0×10^{-9}	7.4×10^{-9}	7.4×10^{-8}
*rs*1329427	1	*CFH*	0.28	0.64	4.4×10^{-9}	8.3×10^{-9}	8.1×10^{-8}
*rs*10801559	1	*CFH*	0.28	0.64	4.8×10^{-9}	9.3×10^{-9}	8.8×10^{-8}
*rs*1410996	1	*CFH*	0.28	0.64	5.3×10^{-9}	1.1×10^{-8}	1.0×10^{-7}
*rs*12708701	16	*ATF7IP2*	0.13	1.62	1.1×10^{-7}	2.5×10^{-7}	7.0×10^{-7}
*rs*28368872	16	*ATF7IP2*	0.13	1.62	1.3×10^{-7}	4.3×10^{-7}	8.7×10^{-7}

has the great advantage in computation and it is applicable to analyze large data sets and to perform a large number of tests. All the methods have been built into an R package {CopulaCenR} [Sun and Ding, 2020a,b], which includes a variety of copula functions (e.g., Copula2, Clayton, Gumbel, Frank, Joe, AMH) and is available on CRAN at https://cran.r-project.org/package=CopulaCenR.

Several model selection procedures have been proposed for copula-based methods. For example, the AIC is widely used for model selection purpose in copula models. Wang and Wells [2000] proposed a model selection procedure based on the nonparametric estimation of the bivariate joint survival function within Archimedean copulas. For model diagnostics, Chen et al. [2010] proposed a penalized pseudo-likelihood ratio test for copula models in complete data. Recently, Zhang et al. [2016] developed a goodness-of-fit test for copula models using the pseudo in-and-out-of-sample method.

To the best of our knowledge, there is no existing goodness-of-fit test for copula models of bivariate interval-censored data. In our real data analysis, we used AIC to guide the model selection for simplicity. However, a formal test for goodness-of-fit is desirable, especially for bivariate interval-censored data under the regression setting. It is worthwhile to investigate it as a future research direction.

We applied our method to a GWAS of AMD progression and successfully identified variants from two known AMD risk regions (*CFH* on chromosome 1 and *PLEKHA1–ARMS2–HTRA1* on chromosome 10) being significantly associated with AMD progression. Moreover, we also discovered variants from a region (*ATF7IP2* on chromosome 16), which has not been reported before, showing moderate to strong association with AMD progression. On the contrary, we found that some known AMD risk loci (e.g., $rs12357257$ from *ARHGAP21* on chromosome 10, $p = 0.12$) are not associated with AMD progression. Therefore, the variants associated with risks of having AMD may not be necessarily associated with the disease progression; while some variants may be only associated with AMD progression but not with the disease onset. Our work is the first research on AMD progression which adopts a solid statistical model that appropriately handles bivariate interval-censored data. Our findings provided new insights into the genetic causes on AMD progression, which are critical for establishing novel and reliable predictive models of AMD progression to identify high-risk patients at an early stage accurately. Our proposed method applies to general bilateral diseases and complex diseases with co-primary endpoints.

2.7 Software development

2.7.1 Existing R packages implementing copula-based models

To the best of our knowledge, there exists no R package for fitting copula-based regression models for both bivariate right-censored and interval-censored data. The existing copula packages for bivariate data handle either the non-censoring (i.e., complete data) or the right-censoring situation. In the non-censoring situation, the package copula [Hofert et al., 2018] by Yan [2007] and Kojadinovic and Yan [2010] implements multivariate copula models without covariates for complete data and obtains the maximum likelihood estimator for the copula dependence parameter. It gives useful codes for implementing regression models in bivariate complete data in the appendix of Yan [2007]. It also provides copula goodness-of-fit tests for model selection purpose. The package VineCopula [Schepsmeier et al., 2018] can also model bivariate or multivariate complete data without covariates through the vine copula models [Aas et al., 2009]. Packages such as CopulaRegression [Nicole Kraemer, 2014] and gcmr [Masarotto and Varin, 2017] can provide copula-based regression models with parametric margins for bivariate or multivariate complete data and provide maximum likelihood estimators for model parameters. The package gamCopula [Nagler and Vatter, 2020] implements a generalized additive model that can take into account the effect of the predictors on the dependence structure of bivariate and vine copula models [Vatter and Chavez-Demoulin, 2015]. For the right-censoring situation, the Copula.surv package [Emura, 2018] can estimate the Clayton copula dependence parameter in bivariate right-censored data without covariates and also perform a goodness-of-fit test for a fitted Clayton model [Emura et al., 2010]. The Sunclarco package [Prenen et al., 2017b] provides Clayton or Gumbel copula-based regression models with parametric (Weibull and piecewise constant) or Cox semiparametric margins for multivariate right-censored data [Prenen et al., 2017a]. The package GJRM

[Marra and Radice, 2020] can fit both marginal and copula regression models in complete and right-censored data [Marra and Radice, 2017, 2019, Marra et al., 2017]. By far, there is no copula-based R package for bivariate interval-censored data.

2.7.2 CopulaCenR package

We develop the CopulaCenR package [Sun and Ding, 2020a,b], which fits copula-based regression models for both bivariate right-censored and interval-censored data. The package is available from the Comprehensive R Archive Network (CRAN) at `https://CRAN.R-project.org/package=CopulaCenR`.

The main advantage of CopulaCenR relies on the diverse choice of copula and marginal models for both bivariate right-censored and interval-censored data. Specifically, it provides a class of Archimedean copulas that correspond to a variety of dependence structures, as illustrated in Table 2.7.1. In particular, in addition to these frequently used one-parameter Archimedean copulas, a two-parameter copula function (Copula2) is also included. Furthermore, CopulaCenR implements a list of parametric and semiparametric marginal regression models, as illustrated in Table 2.7.2. For parameter estimation, the package utilizes a novel two-step procedure that is computationally stable and efficient. For the inference of regression parameters, three likelihood-based tests such as Wald, generalized score and likelihood ratio tests are provided.

To fit a copula-based regression model, one also needs to choose a regression model for the margins. Table 2.7.2 lists the available marginal models in CopulaCenR. For bivariate right-censored data, users can fit either a parametric marginal model via the function rc_par_copula or a semiparametric Cox PH model via the function rc_spCox_copula [Sun et al., 2019]. Specifically, the parametric models incorporate both the PH (e.g., Weibull, Gompertz) and the PO (e.g., Loglogistic) models. For bivariate interval-censored data, one can choose to fit a parametric marginal model via the function ic_par_copula. Moreover, the package can also fit a semiparametric transformation model via the function ic_spTran_copula [Sun and Ding, 2019].

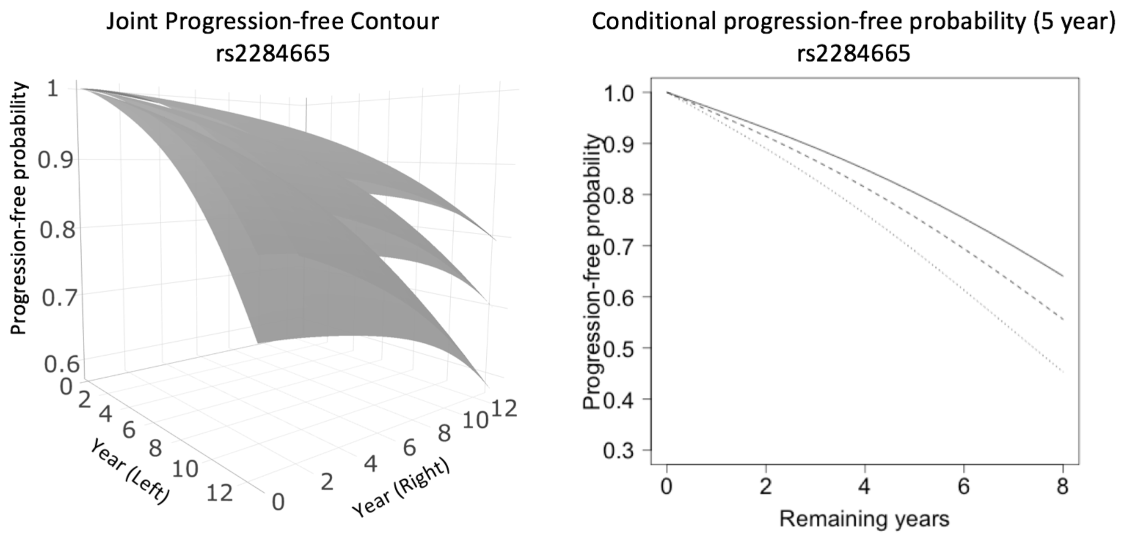

Figure 2.5.2: Estimated progression-free probabilities for subjects with different genotypes of $rs2284665$.

Table 2.7.1: Summary of implemented Archimedean copula families.

Family	Parameter Space	Generator $\phi_\eta(t), t \in [0, \infty)$	Generator Inverse $\phi_\eta^{-1}(s), s \in (0, 1]$	τ_L	τ_U	Kendall's τ
Clayton	$\eta > 0$	$(1+t)^{-1/\eta}$	$s^{-\eta} - 1$	$2^{-1/\eta}$	0	$\eta/(2+\eta)$
Gumbel	$\eta \geq 1$	$\exp(-t^{1/\eta})$	$(-\log s)^\eta$	0	$2 - 2^{1/\eta}$	$1 - 1/\eta$
Frank	$\eta \geq 0$	$-\eta^{-1} \log\{1 + e^{-t}(e^{-\eta} - 1)\}$	$-\log\{(e^{-\eta s} - 1)/(e^{-\eta} - 1)\}$	0	0	$1 + 4\{D_1(\eta) - 1\}/\eta$
AMH	$\eta \in [0, 1)$	$(1-\eta)/(e^t - \eta)$	$\log[\{1 + \eta(s-1)\}/s]$	0	0	$1 - 2\{(1-\eta)^2 \log(1-\eta) + \eta\}/(3\eta^2)$
Joe	$\eta \geq 1$	$1 - (1 - e^{-t})^{1/\eta}$	$-\log\{1 - (1-s)^\eta\}$	0	$2 - 2^{1/\eta}$	$1 - 4\sum_{k=1}^{\infty} 1/\{k(\eta k + 2)[\eta(k-1) + 2]\}$
Copula2	$\alpha \in (0, 1], \kappa > 0$	$\{1/(1+t^\alpha)\}^\kappa$	$(s^{-1/\kappa} - 1)^{1/\alpha}$	$2^{-\alpha\kappa}$	$2 - 2^\alpha$	$1 - 2\alpha\kappa/(2\kappa + 1)$

τ_L and τ_U are the lower and upper tail dependence measures.
$D_1(\cdot)$ is the Debye function written as $D_1(\eta) = \frac{1}{\eta} \int_0^\eta \frac{t}{e^t - 1} dt$.

Table 2.7.2: Summary of implemented marginal models.

Type	Models	Survival Distributions $S(t)$	Corresponding R Functions
	Weibull	$\exp\{-(t/\lambda)^k e^{Z^\top \beta}\}$	
Parametric	Gompertz	$\exp\{-\frac{b}{a}(e^{at}-1)e^{Z^\top \beta}\}$	rc_par_copula, ic_par_copula
	Loglogistic	$\{1+(t/\lambda)^k e^{Z^\top \beta}\}^{-1}$	
Semiparametric	Cox	$\exp\{-\Lambda(t)e^{Z^\top \beta}\}$	rc_spCox_copula
	Transformation	$\exp[-G\{\Lambda(t)e^{Z^\top \beta}\}]$	ic_spTran_copula

After a copula model has been fitted, fitted values (i.e., linear predictors, survival probabilities) can be extracted by the general S3 function *fitted*. For new observations, the linear predictors and survival probabilities can be obtained using the function *predict*. Moreover, the user can plot three types of distributions (joint, conditional and marginal) using the general functions plot and lines. In particular, an interactive 3D contour will be plotted to visualize the joint distribution. Besides, the package provides a bivariate event time gener-ating function data sim copula, which can generate random bivariate event times based on a specified copula function, a marginal distribution, and covariate values. More illustration examples can be found in our paper [Sun and Ding, 2020b].

3.0 An Information Ratio based Goodness-of-fit Test for Copula Models on Censored Data

3.1 Motivation

Although our proposed two-parameter copula semiparametric model can flexibly account for dependency at two tails, a formal goodness-of-fit test for copula-based survival models is still highly desired.

As summarized in Section 1.7, there are different approaches for testing copula-based survival models under right-censoring. Shih [1998] and Emura et al. [2010] developed tests for one-parameter Archimedean copulas based on the discrepancy between the unweighted and weighted copula dependence estimators. Their tests depend on modeling orderable pairs of the bivariate event times, which is difficult for multivariate interval-censored data where no exact event times are observed. Wang and Wells [2000] proposed a model selection method in Archimedean copulas based on a truncated Kendall process introduced by Genest and Rivest [1993]. Lakhal-Chaieb [2010] further extended Wang and Wells [2000] by developing an inverse probability censoring weighted estimator for Kendall's distribution. The Andersen test statistic [Andersen et al., 2005] was built on comparing the parametrically estimated bivariate joint distribution with the non-parametric bivariate empirical copula. Chen et al. [2010] proposed a penalized pseudo-likelihood ratio statistic for examining whether the assumed copula fits data better than a group of other copula models. Wang [2010] proposed a Fisher's Z test statistic based on the correlation between two random quantities that are shown to be independent under Archimedean copulas by [Genest and Rivest, 1993]. More recently, Mei [2016] proposed a likelihood-based pseudo-in-and-out-of-sample (PIOS) test, similar to Zhang et al. [2016] for the complete data. Very recently, Lin and Wu [2020] developed a smooth test for copula specification in modeling right-censored data.

To the best of our knowledge, there is no formal statistical test for copula specification under interval censoring. In this chapter, we develop a novel information ratio (IR)-based goodness-of-fit test for diagnosing copula-based survival models under interval- or right-censoring. The test procedure applies to any copula function with a parametric form, including Archimedean (i.e., Clayton, Gumbel, Frank) and non-Archimedean (e.g., Gaussian, Plackett) copula families. For ease of notations, we will illustrate our method in the bivariate data case, and the generalization to multivariate cases is relatively straightforward.

3.2 Methods

3.2.1 Copula model for bivariate censored data

Let (T_{i1}, T_{i2}) be the true bivariate event times for subject i, with marginal survival functions $S_j(t_{ij}) = Pr(T_{ij} > t_{ij}),\ j = 1, 2$, and joint survival function $S(t_{i1}, t_{i2}) = Pr(T_{i1} > t_{i1}, T_{i2} > t_{i2})$. Assume there are n independent subjects in a study. When (T_{i1}, T_{i2}) are under interval-censoring, we observe $D_i = \{(L_{ij}, R_{ij}), j = 1, 2\}$ for subject i, where $(L_{ij}, R_{ij}]$ is the time interval that T_{ij} lies in. When $R_{ij} = \infty$, T_{ij} is under right-censoring. When (T_{i1}, T_{i2}) are subject to right-censoring, for subject $i = 1 \cdots n$, we observe $D_i = \{(Y_{ij}, \Delta_{ij}) : Y_{ij} = \min(T_{ij}, C_{ij}), \Delta_{ij} = I(T_{ij} \leq C_{ij}), j = 1, 2\}$, where C_{ij} is the censoring time of T_{ij}, Δ_{ij} is the censoring indicator.

By the Sklar's theorem (Sklar, 1959), so long as the marginal survival functions S_j are continuous, there exists a unique function C_η that connects two marginal survival functions into the joint survival function: $S(t_1, t_2) = C_\eta\{S_1(t_1), S_2(t_2)\},\ t_1, t_2 \geq 0$. Here, the function C_η is called a copula, and its parameter η measures the dependence between the two margins. A signature feature of the copula is that it allows the dependence to be modeled separately from the marginal distributions.

3.2.2 Joint likelihood functions for bivariate censored data

In this section, we present the joint likelihood functions for bivariate interval-censored data and bivariate right-censored data under the copula framework by using the notations introduced in Section 3.2.1, respectively.

When data are bivariate interval-censored, denote $\{S_1(L_{i1}), S_1(R_{i1})\}$ by S_{i1} and similarly denote $\{S_2(L_{i2}), S_2(R_{i2})\}$ by S_{i2}. Then the joint likelihood function for n independent subjects can be written as

$$\begin{aligned} L_n(S_1, S_2; \eta) &= \prod_{i=1}^{n} Pr(L_{i1} < T_{i1} \le R_{i1}, L_{i2} < T_{i2} \le R_{i2}) \\ &= \prod_{i=1}^{n} \Big[Pr(T_{i1} > L_{i1}, T_{i2} > L_{i2}) - Pr(T_{i1} > L_{i1}, T_{i2} > R_{i2}) \\ &\qquad - Pr(T_{i1} > R_{i1}, T_{i2} > L_{i2}) + Pr(T_{i1} > R_{i1}, T_{i2} > R_{i2}) \Big] \\ &= \prod_{i=1}^{n} \Big[C_\eta\{S_1(L_{i1}), S_2(L_{i2})\} - C_\eta\{S_1(L_{i1}), S_2(R_{i2})\} \\ &\qquad - C_\eta\{S_1(R_{i1}), S_2(L_{i2})\} + C_\eta\{S_1(R_{i1}), S_2(R_{i2})\} \Big]. \qquad (3.2.1) \end{aligned}$$

The right interval R_{ij} can take values in $(0, \infty]$. For a subject i, if $R_{ij} = \infty$ (i.e., T_{ij} is right-censored), then any term involving R_{ij} becomes 0, and the joint survival function for subject i reduces to only one (if both R_{i1} and R_{i2} are ∞) or two (if one R_{ij} is ∞) terms. The special case of bivariate current status data (i.e., only one assessment time for each subject) can also fit into this framework, where for each T_{ij}, either $L_{ij} = 0$ (T_{ij} is smaller than the assessment time, which is R_{ij} in this case) or $R_{ij} = \infty$ (T_{ij} is greater than the assessment time, which is L_{ij} in this case). Therefore, the likelihood function (3.2.1) can handle the bivariate data under general interval-censoring.

For bivariate right-censored data, denote $\{S_1(Y_{i1}), S_2(Y_{i2})\}$ by (S_{i1}, S_{i2}). Let the density function for copula $C_\eta(u, v)$ be $c_\eta(u, v) = \partial^2 C_\eta(u, v)/\partial u \partial v$. Then, the joint likelihood under right censoring can be written as

$$\begin{aligned} L_n(S_1, S_2; \eta) = &\prod_{i=1}^{n} c_\eta(S_{i1}, S_{i2})^{\delta_{i1}\delta_{i2}} \left[\frac{\partial\, C_\eta(S_{i1}, S_{i2})}{\partial S_{i1}} \right]^{\delta_{i1}(1-\delta_{i2})} \\ &\times \left[\frac{\partial\, C_\eta(S_{i1}, S_{i2})}{\partial S_{i2}} \right]^{(1-\delta_{i1})\delta_{i2}} C_\eta(S_{i1}, S_{i2})^{(1-\delta_{i1})(1-\delta_{i2})}, \end{aligned} \qquad (3.2.2)$$

where $(\delta_{i1}, \delta_{i2}) \in \{(0,0), (0,1), (1,0), (1,1)\}$.

3.2.3 Test hypothesis and Information Ratio (IR) statistic

Suppose C_0 is the true and unknown copula model, and we fit a copula model $C_\eta(\cdot)$. We are interested in testing $\mathbf{H}_0$: $C_0 \in \mathbb{C} = \{C_\eta(\cdot) : \eta \in \Theta\}$ versus $\mathbf{H}_A$: $C_0 \notin \mathbb{C}$, where Θ is the parameter space of the copula parameter η with $p = \dim(\Theta)$.

Let $l(S_1, S_2; \eta) = \log L(S_1, S_2; \eta)$ be the copula-based survival model log-likelihood, where S_1 and S_2 are the two marginal survival functions. Under the null hypothesis, we define two types of Fisher information matrices [Song and Song, 2007]: the negative sensitivity matrix as

$$S(\eta) = -P_0\{\ddot{l}_{\eta\eta}(S_1, S_2; \eta)\},$$

and the variability matrix as

$$V(\eta) = P_0\{\dot{l}_\eta(S_1, S_2; \eta)\dot{l}_\eta^T(S_1, S_2; \eta)\},$$

where $\dot{l}_\eta(S_1, S_2; \eta) = \frac{\partial}{\partial\eta} l(S_1, S_2; \eta)$, $\ddot{l}_{\eta\eta}(S_1, S_2; \eta) = \frac{\partial^2}{\partial\eta\partial\eta^T} l(S_1, S_2; \eta)$, and $P_0(\cdot)$ represents the probability measure under the true copula C_0.

Then, we define the Information Ratio (IR) statistic as follows

$$\begin{aligned} IR =&: P_0\{\dot{l}_\eta^T(S_1, S_2; \eta^*)S^{-1}(\eta^*)\dot{l}_\eta(S_1, S_2; \eta^*)\} \\ &= tr\{S^{-1}(\eta^*)V(\eta^*)\} \\ &= p, \end{aligned}$$

where $\eta^* \in \Theta$ is the true η under the null hypothesis, and $tr(A)$ denotes the trace of a matrix A. The last equation holds because $S(\eta^*) = V(\eta^*)$ under the null hypothesis that copula is correctly specified [White, 1982], and p is the trace of the p-dimensional identity matrix. The key idea behind the information ratio-based goodness-of-fit test is that the ratio between the two types of Fisher information equals the dimension of the copula parameter if the null hypothesis is true.

3.2.4 IR statistic estimation

We propose an IR statistic estimator denoted as $\widehat{IR}_n$ under the null hypothesis of the assumed copula model being the correct model, which can be written as

$$\begin{aligned}\widehat{IR}_n =&: \frac{1}{n}\sum_{i=1}^{n} \dot{l}_\eta^T(\tilde{S}_{i1}, \tilde{S}_{i2}; \hat{\eta})\tilde{S}^{-1}(\hat{\eta})\dot{l}_\eta(\tilde{S}_{i1}, \tilde{S}_{i2}; \hat{\eta}) \\ &= tr\{\tilde{S}^{-1}(\hat{\eta})\tilde{V}(\hat{\eta})\},\end{aligned}$$

where $\tilde{S}(\hat{\eta}) = -\frac{1}{n}\sum_{i=1}^{n} \ddot{l}_{\eta\eta}(\tilde{S}_{i1}, \tilde{S}_{i2}; \hat{\eta})$ and $\tilde{V}(\hat{\eta}) = \frac{1}{n}\sum_{i=1}^{n} \dot{l}_\eta(\tilde{S}_{i1}, \tilde{S}_{i2}; \hat{\eta})\dot{l}_\eta^T(\tilde{S}_{i1}, \tilde{S}_{i2}; \hat{\eta})$ are consistent estimators for $S(\eta)$ and $V(\eta)$. $\dot{l}_\eta$ and $\ddot{l}_{\eta\eta}$ are the first- and second-order derivatives of the log-likelihood function. This IR statistic is similar to the information ratio test statistic proposed by Zhou et al. [2012] for cross-sectional and longitudinal data, which was later extended to test copula specification in uncensored complete data in Zhang et al. [2016].

Due to complex structures of the log-likelihood function, obtaining the analytical forms for $\dot{l}_\eta$ and $\ddot{l}_{\eta\eta}$ for various types of copula models is tedious. We propose to use the Richardson's extrapolation method [Lindfield and Penny, 1989] to approximate these first- and second-order derivatives for all copula models. $\tilde{S}_{i1}$ and $\tilde{S}_{i2}$ are the consistent estimators for the marginal survival probability of subject i, which can be obtained from the non-parametric distribution estimators in the absence of covariates (e.g., Kaplan-Meier estimator [Kaplan and Meier, 1958] under right censoring or Turnbull estimator [Turnbull, 1976] under interval censoring) or fitting a marginal regression model in the presence of covariates. For this work, we use non-parametric estimators. In addition to $\tilde{S}_{i1}$ and $\tilde{S}_{i2}$, $\hat{\eta}$ is a consistent estimator of η under the null hypothesis such that $\hat{\eta} \rightarrow_p \eta^*$, and it can be obtained by following the two-stage pseudo-likelihood estimation procedures in Shih and Louis [1995] and Sun et al. [2006] under the right- and interval-censoring, respectively. The asymptotic consistency of the resulting $\hat{\eta}$ has been proved by Shih and Louis [1995] and Sun et al. [2006] when the null hypothesis copula function is correctly specified. One big advantage of our proposed IR estimator is its computational simplicity and flexibility in handling any copula with a parametric form under various types of censoring.

3.2.5 Asymptotic properties

Under the regularity conditions presented in Appendix B, we have the following asymptotic properties for $\widehat{IR}_n$.

Theorem 3.2.1. *(Consistency) Under Conditions 1 and 2, we have*

$$\widehat{IR}_n \xrightarrow{p} tr\{S^{-1}(\eta^*)V(\eta^*)\} = p,$$

where η^* is the limiting value of $\hat{\eta}$, p is the dimension of the assumed copula function parameter.

Theorem 3.2.2. *(Normality under null) Under the null hypothesis and Conditions 2-4, we have*

$$n^{1/2}(\widehat{IR}_n - p) \xrightarrow{d} N(0, \sigma^2_{IR}),$$

where σ^2_{IR} is the asymptotic variance of $n^{1/2}(\widehat{IR}_n - p)$.

Remark. *These theorems are the theoretical bases for our proposed goodness-of-fit test procedures. To implement our test, σ^2_{IR} needs to be estimated. Since the analytical form of σ^2_{IR} is intractable, we propose a parametric bootstrap procedure to estimate σ^2_{IR}.*

3.2.6 IR-based goodness-of-fit test procedures

We construct the IR-based goodness-of-fit procedure for copula specification based on the asymptotic properties of $\widehat{IR}_n$. Due to the absence of the analytical form of σ^2_{IR}, we propose a novel parametric bootstrap procedure for estimating σ^2_{IR}. In brief, we first fit the null hypothesis copula model, calculate $\widehat{IR}_n$ from the original data, then perform parametric bootstrap sampling based on the fitted model, and finally calculate the asymptotic or empirical p-value of the IR-based goodness-of-fit test. We also implement parallel computing into our bootstrap procedures, which significantly enhances computational efficiency, as illustrated in our simulations. We propose two separate bootstrap procedures to handle bivariate interval- or right-censored data, respectively.

Scenario I: bivariate interval-censored data

Step 1: Obtain the non-parametric Turnbull estimators $\tilde{S}_{i1}, \tilde{S}_{i2}$;

Step 2: Obtain the pseudo maximum likelihood estimator $\hat{\eta}$ based on the estimation procedure in Sun et al. [2006], and calculate $\widehat{IR}_n$ based on $\hat{\eta}, \tilde{S}_{i1}, \tilde{S}_{i2}$;

Step 3: First, generate the b^{th} bootstrap bivariate true event times $\{t^b_{i1}, t^b_{i2}, i = 1, \ldots, n\}$ via parametric bootstraps by following the similar bivariate data generation procedure as in Sun et al. [2019]. Specifically, we obtain u^b_i and w^b_i from two independent $U[0, 1]$ uniform distributions. Let $w^b_i = h_{\hat{\eta}}(u^b_i, v^b_i) = \partial C_{\hat{\eta}}(u^b_i, v^b_i)/\partial u^b_i$ and solve for v^b_i from the inverse of $h_{\hat{\eta}}$ function $h^{-1}_{\hat{\eta}}$. We further obtain $\{t^b_{i1}, t^b_{i2}\}$ by applying $t^b_{i1} = \tilde{S}^{-1}_1(u^b_i) = \min\{t : \tilde{S}_1(t) \leq u^b_i\}$ and $t^b_{i2} = \tilde{S}^{-1}_2(v^b_i) = \min\{t : \tilde{S}_2(t) \leq v^b_i\}$. Next, we estimate the length between intermittent assessment times using the average interval length of the original data, and adjust the number of assessments K^b so that the right censoring rate in the b^{th} bootstrap is closest to the right censoring rate in the original data. Eventually, the observed times for the b^{th} bootstrap are $\{L^b_{i1}, R^b_{i1}, L^b_{i2}, R^b_{i2}, i = 1, \ldots, n\}$, where $(L^b_{ij}, R^b_{ij}]$ is the smallest interval that bounds $t^b_{ij}, j = 1, 2$;

Step 5: Repeat Steps 3-4 for $B = 200$ times and obtain a set of test statistics $\{\widehat{IR}^b_n, b = 1, \ldots, B\}$;

Step 6: Compute the empirical p-value as $p_e = \frac{1}{B}\sum_{b=1}^{B} I(|\widehat{IR}^b_n| \geq |\widehat{IR}_n|)$ or the asymptotic p-value as $p_a = 2\{1 - \phi(\frac{|\widehat{IR}_n - p|}{\hat{\sigma}_{IR}})\}$, where $\hat{\sigma}_{IR}$ is calculated using the bootstrap IR statistics $\{\widehat{IR}^b_n\}$.

Scenario II: bivariate right-censored data

Step 1: Calculate the Kaplan-Meier estimators $\tilde{S}_{i1}, \tilde{S}_{i2}$.

Step 2: Obtain the pseudo maximum likelihood estimator $\hat{\eta}$ based on the estimation procedure in Shih and Louis [1995], and calculate $\widehat{IR}_n$ based on $\hat{\eta}, \tilde{S}_{i1}, \tilde{S}_{i2}$;

Step 3: Use the same procedure as the Step 3 in Scenario I to generate the b^{th} bootstrap true event times $\{t^b_{i1}, t^b_{i2}, i = 1, \ldots, n\}$ based on the $\hat{\eta}, \tilde{S}_{i1}, \tilde{S}_{i2}$. Next, we obtain the censoring times c^b_{i1} and c^b_{i2} by sampling from the non-parametrically estimated censoring distributions. Finally, we calculate the observed times and censoring status for the b^{th} bootstrap $\{y^b_{i1}, y^b_{i2}, \delta^b_{i1}, \delta^b_{i2}, i = 1, \ldots, n\}$, where $y^b_{ij} = \min(t^b_{ij}, c^b_{ij})$ and $\delta^b_{ij} = I(t^b_{ij} \leq c^b_{ij})$, $j = 1, 2$;

Step 4: Based on $\{y^b_{i1}, y^b_{i2}, \delta^b_{i1}, \delta^b_{i2}, i = 1, \ldots, n\}$ from Step 3, obtain the copula parameter estimator η^b under the null hypothesis and calculate the b^{th} bootstrap IR test statistic $\widehat{IR}^b_n$; Steps 4-6 are the same as in Scenario I.

In practice, we find the empirical and asymptotic p-values are similar in most cases, and we will report the results from empirical p-values in the following simulation studies and real applications.

3.3 Simulations

In this section, we assess the type-I error control and power performance of the proposed IR goodness-of-fit test for copula specification. We also evaluate the computing speed of our proposed test.

3.3.1 Data generation

The data are generated from various Archimedean copula models (i.e., Clayton, Gumbel, Frank) with the marginal survival times following Weibull distributions (scale $\lambda = 0.1$ and shape $k = 2$ for both margins). The sample size is 300, and the number of replications is $1,000$. Data are also generated under different dependent strengths between two margins, as measured by Kendall's τ (0.3 or 0.6), as well as under different right censoring rates (0%, 25%, or 50%). The following paragraphs explicitly explain the data generation processes under interval and right censoring, respectively.

Scenario I: Bivariate interval-censored data: Recall that the bivariate joint survival function under a copula model is $S(t_1, t_2) = C_\eta\{S_1(t_1), S_2(t_2)\}$, where $U = S_1(T_1)$, $V = S_2(T_2)$ each follows a uniform distribution $U[0, 1]$. Define $W_u(v) = h(u, v) = P(V \leq v|U = u)$, which equals to $\partial C_\eta(u, v)/\partial u$. To generate bivariate survival data (t_{i1}, t_{i2}), $i = 1, .., n$, by following the data generation procedures in Sun et al. [2019], we first generate u_i and w_i from two independent $U[0, 1]$ distributions. Then let $w_i = h(u_i, v_i)(= C_\eta(u_i, v_i)/\partial u_i)$ and solve for v_i from the inverse of h function h^{-1}. Finally, we obtain t_{1i} and t_{2i} from $S_1^{-1}(u_i)$ and $S_2^{-1}(v_i)$, respectively. To obtain bivariate interval-censored data, we follow the similar censoring procedure as in Sun and Ding [2019]. Specifically, we assume each subject was assessed for K times with the length between two adjacent assessment times following an

Exponential distribution. The value of K controls the size of the right censoring rate. In the end, for each subject i, L_{ij} is defined as the last assessment time before T_{ij}, and R_{ij} is the first assessment time after T_{ij}. When $R_{ij} = \infty$, T_{ij} is right-censored.

Scenario II: Bivariate right-censored data: We first generate bivariate true event times T_{ij} using the conditioning approach described as in the interval censoring case above. Then, we generate the censoring times C_{ij} from an Exponential distribution. In the end, the observed data are $Y_{ij} = \min(T_{ij}, C_{ij})$ and $\Delta_{ij} = I(T_{ij} \leq C_{ij})$.

3.3.2 Test performance in bivariate interval-censored data

We first evaluate the type-I error control and power performance of our proposed IR test under interval censoring. Specifically, we assess the test under different copula dependence strengths (Kendall's $\tau = 0.3, 0.6$) and various right censoring rates $(0\%, 25\%, 50\%)$. True data are generated from Clayton, Gumbel, and Frank copulas. We examine the goodness of a series of copulas under each scenario, including a flexible Archimedean two-parameter copula model denoted as "copula2" [Sun and Ding, 2019] that incorporates Clayton and Gumbel as special cases. Results are summarized in Table 3.3.1, which shows good type-I error control in all scenarios. Our test also presents good power performance in general, and the power increases as the dependence strength increases. In general, the power decreases as the right censoring rate increases, especially when fitting Gumbel copula to Clayton data or fitting Clayton to Frank data.

3.3.3 Test performance in bivariate right-censored data

We further evaluate the type-I error control and power performance of our proposed IR test under right censoring. The settings are similar to the settings in Table 3.3.1. We illustrate the results in Table 3.3.2, which suggests good type-I error control in all scenarios. We also compare with an existing method [Emura et al., 2010], which can test Archimedean copula specification (such as Clayton, Gumbel, Frank) under right-censoring. However, its associated R package [Emura, 2018] on CRAN can only test for the Clayton copula. Therefore, we only include its results for testing Clayton copula (denoted as "Emura"). Due

Table 3.3.1: Type-I errors and powers of IR tests under interval censoring (IC).

True Copula:		Clayton			Gumbel			Frank		
RC rate	τ	Clayton	Copula2	Gumbel	Gumbel	Copula2	Clayton	Frank	Copula2	Clayton
0%	0.3	**0.045**	**0.032**	0.857	**0.049**	**0.043**	0.602	**0.056**	0.520	0.665
	0.6	**0.045**	**0.053**	1.000	**0.047**	**0.038**	0.983	**0.046**	0.878	0.999
25%	0.3	**0.052**	**0.049**	0.803	**0.061**	**0.026**	0.614	**0.045**	0.414	0.296
	0.6	**0.052**	**0.049**	0.861	**0.055**	**0.024**	0.910	**0.060**	0.793	0.737
50%	0.3	**0.060**	**0.060**	0.337	**0.059**	**0.026**	0.683	**0.051**	0.267	0.241
	0.6	**0.057**	**0.058**	0.492	**0.061**	**0.032**	0.830	**0.041**	0.415	0.339

to the slow computing speed of the Emura method, we only perform 100 replications, which was also used in the original paper [Emura et al., 2010]. We find that the Emura test is more powerful than ours when the dependence (as measured by Kendall's τ) is relatively small; when Kendall's τ becomes larger, both tests exhibit satisfactory and comparable power performance.

Table 3.3.2: Type-I error rates and power of IR tests under right censoring (RC).

True Copula:		Clayton				Gumbel				Frank			
RC rate	τ	Clayton	Copula2	Gumbel	Emura	Gumbel	Copula2	Clayton	Emura	Frank	Copula2	Clayton	Emura
0%	0.3	**0.028**	**0.022**	0.722	**0.09**	**0.047**	**0.021**	0.423	1.00	**0.051**	0.225	0.395	0.99
	0.6	**0.023**	**0.020**	1.000	**0.05**	**0.059**	**0.012**	0.969	1.00	**0.048**	0.675	0.997	1.00
25%	0.3	**0.052**	**0.053**	0.610	**0.04**	**0.055**	**0.039**	0.451	1.00	**0.042**	0.230	0.358	0.95
	0.6	**0.047**	**0.033**	0.978	**0.05**	**0.044**	**0.022**	0.946	1.00	**0.038**	0.642	0.975	1.00
50%	0.3	**0.051**	**0.051**	0.401	**0.02**	**0.047**	**0.028**	0.469	0.98	**0.044**	0.213	0.282	0.80
	0.6	**0.041**	**0.040**	0.806	**0.06**	**0.042**	**0.017**	0.882	1.00	**0.050**	0.558	0.870	0.97

Table 3.3.3: Computing time (in minutes) of IR tests when sample size ranges from 100 to 1,000 under the right (RC) or interval censoring (IC).

Censoring type	Time (mins)	n = 100	n = 300	n = 500	n = 1,000
RC	Emura	4	142	1,385	10,237
	IR, 1 core	5	16	23	46
	IR, 10 cores	1	2	4	7
	IR, 20 cores	0.5	1	2	3
IC	IR, 1 core	13	31	51	103
	IR, 10 cores	2	4	9	16
	IR, 20 cores	1	2	4	8

3.3.4 Test computing time

We compare the computing speed of our proposed IR test and the Emura test. Data are generated from the Clayton model with Weibull margins, with Kendall's $\tau = 0.6$ and the right censoring rate = 25%. Sample sizes are set as $n = 100, 300, 500, 1000$. The fitted copula model is Clayton. The number of replication is set at ten since in practice, one typically only needs to test multiple copula models for one real dataset. For our proposed method, we illustrate the computing efficacy by computing the tests in parallel across multiple cores (e.g., $1, 10, 20$). As shown in Table 3.3.3, we find that under right censoring with one core, our IR test is about ten times faster than the Emura test when sample size is 300, and the speed difference becomes more dramatic when n becomes larger. In fact, the computing time using our method increases linearly with the sample size, whereas the relationship seems to be exponential for the Emura test. The performance of IR test is further enhanced by parallel computing, which can complete ten tests within minutes. Similarly, under interval censoring, our IR test can also complete within a reasonable amount of minutes.

3.4 Real data analysis

We present the IR test results in multiple real data examples under various censoring types. We evaluate several copula functions such as Archimedean (i.e., Copula2, Clayton, Gumbel, Frank), Gaussian and Plackett copulas. The results are summarized in Table 3.4.1. Besides, the Emura test p-value and an estimated Kendall's τ for fitting Clayton copula are included under right censoring.

The first dataset, denoted as "AREDS-sub" contain 629 (moderate to severe) from a clinical trial study called Age-Related Eye Disease Study (AREDS) [AREDS Group, 1999], which examines the time to the late-stage Age-related Macular Degeneration (AMD) in left and right eyes of the AMD patients. Due to the intermittent assessment design, the exact time to late AMD for each eye was interval-censored. As shown in Table 3.4.1, the IR test rejects almost all the tested copulas (Clayton, Gumbel, Frank, Gaussian) except Copula2 and Plackett. The corresponding estimated Kendall's τ is 0.52 under Copula2. Interestingly, we observe that although Clayton has a very close AIC value (4364.51) to Copula2 (4363.87), Clayton is still rejected by our method. It suggests that AIC might not be an ideal criterion in selecting a proper copula model in the real data analysis.

The next dataset, named as "Tandmob", comes from a longitudinal prospective dental study called the Signal-Tandmobiel project [Vanobbergen et al., 2000], which was performed in Flanders (North of Belgium) in 1996-2001. The cohort of $4,468$ randomly sampled children who attended the first year of the basic school at the beginning of the study was annually dental examined by one of the 16 trained dentists. Among these $4,468$ children, 38 children did not come to any of the designed dental examinations, resulting in $n = 4,430$ in the final dataset for analysis. The dataset contains the information on the emergence times of teeth, which are interval-censored due to the intermittent assessments. Bogaerts and Lesaffre [2008] extracted the bivariate emergence times of the maxillar first premolars (contralateral teeth 14 and 24), and fitted the data with three different copula models (i.e., Clayton, Gaussian, Plackett). The paper shows that the Plackett copula has the smallest AIC value, and concludes that Plackett is the best choice. We also notice that Bogaerts and Lesaffre [2008] models the copula dependence parameter using a covariate.

For right censoring case, the dataset "LOSS-ALAE" is a well-known insurance dataset on losses and allocated loss adjusted expenses (ALAE), which are collected by the US Insurance Service Office. It consists of $1,500$ general liability claims, and each claim includes an indemnity payment (i.e., LOSS) and an allocated loss adjusted expense (e.g., ALAE). For the LOSS data, 34 observations are right-censored due to late settlement lags. For the ALAE data, all claims are uncensored. Zhang et al. [2016] tests copula specification using the $1,466$ complete observations and finds that the Gumbel copula is the most suitable model, whereas Clayton and Gaussian are rejected. Lin and Wu [2020] also fails to reject the Gumbel copula and rejects Gaussian by examining the entire $1,500$ observations. To be consistent with Zhang et al. [2016] and Lin and Wu [2020], we use the estimated cumulative distribution functions for the two margins, instead of the estimated survival functions (with censoring taken into account). In Table 3.4.1, we find that Copula2 and Gumbel appear to be the most suitable models with the highest p-values, while Clayton and Gaussian are rejected. The Kendall's τ estimates are 0.31 under both Copula2 and Gumbel. Both Copula2 and Gumbel also have the smallest AIC values. Our conclusions are consistent with the findings from Zhang et al. [2016] and Lin and Wu [2020]. The Emura method cannot be performed since the extensive computing time makes it infeasible to test such a large data.

The second dataset, which is a subset of the Diabetic Retinopathy Study (DRS), includes 83 patients, who have both onset times for diabetes and diabetic retinopathy. The study examines whether laser photocoagulation is effective in delaying the onset of blindness in diabetic retinopathy patients. Among these 83 patients, four experienced failure only in the right eye, 36 experienced failure in the left eye, and 14 experienced failure in both eyes. This dataset has been used for examining copula goodness-of-fit in several previous works. For example, Manatunga and Oakes [1999] suggests that the Clayton model fits this dataset well using the diagnostic plot from Oakes [1989]. Later, Wang [2010] shows that the Clayton and Frank copula models are not rejected with p-values > 0.05. As shown in Table 3.4.1, both Frank and Clayton copulas are not rejected using our method, with estimated Kendall's $\tau = 0.28$ and 0.35, respectively. In addition, the Emura method does not reject the Clayton model either, and reports a similar estimated Kendall's τ.

The third dataset, the Kidney study, records the recurrence times to infection at the point of insertion of the catheter in 38 kidney patients using portable dialysis equipment [McGilchrist and Aisbett, 1991]. The catheter may be removed for reasons other than in-fection leading to right-censored observations. Each patient has exactly two observations. Among the 38 patients, 9 have exact event times for the first infection only, 3 have exact event times for the second infection, 23 have exact event times for both infections, and the rest 3 are right-censored for both infections. This dataset is also used in Emura et al. [2010], in which the method does not reject Gumbel, Clayton, and Frank copulas. Table 3.4.1 con-veys the similar message. Note that the Emura test using the public R package [Emura, 2018] gives a $p-$value $= 0.473$, whereas the original paper reports 0.189.

The fourth dataset "Ovarian" contains 1, 192 advanced ovarian cancer patients from four randomized multi-center clinical trials. Each patient has two endpoints: the progression-free survival time and the overall survival time. After a minimum follow-up of 10 years in all four trials, either one of the two events has occurred for most patients (80%). Burzykowski et al. [2001] proposed to use the copula model (i.e., Clayton and Gumbel) to account for the dependence between the two events, and reported strong dependence strength (Kendall's τ at about 0.8). Our IR test suggests that Clayton and Copula2 are likely the proper models. We do not report the Emura method due to the large sample size.

The last dataset "Gastadv" contains individual data (overall and progression-free survival) of 4, 069 patients with advanced or recurrent gastric cancer from 20 randomized trials of chemotherapy. Rotolo et al. [2018] used this dataset to illustrate the method proposed in Burzykowski et al. [2001] by fitting three types of copula models (i.e., Clayton, Gumbel, and Plackett). It reported similar Kendall's τ estimates under Clayton and Plackett, and a lower τ estimate under Gumbel. Our IR test rejects Clayton, and indicates Copula2 and Plackett are suitable for this dataset. We do not report the Emura method due to the large sample size.

Table 3.4.1: Performance of IR tests in real datasets under interval censoring (IC) or right censoring (RC).

			IR						Emura
Type	Datasets		Copula2	Clayton	Gumbel	Frank	Gaussian	Plackett	Clayton
IC	AREDS-sub	p-value	0.225	< 0.005	0.020	0.010	< 0.005	0.300	NA
		$\hat{\tau}$	0.52	0.52	0.40	0.45	0.45	0.45	
		AIC	4363.87	4364.51	4406.16	4368.93	4385.18	4366.55	
		loglik	-2179.93	-2181.25	-2202.08	-2183.47	-2191.59	-2182.28	
	Tandmob	p-value	< 0.005	< 0.005	< 0.005	< 0.005	< 0.005	< 0.005	NA
		$\hat{\tau}$	0.64	0.63	0.57	0.64	0.58	0.66	
		AIC	19693.41	19842.20	20116.81	19727.00	20041.38	19441.68	
		loglik	-9844.70	-9920.10	-10057.41	-9862.50	-10019.69	-9719.84	
RC	LOSS-ALAE	p-value	0.940	< 0.005	0.875	0.710	< 0.005	0.940	-
		$\hat{\tau}$	0.31	0.21	0.31	0.31	0.31	0.31	-
		AIC	-360.62	-171.19	-362.62	-309.30	-323.78	-310.71	-
		loglik	182.31	86.59	182.31	155.65	162.89	156.35	-
	DRS	p-value	0.080	0.170	0.160	0.405	0.230	0.970	0.325
		$\hat{\tau}$	0.30	0.35	0.24	0.28	0.28	0.28	0.34
		AIC	90.77	89.68	88.96	89.22	89.35	89.10	-
		loglik	-43.39	-43.84	-43.48	-43.61	-43.68	-43.55	-
	Kidney	p-value	0.530	0.105	0.270	0.355	0.160	0.240	0.473
		$\hat{\tau}$	0.27	0.19	0.28	0.24	0.25	0.25	0.14
		AIC	13.50	13.93	11.50	12.67	12.79	12.51	-
		loglik	-4.75	-5.96	-4.75	-5.33	-5.40	-5.25	-
	Ovarian	p-value	0.800	0.900	< 0.005	< 0.005	< 0.005	0.070	-
		$\hat{\tau}$	0.82	0.80	0.77	0.80	0.79	0.78	-
		AIC	-1277.31	-1100.29	-982.67	-1042.49	-980.25	-1156.15	-
		loglik	640.66	551.15	492.33	522.24	491.12	579.08	-
	Gastadv	p-value	1.00	< 0.005	0.15	0.14	< 0.005	0.600	-
		$\hat{\tau}$	0.59	0.51	0.57	0.57	0.59	0.55	-
		AIC	-1668.48	-774.44	-1453.71	-998.50	-1419.14	-1130.75	-
		loglik	836.24	388.22	727.85	500.25	710.57	566.38	-

3.5 Conclusion

We have proposed a novel IR-based goodness-of-fit copula specification test for multivariate uncensored, interval-censored, and right-censored data. We have established the asymptotic consistency and normality of our proposed IR statistic estimator, which is further used to construct the goodness-of-fit test procedures. To the best of our knowledge, this is the first method that can be flexibly used for testing copula specifications under both interval and right censoring. The method can test any copula model with an analytical form (including Archimedean, Gaussian and Plackett copula families). Our method applies to copula models with more than one dependence parameters (i.e., $p \geq 2$, like Copula2). Our test statistic is simple to calculate and straightforward to implement. The test procedure enjoys high computational efficiency through parallel computing, as comparedto the Emura method. The simulations show that our method can control type-I errors well under both interval and right censoring. Our method also achieves satisfactory power performance when Kendall's τ is moderately high. We demonstrate the strong differentiating power of the proposed IR test when applied to two interval-censored real datasets. Its applications to several right-censored real datasets also reveal consistent findings as compared with the existing methods for right-censored data.

4.0 GWAS-based deep learning for survival prediction

4.1 Introduction

Accurate 'time-to-event' data based survival prediction is fundamental to effective clinical management and precision medicine of human diseases [Chin et al., 2011, Compton, 2018]. It relies on a survival model to predict the dynamic risk profile of a future event over time (e.g., disease onset, recurrence, progression, or death) based on the individual's current status, such as clinical characteristics, genetic information and medical images. Most importantly, such a prediction addresses the patient's key concern regarding the disease progression pattern in the future and shapes the physician's decision making for the treatment or clinical management strategy. It is to be noted that survival prediction is fundamentally different from typical prediction models that predict a future event (whether occurs or not) by fixing the time of interest through a binary classification [Castro-Rodríguez et al., 2000, Chi et al., 2007]. Despite its essential role in precision medicine, survival prediction remains a challenging task [Abrams et al., 2014, Barillot et al., 2012, Schumacher et al., 2012], largely due to the complex nature of diseases and heterogeneity between patients. Therefore, there is an urgent need for developing accurate and personalized survival prediction models with improved capacity in learning the complex structures and interplays among predictors. Recent advances in high-throughput technologies have generated large volumes of molecular profiling data for each patient, which provides unprecedented opportunities in identifying potential biomarkers and further establishing accurate survival prediction models [Chen et al., 2019, Collins and Varmus, 2015, Sarnowski et al., 2018]. In particular, several national-wide large-scale longitudinal studies, such as the Trans-Omics for Precision Medicine (TOPMed) and All of Us, are underway using whole-genome sequencing and other omics technologies, with the ultimate goal of accelerating precision medicine. However, how to effectively utilize the wealthy amount of data is challenging. The first challenge comes from how to connect high-dimensional predictors with the outcome of interest. This problem is particularly difficult in survival prediction because the events of interest are often censored due to either a short study period or loss of follow-up during the study. The second challenge is how to model the complex structure among numerous biomarkers, where the specific structure is largely unknown. The third challenge is given the heterogeneity of patients how to interpret the importance of each predictor for each patient and further how to identify patient subgroups to provide personalized prevention or treatment strategies.

The recent advances in multi-layer deep neural network models have made extraordinary achievements in providing new effective risk prediction models from complex and high dimensional biomedical data, such as omics and biomedical imaging [Grassmann et al., 2018, Min et al., 2016, Miotto et al., 2017, Poplin et al., 2018]. However, the application of deep learning in survival prediction is still limited. Faraggi and Simon [1995] proposed a single-layer neural network based on the Cox proportional hazards (PH) model. However, its performance did not exceed the regular Cox model in a prostate cancer survival data set with 475 patents and only 4 clinical predictors. More recently, multiple efforts have been devoted to evaluating Cox-based neural network survival models using larger data sets with omic biomarkers. For example, Katzman et al. [2018] demonstrated that a single hidden layer neural network survival model performed marginally better than the Cox model and random survival forest (RSF) model in a breast cancer survival data set with $1,980$ patients and 9 predictors. In another study, Ching et al. [2018] applied a single hidden layer neural network survival model to 10 TCGA cancer survival data sets (sample sizes range from 302 to $1,077$) with high-throughput gene expression biomarkers, from which the neural network survival models resulted in comparable or better performance than the Cox model, the penalized Cox models such as Cox-LASSO and the RSF model. In another study [Yousefi et al., 2017] that also used TCGA cancer survival data sets (sample sizes range from 194 to $1,092$ with up to $17,000$ gene expression biomarkers), the neural network survival models yielded comparable performance to the penalized Cox model and better performance than the RSF model. Hao et al. [2018] developed a pathway-based neural network survival model and applied it to a TCGA cancer data set (sample size 522 with 860 pathways and $5,567$ genes). However, all these studies have limited sample sizes, particular in the presence of tens of thousands of predictors, and thus may lead to severe model overfitting problems. Moreover, patient-specific predictor importance was not considered in those studies. They also did not carefully account for the scenario of tied events, which is commonly seen in practice, especially when the sample size is large.

In this chapter, we propose and evaluate a multi-hidden-layer Cox-based deep neural network (DNN) survival model to predict the progression of a progressive eye disease, namely, age-related macular degeneration (AMD). The genome-wide association study (GWAS) of AMD is the first and most successful GWAS research, where the massive GWAS data provide unprecedented opportunities to study disease risk and progression. Although some attempts have been tried to predict AMD progression risks using genetic information such as SNPs, most statistical models focus on the structured regression framework, which typically only accounts for (generalized) linear effects of the SNPs and thus have considerable limitations. To the best of our knowledge, there has no existing work on survival prediction using deep learning to effectively extract features from the GWAS data. Therefore, we build an accurate and interpretable DNN survival prediction model for AMD progression.

The rest of the paper is organized as follows. Section 4.2 describes the deep learning survival methods and prediction evaluation procedures. We assess the performance of three machine/deep learning survival prediction models (DNN, Cox-LASSO, RSF) through extensive simulation studies in Section 4.3 and apply them to the GWAS data from two large-scale clinical studies of AMD in Section 4.4. The discussions are presented in Section 4.5.

4.2 Methods

For each subject $i \in \{1, ..., n\}$, the observations are $\{Y_i, \delta_i, Z_i\}$, where $Y_i = \min(T_i, C_i)$ is the minimum of survival time T_i and censoring time C_i; $\delta_i = I(T_i \leq C_i)$ is the right-censoring indicator; Z_i is the covariate vector.

4.2.1 Cox-based DNN survival model

The Cox PH model is the most popular regression model for censored survival data. It assumes that the hazard function of survival time T takes the form $h(t|Z_i) = h_0(t)\exp(Z_i^T\theta)$, where $h_0(t)$ is the unspecified baseline hazard function at time t and θ is a vector of covariate effects. The term $Z_i^T\theta$ is called the linear predictor or prognostic index. On the other hand, the deep neural network model is well known for its capacity in learning complex covariate structures (i.e., non-linearity, interactions) [LeCun et al., 2015]. By the Universal Approximation Theorem [Cybenko, 1989, Hornik et al., 1989], for any continuous function $g(Z;\theta)$, it is guaranteed to exist a neural network that approximates this function. Moreover, this theorem holds even if we restrict the neural networks to have just one single hidden layer. Therefore, even very simple neural network architecture can be extremely powerful. The synergy of the powerful DNN and the popular Cox model leads us to build our Cox-based DNN survival model and apply it to AMD progression prediction.

Assumption and loss function of DNN survival model

The DNN survival model we consider here can be written as $h(t|Z_i) = h_0(t)e^{g(Z_i;\theta)}$. The major difference between this DNN model and the regular Cox model is that DNN takes the prognostic index $g(Z_i;\theta)$ as an unknown function with parameters θ, instead of assuming a simple linear relationship. In this way, the DNN model can approximate various non-linear covariate structures by estimating $g(Z_i;\theta)$. We will employ a feedforward DNN with multiple hidden layers to estimate the unspecified $g(Z;\theta)$. In fact, one can regard the regular Cox model as a special case of DNN when $g(Z_i;\theta) = Z_i^T\theta$.

In large-scale studies, it is quite common that more than one observations develop events at the same time. Such events are called tied events. To handle this scenario, we approximate the partial likelihood via Efron's approach [Efron, 1977]. Moreover, to deal with high-dimensional covariates, we introduce the L_1 penalty to the DNN loss function $-l(\theta;Z) + \lambda||\theta||_1$, where $l(\theta;Z)$ is the Efron approximation of log partial likelihood:

$$l(\theta;Z) = \frac{1}{N_D}\sum_{j\in D}\left\{\sum_{i\in H_j} g(Z_i;\theta) - \sum_{l=0}^{m_j-1}\log\left(\sum_{i\in R_j} e^{g(Z_i;\theta)} - \frac{l}{m_j}\sum_{i\in H_j} e^{g(Z_i;\theta)}\right)\right\}, \qquad (4.2.1)$$

where D is the set of all events with size N_D and $\{t_j\}$ is the set of unique event times; H_j is the set of subjects $\{i\}$ such that $Y_i = t_j$ and $\delta_i = 1$ and m_j is the size of H_j; and R_j is the risk set satisfying $Y_i \geq t_j$.

DNN architecture

First, we introduce the general form of an L-hidden-layer feedforward DNN, which is composed of one input layer, L hidden layers and one output layer (with one node in our case). For each subject, DNN inputs the vector of covariates Z into its input layer and output a scalar prognostic index $g(Z;\theta)$. For each hidden layer $l \in \{1,...,L\}$ with n_l number of nodes, it takes the input $n_{l-1}-$dimensional $\mathbf{a}^{(l-1)}$ from the $(l-1)^{th}$ layer and outputs n_l-dimensional $\mathbf{a}^{(l)}$ through a n_l-dimensional activation function $\mathbf{f}^l$. Mathematically, the l^{th} hidden layer model can be written as $\mathbf{a}^{(l)} = \mathbf{f}^{(l)}(\mathbf{W_0}^{(l)} + \mathbf{W}^{(l)}\mathbf{a}^{(l-1)})$, where $\mathbf{W_0}^{(l)}$ is the bias vector with length n_l; $\mathbf{W}^{(l)}$ is an $n_l \times n_{l-1}$ weight matrix. $\mathbf{f}^{(l)}(\cdot)$ is a vector of activation functions $f^{(l)}(\cdot)$. Often a common $f^{(l)}(\cdot)$ function is assumed for all the nodes in the l^{th} hidden layer and it is usually a non-linear function, such as the sigmoid [Hornik et al., 1989] $f(x) = \frac{1}{1+e^{-x}}$, the tangent $f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$, the rectified linear unit (ReLU) [Sutskever et al., 2013] $f(x) = \max(0,x)$, and the scaled exponential linear units (SeLU) [Klambauer et al., 2017] $f(x) = \lambda \times ReLU(x) + \lambda I(x<0)\alpha(e^x - 1)$, where λ and α are constants. The final or output layer also has weights and an output function f^{out}, which is an identity function.

Take a simple one-hidden layer neural network for example. We have p-dimensional input covariates z_i from the i^{th} subject, n_1 number of hidden nodes with $k = 1,...,n_1$ and one single output node. For the k^{th} hidden node, we have $a_k^{(1)} = f_k^{(1)}(w_{k0}^{(1)} + \sum_{j=1}^{p} w_{kj}^{(1)} z_{ij})$. Similarly, the output node is $o_i = f^{out}(w_0^{(2)} + \sum_{k=1}^{n_1} w_k^{(2)} a_k^{(1)}) = w_0^{(2)} + \sum_{k=1}^{n_1} w_k^{(2)} a_k^{(1)}$ by assuming f^{out} is an identity function. Typically we have $o_i = g(z_i;\theta)$. The full parameter set θ is composed of $\{w_{k0}^{(1)}, k=1,...,n_1\}$, $\{w_{kj}^{(1)}, k=1,...,n_1, j=1,...,p\}$, $w_0^{(2)}$ and $\{w_k^{(2)}, k=1,...,n_1\}$.

DNN optimization and survival prediction

To solve for $\hat{\theta}$, we use the mini-batch stochastic gradient descent algorithm [Hinton et al., 2012] to minimize the loss function in equation (4.2.1). Comparing with the standard stochastic gradient descent that uses all samples for each iteration, the mini-batch algorithm is much faster. Specifically, we randomly divide all observations into mini-batches with size N_B and update $\hat{\theta}$ by adding the gradient contributed by each mini-batch. In particular, the loss function for the r^{th} batch is

$$\begin{aligned}-l^r(\theta;Z)+\lambda||\theta||_1 = &-\frac{1}{N_D^r}\sum_{j\in D^r}\Bigg\{\sum_{i\in H_j^r} g(Z_i;\theta)-\sum_{l=0}^{m_j^r-1}\log\Bigg(\sum_{i\in R_j^r} e^{g(Z_i;\theta)}-\frac{l}{m_j^r}\sum_{i\in H_j^r} e^{g(Z_i;\theta)}\Bigg)\Bigg\}\\ &+\lambda||\theta||_1,\end{aligned}$$

where N_D^r, D^r, H_j^r, m_j^r and R_j^r are the corresponding terms for the r^{th} batch similar to those defined in equation (4.2.1). Then we update θ by adding the gradient contributed by the r^{th} batch through:

$$\begin{aligned}&\Delta_r = -\bigtriangledown_\theta\, l^r(\theta;Z)+\lambda\bigtriangledown_\theta ||\theta||_1\\ &\theta \leftarrow \theta-\gamma\Delta_r,\end{aligned}$$

where γ is the learning rate (also called step size). This process will be repeated for N_E times (also called epochs) before convergence. We employ the Glorot uniform initializer [Glorot and Bengio, 2010] to randomly select initial values. Once we get $\hat{g}(Z_i;\hat{\theta})$, we can obtain the predicted survival probability for subject i at time t through $\hat{S}(t|Z_i)=\exp\{-\hat{H}_0(t)e^{\hat{g}(Z_i;\hat{\theta})}\}$.

DNN hyperparameters

To perform the survival prediction based on the DNN survival model, we need to select the DNN hyperparameters. The main hyperparameters include the number of hidden layers, number of nodes per hidden layer, choice of activation function, the L_1 penalty parameter, batch size, epoch size, and learning rate. In this work, we perform cross-validations in the training data and select the combination of hyperparameters that lead to the most optimal prediction performance based on the validation results. Specifically, in all simulations we use the following hyperparameter setting: 2 hidden layers, 30 nodes per hidden layer, activation function SeLU (parameters $\alpha = 1.6732$ and $\lambda = 1.0507$ by default), L_1 penalty $= 0.1$, batch size $N_B = 50$, epoch size $N_E = 1,000$ and learning rate $\gamma = 0.01$ (for sparse signals) or $\gamma = 0.0001$ (for weak signals). For the real data analysis, we select the following hyperparameters: includes 2 hidden layers, 300 nodes per hidden layer, activation function SeLU (α and λ by default), L_1 penalty $= 0.01$, batch size $N_B = 50$, epoch size $N_E = 1,000$ and learning rate $\gamma = 0.00001$.

DNN interpretation

It is important to understand and interpret the fitted neural network prediction model. One way is to export feature (i.e., predictor) importance measures that decide the top important features in a prediction model. The Local Interpretable Model-Agnostic Explanation (LIME) method [Ribeiro et al., 2016] provides prediction importance of each predictor for each subject in the model by perturbing the feature values and evaluating how the prediction results change. Specifically, for one feature in a specific subject, the method trains an interpretable model (e.g., linear regression) on new data sets based on small perturbations (e.g., adding noises) of the particular feature in this subject. LIME has been widely applied to neural network models with continuous or categorical outcomes, but not with censored survival outcomes yet. In this paper, we apply the LIME method to the neural network survival model and produce subject-specific predictor importance measures with meaningful interpretations.

4.2.2 Evaluation metrics for survival prediction performance

We calculate the Harrell's concordance index (c-index) [Harrell et al., 1996] to measure the proportion of concordant pairs (i.e., the predicted and observed outcomes are concordant) among all comparable pairs (i.e., the true progression statuses can be ordered for two observations within one pair). Pairs are not comparable if both are censored, or one is censored at time c_1 and the other event occurs at time t_2 with $t_2 > c_1$. The c-index is between 0 and 1 with a larger value indicating a better prediction model, which can be estimated by $\widehat{C} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} \delta_i I(Y_i<Y_j) I(\hat{g}(Z_i;\hat{\theta})>\hat{g}(Z_j;\hat{\theta}))+0.5*I(\hat{g}(Z_i;\hat{\theta})=\hat{g}(Z_j;\hat{\theta}))}{\sum_{i=1}^{n}\sum_{j=1}^{n} \delta_i I(Y_i<Y_j)+I(\hat{g}(Z_i;\hat{\theta})=\hat{g}(Z_j;\hat{\theta}))}$.

We also obtain the time-dependent Brier score [Gerds and Schumacher, 2006, Graf et al., 1999]. At a specific time point t, the Brier score measures the mean squared error between the observed progression status at time t (i.e., $Y_i(t) = I(Y_i \geq t)$) and the predicted survival probability (i.e., $\hat{S}(t|Z_i)$). A lower Brier score indicates a better prediction model. A Brier score of 33% corresponds to predicting the risk by a random number drawn from Uniform $[0, 1]$ and 25% corresponds to predicting 50% risk for every observation. The estimated Brier score is expressed as $\widehat{BS}(t, \widehat{S}) = \frac{1}{M}\sum_{i \in D_M} \widehat{W}_i(t)\{Y_i(t) - \hat{S}(t|Z_i)\}^2$, where D_M is the test data set with size M, $\hat{S}(t|Z_i)$ is estimated using the training data, and $\widehat{W}_i(t) = \frac{(1-Y_i(t))\delta_i}{\hat{G}(Y_i-)} + \frac{Y_i(t)}{\hat{G}(t)}$ is the inverse probability of censoring weights with $\hat{G}(t) = \hat{P}(C > t)$ [Gerds and Schumacher, 2006].

We also obtain the time-dependent ROC curve and the associated area under the curve (AUC) [Heagerty et al., 2000]. The AUC measures the discrimination capability of $\hat{g}(Z; \hat{\theta})$. It ranges between 0 and 1, with higher AUC indicating better discrimination ability. Specifically, we first derive the time-dependent sensitivity and specificity

$$\text{sensitivity}(c, t) = P\{\hat{g}(Z; \hat{\theta}) > c | T \leq t\},$$
$$\text{specificity}(c, t) = P\{\hat{g}(Z; \hat{\theta}) \leq c | T > t\},$$

where c is some arbitrary cut-off. For a given t, sensitivity(c, t) and specificity(c, t) determine the ROC curve profile and the associated AUC at time t.

4.2.3 K-fold cross-validations

Over-fitting is a common issue for all machine learning models. One way to alleviate the issue is to perform K-fold cross-validation. Specifically, the original data D_N are split into K subsets $D_k, k = 1, ..., K$, accounting for the censoring proportions. For the k^{th} cross-validation, models are trained in the samples $D_N \backslash D_k$ (original data without k^{th} subset) and then validated in the test samples D_k. Finally, the K-fold cross-validation estimates (i.e., c-index and Brier score) are calculated by averaging over the test data results, as shown below

$$\widehat{CvBS}(t, \hat{S}) = \frac{1}{K}\sum_{k=1}^{K} \frac{1}{M_k} \sum_{i \in D_k} \hat{W}_i(t)\{Y_i(t) - \hat{S}_k(t|Z_i)\}^2,$$
$$\widehat{CvC} =$$
$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{M_K}\frac{\sum\limits_{i \in D_k}\sum\limits_{j \in D_k} \delta_i I(Y_i < Y_j) I(\hat{g}_b(Z_i; \hat{\theta}_k) > \hat{g}_k(Z_j; \hat{\theta}_k)) + 0.5 * I(\hat{g}_k(Z_i; \hat{\theta}_k) = \hat{g}_k(Z_j; \hat{\theta}_k))}{\sum\limits_{i \in D_k}\sum\limits_{j \in D_k} \delta_i I(Y_i < Y_j) + I(\hat{g}_k(Z_i; \hat{\theta}_k) = \hat{g}_b(Z_j; \hat{\theta}_k))},$$

where M_k is the sample size of the k^{th} subset.

4.2.4 Implementation

Our DNN survival model is built with Keras [Chollet et al., 2015] and Tensorflow [Abadi et al., 2016] to ensure computational stability and efficiency. Keras is a deep learning framework that provides a convenient way to define and train deep learning models. It provides high-level building blocks for deep learning models [Chollet and Allaire, 2018]. For example, one can define a neural network model with a few lines of codes in Keras. We use Tensorflow for low-level operations such as differentiation, which serves as the backend engine of Keras. Via Keras and Tensorflow, our DNN survival model is compatible with both GPUs and CPUs.

4.3 Simulation studies

We use simulations to evaluate the prediction performance of DNN and compare it with Cox-LASSO (abbreviated as LASSO) [Tibshirani, 1996] and RSF [Ishwaran and Kogalur, 2007, Ishwaran et al., 2008]. Two main simulation settings are considered. In the first setting, data are generated with sparse signals (i.e., only a few predictors with non-zero effects on the survival outcome). In the second setting, all predictors have non-zero but weak signals, which is common in settings with genetics or genomics predictors. Within each simulation setting, we generate multiple scenarios with different structures in predictors' effects. For each scenario, we train the models in a training data set, and then test them in an independent test data set and summarize the results across 200 replications. The sample sizes for both training and test data sets are $1,000$.

All three models involve the selection of tuning parameters. For LASSO, we use 5-fold cross-validation to select the tuning parameter in the L_1 penalty using the training data. After the tuning parameter is determined, we then train the LASSO model using the entire training data and finally validate the model in the test data. For RSF, we train the model by utilizing the default setting of $1,000$ trees and $\sqrt{p}$ number of randomly selected predictors at each split. In the case of DNN, it is widely known for its exhaustive process in selecting optimal tuning parameters since there are many tuning parameters to consider. The tuning process is even more time consuming given that we have multiple simulation scenarios. Therefore, for all simulation scenarios, we fix one set of tuning parameters for DNNs. More details about the tuning parameters are discussed in Section 4.2.

4.3.1 Simulation I: survival data with sparse signals

We consider five scenarios of predictor effects following Mi et al. [2019], which includes linear effects only (scenario 1) and linear effects together with non-linear effects (scenario 2) or with interactions (scenario 3) or with both non-linear and interaction effects (scenario 4) or with non-linear, interaction and threshold effects (scenario 5). The total number of predictors is set at $p = 10, 50, 100, 500$, respectively. The true models for these five scenarios are illustrated as follows:

Scenario 1 : $h(t|Z_i) = h_0(t)\exp(\sum_{j=1}^{5} Z_{ij})$,

Scenario 2 : $h(t|Z_i) = h_0(t)\exp(\sum_{j=1}^{5} Z_{ij} + Z_{i6}^2 + Z_{i7}^2)$,

Scenario 3 : $h(t|Z_i) = h_0(t)\exp(\sum_{j=1}^{5} Z_{ij} + Z_{i6} + Z_{i7} + 5Z_{i6}Z_{i7})$,

Scenario 4 : $h(t|Z_i) = h_0(t)\exp(\sum_{j=1}^{5} Z_{ij} + Z_{i6} + Z_{i7} + 5Z_{i6}Z_{i7} + Z_{i8}^2 + Z_{i9}^2)$,

Scenario 5 : $h(t|Z_i) = h_0(t)\exp(\sum_{j=1}^{5} Z_{ij} + Z_{i6} + Z_{i7} + 5Z_{i6}Z_{i7} + I(Z_{i8} < -0.5 \cup Z_{i9} < -0.5) - I(Z_{i8} \geq -0.5 \cap Z_{i9} \geq -0.5))$, where $h_0(t) = k\lambda^k t^{k-1}$ is the baseline Weibull hazard function with $\lambda = 0.1, k = 2$. For $Z_i = (Z_{i1}, ..., Z_{ip})$, we first generate Z_i from $MVN(0, \Sigma)$ with $\Sigma = \{\sigma_{jj'} = e^{-|j-j'|}, 1 \leq j, j' \leq p\}$ and then transform Z_{i4} into a binary predictor through $I(Z_{i4} > 0)$ and Z_{i5} into a multinomial predictor through $I(Z_{i5} > -0.5) + I(Z_{i5} > 0.5)$.

In Table 4.3.1, we compare the prediction accuracy of the DNN, RSF, LASSO and true models under the five simulation scenarios by summarizing the c-index, which is a predictive metric measuring the concordance between observed and predicted values. LASSO performs the best in scenario 1 where all predictor effects are linear, but its performance declines in the other 4 scenarios. RSF generally has higher c-index than LASSO in non-linear scenarios. For our proposed method, its performance is lower than LASSO as expected in scenario 1 but is better than RSF, while it outperforms both LASSO and RSF in all non-linear scenarios. The last column gives the c-index values that are obtained from fitting the true underlying model. It can be seen that when p is small, DNN produces c-index values that are very close to the truth for all five scenarios.

4.3.2 Simulation II: survival data with weak signals

In genetics and genomics data, we often observe that many predictors have (non-zero) weak effects due to correlations among SNPs or genes. Moreover, there are various types of omics predictors, such as gene expressions (i.e., continuous), mutations (i.e., binary) and SNPs (i.e., multinomial). Therefore, we generate data that include various types of predictors with weak effects. The total number of predictors is set as $p = 20, 50, 100, 500$ and we consider the following five scenarios:

Scenario 1 : $h(t|Z_i) = h_0(t)\exp(\sum_{j=1}^{p}\beta_j Z_{ij})$,

Scenario 2 : $h(t|Z_i) = h_0(t)\exp(\sum_{j=1}^{p}\beta_j Z_{ij} + Z_{i1}^2 + Z_{i2}^2)$,

Scenario 3 : $h(t|Z_i) = h_0(t)\exp(\sum_{j=1}^{p}\beta_j Z_{ij} + Z_{i3}Z_{i4})$,

Scenario 4 : $h(t|Z_i) = h_0(t)\exp(\sum_{j=1}^{p}\beta_j Z_{ij} + Z_{i1}^2 + Z_{i2}^2 + Z_{i3}Z_{i4})$,

Scenario 5 : $h(t|Z_i) = h_0(t)\exp(\sum_{j=1}^{p}\beta_j Z_{ij} + I(Z_{i1} < -0.5 \cup Z_{i2} < -0.5) - I(Z_{i1} \geq -0.5 \cap Z_{i2} \geq -0.5) + Z_{i3}Z_{i4})$,

where $h_0(t)$ is the baseline Weibull hazard function with $\lambda = 0.01, k = 10$. Similarly to the first simulation setting, we first generate Z_i from a multivariate normal distribution $MVN(0, \Sigma)$ with $\Sigma = \{\sigma_{jj'} = e^{-|j-j'|}, 1 \leq j, j' \leq p\}$. Then the first 20% Z_{ij} remain continuous, the second 20% Z_{ij} are transformed into binary predictors through $I(Z_{ij} > 0)$ and the rest 60% Z_{ij} are transformed into multinomial predictors through $I(Z_{ij} > -0.5) + I(Z_{ij} > 0.5)$. For predictor effects, we set $\beta_j = 0.2$ for continuous and binary predictors. For multinomial predictors, we mimic the linkage disequilibrium effect in SNP data by generating β_j from $MVN(0.2, 0.01 \times \Sigma)$ with the same Σ.

Table 4.3.1: The C-index values (range from 0 to 100) over 200 replications for DNN, RSF, LASSO and true models using sparse signals.

	p	DNN	RSF	LASSO	True
scenario 1	10	88.0 (0.7)	82.9 (1.0)	88.2 (0.6)	87.4 (0.6)
	50	85.7 (1.0)	82.8 (1.0)	88.2 (0.6)	
	100	83.2 (1.0)	82.4 (1.2)	88.2 (0.6)	
	500	82.2 (1.0)	81.1 (1.1)	88.0 (0.7)	
scenario 2	10	88.7 (0.9)	80.9 (1.2)	80.0 (1.0)	89.8 (0.5)
	50	84.2 (1.6)	80.2 (1.1)	80.0 (1.0)	
	100	80.6 (2.0)	79.5 (1.1)	79.9 (0.9)	
	500	74.3 (3.1)	77.9 (1.0)	79.9 (1.0)	
scenario 3	10	93.1 (0.6)	79.7 (1.8)	74.0 (1.4)	94.0 (0.4)
	50	91.4 (0.7)	75.6 (1.5)	73.9 (1.5)	
	100	89.8 (0.8)	74.4 (1.6)	73.9 (1.4)	
	500	81.6 (1.8)	72.0 (1.5)	73.7 (1.4)	
scenario 4	10	92.1 (0.8)	80.1 (1.8)	71.4 (1.3)	94.4 (0.4)
	50	88.9 (1.5)	75.6 (1.5)	71.3 (1.4)	
	100	84.5 (2.0)	74.2 (1.6)	71.4 (1.3)	
	500	76.3 (1.8)	71.4 (1.4)	71.1 (1.4)	
scenario 5	10	92.4 (0.6)	79.4 (1.7)	73.3 (1.3)	94.0 (0.4)
	50	90.4 (0.8)	75.2 (1.6)	73.1 (1.4)	
	100	88.6 (0.8)	74.0 (1.4)	73.0 (1.4)	
	500	80.4 (2.0)	71.4 (1.5)	72.7 (1.3)	

Table 4.3.2 summarizes the prediction performance results under the five simulation scenarios. As the size of p increases, our proposed method improves in all scenarios. In particular, when p is large (e.g., $p = 500$), our proposed method outperforms the other two models significantly in all simulation settings. The c-index of LASSO also increases as p gets larger, but remains unchanged or even slightly decreases when p goes up from 100 to 500. RSF also improves with larger p but its performance is generally lower than the other two methods.

4.3.3 Simulation III: sample size effect on prediction performance

We also evaluate the effect of sample sizes on the prediction performance of the DNN survival model in the presence of large-dimensional predictors. We choose the scenarios 4 and 5 with $p = 100, 500$ under the sparse signal setting from Section 4.3.1. Table 4.3.3 presents the c-index values for each scenario. Overall, for both scenarios, the c-index increases as the sample size increases, and the increment is more dramatic between smaller sample sizes such as from $n = 200$ to 500 or $n = 500$ to $1,000$. This demonstrates that the DNN survival model requires a moderately large sample size (i.e., $n = 1,000$ at least) to achieve satisfactory prediction performance when $p \geq 100$.

4.4 Application to AREDS data

4.4.1 Study population

We apply the three machine learning models for predicting AMD progression using genetic and clinical variables. Data are from the Age-Related Eye Disease Studies (AREDS), which is composed of the first study AREDS [AREDS Group, 1999] and the subsequent study AREDS2 [Chew et al., 2012] (with independent participants), designed to assess risk

Table 4.3.2: The C-index values (range from 0 to 100) over 200 replicates for DNN, RSF and LASSO using weak signals.

	p	DNN	RSF	LASSO
scenario 1	20	66.8 (1.3)	55.5 (5.8)	67.0 (1.4)
	50	73.7 (1.3)	58.2 (7.6)	74.0 (1.2)
	100	78.2 (1.2)	60.8 (7.6)	78.6 (1.1)
	500	82.1 (1.4)	62.9 (4.8)	75.9 (1.5)
scenario 2	20	64.6 (1.6)	53.7 (4.9)	63.2 (1.4)
	50	71.3 (1.3)	57.1 (7.3)	71.8 (1.2)
	100	76.6 (1.2)	60.1 (7.3)	76.9 (1.1)
	500	81.5 (1.3)	62.5 (5.0)	75.9 (1.5)
scenario 3	20	67.4 (1.3)	55.2 (5.9)	67.5 (1.3)
	50	73.2 (1.2)	56.6 (7.7)	73.6 (1.2)
	100	77.7 (1.2)	60.0 (7.8)	78.2 (1.1)
	500	81.8 (1.4)	62.7 (5.0)	75.6 (1.5)
scenario 4	20	65.5 (1.5)	53.5 (5.1)	63.9 (1.4)
	50	71.0 (1.3)	56.3 (7.5)	71.4 (1.2)
	100	76.0 (1.2)	59.3 (8.0)	76.5 (1.2)
	500	81.3 (1.4)	62.4 (5.1)	75.6 (1.5)
scenario 5	20	64.8 (1.3)	54.3 (4.8)	64.8 (1.4)
	50	72.0 (1.2)	56.5 (7.6)	72.4 (1.2)
	100	77.1 (1.2)	59.6 (7.9)	77.6 (1.2)
	500	81.8 (1.4)	62.7 (5.1)	75.1 (1.5)

Table 4.3.3: Effect of sample sizes n on DNN c-index performance in presence of high-dimensional predictors. Both scenarios 4 and 5 are from the sparse signal setting. The predictor sizes are set at $p = 100$ or 500.

		sample size n				
	p	200	500	1,000	1,500	2,000
scenario 4	100	65.4 (4.3)	78.9 (2.0)	84.5 (2.0)	87.8 (1.7)	89.0 (1.6)
	500	57.2 (3.2)	62.3 (3.0)	76.3 (1.8)	79.9 (1.7)	82.4 (1.1)
scenario 5	100	66.7 (4.7)	82.8 (1.8)	88.6 (0.8)	89.6 (0.6)	90.5 (0.5)
	500	58.2 (3.3)	63.2 (3.4)	80.4 (2.0)	86.0 (2.4)	87.8 (1.0)

factors and effects of various supplements for AMD development and progression. Both studies collected DNA samples of consenting participants [Fritsche et al., 2016]. The two studies are combined for the following model development and analysis.

4.4.2 Survival outcome and baseline predictors

To measure the disease progression, a severity score, scaled from 1 to 12 (with a larger value indicating more severe AMD), (with a larger value indicating more severe AMD), is determined for each eye at every examination during study follow-up. In this chapter, our outcome of interest is time-to-late-AMD from the baseline visit, where "late-AMD" is defined as the stage with severity score ≥ 9. There are 30% of subjects progressed to late-AMD before the study ends. We develop prediction models on the individual eye level. There are a total of $7,803$ eyes free of late-AMD at baseline. We include a list of potentail predictors, including age at baseline, smoking status (never, former or current smoker), education status ($\leq$ or $>$ high school) and top SNPs that have been reported to be associated with AMD progression (identified in Yan et al. [2018] with various $p-$ value cut-offs and MAF $> 5\%$). Table 4.4.1 summarizes the baseline characteristics of the study samples. We also pre-process the continuous predictors, for example, dividing age by 100 to scale it within $(0, 1)$ and dividing SNP data (originally coded between $[0, 2]$) by 2 to make them within $[0, 1]$, as we find such a scaling procedure enhances the prediction performance in survival machine learning models.

4.4.3 Model development and evaluation

We perform 10-fold cross-validation in the combined AREDS and AREDS2 data. The splitting is stratified based on the censoring status and study source. For LASSO and RSF, we use the same tuning procedure as in the simulations. For DNN, we first perform a grid search for tuning parameters and select the set of hyperparameters that gives the best average prediction performance (i.e., c-index) across the 10 test validations. The final choice of DNN hyperparameters is described in Section 4.2. We also include a benchmark Genetic Risk Score (GRS) model, which is a regular Cox PH model using age, smoking status, education status and an AMD genetic risk score from Ding et al. [2017].

We first examine the prediction performance, measured by c-index (×100), employing various numbers of top genetic variants across different models. We choose various $p-$value cut-offs from the first AMD progression GWAS paper [Yan et al., 2018] (i.e., $p < 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}$) to obtain different numbers of top variants, as shown in Table 4.4.2. In general, the averaged c-index increases with the number of predictors, and becomes relatively stable when $p < 10^{-5}$, which corresponds to 663 SNPs (plus 3 clinical predictors). We also include in the last column of Table 4.4.2 the DNN model computing time for fitting the complete data once. It can be seen that the computing time does not increase much when the total number of predictors increases. On average, it takes about one hour in the presence of $8,000$ observations and $1,000$ predictors.

Then, we report in Table 4.4.3 c-index, 10-year AUC, and 10-year Brier score (a predictive error measurement) at the cutoff $p < 10^{-5}$. DNN achieves higher c-index (76.1) and AUC (81.8) as well as lower Brier score (0.136) than all the other models.

Table 4.4.1: Baseline features of study samples (eye level) in AREDS and AREDS2.

N=7803		n	Mean (SD) or %
Age			69.5 (6.2)
Gender			
	Female	4466	57%
	Male	3337	43%
Education			
	<= high	2369	30%
	>high	5434	70%
Smoke			
	never	3623	46%
	former	3752	48%
	current	428	6%
Baseline severity score			4.2 (2.5)

Table 4.4.2: The 10-fold cross-validation c-index (×100) from five survival models (GRS, LASSO, Ridge, RSF, DNN) using different $p-$value cutoffs in the AREDS data. The last column shows the DNN computing time by running on the complete dataset.

	Number of predictors	GRS*	LASSO	Ridge	RSF	DNN	Time (minutes)
$p < 10^{-7}$	92	73.2 (1.6)	72.4 (1.7)	72.3 (1.7)	68.5 (1.4)	72.2 (1.8)	49
$p < 10^{-6}$	165	73.2 (1.6)	72.6 (1.5)	72.6 (1.5)	68.2 (1.3)	72.6 (1.6)	47
$p < 10^{-5}$	666	73.2 (1.6)	74.4 (1.3)	74.3 (1.3)	70.1 (1.8)	76.1 (1.2)	62
$p < 10^{-4}$	1500	73.2 (1.6)	75.2 (1.1)	74.8 (1.0)	71.1 (1.7)	76.5 (1.4)	77

*GRS is invariant to the choices of $p-$value cutoffs.

Table 4.4.3: The 10-fold cross-validation c-index (×100), 10-year AUC (×100) and 10-year Brier score from five survival models (GRS, LASSO, Ridge, RSF, DNN) in the AREDS data.

	GRS	LASSO	Ridge	RSF	DNN
c-index (SD)	73.2 (1.6)	74.4 (1.3)	74.3 (1.9)	70.1 (1.8)	76.1 (1.2)
10year-AUC (SD)	78.2 (2.1)	79.5 (1.6)	78.7 (1.5)	74.3 (2.1)	81.8 (2.1)
10year-BrS (SD)	0.151 (0.005)	0.146 (0.006)	0.147 (0.005)	0.170 (0.008)	0.136 (0.011)

Figure 4.4.1 presents the time-dependent Brier scores for the test data under each prediction model. The Brier score profile from our DNN survival model is consistently lower than all the other models across most time points, demonstrating its better performance than the other models. Figure 4.4.2 presents the time-dependent AUC for the test data under each model, as an additional matric to evaluate the model prediction performance. Similar to the time-dependent Brier scores, the AUC profile from our DNN survival model is consistently higher than AUCs of the other models across all time points.

4.4.4 DNN interpretation and subgroup identification

To interpret the DNN-based prediction, we obtain the prediction importance measure for the test data subjects using the LIME method under our DNN survival model. We use 9-folds data to train a DNN model and then interpret the model in the rest 1-fold test data. One big advantage of the LIME method is that it provides a subject-specific interpretation of predictor importance. Figure 4.4.3 illustrates the top clinical and genetic predictors (named by their corresponding gene names). Among the top predictors, (older) age and smoking are harmful (colored in red) to AMD progression, whereas genetic variants (carrying minor alleles) can be either harmful (red) or protective (green). For example, the minor allele of

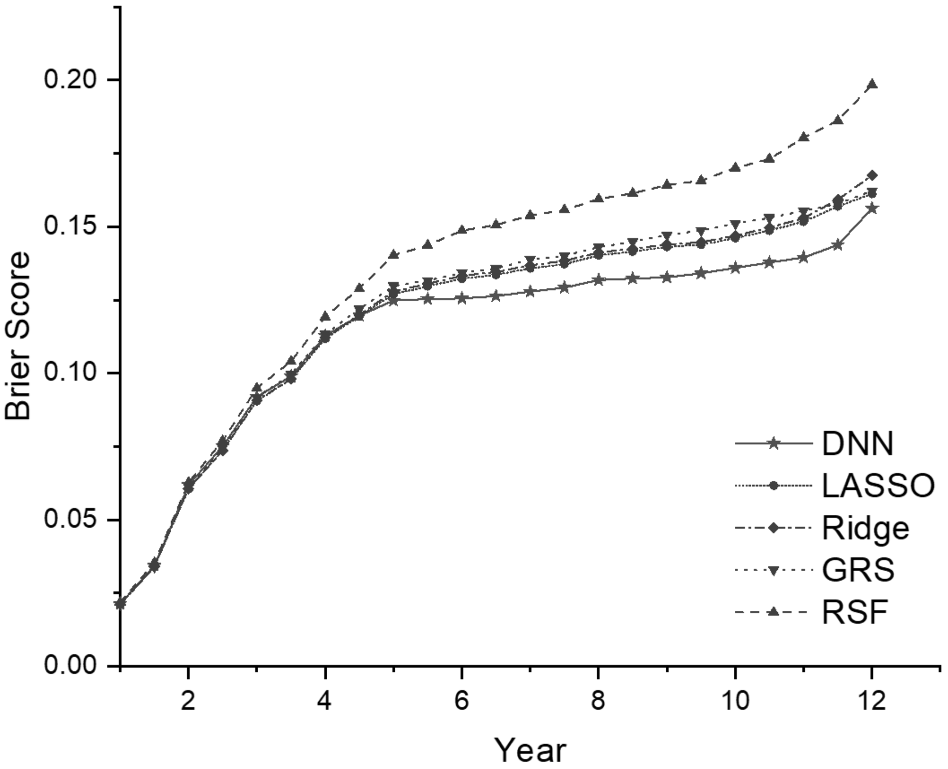

Figure 4.4.1: The time-dependent Brier scores (predictive errors) in the test data from four survival prediction models (GRS, LASSO, Ridge, RSF, DNN).

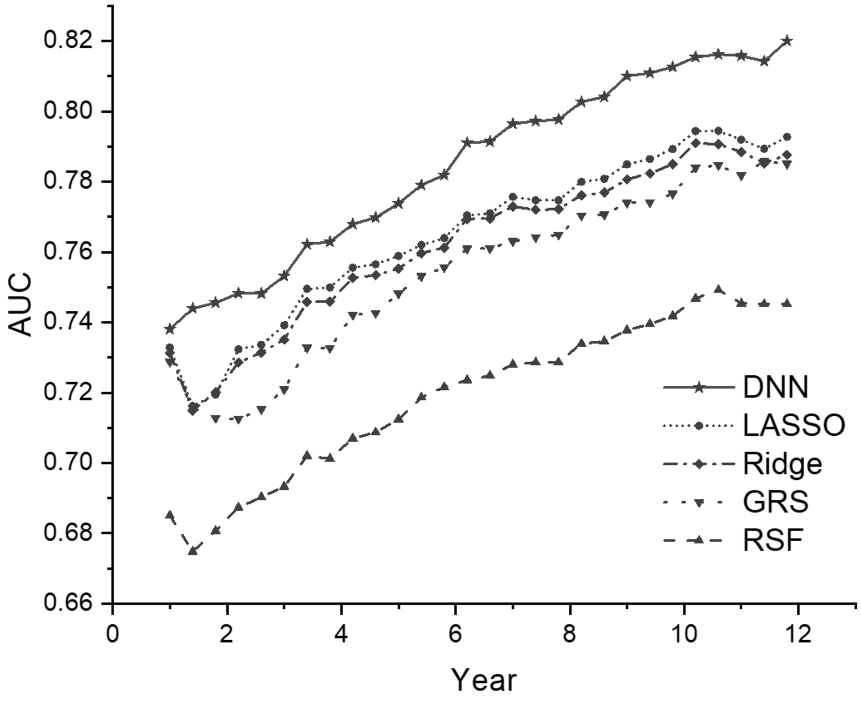

Figure 4.4.2: The time-dependent AUC values in the test data from four survival models (GRS, LASSO, Ridge, RSF, DNN).

$rs10922098$ in the *CFH* gene region shows a protective effect for AMD progression; while the minor allele of $rs12987936$ in the *CROCC2* gene region shows a harmful effect for AMD progression. Moreover, we notice that one predictor could be important for some subjects but may not be crucial for others (visualized by different vertical color bands within each predictor), which suggests there are possible heterogeneous subgroups in this population.

Motivated by the heterogeneity across subjects shown in Figure 4.4.3, we further identify two distinct subgroups of AMD patients by performing the Gaussian Mixture Model on the predicted risk function $\hat{g}$ (output from the DNN model), as illustrated in the histogram of Figure 4.4.4. The corresponding Kaplan-Meier plot on progression probability indicates significantly different progression profiles between the two subgroups (namely, the low-risk and high-risk subgroups), with a very significant log-rank test result ($p = 4.1 \times 10^{-166}$). Further, we find significant differences between the two subgroups in terms of age, smoking status, education level and most top genetic variants in Figure 4.4.3. The comparison results are summarized in Table 4.4.4. On average, the high-risk patients are older, with more smokers and lower education level compared to the low-risk patients. The high-risk patients also carry more AMD progression risk alleles compared to the low-risk patients (e.g., GRS is 1.07 vs 0.94). Moreover, as shown in Figure 4.4.5, the separate LIME plots for the two subgroups also demonstrate that the individual predictors' importance measures are different between the two subgroups. In particular, the harmful predictors generally have stronger influence (darker in red) in the high-risk subgroup than in the low-risk subgroup; whereas the protective predictors show stronger impacts (darker in green) in the low-risk subgroup than the high-risk subgroup. These results provide potentially useful insights for the early prevention and tailored clinical management for the AMD patients.

4.4.5 Data availability

Both phenotype and genotype data of AREDS and AREDS2 are available from the online repository dbGap (accession: $phs000001.v3.p1$, and $phs001039.v1.p1$, respectively).

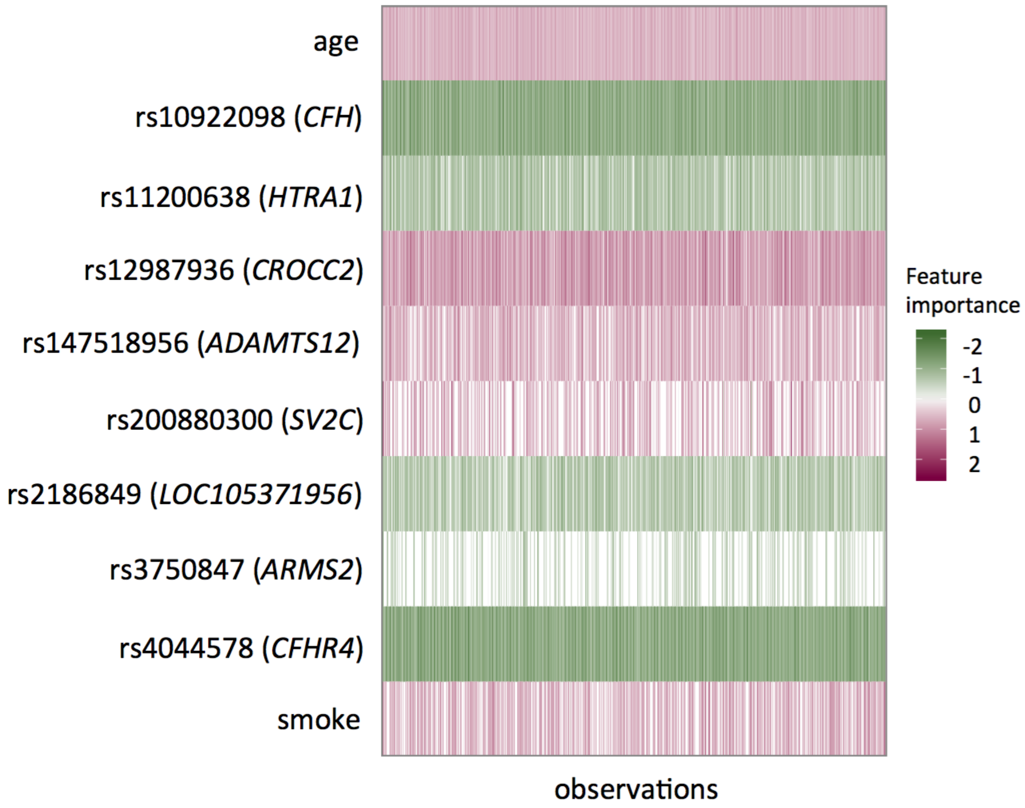

Figure 4.4.3: The representation of importance measures for the top predictors in the test data.

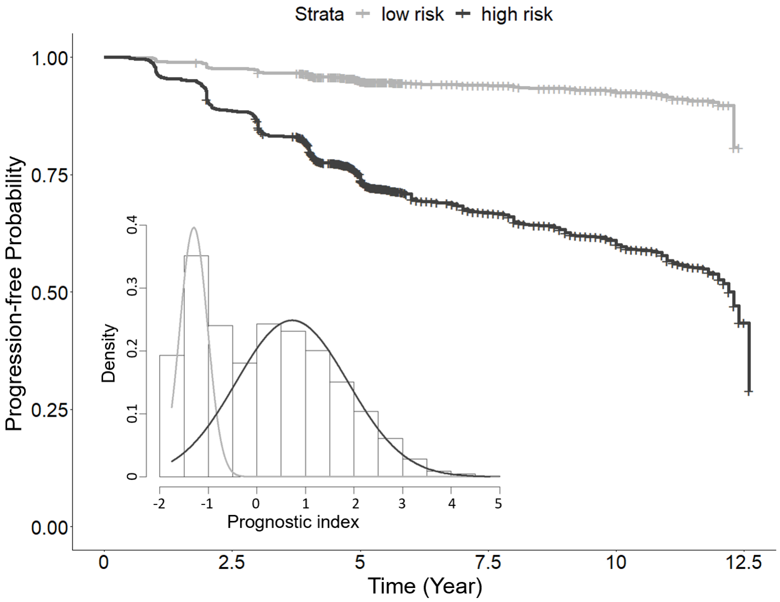

Figure 4.4.4: The Kaplan-Meier (KM) estimated survival profiles for the two identified subgroups in the AREDS and AREDS2 test data.

4.5 Discussion and conclusion

In this work, we implement a multi-layer DNN survival model and successfully apply it on a real data set with both large n and large p to examine and evaluate its effectiveness in making accurate dynamic survival predictions and detecting clinically meaningful subgroups. To open up the "black-box" of DNN, a novel LIME algorithm is implemented to calculate an importance measure of each predictor for each observation. Moreover, our work demonstrates the power of DNN in the presence of various types of complex non-linear structures in the predictors through extensive simulation studies. As we did not perform hyperparameter tuning separately for each scenario, further enhanced performance of DNN would be expected if separate tuning was performed. Our work presents the first deep learning survival prediction model for AMD progression prediction and the model framework can be readily applied to other progressive disorders where large GWAS or omics data are collected.

We evaluate survival models based on the pooled data set of AREDS and AREDS2,

Table 4.4.4: Comparison between the low-risk ($n = 2,516$) and high-risk ($n = 5,287$) subgroups identified by DNN in AREDS and AREDS2 data.

		Low-risk subgroup Mean (sd) or n (%) or risk allele frequency	High-risk subgroup Mean (sd) or n (%) or risk allele frequency	p-values
Predictors:				
Age		66.1 (5.4)	71.1 (5.9)	$< 2.2 \times 10^{-16}$
Smoke				$< 2.2 \times 10^{-16}$
	never	1343 (53%)	2280 (43%)	
	former	1088 (43%)	2664 (50%)	
	current	85 (3%)	343 (6%)	
Education				
	<=high school	625 (25%)	1744 (33%)	3.2×10^{-13}
	>high school	1891 (75%)	3543 (67%)	
rs10922098 (*CFH*)		0.34	0.61	$< 2.2 \times 10^{-16}$
rs11200638 (*HTRA1*)		0.17	0.39	$< 2.2 \times 10^{-16}$
rs12987936 (*CROCC2*)		0.18	0.18	0.35
rs147518956 (*ADAMTS12*)		0.27	0.32	1.8×10^{-13}
rs200880300 (*SV2C*)		0.04	0.06	1.0×10^{-3}
rs2186849 (*LOC105371956*)		0.47	0.50	1.0×10^{-3}
rs3750847 (*ARMS2*)		0.17	0.40	$< 2.2 \times 10^{-16}$
rs4044578 (*CFHR4*)		0.33	0.62	$< 2.2 \times 10^{-16}$
Other characteristics:				
GRS		0.94 (0.13)	1.07 (0.13)	$< 2.2 \times 10^{-16}$
Gender				
	Female	1439 (57)	3027 (57)	0.98
	Male	1077 (43)	2260 (43)	
Baseline severity		3.0 (2.3)	4.7 (2.4)	$< 2.2 \times 10^{-16}$

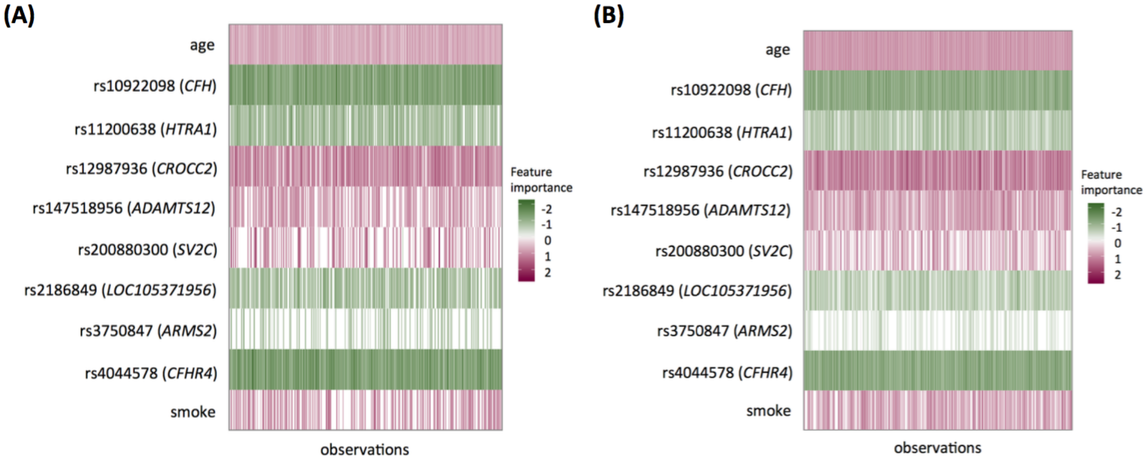

Figure 4.4.5: The representation of importance measures for the top predictors in the low-risk (A) and high-risk (B) subgroups, respectively.

whereas Ding et al. [2017] used AREDS as the training data and AREDS2 as the test data. However, as noted by Ding et al. [2017], AREDS and AREDS2 populations are different in multiple aspects such as disease severity and age (at enrollment). As a result, the top significant SNPs identified by GWAS are largely non-overlapping between the two studies [Yan et al., 2018]. As expected, in Ding et al. [2017], the GRS-based Cox model trained in AREDS achieves a c-index of 0.75 in AREDS but drops to 0.63 in AREDS2. To establish a prediction model that is generalizable to a broader AMD population, we pooled them together. Unsurprisingly, the benchmark GRS model performance improves to 0.73 in terms of c-index, as shown in Table 4.4.3.

We predict disease progression on the eye level by assuming that the two eyes are independent of each other in one individual. Potential future extensions include using a copula model to take the dependence between the two eyes from the same subject into account for the deep learning survival model and predicting the joint progression profiles of the two eyes [Sun and Ding, 2019, Sun et al., 2019].

One potential limitation of the DNN survival model is that it involves tuning of multiple hyperparameters, which is usually computationally expensive. According to our real data analysis and simulations, we could heuristically start from a two-hidden-layer DNN and perform a grid search for the other tuning parameters such as the optimal node size. In general, the DNN model size should be moderate to avoid overfitting. Moreover, the utilization of GPUs could significantly boost the computing speed of our DNN survival model. To further improve the DNN survival model, there are multiple future directions. For example, one may first obtain low-dimensional signals by performing unsupervised feature extraction such as autoencoder [Vincent et al., 2010] and then use the extracted signals as predictors. In this way, the noises in the original data can be greatly reduced. Another possible extension is to build a DNN survival model based on the Bayesian approach [Liang et al., 2018], which could perform variable selection to identify relevant predictors under the high-dimensional non-linear setting.

5.0 Summary and Future Direction

In Chapter 2, I have developed a flexible copula-based semiparametric transformation model for bivariate interval-censored data. In many applications, there are more than two correlated events. I am interested in extending my current work to modeling multivariate interval-censored data using copula models. I am also interested in developing methods for modeling covariate-varying dependence parameters in copula. For example, the dependence strength between two events can vary by age. More excitingly, I am interested in developing a fully semiparametric copula framework with unspecified dependence structure, instead of a specific parametric copula function. For the IR-based goodness-of-fit test proposed in Chapter 3, I plan to extend it from bivariate to multivariate censored data.

In Chapter 4, we have developed and implemented a deep-learning-based prediction model for survival outcomes using genetic data. In the future, I am planning to incorporate different types of high-dimensional features (including genetic, genomic, and imaging data) into the current deep learning framework. I am also interested in predicting various progressive diseases (e.g., cancer, Alzheimers disease) by utilizing wealthy public repositories (e.g., UK Biobank, dbGap, GEO, ADNI, TopMed, All of Us). I also want to extend deep learning methodology to predict other commonly seen settings, such as interval censoring, correlated survival outcomes, and competing risk events. Moreover, I am also interested in developing deep-learning-based methods for identifying novel predictive biomarkers, which interact with treatments. All these new methods have the high potential to facilitate early prevention and precision medicine.

Appendix A

Supplementary materials for Chapter 2

A.1 Additional simulation studies

In the real setting, the value of the transformation function parameter r is often unknown. Therefore, we examined our methods in estimating the transformation function parameter r together with the other parameters in our proposed model. We used the logarithmic transformation function for the Clayton Loglogistic data and the Box-Cox transformation function for the Clayton Weibull data. Table A.1.1 shows satisfactory estimation performance for all parameters including the transformation parameter in both proportional hazards and proportional odds settings.

We also simulated bivariate current status data by setting $K = 1$ to examine how the proposed method works in the particular case of case 1 interval censoring. As shown in Table A.1.2, Copula2-S works as well as the true model in this setting too. The larger standard errors are due to less information in current status data as compared to the standard case 2 interval censoring case in Table 2.4.1 of the main text.

A.2 Secondary real data analysis

We performed the secondary real data analysis that models the Group A subjects (i.e. subjects who were free of late-AMD at time 0) by a two-parameter copula sieve model with the same settings ($m = 4, r = 3$) as the two-part model from the main text. The covariates are the baseline age, severity score and smoking status. We performed genome-wide association tests and summarized the top identified SNPs in the Table A.2.1. The top identified gene regions are consistent with the ones identified from the two-part model.

Table A.1.1: Estimation results using the proposed model when the transformation function parameter r is unspecified.

	proportional odds			proportional hazards		
Param	Bias	SE	SEE (CP)	Bias	SE	SEE (CP)
β_{ng1}	-0.0011	0.0169	0.0159 (0.934)	-0.0004	0.0109	0.0104 (0.942)
β_{ng2}	0.0111	0.1289	0.1278 (0.938)	0.0032	0.0816	0.0806 (0.945)
β_g	-0.0030	0.0902	0.0923 (0.961)	0.0035	0.0584	0.0582 (0.952)
r	-0.0384	0.1294	0.1274 (0.919)	0.0463	0.1532	0.1574 (0.952)
τ	-0.0008	0.0224	0.0216 (0.944)	-0.0006	0.0225	0.0227 (0.946)

Table A.1.2: Estimation results for bivariate current status data.

	True			Copula2-S			Robust-S		
Param	Bias	SE	SEE (CP)	Bias	SE	SEE (CP)	Bias	SE	SEE (CP)
proportional odds									
β_{ng1}	0.0031	0.0399	0.0394 (0.956)	0.0033	0.0400	0.0393 (0.958)	0.0002	0.0536	0.0538 (0.942)
β_{ng2}	-0.0203	0.2563	0.2516 (0.948)	-0.0219	0.2563	0.2527 (0.946)	-0.0249	0.2646	0.2608 (0.946)
β_g	0.0002	0.1819	0.1816 (0.947)	-0.0001	0.1808	0.1822 (0.944)	-0.0008	0.1855	0.1938 (0.948)
τ	-0.0038	0.0587	0.0575 (0.947)	0.0025	0.0680	0.0660 (0.939)	NA	NA	NA
proportional hazards									
β_{ng1}	-0.0008	0.0330	0.0321 (0.947)	-0.0005	0.0333	0.0322 (0.946)	0.0003	0.0397	0.0420 (0.944)
β_{ng2}	0.0001	0.1785	0.1813 (0.949)	0.0025	0.1831	0.1830 (0.945)	0.0030	0.1857	0.1958 (0.956)
β_g	-0.0043	0.1298	0.1312 (0.959)	-0.0052	0.1319	0.1324 (0.959)	-0.0038	0.1346	0.1406 (0.956)
τ	-0.0013	0.0637	0.0628 (0.943)	0.0029	0.0682	0.0665 (0.931)	NA	NA	NA

Table A.2.1: The top identified SNPs from the secondary real data analysis.

SNP	Chr	Gene	MAF	OR	p (Copula2-S)	p (Frailty-S)	p (Robust-S)
rs10922109	1	*CFH*	0.28	0.61	7.6×10^{-9}	5.6×10^{-8}	3.6×10^{-7}
rs1329427	1	*CFH*	0.28	0.61	8.4×10^{-9}	6.3×10^{-8}	4.0×10^{-7}
rs10801559	1	*CFH*	0.28	0.61	9.4×10^{-9}	6.9×10^{-8}	4.4×10^{-7}
rs1410996	1	*CFH*	0.28	0.62	1.1×10^{-8}	7.6×10^{-8}	5.0×10^{-7}
rs2284665	10	*HTRA1*	0.33	1.54	1.6×10^{-8}	5.3×10^{-7}	4.3×10^{-6}
rs2293870	10	*ARMS2-HTRA1*	0.33	1.53	1.6×10^{-8}	4.3×10^{-7}	7.5×10^{-6}
rs61871744	10	*ARMS2-HTRA1*	0.33	1.54	1.6×10^{-8}	1.7×10^{-7}	3.4×10^{-6}
rs3763764	10	*ARMS2-HTRA1*	0.34	1.53	1.7×10^{-8}	4.0×10^{-7}	8.5×10^{-6}
rs28368872	16	*ATF7IP2*	0.13	1.70	1.2×10^{-7}	1.4×10^{-5}	1.9×10^{-5}
rs12708701	16	*ATF7IP2*	0.13	1.71	1.5×10^{-7}	1.2×10^{-5}	1.5×10^{-6}

Although the p-values are all larger compared to the primary analysis result through the two-part model (due to smaller sample size), our Copula2-S method still yielded smaller p-values compared to the other two methods.

A.3 Similarities and differences between copula and frailty models

Both copula and frailty methods are popular for modeling bivariate survival data. The two models sometimes share similarities in their mathematical expressions. For example, the joint survival function under the Clayton copula is written as $S_c(t_1, t_2) = [\{S_{1,c}(t_1)\}^{-\eta} + \{S_{2,c}(t_2)\}^{-\eta} - 1]^{-1/\eta}$, where $S_{j,c}(t_j)$ is the marginal survival function for margin j. On the other hand, the Gamma frailty joint survival function has a similar form:

$$S_f(t_1, t_2) = \mathcal{L}_\rho[\mathcal{L}_\rho^{-1}\{S_{1,f}(t_1)) + \mathcal{L}_\rho^{-1}(S_{1,f}(t_2)\}] = [\{S_{1,f}(t)\}^{-\rho} + \{S_{2,f}(t)\}^{-\rho} - 1]^{-1/\rho},$$

where $\mathcal{L}_\rho(s) = (1 + \rho s)^{-1/\rho}$ is the Laplace function on density function of the random frailty term, which follows an exponential distribution with unit mean and variance ρ; $S_{j,f}(t_j)$ is the marginal survival function under the Gamma frailty setting. Assuming a Weibull margin as an example, then the marginal survival function under Clayton copula model is

$$S_{j,c}(t) = \exp\{-(t/\lambda_j)^{k_j} e^{z_j\beta_j}\},$$

while the marginal survival function under Gamma frailty model becomes

$$S_{j,f}(t) = \mathcal{L}_\rho((t/\tilde{\lambda}_j)^{\tilde{k}_j} e^{z_j\tilde{\beta}_j}) = \{1 + \rho(t/\tilde{\lambda}_j)^{\tilde{k}_j} e^{z_j\tilde{\beta}_j}\}^{-1/\rho}.$$

It can be seen that $S_c(t_1, t_2) = S_f(t_1, t_2)$ only when both η and $\rho \to 0$. Therefore, the joint survival functions of Clayton-Weibull and Gamma-frailty-Weibull models are actually different due to their distinct marginal survival functions.

For another example, the Gumbel copula has the joint survival function as $S_c(t_1, t_2) = \exp\{-[(-\log\{S_{1,c}(t_1)\})^\eta + (-\log\{S_{2,c}(t_2)\})^\eta]^{1/\eta}\}$, whereas the Positive Stable frailty has the joint survival function as

$$\begin{aligned} S_f(t_1, t_2) &= \mathcal{L}_\rho[\mathcal{L}_\rho^{-1}\{S_{1,f}(t_1)) + \mathcal{L}_\rho^{-1}(S_{1,f}(t_2)\}] \\ &= \exp(-[\{-\log S_{1,f}(t_1)\}^\rho + \{-\log S_{2,f}(t_2)\}^\rho]^{1/\rho}), \end{aligned}$$

where $\mathcal{L}_\rho(s) = \exp(-s^{1/\rho})$ is the corresponding Laplace function when the random frailty term follows a Positive Stable distribution. If assume Weibull for its conditional hazard function, then the marginal survival function becomes

$$S_{j,f}(t) = \mathcal{L}_\rho((t/\tilde{\lambda}_j)^{\tilde{k}_j} e^{z_j\tilde{\beta}_j}) = \exp\{-(t/\tilde{\lambda}_j)^{\tilde{k}_j/\rho} e^{z_j\tilde{\beta}_j/\rho}\},$$

which follows a new Weibull distribution. The marginal survival function under the Gumbel copula is still $S_{j,c}(t) = \exp\{-(t/\lambda_j)^{k_j} e^{z_j\beta_j}\}$. It is easy to observe that the Gumbel-Weibull and Positive-Stable-Weibull models are equivalent given $\lambda_j = \tilde{\lambda}_j, k_j = \tilde{k}_j/\rho, \beta_j = \tilde{\beta}_j/\rho, \eta = \rho$. More discussions about the similarities and differences between the copula and frailty models can be found in Goethals et al. [2008] and Wienke [2010].

A.4 Asymptotic properties

This section presents the regularity conditions and the proofs of Theorems 2.3.1 and 2.3.2 being shown in the main paper. The proofs make use of six lemmas and one general theorem, which are stated and proved in Appendix A.5 and A.6, respectively.

First, we state the regularity conditions needed for Theorems 2.3.1 and 2.3.2.

Condition 1. (i) There exists $\tau_j > 0$ such that $pr(R_j - L_j \geq \tau_j) = 1, j = 1, 2$; (ii) The union of the supports for distributions of $L'_{ij}s$ and $R'_{ij}s$ is an interval $[c, u]$ with $0 < c < u < \infty$.

Condition 2. The distribution of covariate Z_j has a bounded support and is not concentrated on any proper subspace of $\mathbb{R}^{p_j}$, $j = 1, 2$, p_j is dimension of Z_j.

Condition 3. Let $L(\beta, \alpha, \kappa, y_1, y_2)$ be the likelihood function with Λ_j being substituted by y_j. Define

$$v^T \dot{L}(\beta, \alpha, \kappa, y_1, y_2) = v_1^T \frac{\partial L}{\partial \beta} + v_2 \frac{\partial L}{\partial \alpha} + v_3 \frac{\partial L}{\partial \kappa} + v_4 \frac{\partial L}{\partial y_1} + v_5 \frac{\partial L}{\partial y_2},$$

with $v = (v_1^T, v_2, v_3, v_4, v_5)^T$. There exist $l_j^*, r_j^* \in [c, u]$ for which there are $p + 4$ different sets of (z_1, z_2) such that if $v^T \dot{L}(\beta_0, \alpha_0, \kappa_0, \Lambda_{10}, \Lambda_{20}; D^*) = 0$ with $D^* = \{l_j^*, r_j^*, z_j\}$ for each of these $p + 4$ sets of values, then $v = 0_{(p+4)\times 1}$.

Condition 4. (i) The function Λ_{j0} is continuously differentiable up to order q with $q \geq 3$ in $[c, u]$ and satisfies $\xi^{-1} < \Lambda_{j0}(c) < \Lambda_{j0}(u) < \xi$ for some positive constant $\xi, j = 1, 2$. Also $(\beta_0^T, \alpha_0, \kappa_0)^T$ is an interior point of $\mathcal{B} \subseteq \mathbb{R}^p \times \mathbb{R}^{(0,1]} \times \mathbb{R}^+$. (ii) The transformation G_j is a strictly increasing function with $G_j(0) = 0$ and is three-times continuously differentiable in $[0, u], j = 1, 2$.

Condition 5. For every θ in a neighborhood of θ_0, $P\{l(\theta; D) - l(\theta_0; D)\} \lesssim -d^2(\theta, \theta_0)$, where $l(\theta; D)$ being the log-likelihood function given in Section 2.3 and $\lesssim$ means that "the left-hand side is bounded above by a constant times the right-hand side".

Remark. *Conditions 1, 2, 4(i) and 5 are commonly used in the studies of interval-censored data (e.g., Huang and Rossini, 1997, Wen and Chen, 2013, Zhou et al., 2017). Condition 4(ii) comes from the definition of the linear transformation model (Cheng et al., 1995).*

Condition 3 ensures both the identifiability of the parameters and the positivity of the efficient Fisher information matrix (Chang et al., 2007, Wen and Chen, 2013, Zhou et al., 2017).

Remark. *Before the proofs, we need to remark that although our Theorems 2.3.1 and 2.3.2 may look familiar among existing literature, we have to deal with challenges from the complicated data structure and likelihood function. For instance, unlike the right-censored data, the modeling of interval-censored data is more difficult due to no exact observed times and thus common theoretical tools, like the counting process theory, cannot be applied. As a result, we turn to the modern empirical process theory (such as van der Vaart and Wellner [1996]) for proving the theorems. In our case, our likelihood involves both the complicated copula dependence structure and two unknown infinite dimensional nuisance parameters. For the convergence rate, we prove it by verifying the conditions in Theorem 1 from Shen and Wong [1994]. To establish the asymptotic normality, we first prove a general theorem on the asymptotic normality of semiparametric M-estimators with two nuisance parameters, which is completely novel as shown in Appendix A.6. Then we prove the asymptotic normality of our sieve estimators through verifying the conditions of the general theorem. This general theorem can be readily extended to the case with more than two nuisance parameters. Furthermore, our proof procedure for convergence rate and asymptotic normality applies to any Archimedean family copula, not limited to the two-parameter copula model as we consider in this work.*

Proof of Theorem 2.3.1. We will derive the convergence rate by verifying the Conditions C1–C3 of Theorem 1 from Shen and Wong [1994]. Define $\Theta^q = \mathcal{B} \otimes \mathcal{M}^q \otimes \mathcal{M}^q$, where $\mathcal{M}^q$ is the collection of $\Lambda_j, j = 1, 2$ with smoothness q as defined in our Condition 4. Similarly, Θ_n^q is the corresponding sieve space containing $\mathcal{M}_n^q$. Then the Condition C1 automatically holds due to our Condition 5, which states for any $\theta \in \Theta_n^q$, $P\{l(\theta_0; D) - l(\theta; D)\} \gtrsim d^2(\theta, \theta_0)$. Next we verify the Condition C2 of Shen and Wong [1994]. Similar to the arguments in the proof of Lemma A3 (using our Conditions 1, 2, 4), we can show that for any $\theta \in \Theta_n^q$,

$$\begin{aligned} |l(\theta; D) - l(\theta_0; D)| &\lesssim |b - b_0| + |\Lambda_1(L_1) - \Lambda_{10}(L_1)| + |\Lambda_1(R_1) - \Lambda_{10}(R_1)| \\ &+ |\Lambda_2(L_2) - \Lambda_{20}(L_2)| + |\Lambda_2(R_2) - \Lambda_{20}(R_2)|, \end{aligned}$$

where $D = (L_j, R_j, Z_j), j = 1, 2$. Then, it follows that for any $\theta \in \Theta_n^q$

$$P\{l(\theta; D) - l(\theta_0; D)\}^2 \lesssim |b - b_0|^2 + P\big[\{\Lambda_1(l_1) - \Lambda_{10}(l_1)\}^2 + \{\Lambda_1(r_1) - \Lambda_{10}(r_1)\}^2\big] \\ + P\big[\{\Lambda_2(l_2) - \Lambda_{20}(l_2)\}^2 + \{\Lambda_2(r_2) - \Lambda_{20}(r_2)\}^2\big] = d^2(\theta, \theta_0).$$

It implies that

$$\sup_{\{d(\theta,\theta_0)\leq\epsilon, \theta\in\Theta_n^q\}} var\{l(\theta_0; D) - l(\theta; D)\} \leq \sup_{\{d(\theta,\theta_0)\leq\epsilon, \theta\in\Theta_n^q\}} P\{l(\theta_0; D) - l(\theta; D)\}^2 \lesssim \epsilon^2.$$

Thus, Condition C2 from Shen and Wong [1994] holds (with $\beta = 1$ in their notation). Finally, we verify the Condition C3 in Shen and Wong [1994]. By Lemma A3, for $\mathcal{F}_n = \{l(\theta; D) - l(\theta_{0,n}; D) : \theta \in \Theta_n^q\}$, we have $N_{[\,]}(\epsilon, \mathcal{F}_n, \|\cdot\|_\infty) \lesssim (1/\epsilon)^{cm_n+d}$, with $d = p+2$ being the dimensionality for $b = (\beta^T, \alpha, \kappa)^T$. Using the fact that the covering number is bounded by the bracketing number, it follows that

$$H(\epsilon, \mathcal{F}_n, \|\cdot\|_\infty) = \log N_{[\,]}(\epsilon, \mathcal{F}_n, \|\cdot\|_\infty) \lesssim (cm_n + d)\log(1/\epsilon) \lesssim n^\nu \log(1/\epsilon).$$

Hence, the Condition C3 of Shen and Wong [1994] in page 583 holds when the constants $2r_0 = \nu$ and $r = 0^+$ in their notations.

Therefore, the constant τ in Theorem 1 of Shen and Wong [1994] on page 584 is $(1 - \nu)/2 - \{\log(\log n)\}(2\log n)^{-1}$. Since $\{\log(\log n)\}(2\log n)^{-1} \to 0$ as $n \to 0$, we can pick a $\tilde{\nu}$ slightly larger than ν such that $(1 - \tilde{\nu})/2 \leq (1 - \nu)/2 - \{\log(\log n)\}(2\log n)^{-1}$ for large n. We still denote $\tilde{\nu}$ as ν so that $\tau = (1 - \nu)/2$. Since $\hat{\theta}_n$ maximizes $\mathbb{P}_n l(\theta; D)$ over Θ_n^q, so $\hat{\theta}_n$ satisfies the inequality (1.1) in Shen and Wong [1994] when $\eta_n = 0$ in their notation. By Lemma A2, there exists a $\Lambda_{j0,n} \in \mathcal{M}_n^q$ such that $\|\Lambda_{j0,n} - \Lambda_{j0}\|_\infty = O(n^{-q\nu/2})$. Thus, the sieve approximation error $\rho(\pi_n\theta_0, \theta_0)$ in Shen and Wong [1994] is $O(n^{-q\nu/2})$. In addition, since $Pl(\theta; D)$ is maximized at θ_0, its first derivative at θ_0 is equal to 0. Then, applying the Taylor expansion for $P\{l(\theta_0; D) - l(\theta; D)\}$ around θ_0 and plugging in $\theta = \theta_{0n} =$

$(\beta_0^T, \alpha_0, \kappa_0, \Lambda_{10n}, \Lambda_{20n})^T$, the Kullback-Leilber pseudodistance of $\theta_0 = (\beta_0^T, \alpha_0, \kappa_0, \Lambda_{10}, \Lambda_{20})^T$ and $\theta_{0,n} = (\beta_0^T, \alpha_0, \kappa_0, \Lambda_{10,n}, \Lambda_{20,n})^T$ follows

$$\begin{aligned}
&K(\theta_0, \theta_{0,n}) = -P\{l(\theta_{0,n}; D) - l(\theta_0; D)\} \\
=& -\frac{1}{2}P\big\{\ddot{l}_{\Lambda_1\Lambda_1}(\theta_0; D)[\Lambda_{10,n} - \Lambda_{10}, \Lambda_{10,n} - \Lambda_{10}] + \ddot{l}_{\Lambda_2\Lambda_2}(\theta_0; D)[\Lambda_{20,n} - \Lambda_{20}, \Lambda_{20,n} - \Lambda_{20}] \\
&+ 2\ddot{l}_{\Lambda_1\Lambda_2}(\theta_0; D)[\Lambda_{10,n} - \Lambda_{10}, \Lambda_{20,n} - \Lambda_{20}]\big\} + o(d^2(\theta_{0,n}, \theta_0)) \\
\lesssim& \|\Lambda_{10,n} - \Lambda_{10}\|_2^2 + \|\Lambda_{20,n} - \Lambda_{20}\|_2^2 + o(\|\Lambda_{10,n} - \Lambda_{10}\|_2^2 + \|\Lambda_{20,n} - \Lambda_{20}\|_2^2) \lesssim O(n^{-q\nu}).
\end{aligned}$$

The second last inequality holds due to boundness of all second order derivatives of log-likelihood in Lemma A1 as well as derivations similar to Lemma A3. The last inequality holds since $\|\Lambda_{j0,n} - \Lambda_{j0}\|_2 \leq \|\Lambda_{j0,n} - \Lambda_{j0}\|_\infty = O(n^{-q\nu/2}), j = 1, 2$.
Therefore, $K^{1/2}(\theta_0, \theta_{0n}) = O(n^{-q\nu/2})$. Hence, by Theorem 1 of Shen and Wong [1994], we obtain the convergence rate for $\hat{\theta}_n$ as

$$d(\hat{\theta}_n, \theta_0) = O_p\big\{\max(n^{-(1-\nu)/2}, n^{-q\nu/2}, n^{-q\nu/2})\big\} = O_p\big(n^{-\min\{q\nu/2,(1-\nu)/2\}}\big).$$

This completes the proof of our Theorem 2.3.1. □

Proof of Theorem 2.3.2. We will prove the theorem by checking assumptions A1-A6 of the general theorem in the Appendix A.6. We can verify the assumption A1 by applying our Theorem 2.3.1 with $\gamma = \min\{q\nu/2, (1-\nu)/2\}$ and the L_2 norm. A2 also automatically holds under the model assumption. For the assumption A3, we need to verify both existence of h_j^* and nonsingularity of the matrix A. Following similar arguments as in Wen and Chen [2013] (page 402-405), the existence of h_j^* can be verified, which satisfies that for all $h_j \in \mathcal{M}^{q-1}$,

$$\begin{aligned}
&P\{\ddot{l}_{b\Lambda_1}(b_0, \Lambda_{10}, \Lambda_{20})[h_1] + \ddot{l}_{b\Lambda_2}(b_0, \Lambda_{10}, \Lambda_{20})[h_2] - \ddot{l}_{\Lambda_1\Lambda_1}(b_0, \Lambda_{10}, \Lambda_{20})[h_1^*, h_1] \\
&- \ddot{l}_{\Lambda_1\Lambda_2}(b_0, \Lambda_{10}, \Lambda_{20})[h_1^*, h_2] - \ddot{l}_{\Lambda_2\Lambda_2}(b_0, \Lambda_{10}, \Lambda_{20})[h_2^*, h_2] - \ddot{l}_{\Lambda_2\Lambda_1}(b_0, \Lambda_{10}, \Lambda_{20})[h_2^*, h_1]\} = 0.
\end{aligned}$$

For any $h_j, \tilde{h}_j \in \mathcal{M}^{q-1}$, we have the following results:

$$\begin{aligned}
&P\ddot{l}_{bb}(b, \Lambda_1, \Lambda_2; D) = -P\{\dot{l}_b(b, \Lambda_1, \Lambda_2; D)\dot{l}_b^T(b, \Lambda_1, \Lambda_2; D)\}, \\
&P\ddot{l}_{b\Lambda_j}(b, \Lambda_1, \Lambda_2; D)[h_j] = P\ddot{l}_{\Lambda_j b}(b, \Lambda_1, \Lambda_2; D)[h_j] = -P\{\dot{l}_b(b, \Lambda_1, \Lambda_2; D)\dot{l}_{\Lambda_j}(b, \Lambda_1, \Lambda_2; D)[h_j]\}, \\
&P\ddot{l}_{\Lambda_j\Lambda_j}(b, \Lambda_1, \Lambda_2; D)[h_j, \tilde{h}_j] = -P\{\dot{l}_{\Lambda_j}(b, \Lambda_1, \Lambda_2; D)[h_j]\dot{l}_{\Lambda_j}(b, \Lambda_1, \Lambda_2; D)[\tilde{h}_j]\}, \\
&P\ddot{l}_{\Lambda_j\Lambda_{j'}}(b, \Lambda_1, \Lambda_2; D)[h_j, h_{j'}] = -P\{\dot{l}_{\Lambda_j}(b, \Lambda_1, \Lambda_2; D)[h_j]\dot{l}_{\Lambda_{j'}}(b, \Lambda_1, \Lambda_2; D)[h_{j'}]\},
\end{aligned}$$

where $j, j' \in \{1, 2\}$, and all first and second order derivatives are defined in the Lemma A1. Then, together with assumption A3 and the above equations, the matrix A follows

$$\begin{aligned}
A &= -P\{\ddot{l}_{bb}(b_0, \Lambda_{10}, \Lambda_{20}; D) - \ddot{l}_{\Lambda_1 b}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_1^*] - \ddot{l}_{\Lambda_2 b}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_2^*]\} \\
&= P\{\dot{l}_b(b_0, \Lambda_{10}, \Lambda_{20}; D) - \dot{l}_{\Lambda_1}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_1^*] - \dot{l}_{\Lambda_2}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_2^*]\}^{\otimes 2} \\
&= Pl^*(b_0, \Lambda_{10}, \Lambda_{20}; D)^{\otimes 2},
\end{aligned}$$

where $l^*(b_0, \Lambda_{10}, \Lambda_{20}; D) = \dot{l}_b(b_0, \Lambda_{10}, \Lambda_{20}; D) - \dot{l}_{\Lambda_1}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_1^*] - \dot{l}_{\Lambda_2}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_2^*]$. Therefore, the matrix A is the same as the matrix B in the general theorem stated in the Appendix A.6. Now, we will show A is non-singular, which is equivalent to show if $v^T A v = v^T P l^*(b_0, \Lambda_{10}, \Lambda_{20}; D)^{\otimes 2} v = 0$ for some $v = (v_1^T, v_2, v_3)^T \in \mathbb{R}^d$ then $v = 0$. Further, it is sufficient to showing if $v^T l^*(b_0, \Lambda_{10}, \Lambda_{20}; D) = 0$, then $v = 0$. It follows that

$$\begin{aligned}
0 =& v^T\big\{\dot{l}_b(b_0, \Lambda_{10}, \Lambda_{20}; D) - \dot{l}_{\Lambda_1}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_1^*] - \dot{l}_{\Lambda_2}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_2^*]\big\} \\
=& \Big\{v_1^T \dot{L}_\beta(b_0, \Lambda_{10}, \Lambda_{20}; D) + v_2 \dot{L}_\alpha(b_0, \Lambda_{10}, \Lambda_{20}; D) + v_3 \dot{L}_\kappa(b_0, \Lambda_{10}, \Lambda_{20}; D) \\
&+ \dot{L}_{\Lambda_1}(b_0, \Lambda_{10}, \Lambda_{20}; D)[-v^T h_1^*] + \dot{L}_{\Lambda_2}(b_0, \Lambda_{10}, \Lambda_{20}; D)[-v^T h_2^*]\Big\} \frac{1}{L(b_0, \Lambda_{10}, \Lambda_{20}; D)}.
\end{aligned}$$

By our Condition 3, we get $v = 0$. Therefore, the matrix A is non-singular. This completes the verification of assumption A3.

To verify A4, we first note that $\mathbb{P}_n \dot{l}_b(\hat{b}_n, \hat{\Lambda}_{1,n}, \hat{\Lambda}_{2,n}) = o_p(n^{-1/2})$ automatically holds since $(\hat{b}_n, \hat{\Lambda}_{1,n}, \hat{\Lambda}_{2,n})$ are the sieve maximum likelihood estimators and satisfy
$\mathbb{P}_n \dot{l}_b(\hat{b}_n, \hat{\Lambda}_{1,n}, \hat{\Lambda}_{2,n}) = 0$. Now we need to show $\mathbb{P}_n \dot{l}_{\Lambda_j}(\hat{b}_n, \hat{\Lambda}_{1,n}, \hat{\Lambda}_{2,n})[h_{jl}^*] = o_p(n^{-1/2})$, where h_{jl}^* is an element in h_j^*, $l = 1, \cdots, d$. According to Lemma A4, there exists an $h_{jl,n}^* \in \mathcal{M}_n^2$ such that $\|h_{jl}^* - h_{jl,n}^*\|_\infty = O(n^{-\nu})$ and $\mathbb{P}_n \dot{l}_{\Lambda_j}(\hat{b}_n, \hat{\Lambda}_{1,n}, \hat{\Lambda}_{2,n})[h_{jl,n}^*] = 0$. Thus, we want to show

$$I_{j,n} = \mathbb{P}_n \dot{l}_{\Lambda_j}(\hat{b}_n, \hat{\Lambda}_{1,n}, \hat{\Lambda}_{2,n}; D)[h_{jl}^* - h_{jl,n}^*] = o_p(n^{-1/2}).$$

Further, $I_{j,n}$ can be decomposed into summation of two parts $I_{j1,n}$ and $I_{j2,n}$, where

$$\begin{aligned}
I_{j1,n} &= (\mathbb{P}_n - P)\dot{l}_{\Lambda_j}(\hat{b}_n, \hat{\Lambda}_{1,n}, \hat{\Lambda}_{2,n}; D)[h_{jl}^* - h_{jl,n}^*], \\
I_{j2,n} &= P\big\{\dot{l}_{\Lambda_j}(\hat{b}_n, \hat{\Lambda}_{1,n}, \hat{\Lambda}_{2,n}; D)[h_{jl}^* - h_{jl,n}^*] - \dot{l}_{\Lambda_j}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_{jl}^* - h_{jl,n}^*]\big\}.
\end{aligned}$$

The decomposition holds since $P\big\{\dot{l}_{\Lambda_j}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h^*_{jl} - h^*_{jl,n}]\big\} = 0$. Next, we will show that $I_{j1,n}$ and $I_{j2,n}$ are both $o_p(n^{-1/2})$.

By Lemma A5, the ϵ-bracketing number associated with $\|\cdot\|_\infty$ norm for the class $\mathcal{F}_{n,jl}(\eta)$, defined as $\mathcal{F}_{n,jl}(\eta) = \big\{\dot{l}_{\Lambda_j}(\theta; D)[h^*_{jl} - h] : \theta \in \Theta^q_n, h \in \mathcal{M}^2_n \text{ and } \|h^*_{jl} - h\|_\infty \leq \eta\big\}$, is bounded by $(\eta/\epsilon)^{cm_n+d}$. The lemma implies that

$$\begin{aligned} J_{[\,]}\{\eta, \mathcal{F}_{n,jl}(\eta), L_2(P)\} = \int_0^\eta [1 + \log N_{[\,]}\{\epsilon, \mathcal{F}_{n,jl}(\eta), L_2(P)\}]^{\frac{1}{2}} d\epsilon &\leq \int_0^\eta \{1 + m_n \log(\eta/\epsilon)\}^{\frac{1}{2}} d\epsilon \\ &\lesssim \int_0^\eta \{m_n(\eta/\epsilon)\}^{\frac{1}{2}} d\epsilon = m_n^{\frac{1}{2}}\eta. \end{aligned}$$

Pick $\eta = \eta_n = O(n^{-\min\{\nu,(1-\nu)/2\}})$ and apply Lemma A4, then $\|h^*_{jl} - h^*_{jl,n}\|_\infty = O(n^{-\nu}) \leq O(n^{-\min\{\nu,(1-\nu)/2\}}) = \eta_n$. And since $q \geq 3$, we have $d(\hat{\theta}_n, \theta_0) = O_p(n^{-\min\{q\nu/2,(1-\nu)/2\}}) \leq O(n^{-\min\{\nu,(1-\nu)/2\}}) = \eta_n$. Therefore, we have $\dot{l}_{\Lambda_j}(\hat{b}_n, \hat{\Lambda}_{1,n}, \hat{\Lambda}_{2,n}; D)[h^*_{jl} - h^*_{jl,n}] \in \mathcal{F}_{n,jl}(\eta_n)$. According to Lemma A1, $\|\dot{l}_{\Lambda_j}(\theta; D)[h^*_{jl} - h]\|_\infty$ is bounded by some constant $M > 0$ and $P\{\dot{l}_{\Lambda_j}(\theta; D)[h^*_{jl} - h]\}^2 \lesssim \|h^*_{jl} - h\|^2_\infty \leq \eta^2_n$. Applying the maximal inequality of Lemma 3.4.2 of van der Vaart and Wellner [1996], we have

$$\begin{aligned} E_p\|\mathbb{G}_n\|_{\mathcal{F}_{n,jl}(\eta_n)} &\lesssim J_{[\,]}\{\eta_n, \mathcal{F}_{n,jl}(\eta_n), L_2(P)\}\Big[1 + \frac{J_{[\,]}\{\eta_n, \mathcal{F}_{n,jl}(\eta_n), L_2(P)\}}{\eta^2_n n^{1/2}} M\Big] \\ &\lesssim m_n^{\frac{1}{2}}\eta_n + m_n n^{-\frac{1}{2}} = O\big\{n^{-\min(\frac{1-\nu}{2}, 1-2\nu)}\big\} = o(1), \end{aligned}$$

where $\mathbb{G}_n = n^{1/2}(\mathbb{P}_n - P)$ and the equality holds due to $\nu < 1/2$. Therefore, we have

$$I_{j1,n} = (\mathbb{P}_n - P)\dot{l}_{\Lambda_j}(\hat{\theta}_n; D)[h^*_{jl} - h^*_{jl,n}] = n^{-1/2}\mathbb{G}_n\dot{l}_{\Lambda_j}(\hat{\theta}_n; D)[h^*_{jl} - h^*_{jl,n}] = o_p(n^{-\frac{1}{2}}).$$

Then, for $I_{j2,n} = o_p(n^{-1/2})$, applying Taylor expansion for $\dot{l}_{\Lambda_j}(\hat{\theta}_n; D)[h^*_{jl} - h^*_{jl,n}]$ at θ_0 gives

$$\begin{aligned} \dot{l}_{\Lambda_j}(\hat{\theta}_n; D)[h^*_{jl} - h^*_{jl,n}] - \dot{l}_{\Lambda_j}(\theta_0; D)[h^*_{jl} - h^*_{jl,n}] = (\hat{b}_n - b_0)^T \ddot{l}_{\Lambda_j b}(\tilde{\theta}_n; D)[h^*_{jl} - h^*_{jl,n}] \\ + \ddot{l}_{\Lambda_j\Lambda_j}(\tilde{\theta}_n; D)[h^*_{jl} - h^*_{jl,n}, \hat{\Lambda}_{j,n} - \Lambda_{j0}] + \ddot{l}_{\Lambda_j\Lambda_{j'}}(\tilde{\theta}_n; D)[h^*_{jl} - h^*_{jl,n}, \hat{\Lambda}_{j',n} - \Lambda_{j'0}], \end{aligned}$$

where $j, j' \in \{1, 2\}$ and $\tilde{\theta}_n$ lies between θ_0 and $\hat{\theta}_n$. Using similar procedures and applying Lemma A1 and A4, we get $\big|\ddot{l}_{\Lambda_j b}(\tilde{\theta}_n; D)[h^*_{jl} - h^*_{jl,n}]\big| \lesssim \|h^*_{jl} - h^*_{jl,n}\|_\infty = O(n^{-\nu})$. Thus,

$$(\hat{b}_n - b_0)^T \ddot{l}_{\Lambda_j b}(\tilde{\theta}_n; D)[h^*_{jl} - h^*_{jl,n}] \lesssim O_p\big\{n^{-\min(\frac{q\nu}{2}, \frac{1-\nu}{2})}\big\}O(n^{-\nu}) = O_p\big\{n^{-\min(\frac{(q+2)\nu}{2}, \frac{1+\nu}{2})}\big\},$$

$$\ddot{l}_{\Lambda_j\Lambda_j}(\tilde{\theta}_n; D)[h^*_{jl} - h^*_{jl,n}, \hat{\Lambda}_{j,n} - \Lambda_{j0}] \lesssim O(n^{-\nu})O_p\{n^{-\min(\frac{q\nu}{2}, \frac{1-\nu}{2})}\} = O_p\big\{n^{-\min(\frac{(q+2)\nu}{2}, \frac{1+\nu}{2})}\big\},$$

and similarly $\ddot{l}_{\Lambda_j\Lambda_{j'}}(\tilde{\theta}_n; D)[h^*_{jl} - h^*_{jl,n}, \hat{\Lambda}_{j',n} - \Lambda_{j'0}] = O_p\big\{n^{-\min(\frac{(q+2)\nu}{2}, \frac{1+\nu}{2})}\big\}$. Finally, since $\nu > (2+q)^{-1}$ we have $I_{j2,n} = O_p\big\{n^{-\min(\frac{(q+2)\nu}{2}, \frac{1+\nu}{2})}\big\} = o(n^{-1/2})$. Therefore, $I_{j,n} = I_{j1,n} + I_{j2,n} = o_p(n^{-1/2})$ for $j = 1, 2$ and assumption A4 holds.

Now we verify assumption A5. According to Lemma A6, the ϵ-bracketing number associated with $\|\cdot\|_\infty$ norm for the following classes of functions

$$\mathcal{F}^b_{n,l}(\eta) = \big\{\dot{l}_{b_l}(\theta; D) - \dot{l}_{b_l}(\theta_0; D) : \theta \in \Theta^q_n, d(\theta, \theta_0) \leq \eta, j = 1, 2\big\},$$
$$\mathcal{F}^{\Lambda_j}_{n,jl}(\eta) = \big\{\dot{l}_{\Lambda_j}(\theta; D)[h^*_{jl}] - \dot{l}_{\Lambda_j}(\theta_0; D)[h^*_{jl}] : \theta \in \Theta^q_n, d(\theta, \theta_0) \leq \eta\big\},$$

are both bounded by $(\eta/\epsilon)^{cm_n+d}$. The lemma implies that the corresponding ϵ-bracketing integrals are both bounded by $m_n^{1/2}\eta$, that is

$$J_{[\,]}\{\eta, \mathcal{F}^b_{n,l}(\eta), L_2(P)\} \lesssim m_n^{\frac{1}{2}}\eta \text{ and } J_{[\,]}\{\eta, \mathcal{F}^{\Lambda_j}_{n,jl}(\eta), L_2(P)\} \lesssim m_n^{\frac{1}{2}}\eta.$$

Then, for $\dot{l}_{b_l}(\theta; D) - \dot{l}_{b_l}(\theta_0; D) \in \mathcal{F}^b_{n,l}(\eta)$, applying Taylor expansion and Lemma A1 gives

$$\begin{aligned}
&P\{\dot{l}_{b_l}(\theta; D) - \dot{l}_{b_l}(\theta_0; D)\}^2 \\
\lesssim& P\{|b - b_0|_2^2\ \ddot{l}^T_{b_l b}(\tilde{\theta}; D)\ \ddot{l}_{b_l b}(\tilde{\theta}; D)\} + P\{\ddot{l}_{b_l\Lambda_1}(\tilde{\theta}; D)[\Lambda_1 - \Lambda_{10}]\}^2 + P\{\ddot{l}_{b_l\Lambda_2}(\tilde{\theta}; D)[\Lambda_2 - \Lambda_{20}]\}^2 \\
\lesssim& |b - b_0|^2 + \|\Lambda_1 - \Lambda_{10}\|_2^2 + \|\Lambda_2 - \Lambda_{20}\|_2^2 = d^2(\theta, \theta_0) = \eta^2.
\end{aligned}$$

Similarly, we can also obtain $P\{\dot{l}_{\Lambda_j}(\theta; D)[h^*_{jl}] - \dot{l}_{\Lambda_j}(\theta_0; D)[h^*_{jl}]\}^2 \lesssim \eta^2$ for any $\dot{l}_{\Lambda_j}(\theta; D)[h^*_{jl}] - \dot{l}_{\Lambda_j}(\theta_0; D)[h^*_{jl}] \in \mathcal{F}^{\Lambda_j}_{n,jl}(\eta)$. From Lemma A1, we know that $\|\dot{l}_{b_l}(\theta; D) - \dot{l}_{b_l}(\theta_0; D)\|_\infty$ and $\|\dot{l}_{\Lambda_j}(\theta; D)[h^*_{jl}] - \dot{l}_{\Lambda_j}(\theta_0; D)[h^*_{jl}]\|_\infty$ are both bounded. Now pick $\eta = \eta_n = O\{n^{-\min(\frac{q\nu}{2}, \frac{1-\nu}{2})}\}$. Similar to the verification of assumption A4, we have

$$E_p\|\mathbb{G}_n\|_{\mathcal{F}^b_{n,l}(\eta)} \lesssim m_n^{\frac{1}{2}}\eta_n + m_n n^{-\frac{1}{2}} = O\big\{n^{-\min(\frac{(q-1)\nu}{2}, \frac{1}{2}-\nu)}\big\} + O(n^{\nu-\frac{1}{2}}) = o(1),$$

where the last equality holds due to $0 < \nu < 1/2$ and $q \geq 3$. Similarly, $E_p\|\mathbb{G}_n\|_{\mathcal{F}^{\Lambda_j}_{n,jl}(\eta)} \lesssim m_n^{1/2}\eta_n + m_n n^{-1/2} = o(1)$. Thus, for $\gamma = \min\{q\nu/2, (1-\nu)/2\}$ and $Cn^{-\gamma} = Cn^{-\min(\frac{q\nu}{2}, \frac{1-\nu}{2})} = \eta_n$, by the Markov's inequality, we get

$$\sup_{d(\theta,\theta_0)\leq Cn^{-\gamma}} \mathbb{G}_n\{\dot{l}_{b_l}(b, \Lambda_1, \Lambda_2; D) - \dot{l}_{b_l}(b, \Lambda_{10}, \Lambda_{20}; D)\} = o_p(1),$$
$$\sup_{d(\theta,\theta_0)\leq Cn^{-\gamma}} \mathbb{G}_n\{\dot{l}_{\Lambda_j}(b, \Lambda_1, \Lambda_2; D)[h^*_{jl}] - \dot{l}_{\Lambda_j}(b, \Lambda_{10}, \Lambda_{20}; D)[h^*_{jl}]\} = o_p(1).$$

This completes the verification of assumption A5.

Finally, for the two equations in assumption A6, we will verify the second one and the proof of the other equation follows similar steps. In a neighborhood of $(b_0, \Lambda_{10}, \Lambda_{20})$: $\{(b, \Lambda_1, \Lambda_2) : |b - b_0| + \|\Lambda_1 - \Lambda_{10}\|_2 + \|\Lambda_2 - \Lambda_{20}\|_2 \leq Cn^{-\gamma}\}$ with $\gamma = \min(q\nu/2, (1-\nu)/2)$, applying the Taylor expansion for $\dot{l}_{\Lambda_j}(b, \Lambda_1, \Lambda_2; D)[h_j^*]$ yields

$$\begin{aligned}
&P\big\{\dot{l}_{\Lambda_j}(b, \Lambda_1, \Lambda_2; D)[h_j^*] - \dot{l}_{\Lambda_j}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_j^*] - \ddot{l}_{\Lambda_j b}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_j^*](b - b_0)\\
&\quad - \ddot{l}_{\Lambda_j\Lambda_j}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_j^*, \Lambda_j - \Lambda_{j0}] - \ddot{l}_{\Lambda_j\Lambda_{j'}}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_j^*, \Lambda_{j'} - \Lambda_{j'0}]\big\}\\
=&P\big\{\ddot{l}_{\Lambda_j b}(\tilde{b}, \tilde{\Lambda}_1, \tilde{\Lambda}_2; D)[h_j^*] - \ddot{l}_{\Lambda_j b}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_j^*]\big\}(b - b_0)\\
&\quad + P\big\{\ddot{l}_{\Lambda_j\Lambda_j}(\tilde{b}, \tilde{\Lambda}_1, \tilde{\Lambda}_2; D)[h_j^*, \Lambda_j - \Lambda_{j0}] - \ddot{l}_{\Lambda_j\Lambda_j}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_j^*, \Lambda_j - \Lambda_{j0}]\big\}\\
&\quad + P\big\{\ddot{l}_{\Lambda_j\Lambda_{j'}}(\tilde{b}, \tilde{\Lambda}_1, \tilde{\Lambda}_2; D)[h_j^*, \Lambda_{j'} - \Lambda_{j'0}] - \ddot{l}_{\Lambda_j\Lambda_{j'}}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_j^*, \Lambda_{j'} - \Lambda_{j'0}]\big\},
\end{aligned}$$

where $(\tilde{b}, \tilde{\Lambda}_1, \tilde{\Lambda}_2)$ are intermediate values between $(b_0, \Lambda_{10}, \Lambda_{20})$ and $(b, \Lambda_1, \Lambda_2)$. By applying similar arguments that we used for other assumptions, we have

$$\begin{aligned}
&P\big|\ddot{l}_{\Lambda_j\Lambda_{j'}}(\tilde{b}, \tilde{\Lambda}_1, \tilde{\Lambda}_2; D)[h_{jl}^*, \Lambda_{j'} - \Lambda_{j'0}] - \ddot{l}_{\Lambda_j\Lambda_{j'}}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_{jl}^*, \Lambda_{j'} - \Lambda_{j'0}]\big|\\
\lesssim&\big(|\tilde{b} - b_0| + \|\tilde{\Lambda}_1 - \Lambda_{10}\|_2 + \|\tilde{\Lambda}_2 - \Lambda_{20}\|_2\big)\big(\|\Lambda_{j'} - \Lambda_{j'0}\|_2\big)\\
=&O\big\{\big(|\tilde{b} - b_0| + \|\tilde{\Lambda}_1 - \Lambda_{10}\|_2 + \|\tilde{\Lambda}_2 - \Lambda_{20}\|_2\big)^{\alpha}\big\} = O(n^{-\alpha\gamma}),
\end{aligned}$$

where the inequality holds due to the Hölder's inequality and Cauchy-Schwarz inequality; the two equalities hold for some $\alpha > 1$ and $\gamma > 0$ as defined in assumption A6. Further, $O(n^{-\alpha\gamma}) = \big(|\tilde{b} - b_0| + \|\tilde{\Lambda}_1 - \Lambda_{10}\|_2 + \|\tilde{\Lambda}_2 - \Lambda_{20}\|_2\big)\big(\|\Lambda_{j'} - \Lambda_{j'0}\|_2\big)$ $= O(n^{-\min(\frac{q\nu}{2}, \frac{1-\nu}{2})})O(n^{-\min(\frac{q\nu}{2}, \frac{1-\nu}{2})}) = O(n^{-\min(q\nu, 1-\nu)}) = o(n^{-1/2})$ due to $(2+q)^{-1} < \nu < 1/2$. Hence, we get $\alpha\gamma > 1/2$. Similarly, for some $\alpha > 1$ satisfying $\alpha\gamma > 1/2$, we can have

$$\begin{aligned}
&P\big|\ddot{l}_{\Lambda_j\Lambda_j}(\tilde{b}, \tilde{\Lambda}_1, \tilde{\Lambda}_2; D)[h_{jl}^*, \Lambda_j - \Lambda_{j0}] - \ddot{l}_{\Lambda_j\Lambda_j}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_{jl}^*, \Lambda_j - \Lambda_{j0}]\big| = O(n^{-\alpha\gamma}),\\
&P\big|\ddot{l}_{\Lambda_j b}(\tilde{b}, \tilde{\Lambda}_1, \tilde{\Lambda}_2; D)[h_{jl}^*] - \ddot{l}_{\Lambda_j b}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_{jl}^*]\big|(b - b_0) = O(n^{-\alpha\gamma}).
\end{aligned}$$

Putting all together, we obtain that for some $\alpha > 1$ satisfying $\alpha\gamma > 1/2$,

$$\begin{aligned}
&\big|P\big\{\dot{l}_{\Lambda_j}(b, \Lambda_1, \Lambda_2; D)[h_j^*] - \dot{l}_{\Lambda_j}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_j^*] - \ddot{l}_{\Lambda_j b}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_j^*](b - b_0)\\
&- \ddot{l}_{\Lambda_j\Lambda_j}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_j^*, \Lambda_j - \Lambda_{j0}] - \ddot{l}_{\Lambda_j\Lambda_{j'}}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_j^*, \Lambda_{j'} - \Lambda_{j'0}]\big\}\big| = O(n^{-\alpha\gamma}).
\end{aligned}$$

Therefore, we have verified all assumptions A1–A6 of the general theorem in the Appendix A.6 and thus we get

$$n^{1/2}(\hat{b}_n - b_0) = A^{-1}n^{1/2}\mathbb{P}_n l^*(b_0, \Lambda_{10}, \Lambda_{20}; D) + o_p(1) \rightarrow_d N\{0, A^{-1}B(A^{-1})^T\},$$

where $l^*(b_0, \Lambda_{10}, \Lambda_{20}; D) = \dot{l}_b(b_0, \Lambda_{10}, \Lambda_{20}; D) - \dot{l}_{\Lambda_1}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_1^*] - \dot{l}_{\Lambda_2}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_2^*]$. Since $A = B = P\{l^*(b_0, \Lambda_{10}, \Lambda_{20}; D)\}^{\otimes 2}$, as shown in the verification of assumption A3, therefore $A^{-1}B(A^{-1})^T = A^{-1} := I^{-1}(b_0)$, thus $n^{1/2}(\hat{b}_n - b_0) \rightarrow_d N\{0, I^{-1}(b_0)\}$, where $I(b_0) = P\{l^*(b_0, \Lambda_{10}, \Lambda_{20}; D)\}^{\otimes 2}$ with $l^*(b_0, \Lambda_{10}, \Lambda_{20}; D)$ being the efficient score function of b_0. Now we complete the proof of Theorem 2.3.2. □

A.5 Technical lemmas

Lemma A1. Under Conditions 1, 2 and 4 (in Appendix A.4), the log-likelihood function $l(\theta; D) = \log L(\theta; D) = \log[S(L_1, L_2 \mid z) - S(L_1, R_2 \mid z) - S(R_1, L_2 \mid z) + S(R_1, R_2 \mid z)]$ has bounded and continuous first and second order derivatives with respect to $\theta = (\beta^T, \alpha, \kappa, \Lambda_1, \Lambda_2)^T \in \mathcal{B} \otimes \mathcal{M}^q \otimes \mathcal{M}^q$.

Proof. Since $L(\theta; D)$ is bounded away from 0 according to Condition 1, it is equivalent to show the boundedness and continuity for the first and second order derivatives of $S(t_1, t_2 \mid z)$. We first define some notation. Define $ll(\theta; t_1, t_2, z)$ as

$$\begin{aligned} & ll(\theta; t_1, t_2, z) = \log S(t_1, t_2 \mid z) \\ & \quad = -\kappa \; log\{1 + [(\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1))) - 1)^{1/\alpha} + (\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_2(t_2))) - 1)^{1/\alpha}]^{\alpha}\}. \end{aligned}$$

Denote $A = 1 + [(\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1))) - 1)^{1/\alpha} + (\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_2(t_2))) - 1)^{1/\alpha}]^{\alpha}$. For any fixed $\Lambda_j \in \mathcal{M}^q$, let $\{\Lambda_{j\eta}$: η in a neighborhood of $0 \in \mathbb{R}\}$ be a smooth parametric path in $\mathcal{M}^q$ running through Λ_j at $\eta = 0$ (i.e. $\Lambda_{j\eta} \in \mathcal{M}^q, \Lambda_{j\eta} \mid_{\eta=0} = \Lambda_j$). Let $h_j(t_j) \in$

$H_j = \{h_j : h_j = \frac{\partial \Lambda_{j\eta}}{\partial \eta} \mid_{\eta=0}, \Lambda_{j\eta} \in \mathcal{M}^q\}$, with $h_j(t_j)$ satisfying the Fréchet derivative $\lim_{\eta\to 0} \frac{\Lambda_j(t_j+\eta)-\Lambda_j(t_j)-h_j(t_j)\eta}{\eta} = 0$. Then the first order derivatives of $ll(\theta; t_1, t_2, z)$ are:

$$\begin{aligned}
&\dot{l}l_\beta(\theta; t_1, t_2, z) = \\
&\frac{-1}{A}\Bigg\{\Bigg[1+\Bigg(\frac{\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_2(t_2)))-1}{\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1)))-1}\Bigg)^{1/\alpha}\Bigg]^{\alpha-1}\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1)))\dot{G}(e^{z^T\beta}\Lambda_1(t_1))\Lambda_1(t_1) \\
&+\Bigg[1+\Bigg(\frac{\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1)))-1}{\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_2(t_2)))-1}\Bigg)^{1/\alpha}\Bigg]^{\alpha-1}\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_2(t_2)))\dot{G}(e^{z^T\beta}\Lambda_2(t_2))\Lambda_2(t_2)\Bigg\} \\
&e^{z^T\beta}z,
\end{aligned}$$

$$\begin{aligned}
\dot{l}l_\alpha(\theta; t_1, t_2, z) =& \log\Bigg\{(\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1)))-1)^{1/\alpha}+(\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_2(t_2)))-1)^{1/\alpha}\Bigg\} \\
&\times\Bigg[\big(\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1)))-1\big)^{1/\alpha}log(\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1)))-1) \\
&+\big(\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_2(t_2)))-1\big)^{1/\alpha}log(\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_2(t_2)))-1)\Bigg]\frac{\kappa}{\alpha^2}(1-\frac{1}{A}),
\end{aligned}$$

$$\begin{aligned}
\dot{l}l_\kappa(\theta; t_1, t_2, z) =& -log(A) \\
&+\frac{1}{\kappa A}\Bigg\{\Bigg[1+\Bigg(\frac{\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_2(t_2)))-1}{\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1)))-1}\Bigg)^{1/\alpha}\Bigg]^{\alpha-1}\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1)))G(e^{z^T\beta}\Lambda_1(t_1)) \\
&+\Bigg[1+\Bigg(\frac{\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1)))-1}{\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_2(t_2)))-1}\Bigg)^{1/\alpha}\Bigg]^{\alpha-1}\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_2(t_2)))G(e^{z^T\beta}\Lambda_2(t_2))\Bigg\},
\end{aligned}$$

$$\begin{aligned}
&\dot{l}l_{\Lambda_j}(\theta; t_1, t_2, z)[h_j] = \\
&\frac{-1}{A}\Bigg[1+\Bigg(\frac{\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_{j'}(t_{j'})))-1}{\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_j(t_j)))-1}\Bigg)^{1/\alpha}\Bigg]^{\alpha-1}\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_j(t_j)))\dot{G}(e^{z^T\beta}\Lambda_j(t_j))e^{z^T\beta}h_j(t_j),
\end{aligned}$$

with $j = 1, 2$, $\Lambda_j \in \mathcal{M}^q$, and $h_j(t_j) \in \mathcal{M}^{q-1}$.

The second order derivatives of $ll(\theta; t_1, t_2, z)$ can be written as

$$
\begin{aligned}
&\ddot{l}l_{\Lambda_1\Lambda_1}(\theta; t_1, t_2, z)[h_1, \tilde{h}_1] = \\
&\Bigg\{ \frac{-1}{A}\Bigg[1+\Bigg(\frac{\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_2(t_2)))-1}{exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1)))-1}\Bigg)^{1/\alpha}\Bigg]^{\alpha-1} exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1)))\ddot{G}(e^{z^T\beta}\Lambda_1(t_1))e^{2z^T\beta} \\
&+\frac{-1}{A}\Bigg[1+\Bigg(\frac{\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_2(t_2)))-1}{\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1)))-1}\Bigg)^{1/\alpha}\Bigg]^{\alpha-1} \exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1)))\big(\dot{G}(e^{z^T\beta}\Lambda_1(t_1))\big)^2e^{2z^T\beta}\frac{1}{\kappa} \\
&+\frac{1}{A}\Bigg[1+\Bigg(\frac{\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_2(t_2)))-1}{\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1)))-1}\Bigg)^{1/\alpha}\Bigg]^{\alpha-2} \exp(\frac{2}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1)))\big(\dot{G}(e^{z^T\beta}\Lambda_1(t_1))\big)^2e^{2z^T\beta}\frac{1}{\kappa} \\
&\quad\times(1-\frac{1}{\alpha})\big(\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_2(t_2)))-1\big)^{1/\alpha}\big(\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1)))-1\big)^{-1-1/\alpha} \\
&+\frac{1}{A^2}\Bigg[1+\Bigg(\frac{\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_2(t_2)))-1}{\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1)))-1}\Bigg)^{1/\alpha}\Bigg]^{2\alpha-2} \exp(\frac{2}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1)))\big(\dot{G}(e^{z^T\beta}\Lambda_1(t_1))\big)^2e^{2z^T\beta}\Bigg\} \\
&\times h_1(t_1)\tilde{h}_1(t_1),
\end{aligned}
$$

$$
\begin{aligned}
&\ddot{l}l_{\Lambda_1\Lambda_2}(\theta; t_1, t_2, z)[h_1, h_2] = \ddot{l}l_{\Lambda_2\Lambda_1}(\theta; t_1, t_2, z)[h_2, h_1] \\
&=\Bigg\{ \frac{-1}{A}\Bigg[1+\Bigg(\frac{\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_2(t_2)))-1}{\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1)))-1}\Bigg)^{1/\alpha}\Bigg]^{\alpha-2} \exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_2(t_2)))\dot{G}(e^{z^T\beta}\Lambda_2(t_2))e^{z^T\beta}\frac{1}{\kappa} \\
&\quad\times\Bigg(\frac{\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_2(t_2)))-1}{\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1)))-1}\Bigg)^{-1+1/\alpha}\frac{1}{\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1)))-1}(1-\frac{1}{\alpha}) \\
&\quad+\frac{1}{A^2}\Bigg[1+\Bigg(\frac{\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1)))-1}{\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_2(t_2)))-1}\Bigg)^{1/\alpha}\Bigg]^{\alpha-1} \exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_2(t_2)))\dot{G}(e^{z^T\beta}\Lambda_2(t_2))e^{z^T\beta}\Bigg\} \\
&\quad\times h_1(t_1)h_2(t_2).
\end{aligned}
$$

Similarly, we can derive $\ddot{l}l_{\Lambda_2\Lambda_2}(\theta; t_1, t_2, z)[h_2, \tilde{h}_2]$, $\ddot{l}l_{bb}(\theta; t_1, t_2, z)$ and $\ddot{l}l_{b\Lambda_j}(\theta; t_1, t_2, z)[h_j]$. Under Conditions 1, 2 and 4, t_1, t_2 reside in a closed, bounded and positive interval; covariate z is bounded in $\mathbb{R}^p$; Λ_{j0} has the order of smoothness $q \geq 3$ and it is positive; $G_j(\cdot)$ has an order of smoothness of 3 and it is strictly increasing with $G(0) = 0$. These conditions assure that all the derivatives are continuous and bounded. Therefore, the first and second order derivatives for the log-likelihood function $l(\theta; D)$ are all continuous and bounded. □

Lemma A2. For $\Lambda_{j0} \in \mathcal{M}^q$, $j = 1, 2$, there exists a $\Lambda_{j0,n} \in \mathcal{M}_n^q$ such that

$$\|\Lambda_{j0,n} - \Lambda_{j0}\|_\infty = O(n^{-qv/2}).$$

Proof. This is a direct result according to Theorem 1.6.1 of Lorentz [1986], which indicates there exists a Bernstein polynomial $\Lambda_{j0,n}$ such that $\|\Lambda_{j0,n} - \Lambda_{j0}\|_\infty = O(m_n^{-q/2}) = O(n^{-qv/2})$. □

Lemma A3. Let $\theta_{0,n} = (b_0^T, \Lambda_{10,n}, \Lambda_{20,n})^T = (\beta_0^T, \alpha_0, \kappa_0, \Lambda_{10,n}, \Lambda_{20,n})^T$ with $b_0 \in \mathcal{B}$ and $\Lambda_{10,n}, \Lambda_{20,n} \in \mathcal{M}_n^q, j = 1, 2$. Denote $\mathcal{F}_n = \{l(\theta; D) - l(\theta_{0,n}; D) : \theta \in \Theta_n^q = \mathcal{B} \times \mathcal{M}_n^q \times \mathcal{M}_n^q\}$. Assume Conditions 1, 2 and 4 hold, then the ϵ-bracketing number associated with $\|\cdot\|_\infty$ norm for $\mathcal{F}_n$ is bounded by $(1/\epsilon)^{cm_n+d}$, where $d = p + 2$. That is, for some constant $c > 0$, $N_{[\,]}(\epsilon, \mathcal{F}_n, \|\cdot\|_\infty) \lesssim (1/\epsilon)^{cm_n+d}$.

Proof. We first define some notation:

$$\begin{aligned} B =&(\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_1(t_1))) - 1)^{1/\alpha} + (\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_2(t_2))) - 1)^{1/\alpha}, \\ C_j =&\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_j(t_j))) - 1,\ j = 1, 2. \end{aligned}$$

According to Shen and Wong [1994] on page 597, $\forall \epsilon > 0$, $\exists$ a set of brackets $\{[\Lambda_{ji}^L, \Lambda_{ji}^U] : i = 1, 2, \cdots, \lceil(1/\epsilon)^{c_1 m_n}\rceil, j = 1, 2\}$ such that for any $\Lambda_j \in \mathcal{M}_n^q$, $\Lambda_{ji}^L(t_j) \le \Lambda_j(t_j) \le \Lambda_{ji}^U(t_j)$ for some $1 \le i \le \lceil(1/\epsilon)^{c_1 m_n}\rceil$ and all $t_j \in [c, u]$, and $\|\Lambda_{ji}^L - \Lambda_{ji}^U\|_\infty \le \epsilon$. In other words, we have $N_{[\,]}(\epsilon, \mathcal{M}_n^q, \|.\|_\infty) \le c(1/\epsilon)^{c_1 m_n}$.

Define $\mathcal{B}$ as a compact set, then $\mathcal{B}$ can be covered by $\lceil c_2(1/\epsilon)^d\rceil$ balls with radius ϵ. Thus, for any $b \in \mathcal{B}$, there exists $1 \le s \le \lceil c_2(1/\epsilon)^d\rceil$ such that $|\beta - \beta_s| \le \epsilon$, $|\alpha - \alpha_s| \le \epsilon$ and $|\kappa - \kappa_s| \le \epsilon$. Equivalently, we have $\beta \in [\beta_s - \epsilon, \beta_s + \epsilon]$, $\alpha \in [\alpha_s - \epsilon, \alpha_s + \epsilon]$ and $\kappa \in [\kappa_s - \epsilon, \kappa_s + \epsilon]$. Hence, we can construct a set of brackets $\{[m_{i,s}^L(D), m_{i,s}^U(D)], i = \{1, \cdots, \lceil(1/\epsilon)^{c_1 m_n}\rceil\}, s = \{1, ..., \lceil c_2(1/\epsilon)^d\rceil, \}\}$ so that for any $m(\theta; D) \in \mathcal{F}_n$, there exists a set (i, s) such that for any sample point D, we have $m(\theta; D) \in [m_{i,s}^L(D), m_{i,s}^U(D)]$, where

$$\begin{aligned} m_{i,s}^L(D) &= \log\left[S_{i,s}^L(L_1, L_2|z) - S_{i,s}^U(L_1, R_2|z) - S_{i,s}^U(R_1, L_2|z) + S_{i,s}^L(R_1, R_2|z)\right] - l(\theta_{0,n}; D), \\ m_{i,s}^U(D) &= \log\left[S_{i,s}^U(L_1, L_2|z) - S_{i,s}^L(L_1, R_2|z) - S_{i,s}^L(R_1, L_2|z) + S_{i,s}^U(R_1, R_2|z)\right] - l(\theta_{0,n}; D), \end{aligned}$$

with L and U representing lower and upper bound of each term. To prove this Lemma, we will need to show $\|m^U_{i,s} - m^L_{i,s}\|_\infty \leq \epsilon$. By the Mean Value Theorem,

$$\begin{aligned}
|m^U_{i,s}(D) - m^L_{i,s}(D)| \leq{}& |S^U_{i,s}(L_1, L_2|z) - S^L_{i,s}(L_1, L_2|z)| + |S^U_{i,s}(L_1, R_2|z) - S^L_{i,s}(L_1, R_2|z)| \\
&+ |S^U_{i,s}(R_1, L_2|z) - S^L_{i,s}(R_1, L_2|z)| + |S^U_{i,s}(R_1, R_2|z) - S^L_{i,s}(R_1, R_2|z)|.
\end{aligned}$$

Since $L_j, R_j \in [c, u]$, it suffices to prove that for any $t_1, t_2 \in [c, u]$ and z as defined in Condition 2, $|S^U_{i,s}(t_1, t_2 \mid z) - S^L_{i,s}(t_1, t_2 \mid z)| \leq \epsilon$. Applying the Mean Value theorem again,

$$\begin{aligned}
&|S^U_{i,s}(t_1, t_2 \mid z) - S^L_{i,s}(t_1, t_2 \mid z)| = |e^{\log(S^U_{i,s}(t_1,t_2|z))} - e^{\log(S^L_{i,s}(t_1,t_2|z))}| \\
&\leq |\log(S^U_{i,s}(t_1, t_2 \mid z)) - \log(S^L_{i,s}(t_1, t_2 \mid z))| := |m^U_{i,s}(t_1, t_2, z) - m^L_{i,s}(t_1, t_2, z)|,
\end{aligned}$$

where

$$\begin{aligned}
m^L_{i,s}(t_1, t_2, z) ={}& -(\kappa_s + \epsilon) \log \Big\{1 + \big[(\exp(\frac{1}{\kappa_s + \epsilon} G(e^{z^T(\beta_s+\epsilon)} \Lambda^U_{1i}(t_1))) - 1)^{\frac{1}{\alpha_s - \epsilon}} \\
&+ (\exp(\frac{1}{\kappa_s + \epsilon} G(e^{z^T(\beta_s+\epsilon)} \Lambda^U_{2i}(t_2))) - 1)^{\frac{1}{\alpha_s - \epsilon}}\big]^{\alpha_s - \epsilon}\Big\} := -(\kappa_s + \epsilon) \log A_1
\end{aligned}$$

and

$$\begin{aligned}
m^U_{i,s}(t_1, t_2, z) ={}& -(\kappa_s - \epsilon) \log \Big\{1 + \big[(\exp(\frac{1}{\kappa_s - \epsilon} G(e^{z^T(\beta_s-\epsilon)} \Lambda^U_{1i}(t_1))) - 1)^{\frac{1}{\alpha_s + \epsilon}} \\
&+ (\exp(\frac{1}{\kappa_s - \epsilon} G(e^{z^T(\beta_s-\epsilon)} \Lambda^U_{2i}(t_2))) - 1)^{\frac{1}{\alpha_s + \epsilon}}\big]^{\alpha_s + \epsilon}\Big\} := -(\kappa_s - \epsilon) \log A_2,
\end{aligned}$$

in which A_1 and A_2 represent the corresponding A term (previously defined in Lemma A1) in $m^U_{i,s}(t_1, t_2, z)$ and $m^L_{i,s}(t_1, t_2, z)$, respectively. Similarly, we denote B_1, B_2 corresponding to B as well as C_{j1}, C_{j2} corresponding to C_j in $m^U_{i,s}(t_1, t_2, z)$ and $m^L_{i,s}(t_1, t_2, z)$, respectively. It then follows that

$$\begin{aligned}
|m^U_{i,s}(t_1, t_2, z) - m^L_{i,s}(t_1, t_2, z)| &= |-(\kappa_s - \epsilon) \log A_1 + (\kappa_s + \epsilon) \log A_2| \\
&\leq |\kappa_s + \epsilon| \times |\log A_1 - \log A_2| + 2\epsilon \log A_1 \lesssim |A_1 - A_2| + \epsilon.
\end{aligned}$$

The last inequality holds due to the Mean Value Theorem and Lemma A1.

For $|A_1 - A_2|$, applying the Mean Value Theorem and Lemma A1 again, we have

$$\begin{aligned}
|A_1 - A_2| &= |B_1^{\alpha_s+\epsilon} - B_2^{\alpha-\epsilon}| = |\exp(\log B_1^{\alpha_s+\epsilon}) - \exp(\log B_2^{\alpha_s-\epsilon})| \\
&\le |\log B_1^{\alpha_s+\epsilon} - \log B_2^{\alpha_s-\epsilon}| = |(\alpha_s+\epsilon)\log B_1 - (\alpha_s-\epsilon)\log B_2| \\
&\lesssim |B_1 - B_2| + \epsilon \le |C_{11}^{\frac{1}{\alpha_s+\epsilon}} - C_{12}^{\frac{1}{\alpha_s-\epsilon}}| + |C_{21}^{\frac{1}{\alpha_s+\epsilon}} - C_{22}^{\frac{1}{\alpha_s-\epsilon}}| + \epsilon,
\end{aligned}$$

where

$$\begin{aligned}
&|C_{11}^{\frac{1}{\alpha_s+\epsilon}} - C_{12}^{\frac{1}{\alpha_s-\epsilon}}| = |e^{\log C_{11}^{\frac{1}{\alpha_s+\epsilon}}} - e^{\log C_{12}^{\frac{1}{\alpha_s-\epsilon}}}| \le |\log C_{11}^{\frac{1}{\alpha_s+\epsilon}} - \log C_{12}^{\frac{1}{\alpha_s-\epsilon}}| \\
&\le |\log C_{11}| \times |\frac{1}{\alpha_s+\epsilon} - \frac{1}{\alpha_s-\epsilon}| + |\frac{1}{\alpha_s-\epsilon}| \times |\log C_{11} - \log C_{12}| \lesssim \epsilon + |C_{11} - C_{12}|.
\end{aligned}$$

Further, applying the Mean Value Theorem, Lemma A1 and Condition 4(ii), we have

$$\begin{aligned}
|C_{11} - C_{12}| &= |\exp(\frac{1}{\kappa_s-\epsilon}G(e^{z^T(\beta_s-\epsilon)}\Lambda_{1i}^L(t_1))) - \exp(\frac{1}{\kappa_s+\epsilon}G(e^{z^T(\beta_s+\epsilon)}\Lambda_{1i}^U(t_1)))| \\
&\le |\frac{1}{\kappa_s-\epsilon}G(e^{z^T(\beta_s-\epsilon)}\Lambda_{1i}^L(t_1)) - \frac{1}{\kappa_s+\epsilon}G(e^{z^T(\beta_s+\epsilon)}\Lambda_{1i}^U(t_1))| \\
&\le G(e^{z^T(\beta_s-\epsilon)}\Lambda_{1i}^L(t_1))|\frac{1}{\kappa_s-\epsilon} - \frac{1}{\kappa_s+\epsilon}| \\
&+ |\frac{1}{\kappa_s+\epsilon}| \times |G(e^{z^T(\beta_s-\epsilon)}\Lambda_{1i}^L(t_1)) - G(e^{z^T(\beta_s+\epsilon)}\Lambda_{1i}^U(t_1))| \\
&\lesssim \epsilon + |e^{z^T(\beta_s-\epsilon)}\Lambda_{1i}^L(t_1) - e^{z^T(\beta_s+\epsilon)}\Lambda_{1i}^U(t_1)| \\
&\le \epsilon + \Lambda_{1i}^L(t_1)|e^{z^T(\beta_s-\epsilon)} - e^{z^T(\beta_s+\epsilon)}| + e^{z^T(\beta_s+\epsilon)}|\Lambda_{1i}^U(t_1) - \Lambda_{1i}^L(t_1)| \lesssim \epsilon.
\end{aligned}$$

The last inequality holds due to $\|\Lambda_{1i}^U - \Lambda_{1i}^L\|_\infty = \epsilon$.

Similarly, we can obtain $|C_{21}^{\frac{1}{\alpha_s+\epsilon}} - C_{22}^{\frac{1}{\alpha_s-\epsilon}}| \lesssim \epsilon$. Therefore, $\|m_{i,s}^U - m_{i,s}^L\|_\infty \le \epsilon$ and the ϵ-bracketing number associated with $\|\cdot\|_\infty$ norm for the class $\mathcal{F}_n$ follows

$$N_{[\,]}(\epsilon, \mathcal{F}_n, \|\cdot\|_\infty) \le (1/\epsilon)^{c_1 m_n}(1/\epsilon)^{c_1 m_n} c_2 (1/\epsilon)^d \lesssim (1/\epsilon)^{c m_n + d}.$$

□

Lemma A4. Let $h_{jl}^*, j = 1, 2, l = 1, .., d$, be an element of h_j^* defined in the proof of Theorem 2.3.2. This is the least favorable direction for the score function of Λ_j. Assume Conditions 1, 2, 4 hold, then there exists an $h_{jl,n}^* \in \mathcal{M}_n^2$ such that $\|h_{jl,n}^* - h_{jl}^*\|_\infty = O(n^{-\nu})$.

Proof. We will first show that $h_{jl}^* \in \mathcal{M}^2$, and then apply Theorem 1.6.1 of Lorentz [1986] to complete the proof. By definition of h_j^*, $j = 1, 2$, and the fact that $-P(\dot{l}_\theta(\theta_0; D)\dot{l}_\theta^T(\theta_0; D)) = P(\ddot{l}_{\theta\theta^T}(\theta_0; D))$, for any $h_j \in \mathcal{M}^{q-1}$, we have

$$
\begin{aligned}
&P\Big\{ -\dot{l}_b(b_0, \Lambda_{10}, \Lambda_{20}; D)\dot{l}_{\Lambda_1}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_1] - \dot{l}_b(b_0, \Lambda_{10}, \Lambda_{20}; D)\dot{l}_{\Lambda_2}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_2] \\
&+\dot{l}_{\Lambda_1}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_1^*]\dot{l}_{\Lambda_1}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_1] + \dot{l}_{\Lambda_1}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_1^*]\dot{l}_{\Lambda_2}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_2] \\
&+\dot{l}_{\Lambda_2}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_2^*]\dot{l}_{\Lambda_2}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_2] + \dot{l}_{\Lambda_2}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_2^*]\dot{l}_{\Lambda_1}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_1] \\
&\quad\Big\} = 0.
\end{aligned}
$$

Based on Lemma A1, we can see that $\dot{l}_\beta(b_0, \Lambda_{10}, \Lambda_{20}; D)$ has bounded derivatives upto order 2, which is the order of $\dot{G}_j$ in Condition 4. Also, $\dot{l}_\alpha(b_0, \Lambda_{10}, \Lambda_{20}; D)$ and $\dot{l}_\kappa(b_0, \Lambda_{10}, \Lambda_{20}; D)$ have bounded derivatives upto order $\min\{q, 3\}$, which is the minimum order of Λ_j and $G_j(\cdot)$. Similarly, $\dot{l}_{\Lambda_j}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h], j = 1, 2$ have bounded derivatives upto order of min{order of h_j, order of $G(\cdot) - 1\} = \min\{q - 1, 2\}$. Since $q \geq 3$ based on Condition 4, $\min\{q - 1, 2\} = 2$. Hence, $h_{jl}^* \in \mathcal{M}^2$. Then, applying Theorem 1.6.1 of Lorentz [1986], there exists an $h_{jl,n}^* \in \mathcal{M}_n^2$ such that $\|h_{jl,n}^* - h_{jl}^*\|_\infty = O(m_n^{-2/2}) = O(n^{-\nu})$, where $j = 1, 2$. □

Lemma A5. Let h_{jl}^* be the function defined in Lemma A4, and denote the class of functions $\mathcal{F}_{n,jl}(\eta) = \big\{\dot{l}_{\Lambda_j}(\theta; D)[h_{jl}^* - h] : \theta \in \Theta_n^q, h \in \mathcal{M}_n^2, \|h_{jl}^* - h\|_\infty \leq \eta\big\}$. Assume Conditions 1, 2, 4 hold, then $N_{[\,]}(\epsilon, \mathcal{F}_{n,jl}(\eta), \|\cdot\|_\infty) \lesssim (\eta/\epsilon)^{cm_n+d}$ for some constant $c > 0$.

Proof. First define three classes of functions: $\mathcal{M}_{n,j}^q(\eta) = \big\{\Lambda_j \in \mathcal{M}_n^q, \|\Lambda_j - \Lambda_{j0}\|_2 \leq \eta\big\}$, $\mathcal{M}_{n,jl}^2(\eta) = \big\{h \in \mathcal{M}_n^2, \|h - h_{jl}^*\|_\infty \leq \eta\big\}$ and $\mathcal{B}(\eta) = \big\{b \in \mathcal{B} \subseteq \mathbb{R}^d, |\beta - \beta_0| + |\alpha - \alpha_0| + |\kappa - \kappa_0| \leq \eta\big\}$, where $j = 1, 2$, $l = 1, ..., d$. Then, following Shen and Wong [1994] (page 597) gives $N_{[\,]}(\epsilon, \mathcal{M}_{n,j}^q(\eta), \|\cdot\|_\infty) \leq (\eta/\epsilon)^{c_1 m_n}$ and $N_{[\,]}(\epsilon, \mathcal{M}_{n,jl}^2(\eta), \|\cdot\|_\infty) \leq (\eta/\epsilon)^{c_2 m_n}$ for some constants $c_1, c_2 > 0$. In addition, since $\mathcal{B} \subseteq \mathbb{R}^d$ is compact, the covering number of $\mathcal{B}(\eta)$ follows $N(\epsilon, \mathcal{B}(\eta), \|\cdot\|_\infty) \leq c_3(\eta/\epsilon)^d$.

Similar to the proof of Lemma A3, Λ_{ji}^L and Λ_{ji}^U are functions that bracket Λ_j, with $\|\Lambda_{ji}^U - \Lambda_{ji}^L\|_\infty \leq \epsilon, i \in \{1, \cdots, \lceil(\eta/\epsilon)^{c_1 m_n}\rceil\}$; h_k^L and h_k^U are functions that bracket h, with $\|h_k^U - h_k^L\|_\infty \leq \epsilon, k \in \{1, \cdots, \lceil(\eta/\epsilon)^{c_2 m_n}\rceil\}$; $\beta \in [\beta_s - \epsilon, \beta_s + \epsilon]$, $\alpha \in [\alpha_s - \epsilon, \alpha_s + \epsilon]$ and

$\kappa \in [\kappa_s - \epsilon, \kappa_s + \epsilon], s \in \{1, \cdots, \lceil c_3(\eta/\epsilon)^d \rceil\}$. Then we construct a set of brackets for $\mathcal{F}_{n,jl}(\eta)$ (let $j = 1$ without loss of generality)

$$\big\{[d^{L,1}_{i,k,s}(D), d^{U,1}_{i,k,s}(D)] : 1 \leq i \leq \lceil (\eta/\epsilon)^{c_1 m_n} \rceil; 1 \leq k \leq \lceil (\eta/\epsilon)^{c_2 m_n} \rceil; 1 \leq s \leq \lceil c_3(\eta/\epsilon)^d \rceil \big\}$$

so that for any $\dot{l}_{\Lambda_1}(\theta; D)[h^*_{1l} - h] \in \mathcal{F}_{n,1l}(\eta)$, there exists a triplet (i, k, s) such that $\dot{l}_{\Lambda_1}(\theta; D)[h^*_{1l} - h] \in [d^{L,1}_{i,k,s}(D), d^{U,1}_{i,k,s}(D)]$ for any sample point D. Following similar reasoning in Lemma A1 and Lemma A3, it suffices to show that for any $t_1, t_2 \in [c, u]$ and z as defined in Condition 2,

$$|\dot{l}l^{U,1}_{\Lambda_1,i,k,s}(\theta; t_1, t_2, z) - \dot{l}l^{L,1}_{\Lambda_1,i,k,s}(\theta; t_1, t_2, z)| := |d^{U,1}_{i,k,s}(t_1, t_2, z) - d^{L,1}_{i,k,s}(t_1, t_2, z)| \leq \epsilon,$$

where $\dot{l}l_{\Lambda_1}(\theta; t_1, t_2, z)[h^*_{1l} - h] \in [d^{L,1}_{i,k,s}(t_1, t_2, z), d^{U,1}_{i,k,s}(t_1, t_2, z)]$, with

$$\begin{aligned} d^{L,1}_{i,k,s}(t_1, t_2, z) =& \frac{-1}{A^L_{i,k,s}} \Big[1 + \Big(\frac{\exp(\frac{1}{\kappa_s+\epsilon} G(e^{z^T(\beta_s-\epsilon)} \Lambda^L_{2i}(t_2))) - 1}{\exp(\frac{1}{\kappa_s-\epsilon} G(e^{z^T(\beta_s+\epsilon)} \Lambda^U_{1i}(t_1))) - 1}\Big)^{1/(\alpha_s+\epsilon)}\Big]^{\alpha_s+\epsilon-1} \\ & \times \exp\big(\frac{1}{\kappa_s - \epsilon} G(e^{z^T(\beta_s+\epsilon)} \Lambda^U_{1i}(t_1))\big) \, \dot{G}\big(e^{z^T(\beta_s+\epsilon)} \Lambda^U_{1i}(t_1)\big)\big) \, e^{z^T(\beta_s+\epsilon)} \; (h^*_{1l}(t_1) - h^L_k(t_1)), \end{aligned}$$

$$\begin{aligned} d^{U,1}_{i,k,s}(t_1, t_2, z) =& \frac{-1}{A^L_{i,k,s}} \Big[1 + \Big(\frac{\exp(\frac{1}{\kappa_s+\epsilon} G(e^{z^T(\beta_s-\epsilon)} \Lambda^L_{2i}(t_2))) - 1}{\exp(\frac{1}{\kappa_s-\epsilon} G(e^{z^T(\beta_s+\epsilon)} \Lambda^U_{1i}(t_1))) - 1}\Big)^{1/(\alpha_s+\epsilon)}\Big]^{\alpha_s+\epsilon-1} \\ & \times \exp\big(\frac{1}{\kappa_s - \epsilon} G(e^{z^T(\beta_s+\epsilon)} \Lambda^U_{1i}(t_1))\big) \, \dot{G}\big(e^{z^T(\beta_s+\epsilon)} \Lambda^U_{1i}(t_1)\big)\big) \, e^{z^T(\beta_s+\epsilon)} \; (h^*_{1l}(t_1) - h^U_k(t_1)). \end{aligned}$$

The term $A^L_{i,k,s}$ presents the lower bound of A, which is previously defined in Lemma A1. We notice that $d^{L,1}_{i,k,s}(t_1, t_2, z)$ and $d^{U,1}_{i,k,s}(t_1, t_2, z)$ are only different in the last term because $d^{L,1}_{i,k,s}(t_1, t_2, z) < 0$ whereas $d^{U,1}_{i,k,s}(t_1, t_2, z) > 0$. Thus, due to the boundedness of the first order derivatives, it follows that $|d^{U,1}_{i,k,s}(t_1, t_2, z) - d^{L,1}_{i,k,s}(t_1, t_2, z)| \lesssim \|h^U_k - h^L_k\|_\infty \leq \epsilon$. Similarly, we can show $|d^{U,2}_{i,k,s}(t_1, t_2, z) - d^{L,2}_{i,k,s}(t_1, t_2, z)| \lesssim \epsilon$. Therefore, the ϵ-bracketing number for the class $\mathcal{F}_{n,jl}(\eta), j = 1, 2$, is bounded by $(\eta/\epsilon)^{c_1 m_n} \, (\eta/\epsilon)^{c_2 m_n} \, c_3(\eta/\epsilon)^d$, that is $N_{[\,]}(\epsilon, \mathcal{F}_{n,jl}(\eta), \|\cdot\|_\infty) \lesssim (\eta/\epsilon)^{c_1 m_n + c_2 m_n + d} = (\eta/\epsilon)^{c m_n + d}$. □

Lemma A6. For $l = 1, \cdots, d$, define two classes of functions as

$$\mathcal{F}^{b}_{n,l}(\eta) = \big\{\dot{l}_{b_l}(\theta; D) - \dot{l}_{b_l}(\theta_0; D) : \theta \in \Theta^q_n, d(\theta, \theta_0) \leq \eta, j = 1, 2\big\}, \text{ and}$$
$$\mathcal{F}^{\Lambda_j}_{n,jl}(\eta) = \big\{\dot{l}_{\Lambda_j}(\theta; D)[h^*_{jl}] - \dot{l}_{\Lambda_j}(\theta_0; D)[h^*_{jl}] : \theta \in \Theta^q_n, d(\theta, \theta_0) \leq \eta\big\},$$

where $\dot{l}_{b_l}(\theta; D)$ is the l^{th} element of $\dot{l}_{b_l}(\theta; D) = (\dot{l}^T_\beta(\theta; D), \dot{l}_\alpha(\theta; D), \dot{l}_\kappa(\theta; D))^T$ and h^*_{jl} is the l^{th} element of h^*_j in Lemma A4. Assume Conditions 1-5 hold, then $N_{[\ ]}(\epsilon, \mathcal{F}^b_{n,l}(\eta), \|\cdot\|_\infty) \lesssim (\eta/\epsilon)^{cm_n+d}$ and $N_{[\ ]}(\epsilon, \mathcal{F}^{\Lambda_j}_{n,jl}(\eta), \|\cdot\|_\infty) \lesssim (\eta/\epsilon)^{cm_n+d}$ for some constant $c > 0$.

Proof. First define several classes of functions $\mathcal{M}^q_{n,j}(\eta) = \big\{\Lambda_j \in \mathcal{M}^q_n, \|\Lambda_j - \Lambda_{j0}\|_2 \leq \eta\big\}$ and $\mathcal{B}(\eta) = \big\{b = (\beta^T, \alpha, \kappa)^T \in \mathcal{B} \subseteq \mathbb{R}^d, |\beta - \beta_0| + |\alpha - \alpha_0| + |\kappa - \kappa_0| \leq \eta\big\}$, where $j = 1, 2$. By the same arguments as in Lemma A5, we have $N_{[\ ]}(\epsilon, \mathcal{M}^q_{n,j}(\eta), \|\cdot\|_\infty) \leq (\eta/\epsilon)^{c_1 m_n}$ and $N(\epsilon, \mathcal{B}(\eta), \|\cdot\|_\infty) \leq c_2(\eta/\epsilon)^d$ for some constants $c_1, c_2 > 0$. Similar to Lemma A5, let Λ^L_{ji} and Λ^U_{ji} be the functions that bracket $\Lambda_j \in \mathcal{M}^q_{n,j}$, with $\|\Lambda^U_{ji} - \Lambda^L_{ji}\|_\infty \leq \epsilon, i \in \{1...\lceil(\eta/\epsilon)^{c_1 m_n}\rceil\}$; also for any $b \in \mathcal{B}$, we have $\beta \in [\beta_s - \epsilon, \beta_s + \epsilon]$, $\alpha \in [\alpha_s - \epsilon, \alpha_s + \epsilon]$ and $\kappa \in [\kappa_s - \epsilon, \kappa_s + \epsilon]$, $s \in \{1...\lceil c_2(\eta/\epsilon)^d\rceil\}$.

For $\mathcal{F}^b_{n,l}(\eta)$ we can construct a set of brackets

$$\big\{[u^L_{b_l,i,s}(D), u^U_{b_l,i,s}(D)] : 1 \leq i \leq \lceil(\eta/\epsilon)^{c_1 m_n}\rceil; 1 \leq s \leq \lceil c_2(\eta/\epsilon)^d\rceil\big\},$$

so that for any element in $\mathcal{F}^b_{n,l}(\eta)$, there exists a set (i, s) such that $\dot{l}_{b_l}(\theta; D) - \dot{l}_{b_l}(\theta_0; D) \in [u^L_{b_l,i,s}(D), u^U_{b_l,i,s}(D)]$ for any sample point D. Similar to Lemma A3 and Lemma A5, it suffices to prove that for any $t_1, t_2 \in [c, u]$ and z as defined in Condition 2,

$$|\dot{l}l^U_{b_l,i,s}(\theta; t_1, t_2, z) - \dot{l}l^L_{b_l,i,s}(\theta; t_1, t_2, z)| := |u^U_{b_l,i,s}(t_1, t_2, z) - u^L_{b_l,i,s}(t_1, t_2, z)| \leq \epsilon,$$

where $\dot{l}l_{b_l}(\theta; t_1, t_2, z) - \dot{l}l_{b_l}(\theta_0; t_1, t_2, z) \in [u^L_{b_l,i,s}(t_1, t_2, z), u^U_{b_l,i,s}(t_1, t_2, z)]$.

When $b_l = \alpha$, we need to show $|u^U_{\alpha,i,s}(t_1, t_2, z) - u^L_{\alpha,i,s}(t_1, t_2, z)| \leq \epsilon$. Due to similar structures of $u^L_{\alpha,i,s}(t_1, t_2, z)$ and $u^U_{\alpha,i,s}(t_1, t_2, z)$, we will present the explicit form for the first term only. Also we assume $\exp(\frac{1}{\kappa}G(e^{z^T\beta}\Lambda_j(t_j))) > 2$ without loss of generality. Following similar steps and notations as in Lemma 5, we have

$$
\begin{aligned}
& u^L_{\alpha,i,s}(t_1,t_2,z) = \dot{ll}^L_\alpha(\theta;t_1,t_2,z) - \dot{ll}_\alpha(\theta_0;t_1,t_2,z) = \frac{\kappa_s-\epsilon}{(\alpha_s+\epsilon)^2}\,(1-\frac{1}{A^L_{i,s}})\times \\
& \log\Bigg\{\big[\exp(\frac{1}{\kappa_s+\epsilon}G(e^{z^T(\beta_s-\epsilon)}\Lambda^L_{1i}(t_1))) - 1\big]^{\frac{1}{\alpha_s+\epsilon}} + \big[\exp(\frac{1}{\kappa_s+\epsilon}G(e^{z^T(\beta_s-\epsilon)}\Lambda_{2i}(t_2))) - 1\big]^{\frac{1}{\alpha_s+\epsilon}}\Bigg\} \\
& \times\Bigg\{\big[\exp(\frac{1}{\kappa_s+\epsilon}G(e^{z^T(\beta_s-\epsilon)}\Lambda^L_{1i}(t_1))) - 1\big]^{\frac{1}{\alpha_s+\epsilon}}\,\log\big[\exp(\frac{1}{\kappa_s+\epsilon}G(e^{z^T(\beta_s-\epsilon)}\Lambda^L_{1i}(t_1))) - 1\big] \\
& + \big[\exp[\frac{1}{\kappa_s+\epsilon}G(e^{z^T(\beta_s-\epsilon)}\Lambda^L_{2i}(t_2))) - 1\big]^{\frac{1}{\alpha_s+\epsilon}}\,\log\big[\exp(\frac{1}{\kappa_s+\epsilon}G(e^{z^T(\beta_s-\epsilon)}\Lambda^L_{2i}(t_2))) - 1\big]\Bigg\} \\
& - \dot{ll}_\alpha(\theta_0;t_1,t_2,z).
\end{aligned}
$$

Applying similar arguments in Lemma A3 and A5 gives $\|u^U_{\alpha,i,s} - u^L_{\alpha,i,s}\|_\infty \lesssim \epsilon$. Likewise, we have $\|u^U_{\kappa,i,s} - u^L_{\kappa,i,s}\|_\infty \lesssim \epsilon$ and $\|u^U_{\beta,i,s} - u^L_{\beta,i,s}\|_\infty \lesssim \epsilon$. Therefore, the ϵ-bracketing number associated with $\|\cdot\|_\infty$ for the class $\mathcal{F}^b_{n,l}(\eta)$ is $N_{[\,]}(\epsilon,\mathcal{F}^b_{n,l}(\eta),\|\cdot\|_\infty) \lesssim (\eta/\epsilon)^{cm_n+d}$.

Next, following similar steps we can find a set of brackets for the class $\mathcal{F}^{\Lambda_1}_{n,1l}(\eta)$ as $\big\{[v^{L,1}_{i,s}(D), v^{U,1}_{i,s}(D)] : 1 \le i \le \lceil(\eta/\epsilon)^{c_1m_n}\rceil; 1 \le s \le \lceil c_2(\eta/\epsilon)^d\rceil\big\}$ so that for any element in $\mathcal{F}^{\Lambda_1}_{1l}(\eta)$, there exists a set (i,s) such that $\dot{l}_{\Lambda_1}(\theta;D)[h^*_{1l}] - \dot{l}_{\Lambda_1}(\theta_0;D)[h^*_{1l}] \in [v^{L,1}_{i,s}(D), v^{U,1}_{i,s}(D)]$ for any sample point D. Applying the same arguments again, it suffices to prove that for any $t_1, t_2 \in [c,u]$ and z defined in Condition 2, $\big|\dot{ll}^U_{\Lambda_1,i,s}(\theta;t_1,t_2,z)[h^*_{1l}] - \dot{ll}^L_{\Lambda_1,i,s}(\theta;t_1,t_2,z)[h^*_{1l}]\big| := |v^{U,1}_{i,s}(t_1,t_2,z) - v^{L,1}_{i,s}(t_1,t_2,z)| \le \epsilon$, where $\dot{ll}_{\Lambda_1}(\theta;t_1,t_2,z)[h^*_{1l}] - \dot{ll}_{\Lambda_1}(\theta_0;t_1,t_2,z)[h^*_{1l}] \in [v^{L,1}_{i,s}(t_1,t_2,z), v^{U,1}_{i,s}(t_1,t_2,z)]$, with

$$
\begin{aligned}
v^{L,1}_{i,s}(t_1,t_2,z) =& \frac{-1}{A^L_{i,s}}\left[1+\left(\frac{\exp(\frac{1}{\kappa_s+\epsilon}G(e^{z^T(\beta_s-\epsilon)}\Lambda^L_{2i}(t_2))) - 1}{\exp(\frac{1}{\kappa_s-\epsilon}G(e^{z^T(\beta_s+\epsilon)}\Lambda^U_{1i}(t_1))) - 1}\right)^{1/(\alpha_s+\epsilon)}\right]^{\alpha_s+\epsilon-1} \\
& \times\exp\big(\frac{1}{\kappa_s-\epsilon}G(e^{z^T(\beta_s+\epsilon)}\Lambda^U_{1i}(t_1))\big)\,\dot{G}(e^{z^T(\beta_s+\epsilon)}\Lambda^U_{1i}(t_1)))\,e^{z^T(\beta_s+\epsilon)}\,h^*_{1l}(t_1) \\
& - \dot{ll}_{\Lambda_1}(\theta_0;t_1,t_2,z)[h^*_{1l}],
\end{aligned}
$$

$$
\begin{aligned}
v^{U,1}_{i,s}(t_1,t_2,z) =& \frac{-1}{A^U_{i,s}}\left[1+\left(\frac{\exp(\frac{1}{\kappa_s-\epsilon}G(e^{z^T(\beta_s+\epsilon)}\Lambda^U_{2i}(t_2))) - 1}{\exp(\frac{1}{\kappa_s+\epsilon}G(e^{z^T(\beta_s-\epsilon)}\Lambda^L_{1i}(t_1))) - 1}\right)^{1/(\alpha_s-\epsilon)}\right]^{\alpha_s-\epsilon-1} \\
& \times\exp\big(\frac{1}{\kappa_s+\epsilon}G(e^{z^T(\beta_s-\epsilon)}\Lambda^L_{1i}(t_1))\big)\,\dot{G}(e^{z^T(\beta_s-\epsilon)}\Lambda^L_{1i}(t_1)))\,e^{z^T(\beta_s-\epsilon)}\,h^*_{1l}(t_1) \\
& - \dot{ll}_{\Lambda_1}(\theta_0;t_1,t_2,z)[h^*_{1l}].
\end{aligned}
$$

Applying the similar arguments as before gives $\|v_{i,s}^{U,1} - v_{i,s}^{L,1}\|_\infty \lesssim \epsilon$. Likewise, $\|v_{i,s}^{U,2} - v_{i,s}^{L,2}\|_\infty \lesssim \epsilon$. Hence, the ϵ-bracketing number associated with $\|\cdot\|_\infty$ for the class $\mathcal{F}_{n,jl}^{\Lambda_j}(\eta)$ is $N_{[\,]}(\epsilon, \mathcal{F}_{n,jl}^{\Lambda_j}(\eta), \|\cdot\|_\infty) \lesssim (\eta/\epsilon)^{cm_n+d}$. □

A.6 A general theorem on the asymptotic normality of semiparametric M-estimators with two nuisance parameters

To prove the asymptotic normality for the M-estimator $\hat{b}_n$, we first need to prove a general theorem that is similar to Theorem 6 in Wellner and Zhang [2007] and Theorem 2.1 in Ding and Nan [2011]. Our log-likelihood function is more complicated and involves two (infinite-dimensional) nuisance parameters $\Lambda_j, j = 1, 2$. We first denote

$$\begin{aligned}
&\dot{S}_b(b,\Lambda_1,\Lambda_2) = P\dot{l}_b(b,\Lambda_1,\Lambda_2;D),\ \dot{S}_{b,n}(b,\Lambda_1,\Lambda_2) = \mathbb{P}\dot{l}_b(b,\Lambda_1,\Lambda_2;D),\\
&\dot{S}_{\Lambda_j}(b,\Lambda_1,\Lambda_2)[h_j] = P\dot{l}_{\Lambda_j}(b,\Lambda_1,\Lambda_2;D)[h_j],\ \dot{S}_{\Lambda_j,n}(b,\Lambda_1,\Lambda_2)[h_j] = \mathbb{P}\dot{l}_{\Lambda_j}(b,\Lambda_1,\Lambda_2;D)[h_j],\\
&\ddot{S}_{bb}(b,\Lambda_1,\Lambda_2) = P\ddot{l}_{bb}(b,\Lambda_1,\Lambda_2;D),\\
&\ddot{S}_{b\Lambda_j}(b,\Lambda_1,\Lambda_2)[h_j] = \ddot{S}^T_{\Lambda_j b}(b,\Lambda_1,\Lambda_2)[h_j] = P\ddot{l}_{b\Lambda_j}(b,\Lambda_1,\Lambda_2;D)[h_j],\\
&\ddot{S}_{\Lambda_j\Lambda_j}(b,\Lambda_1,\Lambda_2)[h_j,\tilde{h}_j] = P\ddot{l}_{\Lambda_j\Lambda_j}(b,\Lambda_1,\Lambda_2;D)[h_j,\tilde{h}_j],\\
&\ddot{S}_{\Lambda_j\Lambda_{j'}}(b,\Lambda_1,\Lambda_2)[h_j,h_{j'}] = P\ddot{l}_{\Lambda_j\Lambda_{j'}}(b,\Lambda_1,\Lambda_2;D)[h_j,h_{j'}].
\end{aligned}$$

We list the following assumptions:

A1. (Rate of convergence) $|\hat{b}_n - b_0| + \|\hat{\Lambda}_{1n} - \Lambda_{10}\| + \|\hat{\Lambda}_{2n} - \Lambda_{20}\| = O_p(n^{-\gamma})$ for some $\gamma > 0$ and some norm $\|\cdot\|$.

A2. $\dot{S}_b(b_0,\Lambda_{10},\Lambda_{20}) = 0, \dot{S}_{\Lambda_j}(b_0,\Lambda_{10},\Lambda_{20})[h_j] = 0$, for all $h_j \in H_j$, $j = 1, 2$.

A3. There exists $h_j^* = (h_{j1}^*, \cdot, \cdot, \cdot, h_{jd}^*)^T$, where $h_{jl}^* \in H_j$, $j = 1, 2$, $l = 1, \cdots, d$, so that

$$\begin{aligned}
&\ddot{S}_{b\Lambda_1}(b_0,\Lambda_{10},\Lambda_{20})[h_1] + \ddot{S}_{b\Lambda_2}(b_0,\Lambda_{10},\Lambda_{20})[h_2] - \ddot{S}_{\Lambda_1\Lambda_1}(b_0,\Lambda_{10},\Lambda_{20})[h_1^*, h_1]\\
&- \ddot{S}_{\Lambda_1\Lambda_2}(b_0,\Lambda_{10},\Lambda_{20})[h_1^*, h_2] - \ddot{S}_{\Lambda_2\Lambda_2}(b_0,\Lambda_{10},\Lambda_{20})[h_2^*, h_2] - \ddot{S}_{\Lambda_2\Lambda_1}(b_0,\Lambda_{10},\Lambda_{20})[h_2^*, h_1] = 0
\end{aligned}$$

for all $h_j \in H_j, j = 1, 2$. Moreover, the matrix A is nonsingular with

$$\begin{aligned} A =& - \ddot{S}_{bb}(b_0, \Lambda_{10}, \Lambda_{20}) + \ddot{S}_{\Lambda_1 b}(b_0, \Lambda_{10}, \Lambda_{20})[h_1^*] + \ddot{S}_{\Lambda_2 b}(b_0, \Lambda_{10}, \Lambda_{20})[h_2^*] \\ =& - P\{\ddot{l}_{bb}(b_0, \Lambda_{10}, \Lambda_{20}; D) - \ddot{l}_{\Lambda_1 b}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_1^*] - \ddot{l}_{\Lambda_2 b}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_2^*]\}. \end{aligned}$$

A4. The estimator $(\hat{b}_n, \hat{\Lambda}_{1,n}, \hat{\Lambda}_{2,n})$ satisfies

$$\dot{S}_{b,n}(\hat{b}_n, \hat{\Lambda}_{1,n}, \hat{\Lambda}_{2,n}) = o_p(n^{-1/2}), \text{ and } \dot{S}_{\Lambda_j,n}(\hat{b}_n, \hat{\Lambda}_{1,n}, \hat{\Lambda}_{2,n})[h_j^*] = o_p(n^{-1/2}), \ j = 1, 2.$$

A5. (Stochastic equicontinuity) For any $C > 0$ and $j = 1, 2$,

$$\begin{aligned} &\sup_{d(\theta,\theta_0) \le Cn^{-\gamma}} \left| n^{1/2}(\dot{S}_{b,n} - \dot{S}_b)(b, \Lambda_1, \Lambda_2) - n^{1/2}(\dot{S}_{b,n} - \dot{S}_b)(b_0, \Lambda_{10}, \Lambda_{20}) \right| = o_p(1), \\ &\sup_{d(\theta,\theta_0) \le Cn^{-\gamma}} \left| n^{1/2}(\dot{S}_{\Lambda_j,n} - \dot{S}_{\Lambda_j})(b, \Lambda_1, \Lambda_2)[h_j^*] - n^{1/2}(\dot{S}_{\Lambda_j,n} - \dot{S}_{\Lambda_j})(b_0, \Lambda_{10}, \Lambda_{20})[h_j^*] \right| = o_p(1), \end{aligned}$$

where $d(\theta, \theta_0) = |b - b_0| + \|\Lambda_1 - \Lambda_{10}\| + \|\Lambda_2 - \Lambda_{20}\|$ is the distance between $\theta = (b, \Lambda_1, \Lambda_2)$ and $\theta_0 = (b_0, \Lambda_{10}, \Lambda_{20})$ under some well-defined norm $\| \cdot \|$.

A6. (Smoothness of model) For some $\alpha > 1$ satisfying $\alpha\gamma > \frac{1}{2}$, and for $(b, \Lambda_1, \Lambda_2)$ in a neighborhood of $(b_0, \Lambda_{10}, \Lambda_{20}) : |b - b_0| + \|\Lambda_1 - \Lambda_{10}\| + \|\Lambda_2 - \Lambda_{20}\| \le Cn^{-\gamma}$,

$$\begin{aligned} &\big| S_b(b, \Lambda_1, \Lambda_2) - S_b(b_0, \Lambda_{10}, \Lambda_{20}) \\ &- \ddot{S}_{bb}(b_0, \Lambda_{10}, \Lambda_{20})(b - b_0) - \ddot{S}_{b\Lambda_1}(b_0, \Lambda_{10}, \Lambda_{20})[\Lambda_1 - \Lambda_{10}] - \ddot{S}_{b\Lambda_2}(b_0, \Lambda_{10}, \Lambda_{20})[\Lambda_2 - \Lambda_{20}] \big| \\ &= O\big\{(|b - b_0| + \|\Lambda_1 - \Lambda_{10}\| + \|\Lambda_2 - \Lambda_{20}\|)^\alpha\big\}, \\ &\big| S_{\Lambda_j}(b, \Lambda_1, \Lambda_2)[h_j^*] - S_{\Lambda_j}(b_0, \Lambda_{10}, \Lambda_{20})[h_j^*] - \ddot{S}_{\Lambda_j b}(b_0, \Lambda_{10}, \Lambda_{20})[h_j^*](b - b_0) \\ &- \ddot{S}_{\Lambda_j \Lambda_j}(b_0, \Lambda_{10}, \Lambda_{20})[h_j^*, \Lambda_j - \Lambda_{j0}] - \ddot{S}_{\Lambda_j \Lambda_{j'}}(b_0, \Lambda_{10}, \Lambda_{20})[h_j^*, \Lambda_{j'} - \Lambda_{j'0}] \big| \\ &= O\big\{(|b - b_0| + \|\Lambda_1 - \Lambda_{10}\| + \|\Lambda_2 - \Lambda_{20}\|)^\alpha\big\}, \text{ where } j, j' \in \{1, 2\}. \end{aligned}$$

Theorem. *Suppose that assumptions A1-A6 hold, then*

$$n^{1/2}(\hat{b}_n - b_0) = A^{-1} n^{1/2} \mathbb{P}_n l^*(b_0, \Lambda_{10}, \Lambda_{20}; D) + o_p(1) \to_d N\{0, A^{-1} B (A^{-1})^T\},$$

where $l^*(b_0, \Lambda_{10}, \Lambda_{20}; D) = \dot{l}_b(b_0, \Lambda_{10}, \Lambda_{20}; D) - \dot{l}_{\Lambda_1}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_1^*] - \dot{l}_{\Lambda_2}(b_0, \Lambda_{10}, \Lambda_{20}; D)[h_2^*]$, $B = Pl^*(b_0, \Lambda_{10}, \Lambda_{20}; D)^{\otimes 2} = P\big\{l^*(b_0, \Lambda_{10}, \Lambda_{20}; D) l^*(b_0, \Lambda_{10}, \Lambda_{20}; D)^T\big\}$ *and A is defined in assumption A3.*

Proof. We follow the proof for Theorem 2.1 of Ding and Nan [2011]. Using assumptions A1 and A5, we have $n^{1/2}(\dot{S}_{b,n} - \dot{S}_b)(\hat{b}_n, \hat{\Lambda}_{1,n}, \hat{\Lambda}_{2,n}) - n^{1/2}(\dot{S}_{b,n} - \dot{S}_b)(b_0, \Lambda_{10}, \Lambda_{20}) = o_p(1)$. Due to $\dot{S}_{b,n}(\hat{b}_n, \hat{\Lambda}_{1,n}, \hat{\Lambda}_{2,n}) = o_p(n^{-1/2})$ by assumption A4 and $\dot{S}_b(b_0, \Lambda_{10}, \Lambda_{20}) = 0$ by assumption A2, we get $n^{1/2}\dot{S}_b(\hat{b}_n, \hat{\Lambda}_{1,n}, \hat{\Lambda}_{2,n}) + n^{1/2}\dot{S}_{b,n}(b_0, \Lambda_{10}, \Lambda_{20}) = o_p(1)$. Similarly, $n^{1/2}\dot{S}_{\Lambda_j}(\hat{b}_n, \hat{\Lambda}_{1,n}, \hat{\Lambda}_{2,n}) + n^{1/2}\dot{S}_{\Lambda_j,n}(b_0, \Lambda_{10}, \Lambda_{20}) = o_p(1)$. Combining these equations and assumption A6, we obtain

$$\begin{aligned}
&\ddot{S}_{bb}(b_0, \Lambda_{10}, \Lambda_{20})(\hat{b}_n - b_0) + \ddot{S}_{b\Lambda_1}(b_0, \Lambda_{10}, \Lambda_{20})[\hat{\Lambda}_{1,n} - \Lambda_{10}] + \ddot{S}_{b\Lambda_2}(b_0, \Lambda_{10}, \Lambda_{20})[\hat{\Lambda}_{2,n} - \Lambda_{20}] \\
&\quad + \dot{S}_{b,n}(b_0, \Lambda_{10}, \Lambda_{20}) + O\big\{\big(|\hat{b}_n - b_0| + \|\hat{\Lambda}_{1,n} - \Lambda_{10}\| + \|\hat{\Lambda}_{2,n} - \Lambda_{20}\|\big)^{\alpha}\big\} = o_p(n^{-1/2}), \\
&\ddot{S}_{\Lambda_1 b}(b_0, \Lambda_{10}, \Lambda_{20})[h_j^*](\hat{b}_n - b_0) + \ddot{S}_{\Lambda_j\Lambda_j}(b_0, \Lambda_{10}, \Lambda_{20})[h_j^*, \hat{\Lambda}_{j,n} - \Lambda_{j0}] \\
&\quad + \ddot{S}_{\Lambda_j\Lambda_{j'}}(b_0, \Lambda_{10}, \Lambda_{20})[h_j^*, \hat{\Lambda}_{j',n} - \Lambda_{j'0}] + \dot{S}_{\Lambda_j,n}(b_0, \Lambda_{10}, \Lambda_{20})[h_j^*] \\
&\quad + O\big\{\big(|\hat{b}_n - b_0| + \|\hat{\Lambda}_{1,n} - \Lambda_{10}\| + \|\hat{\Lambda}_{2,n} - \Lambda_{20}\|\big)^{\alpha}\big\} = o_p(n^{-1/2}),
\end{aligned}$$

where $j, j' \in \{1, 2\}$. Since $\alpha > 1$ and $\alpha\gamma > 1/2$, the convergence rate in assumption A1 implies that $n^{1/2}O\big\{\big(|\hat{b}_n - b_0| + \|\hat{\Lambda}_{1,n} - \Lambda_{10}\| + \|\hat{\Lambda}_{2,n} - \Lambda_{20}\|\big)^{\alpha}\big\} = O_p(n^{\frac{1}{2}-\alpha\gamma}) = o_p(1)$. Then, two equations above together with assumption A3 leads to

$$\begin{aligned}
&\big(\ddot{S}_{bb}(b_0, \Lambda_{10}, \Lambda_{20}) - \ddot{S}_{\Lambda_1 b}(b_0, \Lambda_{10}, \Lambda_{20})[h_1^*] - \ddot{S}_{\Lambda_2 b}(b_0, \Lambda_{10}, \Lambda_{20})[h_2^*]\big)(\hat{b}_n - b_0) \\
&= -\big(\dot{S}_{b,n}(b_0, \Lambda_{10}, \Lambda_{20}) - \dot{S}_{\Lambda_1,n}(b_0, \Lambda_{10}, \Lambda_{20})[h_1^*] - \dot{S}_{\Lambda_2,n}(b_0, \Lambda_{10}, \Lambda_{20})[h_2^*]\big) + o_p(n^{-1/2}),
\end{aligned}$$

that is, $-A(\hat{b}_n - b_0) = -\mathbb{P}_n l^*(b_0, \Lambda_{10}, \Lambda_{20}; D) + o_p(n^{-1/2})$. This yields $n^{1/2}(\hat{b}_n - b_0) = A^{-1}n^{1/2}\mathbb{P}_n l^*(b_0, \Lambda_{10}, \Lambda_{20}; D) + o_p(1) \to_d N\{0, A^{-1}B(A^{-1})^T\}$. □

Appendix B

Supplementary materials for Chapter 3

B.1 Regularity conditions

Denote $\|x\|$ as the Euclidean norm of a d-dimensional vector $x = (x_1, ..., x_d)^T \in R^d$, namely $\|x\| = \sqrt{x_1^2 + ... + x_d^2}$. For any $d \times d$ matrix A, define $\|A\| = \sqrt{\sum_{i,j=1}^{d} {A_{ij}}^2}$, where A_{ij} is the $(i, j)th$ element of matrix A. Let $N(\eta^*)$ be an open neighbourhood of η^*. For simplicity of notation, under right censoring, we denote $\ddot{l}_{\eta,j}(S_1, S_2; \eta) = \frac{\partial \dot{l}_\eta(S_1,S_2;\eta)}{\partial S_j}$, $\dddot{l}_{\eta\eta,j}(S_1, S_2; \eta) = \frac{\partial \ddot{l}_{\eta\eta}(S_1,S_2;\eta)}{\partial S_j}, j = 1, 2$. Similarly, under interval censoring, we first denote $\{S_1(L_1), S_1(R_1), S_2(L_2), S_2(R_2)\}$ as $\{y_{11}, y_{12}, y_{21}, y_{22}\}$, and write $\ddot{l}_{\eta,j}(S_1, S_2; \eta) = \frac{\partial \dot{l}_\eta(y_{11},y_{12},y_{21},y_{22};\eta)}{\partial y_{j\cdot}}$, $\dddot{l}_{\eta\eta,j}(S_1, S_2; \eta) = \frac{\partial \ddot{l}_{\eta\eta}(y_{11},y_{12},y_{21},y_{22};\eta)}{\partial y_{j\cdot}}, j = 1, 2$. In the following paragraphs, we state the regularity conditions needed for proving the Theorems.

Condition 1. Matrix $S(\eta^*) = -P_0\{\ddot{l}_{\eta\eta}(S_1, S_2; \eta^*)\}$ is finite and non-singular.

Condition 2. Denote $J_i(S_1, S_2) = \mathit{const} \times \{S_1(1 - S_1)\}^{-\epsilon_{i1}} \times \{S_2(1 - S_2)\}^{-\epsilon_{i2}}$, where $\epsilon_{i1}, \epsilon_{i2} \geq 0, i = 1, 2$, $\epsilon_{i1}, \epsilon_{i2}$ are some constants. Suppose that for all $\eta \in N(\eta^*)$, we have $\|\dot{l}_\eta(S_1, S_2; \eta)\dot{l}_\eta^T(S_1, S_2; \eta)\| \leq J_1(S_1, S_2)$, $\|\ddot{l}_{\eta\eta}(S_1, S_2; \eta)\| \leq J_2(S_1, S_2)$, and $P_0\{J_i^2(S_1, S_2)\} < \infty$.

Condition 3. Suppose that both $\ddot{l}_{\eta,j}(S_1, S_2; \eta)$ and $\dddot{l}_{\eta\eta,j}(S_1, S_2; \eta), j = 1, 2$ exist and are continuous. Denote $\tilde{J}_i^1(S_1, S_2) = \mathit{const} \times \{S_1(1 - S_1)\}^{-\tilde{\epsilon}_{i1}} \times \{S_2(1 - S_2)\}^{-\epsilon_{i2}}$, $\tilde{J}_i^2(S_1, S_2) = \mathit{const} \times \{S_1(1 - S_1)\}^{-\epsilon_{i1}} \times \{S_2(1 - S_2)\}^{-\tilde{\epsilon}_{i2}}$, where $\tilde{\epsilon}_{i1} > \epsilon_{i1}$ and $\tilde{\epsilon}_{i2} > \epsilon_{i2}$ are some constants, such that for $\eta \in N(\eta^*)$, $\|\ddot{l}_{\eta,j}(S_1, S_2; \eta)\| \leq \tilde{J}_1^j(S_1, S_2)$ and $\|\dddot{l}_{\eta\eta,j}(S_1, S_2; \eta)\| \leq \tilde{J}_2^j(S_1, S_2)$, and furthermore, $P_0\{\tilde{J}_i^j(S_1, S_2)\} < \infty, i = 1, 2$, and $j = 1, 2$.

Condition 4. Suppose $\frac{\partial \ddot{l}_{\eta\eta}(S_1,S_2;\eta)}{\partial \eta_k}, k = 1, 2, ..., p$ exist and are continuous with $\eta \in N(\eta^*)$, and there exists an integrable function $G_3(S_1, S_2)$ such that $\|\frac{\partial \ddot{l}_{\eta\eta}(S_1,S_2;\eta)}{\partial \eta_k}\| \leq G_3(S_1, S_2)$ for all $\eta \in N(\eta^*), k = 1, ..., p$.

Remark. *Condition 1 indicates that the sensitivity matrix $S(\eta^*)$ is invertible so that the IR* *statistic is well-defined. Conditions 2 and 3 are similar to the conditions in Chen and Fan [2005], Chen and Fan [2006] and Zhang et al. [2016]. Condition 4 is a common assumption to establish the uniform law of large numbers theorem as in Zhang et al. [2016].*

B.2 Verification of conditions

By following Chen and Fan [2006] Section 5, we verify Conditions 2-4 under specific copula functions. Specifically, under interval censoring, we will show that our log-likelihood function's derivatives can satisfy Conditions 2-4. Under right censoring, due to the presence of a copula density function, the corresponding log-likelihood function is more complicated. In the following section, we will use the Clayton copula as an example to verify that it satisfies Conditions 2-4 under both interval and right censoring.

Scenario I interval censoring: Under interval censoring, the log-likelihood function is denoted as

$$
\begin{aligned}
l(S_1, S_2; \eta) = &\log L(S_1, S_2; \eta) \\
= &\log[C_\eta\{S_1(L_1), S_2(L_2)\} - C_\eta\{S_1(L_1), S_2(R_2)\} \\
&- C_\eta\{S_1(R_1), S_2(L_2)\} + C_\eta\{S_1(R_1), S_2(R_2)\}].
\end{aligned}
$$

Assume that there exists $\tau > 0$ such that $pr(R - L \geq \tau) = 1$, so $L(S_1, S_2; \eta)$ is bounded away from 0. It is equivalent to show the boundedness for the derivatives of $C_\eta(u, v), u, v \in (0, 1), \eta \in N(\eta^*)$, with $\eta^* \in \mathcal{A} = [A^{-1}, A]$ for a large $A > 1$. Define $ll(u, v; \eta)$ as

$$
ll(u, v; \eta) = \log C_\eta(u, v) = -\frac{1}{\eta} \log(u^{-\eta} + v^{-\eta} - 1).
$$

Then, the derivatives of $ll(u, v; \eta)$ are:

$$\begin{aligned}
\dot{ll}_{\eta}(u, v; \eta) =& \frac{\log(u^{-\eta} + v^{-\eta} - 1)}{\eta^2} + \frac{u^{-\eta}\log u + v^{-\eta}\log v}{\eta(u^{-\eta} + v^{-\eta} - 1)}, \\
\ddot{ll}_{\eta\eta}(u, v; \eta) =& \frac{-2\log(u^{-\eta} + v^{-\eta} - 1)}{\eta^3} - \frac{2(u^{-\eta}\log u + v^{-\eta}\log v)}{\eta^2(u^{-\eta} + v^{-\eta} - 1)} \\
& - \frac{u^{-\eta}(\log u)^2 + v^{-\eta}(\log v)^2}{\eta(u^{-\eta} + v^{-\eta} - 1)} + \frac{(u^{-\eta}\log u + v^{-\eta}\log v)^2}{\eta(u^{-\eta} + v^{-\eta} - 1)^2}, \\
\ddot{ll}_{\eta,1}(u, v; \eta) =& \frac{-u^{-(\eta+1)}\log u}{u^{-\eta} + v^{-\eta} - 1} + \frac{u^{-(\eta+1)}(u^{-\eta}\log u + v^{-\eta}\log v)}{(u^{-\eta} + v^{-\eta} - 1)^2}, \\
\ddot{ll}_{\eta,2}(u, v; \eta) =& \frac{-v^{-(\eta+1)}\log v}{u^{-\eta} + v^{-\eta} - 1} + \frac{v^{-(\eta+1)}(u^{-\eta}\log u + v^{-\eta}\log v)}{(u^{-\eta} + v^{-\eta} - 1)^2}, \\
\dddot{ll}_{\eta\eta,1}(u, v; \eta) =& -\frac{\frac{2}{\eta}(u^{-\eta}\log u + v^{-\eta}\log v) - (\log u)^2}{(u^{-\eta} + v^{-\eta} - 1)u^{\eta+1}} - \frac{u^{-\eta}(\log u)^2 + v^{-\eta}(\log v)^2}{(u^{-\eta} + v^{-\eta} - 1)^2 u^{\eta+1}} \\
& + \frac{2(u^{-\eta}\log u + v^{-\eta}\log v)(-\log u + \frac{1}{\eta})}{(u^{-\eta} + v^{-\eta} - 1)^2 u^{\eta+1}} + \frac{2(u^{-\eta}\log u + v^{-\eta}\log v)^2}{(u^{-\eta} + v^{-\eta} - 1)^3 u^{\eta+1}} \\
\dddot{ll}_{\eta\eta,2}(u, v; \eta) =& -\frac{\frac{2}{\eta}(u^{-\eta}\log u + v^{-\eta}\log v) - (\log v)^2}{(u^{-\eta} + v^{-\eta} - 1)v^{\eta+1}} - \frac{u^{-\eta}(\log u)^2 + v^{-\eta}(\log v)^2}{(u^{-\eta} + v^{-\eta} - 1)^2 v^{\eta+1}} \\
& + \frac{2(u^{-\eta}\log u + v^{-\eta}\log v)(-\log v + \frac{1}{\eta})}{(u^{-\eta} + v^{-\eta} - 1)^2 v^{\eta+1}} + \frac{2(u^{-\eta}\log u + v^{-\eta}\log v)^2}{(u^{-\eta} + v^{-\eta} - 1)^3 v^{\eta+1}}, \\
\frac{\partial \ddot{ll}_{\eta\eta}(u, v; \eta)}{\partial \eta} =& \frac{6\log(u^{-\eta} + v^{-\eta} - 1)}{\eta^4} + \frac{6(u^{-\eta}\log u + v^{-\eta}\log v)}{\eta^3(u^{-\eta} + v^{-\eta} - 1)} \\
& - \frac{3(u^{-\eta}\log u + v^{-\eta}\log v)^2}{\eta^2(u^{-\eta} + v^{-\eta} - 1)^2} + \frac{u^{-\eta}(\log u)^2 + v^{-\eta}(\log v)^2}{\eta^2(u^{-\eta} + v^{-\eta} - 1)} \\
& - \frac{3[u^{-\eta}(\log u)^2 + v^{-\eta}(\log v)^2](u^{-\eta}\log u + v^{-\eta}\log v)}{\eta(u^{-\eta} + v^{-\eta} - 1)^2} \\
& + \frac{u^{-\eta}(\log u)^3 + v^{-\eta}(\log v)^3}{\eta(u^{-\eta} + v^{-\eta} - 1)} + \frac{2(u^{-\eta}\log u + v^{-\eta}\log v)^3}{\eta(u^{-\eta} + v^{-\eta} - 1)^3},
\end{aligned}$$

where $\eta = \eta_k$, since the Clayton copula only has one parameter ($k = 1$). By following similar arguments as in Chen and Fan [2006] that there are constants $k_1, k_2 > 0$ and $\epsilon_1, \epsilon_2 > 0$ such that the following inequalities hold for all $u, v \in (0, 1)$ and all $\eta \in \mathcal{A}$:

$$\begin{aligned}
& |\log u| \le k_1 u^{-\epsilon_1}, |\log v| \le k_1 v^{-\epsilon_2}, 0 \le \log(u^{-\eta} + v^{-\eta} - 1) \le k_2(u^{-\epsilon_1} + v^{-\epsilon_2}) \\
& 0 \le \frac{u^{-\eta}}{u^{-\eta} + v^{-\eta} - 1} \le 1, 0 \le \frac{v^{-\eta}}{u^{-\eta} + v^{-\eta} - 1} \le 1.
\end{aligned}$$

Then, we can verify Condition 2 by showing that there are constants ϵ_1, ϵ_2 such that

$$\begin{aligned}\sup_{\eta\in\mathcal{A}} \|\dot{ll}_{\eta}(u,v;\eta)\| \leq & const \times (u^{-\epsilon_1} + v^{-\epsilon_2}) \leq const \times u^{-\epsilon_1} \times v^{-\epsilon_2} \\ \leq & const \times \{u(1-u)\}^{-\epsilon_1} \times \{v(1-v)\}^{-\epsilon_2},\end{aligned}$$

where the last inequality holds due to $(1-u)^{-\epsilon_1} > 1$ and $(1-v)^{-\epsilon_2} > 1$. Likewise, we have

$$\begin{aligned}\sup_{\eta\in\mathcal{A}} \|\ddot{ll}_{\eta\eta}(u,v;\eta)\| \leq & const \times (u^{-\epsilon_1} + v^{-\epsilon_2} + u^{-2\epsilon_1} + v^{-2\epsilon_2} + u^{-\epsilon_1} \times v^{-\epsilon_2}) \\ \leq & const \times \{u(1-u)\}^{-\epsilon_1} \times \{v(1-v)\}^{-\epsilon_2},\end{aligned}$$

which completes the verification of Condition 2. To verify Condition 3, we have

$$\begin{aligned}\sup_{\eta\in\mathcal{A}} \|\ddot{ll}_{\eta,1}(u,v;\eta)\| \leq & const \times (u^{-1}u^{-\epsilon_1} + v^{-\epsilon_2}) \\ \leq & const \times \{u(1-u)\}^{-\tilde{\epsilon}_1} \times \{v(1-v)\}^{-\epsilon_2}, \\ \sup_{\eta\in\mathcal{A}} \|\ddot{ll}_{\eta,2}(u,v;\eta)\| \leq & const \times (u^{-\epsilon_1} + v^{-1}v^{-\epsilon_2}) \\ \leq & const \times \{u(1-u)\}^{\epsilon_1} \times \{v(1-v)\}^{-\tilde{\epsilon}_2},\end{aligned}$$

where $\tilde{\epsilon}_1 = \epsilon_1 + 1 > \epsilon_1, \tilde{\epsilon}_2 = \epsilon_2 + 1 > \epsilon_2$. Similarly,

$$\begin{aligned}\sup_{\eta\in\mathcal{A}} \|\dddot{ll}_{\eta\eta,1}(u,v;\eta)\| \leq & const \times (u^{-1}u^{-\epsilon_1} + v^{-\epsilon_2} + u^{-1} \times u^{-\epsilon_1} \times v^{-\epsilon_2}) \\ \leq & const \times \{u(1-u)\}^{-\tilde{\epsilon}_1} \times \{v(1-v)\}^{-\epsilon_2}, \\ \sup_{\eta\in\mathcal{A}} \|\dddot{ll}_{\eta\eta,2}(u,v;\eta)\| \leq & const \times (u^{-\epsilon_1} + v^{-1}v^{-\epsilon_2} + v^{-1} \times u^{-\epsilon_1} \times v^{-\epsilon_2}) \\ \leq & const \times \{u(1-u)\}^{-\epsilon_1} \times \{v(1-v)\}^{-\tilde{\epsilon}_2},\end{aligned}$$

where $\tilde{\epsilon}_1 = \epsilon_1 + 1 > \epsilon_1, \tilde{\epsilon}_2 = \epsilon_2 + 1 > \epsilon_2$. That completes the verification of Condition 3. To verify Condition 4, given that $\eta \in N(\eta^*)$ with $\eta^* \in [A^{-1}, A]$, $u, v \in (0,1)$, we can see that $\frac{\partial \ddot{ll}_{\eta\eta}(u,v;\eta)}{\partial \eta}$ can be bounded, which completes the verification of Condition 4.

Scenario II right censoring: Similar to interval censoring, the log-likelihood function under right censoring is denoted as

$$
\begin{aligned}
l(S_1, S_2; \eta) =& \log L(S_1, S_2; \eta) \\
=& \delta_1 \delta_2 \log\{c_\eta(S_1, S_2)\} + \delta_1(1-\delta_2) \log\left\{\frac{\partial\, C_\eta(S_1, S_2)}{\partial S_1}\right\} \\
&+ (1-\delta_1)\delta_2 \left\{\frac{\partial\, C_\eta(S_1, S_2)}{\partial S_2}\right\} + (1-\delta_1)(1-\delta_2) C_\eta(S_1, S_2),
\end{aligned}
$$

where $c_\eta(S_1, S_2)$ is the density function of the copula model. Among the four components in the log-likelihood function, the term $\log\{c_\eta(S_1, S_2)\}$ is the most complicated one because it involves the second-order derivative of the copula function whereas the other terms involve at most only the first-order derivative. Therefore, in order to verify the boundedness of the log-likelihood function, it is equivalent to the boundedness of $\log\{c_\eta(u, v)\}, u, v \in (0, 1), \eta \in N(\eta^*)$, with $\eta^* \in \mathcal{A} = [A^{-1}, A]$ for a large $A > 1$. Define $ll(u, v; \eta)$ as

$$
\begin{aligned}
ll(u, v; \eta) =& \log c_\eta(u, v) \\
=& \log(1+\eta) - (\eta+1)\log u - (\eta+1)\log v \\
&- (\eta^{-1}+2)\log(u^{-\eta} + v^{-\eta} - 1).
\end{aligned}
$$

Then, the derivatives of $ll(u, v; \eta)$ are:

$$
\begin{aligned}
\dot{ll}_\eta(u, v; \eta) =& \frac{1}{1+\eta} - \log(uv) + \frac{\log(u^{-\eta} + v^{-\eta} - 1)}{\eta^2} \\
&+ (\eta^{-1}+2)\frac{u^{-\eta}\log u + v^{-\eta}\log v}{u^{-\eta} + v^{-\eta} - 1}, \\
\ddot{ll}_{\eta\eta}(u, v; \eta) =& -\frac{1}{(1+\eta)^2} - \frac{2}{\eta^3}\log(u^{-\eta} + v^{-\eta} - 1) - \frac{2(u^{-\eta}\log u + v^{-\eta}\log v)}{\eta^2(u^{-\eta} + v^{-\eta} - 1)} \\
&+ (\eta^{-1}+2)\left\{\frac{(u^{-\eta}\log u + v^{-\eta}\log v)^2}{(u^{-\eta} + v^{-\eta} - 1)^2} - \frac{u^{-\eta}(\log u)^2 + v^{-\eta}(\log v)^2}{(u^{-\eta} + v^{-\eta} - 1)}\right\} \\
\ddot{ll}_{\eta,1}(u, v; \eta) =& \frac{-1}{u} + \frac{(1+2\eta)\{v^{-\eta}(\log v - \log u) + \log u\} + 2(u^{-\eta} + v^{-\eta} - 1)}{(u^{-\eta} + v^{-\eta} - 1)^2 u^{\eta+1}}, \\
\ddot{ll}_{\eta,2}(u, v; \eta) =& \frac{-1}{v} + \frac{(1+2\eta)\{u^{-\eta}(\log u - \log v) + \log v\} + 2(u^{-\eta} + v^{-\eta} - 1)}{(u^{-\eta} + v^{-\eta} - 1)^2 v^{\eta+1}},
\end{aligned}
$$

$$\begin{aligned}
\dddot{ll}_{\eta\eta,1}(u,v;\eta) =& -\frac{2v^{-\eta}(\log v - \log u) + 2\log u}{\eta(u^{-\eta}+v^{-\eta}-1)^2 u^{\eta+1}} + (\eta^{-1}+2)\left[\frac{2\eta(u^{-\eta}\log u + v^{-\eta}\log v)^2}{(u^{-\eta}+v^{-\eta}-1)^3 u^{\eta+1}}\right. \\
&+ \frac{2(u^{-\eta}\log u + v^{-\eta}\log v)(-\eta\log u + 1) - \eta\{u^{-\eta}(\log u)^2 + v^{-\eta}(\log v)^2\}}{(u^{-\eta}+v^{-\eta}-1)^2 u^{\eta+1}} \\
&\left. - \frac{-\eta(\log u)^2 + 2\log u}{(u^{-\eta}+v^{-\eta}-1)u^{\eta+1}}\right] \\
\dddot{ll}_{\eta\eta,2}(u,v;\eta) =& -\frac{2u^{-\eta}(\log u - \log v) + 2\log v}{\eta(u^{-\eta}+v^{-\eta}-1)^2 v^{\eta+1}} + (\eta^{-1}+2)\left[\frac{2\eta(u^{-\eta}\log u + v^{-\eta}\log v)^2}{(u^{-\eta}+v^{-\eta}-1)^3 v^{\eta+1}}\right. \\
&+ \frac{2(u^{-\eta}\log u + v^{-\eta}\log v)(-\eta\log v + 1) - \eta\{u^{-\eta}(\log u)^2 + v^{-\eta}(\log v)^2\}}{(u^{-\eta}+v^{-\eta}-1)^2 v^{\eta+1}} \\
&\left. - \frac{-\eta(\log v)^2 + 2\log v}{(u^{-\eta}+v^{-\eta}-1)v^{\eta+1}}\right] \\
\frac{\partial \ddot{l}l_{\eta\eta}(u,v;\eta)}{\partial\eta} =& \frac{2(u^{-\eta}\log u + v^{-\eta}\log v)}{\eta^3(u^{-\eta}+v^{-\eta}-1)} - \frac{2(u^{-\eta}\log u + v^{-\eta}\log v)^2}{\eta^2(u^{-\eta}+v^{-\eta}-1)^2} \\
&+ (\eta^{-1}+2)\left[\frac{2(u^{-\eta}\log u + v^{-\eta}\log v)^3}{(u^{-\eta}+v^{-\eta}-1)^3}\right. \\
&- \frac{3(u^{-\eta}\log u + v^{-\eta}\log v)\{u^{-\eta}(\log u)^2 + v^{-\eta}(\log v)^2\}}{(u^{-\eta}+v^{-\eta}-1)^2} \\
&\left. + \frac{u^{-\eta}(\log u)^3 + v^{-\eta}(\log v)^3}{(u^{-\eta}+v^{-\eta}-1)}\right]
\end{aligned}$$

The verification procedures follow similar steps as the verification under interval censoring. We will omit the details.

B.3 Proof for Theorem 3.2.1

Proof for Theorem 3.2.1. Define the rescaled empirical copula of $(\tilde{S}_{i1}, \tilde{S}_{i2}), i = 1, ..., n$ by

$$\tilde{C}(s_1, s_2) = \frac{1}{n+1}\sum_{i=1}^{n} I\{\tilde{S}_{i1} \le s_1, \tilde{S}_{i2} \le s_2\}.$$

For any $\eta \in \Theta$, we can rewrite $S(\eta), \tilde{S}(\eta), V(\eta), \tilde{V}(\eta)$ as follows:

$$S(\eta) = -\int_{s_1,s_2\in[0,1]} \ddot{l}_{\eta\eta}(s_1, s_2; \eta)dC_0(s_1, s_2);$$

$$\tilde{S}(\eta) = -\frac{n+1}{n}\int_{s_1,s_2\in[0,1]} \ddot{l}_{\eta\eta}(s_1, s_2; \eta)d\tilde{C}(s_1, s_2),$$

and

$$V(\eta) = \int_{s_1,s_2\in[0,1]} \dot{l}_\eta(s_1,s_2;\eta)\dot{l}_\eta^T(s_1,s_2;\eta)dC_0(s_1,s_2);$$

$$\tilde{V}(\eta) = \frac{n+1}{n}\int_{s_1,s_2\in[0,1]} \dot{l}_\eta(s_1,s_2;\eta)\dot{l}_\eta^T(s_1,s_2;\eta)d\tilde{C}(s_1,s_2),$$

where $C_0(\cdot)$ and $\tilde{C}(\cdot)$ are the true copula and the rescaled empirical copula.

By Condition 2, applying Lemma 1(c) in Chen and Fan [2005], we have

$$\begin{aligned}&\sup_{\eta\in N(\eta^*)} \|\tilde{S}(\eta) - S(\eta)\| \\ = &\sup_{\eta\in N(\eta^*)} \|\int_{s_1,s_2\in[0,1]} \ddot{l}_{\eta\eta}(s_1,s_2;\eta)d\{\frac{n+1}{n}\tilde{C}(s_1,s_2) - C_0(s_1,s_2)\}\| \xrightarrow{p} 0,\end{aligned}$$

as $n\to\infty$.

Therefore, using the two facts $\|\tilde{S}(\hat{\eta}) - S(\eta^*)\| \le \|\tilde{S}(\hat{\eta}) - S(\hat{\eta})\| + \|S(\hat{\eta}) - S(\eta^*)\|$, and $\hat{\eta}\xrightarrow{p}\eta^*$, we obtain $\tilde{S}(\hat{\eta})\xrightarrow{p} S(\eta^*)$. Applying the same arguments above, we can also show that $\tilde{V}(\hat{\eta})\xrightarrow{p} V(\eta^*)$. Furthermore, by Condition 1 and Slutsky's theorem, we have

$$\widehat{IR}_n = tr\{\tilde{S}^{-1}(\hat{\eta})\tilde{V}(\hat{\eta})\} \xrightarrow{p} tr\{S^{-1}(\eta^*)V(\eta^*)\} = p.$$

□

B.4 Proof for Theorem 3.2.2

Proof for Theorem 3.2.2. First note that, $\hat{\eta}$ is the solution to $\sum_{i=1}^n \dot{l}_\eta(\tilde{S}_{i1},\tilde{S}_{i2};\hat{\eta}) = 0$. Applying the Mean Value Theorem, we have

$$0 = \sum_{i=1}^n \dot{l}_\eta(\tilde{S}_{i1},\tilde{S}_{i2};\eta^*) + \sum_{i=1}^n \ddot{l}_{\eta\eta}(\tilde{S}_{i1},\tilde{S}_{i2};\tilde{\eta})(\hat{\eta}-\eta^*),$$

where $\tilde{\eta}$ lies between η^* and $\hat{\eta}$. Therefore,

$$\hat{\eta}-\eta^* = -\left[\frac{1}{n}\sum_{i=1}^n \ddot{l}_{\eta\eta}(\tilde{S}_{i1},\tilde{S}_{i2};\tilde{\eta})\right]^{-1}\frac{1}{n}\sum_{i=1}^n \dot{l}_\eta(\tilde{S}_{i1},\tilde{S}_{i2};\eta^*).$$

For any $1 \leq k, l \leq p$, expanding $\ddot{l}_{\eta\eta}(\tilde{S}_{i1}, \tilde{S}_{i2}; \hat{\eta})_{kl}$, which is an element in the $p \times p$ matrix $\ddot{l}_{\eta\eta}(\tilde{S}_{i1}, \tilde{S}_{i2}; \hat{\eta})$, around η^* leads to

$$
\begin{aligned}
\tilde{S}(\hat{\eta})_{kl} =& \frac{1}{n}\sum_{i=1}^{n} \ddot{l}_{\eta\eta}(\tilde{S}_{i1}, \tilde{S}_{i2}; \hat{\eta})_{kl} \\
=& \frac{1}{n}\sum_{i=1}^{n} \ddot{l}_{\eta\eta}(\tilde{S}_{i1}, \tilde{S}_{i2}; \eta^*)_{kl} + \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ddot{l}_{\eta\eta}(\tilde{S}_{i1}, \tilde{S}_{i2}; \breve{\eta})_{kl}}{\partial \eta^T}(\hat{\eta} - \eta^*) \\
=& \frac{1}{n}\sum_{i=1}^{n} \ddot{l}_{\eta\eta}(\tilde{S}_{i1}, \tilde{S}_{i2}; \eta^*)_{kl} - \frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ddot{l}_{\eta\eta}(\tilde{S}_{i1}, \tilde{S}_{i2}; \breve{\eta})_{kl}}{\partial \eta^T} \\
& \times \left[\frac{1}{n}\sum_{i=1}^{n} \ddot{l}_{\eta\eta}(\tilde{S}_{i1}, \tilde{S}_{i2}; \tilde{\eta})\right]^{-1} \frac{1}{n}\sum_{i=1}^{n} \dot{l}_{\eta}(\tilde{S}_{i1}, \tilde{S}_{i2}; \eta^*),
\end{aligned}
$$

where $\breve{\eta}$ lies between η^* and $\hat{\eta}$.

By Condition 4, applying again Lemma 1(c) in Chen and Fan [2005], we obtain

$$
\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ddot{l}_{\eta\eta}(\tilde{S}_{i1}, \tilde{S}_{i2}; \breve{\eta})_{kl}}{\partial \eta^T} \rightarrow_{pr} P_0\{\frac{\partial \ddot{l}_{\eta\eta}(S_1, S_2; \eta^*)_{kl}}{\partial \eta^T}\}.
$$

Also, we know $-\frac{1}{n}\sum_{i=1}^{n} \ddot{l}_{\eta\eta}(\tilde{S}_{i1}, \tilde{S}_{i2}; \tilde{\eta}) \rightarrow_{pr} S(\eta^*)$ as $n \rightarrow \infty$. Therefore

$$
\begin{aligned}
\tilde{S}(\hat{\eta})_{kl} =& \frac{1}{n}\sum_{i=1}^{n} \{\ddot{l}_{\eta\eta}(\tilde{S}_{i1}, \tilde{S}_{i2}; \eta^*)_{kl} + M_1^{kl} S^{-1}(\eta^*)\dot{l}_{\eta}(\tilde{S}_{i1}, \tilde{S}_{i2}; \eta^*)\} + o_p(1) \\
\triangleq& \frac{1}{n}\sum_{i=1}^{n} h_S(\tilde{S}_{i1}, \tilde{S}_{i2}; \eta^*)_{kl} + o_p(1),
\end{aligned}
$$

where $M_1^{kl} \triangleq P_0\{\frac{\partial \ddot{l}_{\eta\eta}(S_1,S_2;\eta^*)_{kl}}{\partial \eta^T}\}$ is a $1 \times p$ vector, h_S is a $p \times p$ matrix with element $h_S(\tilde{S}_{i1}, \tilde{S}_{i2}; \eta^*)_{kl}$.

Employing the same arguments above, we have

$$
\begin{aligned}
\tilde{V}(\hat{\eta})_{kl} =& \frac{1}{n}\sum_{i=1}^{n} \{\dot{l}_{\eta}(\tilde{S}_{i1}, \tilde{S}_{i2}; \eta^*)_k \dot{l}_{\eta}(\tilde{S}_{i1}, \tilde{S}_{i2}; \eta^*)_l \\
& + M_2^{kl} S^{-1}(\eta^*)\dot{l}_{\eta}(\tilde{S}_{i1}, \tilde{S}_{i2}; \eta^*)\} + o_p(1) \\
\triangleq& \frac{1}{n}\sum_{i=1}^{n} h_V(\tilde{S}_{i1}, \tilde{S}_{i2}; \eta^*)_{kl} + o_p(1),
\end{aligned}
$$

where $M_2^{kl} = P_0\{\frac{\partial \dot{l}_{\eta}(S_1,S_2;\eta^*)_k}{\partial \eta^T}\dot{l}_{\eta}(S_1, S_2; \eta^*)_l + \frac{\partial \dot{l}_{\eta}(S_1,S_2;\eta^*)_l}{\partial \eta^T}\dot{l}_{\eta}(S_1, S_2; \eta^*)_k\}$, h_V is a $p \times p$ matrix with element $h_V(\tilde{S}_{i1}, \tilde{S}_{i2}; \eta^*)_{kl}$.

Under the null hypothesis of the copula model being correctly specified, by the Bartlett identity, we have $S(\eta^*) = V(\eta^*)$. Moreover, the IR test statistic estimator $\widehat{IR}_n$ can be represented as follows:

$$\begin{aligned}
\sqrt{n}(\widehat{IR}_n - p) =& \sqrt{n} tr\{\tilde{S}^{-1}(\hat{\eta})\tilde{V}(\hat{\eta}) - I_p\} \\
=& \sqrt{n} tr\{\tilde{S}^{-1}(\hat{\eta})\tilde{V}(\hat{\eta}) - S^{-1}(\eta^*)V(\eta^*)\} \\
=& tr\Big[S^{-1}(\eta^*)\sqrt{n}\{\tilde{V}(\hat{\eta}) - V(\eta^*)\}\Big] \\
&+ tr\Big[S^{-1}(\eta^*)\tilde{V}(\hat{\eta})S^{-1}(\eta^*)\sqrt{n}\{S(\eta^*) - \tilde{S}(\hat{\eta})\}\Big] \\
&+ tr\Big[\tilde{S}^{-1}(\hat{\eta})\tilde{V}(\hat{\eta})S^{-2}(\eta^*)\sqrt{n}\{S(\eta^*) - \tilde{S}(\hat{\eta})\}^2\Big].
\end{aligned}$$

Utilizing the asymptotic expansions above, we have

$$\begin{aligned}
&\sqrt{n}\{\tilde{S}(\hat{\eta}) - S(\eta^*)\} \\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{h_S(\tilde{S}_{i1}, \tilde{S}_{i2}; \eta^*) - S(\eta^*)\} + o_p(1) \\
=&\sqrt{n}\int_{s_1, s_2 \in (0,1)} h_S(s_1, s_2; \eta^*) d\{\frac{n+1}{n}\tilde{C}(s_1, s_2) - C_0(s_1, s_2)\} + o_p(1),
\end{aligned}$$

and

$$\begin{aligned}
&\sqrt{n}\{\tilde{V}(\hat{\eta}) - V(\eta^*)\} \\
=&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{h_V(\tilde{S}_{i1}, \tilde{S}_{i2}; \eta^*) - V(\eta^*)\} + o_p(1) \\
=&\sqrt{n}\int_{s_1, s_2 \in (0,1)} h_V(s_1, s_2; \eta^*) d\{\frac{n+1}{n}\tilde{C}(s_1, s_2) - C_0(s_1, s_2)\} + o_p(1).
\end{aligned}$$

By Conditions 2 and 3, employing Lemma 2 in Chen and Fan [2005]. we have $\|\tilde{S}(\hat{\eta} - S(\eta^*))\| = O_p(n^{-1/2})$ and $\|\tilde{V}(\hat{\eta} - V(\eta^*))\| = O_p(n^{-1/2})$. In addition, given these facts: $\sqrt{n}\|\tilde{S}(\hat{\eta}) - S(\eta^*)\|^2 = o_p(1)$, $\tilde{S}(\hat{\eta}) \rightarrow_{pr} S(\eta^*)$ and $\tilde{V}(\hat{\eta}) \rightarrow_{pr} V(\eta^*)$, we reach the following expression:

$$\sqrt{n}(\widehat{IR}_n - p) = \sqrt{n}\int_{s_1, s_2 \in [0,1]} h_R(s_1, s_2; \eta^*) d\{\frac{n+1}{n}\tilde{C}(s_1, s_2) - C_0(s_1, s_2)\} + o_p(1),$$

where $h_R(s_1, s_2; \eta^*) = \sum_{k,l=1}^{p} S^{-1}(\eta^*)_{kl}\{h_S(s_1, s_2; \eta^*)_{lk} + h_V(s_1, s_2; \eta^*)_{lk}\}$.

Again, applying Lemma 2 in Chen and Fan [2005], we have

$$\sqrt{n}(\widehat{IR}_n - p) \rightarrow_d N(0, \sigma_R^2),$$

where $\sigma_R^2 = var_0\{h_R(s_1, s_2; \eta^*) + D(S_1, S_2; \eta^*)\}$, and

$$D(S_1, S_2; \eta^*) = \sum_{j=1}^{2} \int_{s_1, s_2 \in [0,1]} \frac{\partial h_R(s_1, s_2; \eta^*)}{\partial s_j} I(S_j \leq s_j) dC_0(s_1, s_2).$$

Note that the additional term $D(S_1, S_2; \eta^*)$ comes from the uncertainty of the estimator for the marginal distributions $S_1(\cdot)$ and $S_2(\cdot)$. It vanishes when the distributions are known.

The asymptotic variance of σ_R^2 may be consistently estimated by

$$\hat{\sigma}_R^2 = \frac{1}{n} \sum_{i=1}^{n} \left[h_R(\tilde{S}_{i1}, \tilde{S}_{i2}; \hat{\eta}) - \sum_{k,l=1}^{p} \tilde{S}(\hat{\eta})_{kl}^{-1} \tilde{V}(\hat{\eta})_{lk} + D(\tilde{S}_{i1}, \tilde{S}_{i2}; \hat{\eta}) \right]^2.$$

□

Bibliography

Kjersti Aas, Claudia Czado, Arnoldo Frigessi, and Henrik Bakken. Pair-copula constructions of multiple dependence. *Insurance: Mathematics and Economics*, 44(2):182 – 198, 2009.

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation*, pages 265–283, 2016.

Jeffrey Abrams, Barbara Conley, Margaret Mooney, James Zwiebel, Alice Chen, John J Welch, et al. National cancer institute's precision medicine initiatives for the new national clinical trials network. *American Society of Clinical Oncology educational book. American Society of Clinical Oncology. Annual Meeting*, pages 71–76, 2014.

Per K Andersen, Claus T Ekstrøm, John P Klein, Youyi Shu, and Mei-Jie Zhang. A class of goodness of fit tests for a copula based on bivariate right-censored data. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 47(6):815–824, 2005.

AREDS Group. The Age-Related Eye Disease Study (AREDS): design implications. AREDS report no. 1. *Controlled Clinical Trials*, 20(6):573–600, 1999.

Emmanuel Barillot, Laurence Calzone, Andrei Zinovyev, Philippe Hupe, and Jean-Philippe Vert. *Computational systems biology of cancer.* CRC Press, 2012.

Steve Bennett. Analysis of survival data by the proportional odds model. *Statistics in medicine*, 2(2):273–277, 1983.

Rebecca Betensky, Daniel Rabinowitz, and Anastasios Tsiatis. Computationally simple accelerated failure time regression for interval censored data. *Biometrika*, 88(3):703–711, 2001.

Kris Bogaerts and Emmanuel Lesaffre. Modeling the association of bivariate interval-censored data using the copula approach. *Statistics in medicine*, 27(30):6379–6392, 2008.

Tomasz Burzykowski, Geert Molenberghs, Marc Buyse, Helena Geys, and Didier Renard. Validation of surrogate end points in multiple randomized clinical trials with failure time end points. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(4): 405–422, 2001.

José A Castro-Rodríguez, Catharine J Holberg, Anne L Wright, and Fernando D Martinez. A clinical index to define risk of asthma in young children with recurrent wheezing. *American journal of respiratory and critical care medicine*, 162(4):1403–1406, 2000.

I Shou Chang, Chi Chung Wen, and Yuh Jenn Wu. A profile likelihood theory for the correlated gamma-frailty model with current status family data. *Statistica Sinica*, 17(3): 1023–1046, 2007.

Han Chen, Jennifer E Huffman, Jennifer A Brody, Chaolong Wang, Seunggeun Lee, Zilin Li, Stephanie M Gogarten, Tamar Sofer, Lawrence F Bielak, Joshua C Bis, et al. Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. *The American Journal of Human Genetics*, 104(2): 260–274, 2019.

Kani Chen, Zhezhen Jin, and Zhiliang Ying. Semiparametric analysis of transformation models with censored data. *Biometrika*, 89(3):659–668, 2002.

Ling Chen and Jianguo Sun. A multiple imputation approach to the analysis of current status data with the additive hazards model. *Communications in Statistics - Theory and Methods*, 38(7):1009–1018, 2009.

Man-Hua Chen, Xingwei Tong, and Jianguo Sun. The proportional odds model for multivariate interval-censored failure time data. *Statistics in medicine*, 26(28):5147–5161, 2007.

Man-Hua Chen, Xingwei Tong, and Jianguo Sun. A frailty model approach for regression analysis of multivariate current status data. *Statistics in medicine*, 28(27):3424–3436, 2009.

Man-Hua Chen, Xingwei Tong, and Liang Zhu. A linear transformation model for multivariate interval-censored failure time data. *Canadian Journal of Statistics*, 41(2):275–290, 2013.

Man-Hua Chen, Li-Ching Chen, Kuen-Hung Lin, and Xingwei Tong. Analysis of multivariate interval censoring by diabetic retinopathy study. *Communications in Statistics-Simulation and Computation*, 43(7):1825–1835, 2014.

Xiaohong Chen and Yanqin Fan. Pseudo-likelihood ratio tests for semiparametric multivariate copula model selection. *Canadian Journal of Statistics*, 33(3):389–414, 2005.

Xiaohong Chen and Yanqin Fan. Estimation of copula-based semiparametric time series models. *Journal of Econometrics*, 130(2):307–335, 2006.

Xiaohong Chen, Yanqin Fan, Demian Pouzo, and Zhiliang Ying. Estimation and model selection of semiparametric multivariate survival functions under general censorship. *Journal of econometrics*, 157(1):129–142, 2010.

S. C. Cheng, L. J. Wei, and Z. Ying. Analysis of transformation models with censored data. *Biometrika*, 82(4):835–845, 1995.

Emily Y Chew, Traci Clemons, John Paul SanGiovanni, Ronald Danis, Amitha Domalpally, Wendy McBee, Robert Sperduto, Frederick L Ferris, AREDS2 Research Group, et al. The age-related eye disease study 2 (areds2): study design and baseline characteristics (areds2 report number 1). *Ophthalmology*, 119(11):2282–2289, 2012.

Chih-Lin Chi, W Nick Street, and William H Wolberg. Application of artificial neural network-based survival analysis on two breast cancer datasets. In *AMIA Annual Symposium Proceedings*, volume 2007, page 130. American Medical Informatics Association, 2007.

Lynda Chin, Jannik N Andersen, and P Andrew Futreal. Cancer genomics: from discovery science to personalized medicine. *Nature medicine*, 17(3):297, 2011.

Travers Ching, Xun Zhu, and Lana X Garmire. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. *PLoS computational biology*, 14(4):e1006076, 2018.

François Chollet et al. Keras. `https://keras.io`, 2015.

Francois Chollet and J. J. Allaire. *Deep Learning with R*. Manning Publications Co., Greenwich, CT, USA, 1st edition, 2018. ISBN 161729554X, 9781617295546.

David Clayton and Jack Cuzick. Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society: Series A (General)*, 148(2):82–108, 1985.

David G. Clayton. A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65 (1):141–151, 1978.

Francis S Collins and Harold Varmus. A new initiative on precision medicine. *New England journal of medicine*, 372(9):793–795, 2015.

Carolyn Compton. Precision medicine core: Progress in prognostication—populations to patients. *Annals of Surgical Oncology*, 25(2):349–350, 2018.

Richard J. Cook and David Tolusso. Second-order estimating equations for the analysis of clustered current status data. *Biostatistics*, 10(4):756–772, 2009.

Richard J Cook, Leilei Zeng, and Ker-Ai Lee. A multistate model for bivariate interval-censored failure time data. *Biometrics*, 64(4):1100–1109, 2008.

D. R. Cox and D. V. Hinkley. *Theoretical Statistics*. Chapman & Hall/CRC, London, 1979.

David R Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Ying Ding and Bin Nan. A sieve m-theorem for bundled parameters in semiparametric models, with application to the efficient estimation in a linear model for censored data. *Annals of Statistics*, 39(1):2795–3443, 2011.

Ying Ding, Yi Liu, Qi Yan, Lars G. Fritsche, Richard J. Cook, Traci Clemons, Rinki Ratnapriya, Michael L. Klein, Gonçalo R. Abecasis, Anand Swaroop, Emily Y. Chew, Daniel E. Weeks, Wei Chen, and the AREDS2 Research Group. Bivariate analysis of Age-Related Macular Degeneration progression using genetic risk scores. *Genetics*, 206(1): 119–133, 2017.

Bradley Efron. The efficiency of cox's likelihood function for censored data. *Journal of the American statistical Association*, 72(359):557–565, 1977.

Takeshi Emura. *Copula.surv: Association Analysis of Bivariate Survival Data Based on Copulas*, 2018. URL `https://CRAN.R-project.org/package=Copula.surv`. R package version 1.0.

Takeshi Emura, Chien-Wei Lin, and Weijing Wang. A goodness-of-fit test for archimedean copula models in the presence of right censoring. *Computational Statistics & Data Analysis*, 54(12):3033–3043, 2010.

David Faraggi and Richard Simon. A neural network model for survival data. *Statistics in medicine*, 14(1):73–82, 1995.

Jason P Fine and Hongyu Jiang. On association in a copula with time transformations. *Biometrika*, 87(3):559–571, 2000.

JP Fine, Z Ying, and LG Wei. On the linear transformation model for censored data. *Biometrika*, 85(4):980–986, 1998.

Dianne M. Finkelstein. A proportional hazards model for interval-censored failure time data. *Biometrics*, 42(4):845–854, 1986.

Dianne M Finkelstein, William B Goggins, and David A Schoenfeld. Analysis of failure time data with dependent interval censoring. *Biometrics*, 58(2):298–304, 2002.

Lars G. Fritsche, W. IgI, J. N. Bailey, et al. A large genome-wide association study of Age-related Macular Degeneration highlights contributions of rare and common variants. *Nature Genetics*, 48(2):134–143, 2016.

Christian Genest and Louis-Paul Rivest. Statistical inference procedures for bivariate archimedean copulas. *Journal of the American statistical Association*, 88(423):1034–1043, 1993.

Christian Genest, Jean-François Quessy, and Bruno Rémillard. Goodness-of-fit procedures for copula models based on the probability integral transformation. *Scandinavian Journal of Statistics*, 33(2):337–366, 2006.

Thomas A Gerds and Martin Schumacher. Consistent estimation of the expected brier score in general survival models with right-censored event times. *Biometrical Journal*, 48(6): 1029–1040, 2006.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010.

Klara Goethals, Paul Janssen, and Luc Duchateau. Frailty models and copulas: similarities and differences. *Journal of Applied Statistics*, 35(9):1071–1079, 2008.

William B. Goggins and Dianne M. Finkelstein. A proportional hazards model for multivariate interval-censored failure time data. *Biometrics*, 56:940–943, 2000.

Erika Graf, Claudia Schmoor, Willi Sauerbrei, and Martin Schumacher. Assessment and comparison of prognostic classification schemes for survival data. *Statistics in medicine*, 18(17-18):2529–2545, 1999.

Felix Grassmann et al. A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. *Ophthalmology*, 125(9):1410–1420, 2018.

Ming Gao Gu, Liuquan Sun, and Guoxin Zuo. A baseline-free procedure for transformation models under interval censorship. *Lifetime Data Analysis*, 11(4):473–488, 2005.

Emil Julius Gumbel. Bivariate exponential distributions. *Journal of the American Statistical Association*, 55(292):698–707, 1960.

SW Guo and DY Lin. Regression analysis of multivariate grouped survival data. *Biometrics*, pages 632–639, 1994.

Jie Hao, Youngsoon Kim, Tejaswini Mallavarapu, Jung Hun Oh, and Mingon Kang. Coxpasnet: pathway-based sparse deep neural network for survival analysis. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 381–386. IEEE, 2018.

Frank E Harrell, Kerry L Lee, and Daniel B Mark. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in medicine*, 15(4):361–387, 1996.

Patrick J Heagerty, Thomas Lumley, and Margaret S Pepe. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56(2):337–344, 2000.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent (2012). *URL https://www. cs. toronto. edu/˜ tijmen/csc321/slides/lecture_slides_lec6. pdf*, 2012.

Marius Hofert, Ivan Kojadinovic, Martin Maechler, and Jun Yan. *copula: Multivariate Dependence with Copulas*, 2018. URL `https://CRAN.R-project.org/package=copula`. R package version 0.999-19.

Theodore R Holford. Life tables with concomitant information. *Biometrics*, pages 587–597, 1976.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Philip Hougaard. A class of multivanate failure time distributions. *Biometrika*, 73(3):671–678, 1986.

Philip Hougaard. *Analysis of multivariate survival data*. Springer Science & Business Media, 2012.

Tao Hu, Qingning Zhou, and Jianguo Sun. Regression analysis of bivariate current status data under the proportional hazards model. *Canadian Journal of Statistics*, 45(4):410–424, 2017.

Jian Huang and AJ Rossini. Sieve estimation for the proportional-odds failure-time regression model with interval censoring. *Journal of the American Statistical Association*, 92(439): 960–967, 1997.

Jian Huang and Jon A Wellner. Interval censored survival data: a review of recent progress. In *Proceedings of the First Seattle Symposium in Biostatistics*, pages 123–169. Springer, 1997.

Jian Huang et al. Efficient estimation for the proportional hazards model with interval censoring. *The Annals of Statistics*, 24(2):540–568, 1996.

Wanling Huang and Artem Prokhorov. A goodness-of-fit test for copulas. *Econometric Reviews*, 33(7):751–771, 2014.

H. Ishwaran and U.B. Kogalur. Random survival forests for r. *R News*, 7(2):25–31, October 2007. URL `https://CRAN.R-project.org/doc/Rnews/`.

Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 09 2008.

H. Joe. *Multivariate Models and Dependence Concepts*. Chapman & Hall, London, 1997.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*, volume 360. John Wiley & Sons, 2011.

Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):24, 2018.

Kaveh Kiani and Jayanthi Arasan. Simulation of interval-censored data in medical and biological studies. *International Journal of Modern Physics*, 9:112–118, 2012.

Mimi Y Kim and Xiaonan Xue. The analysis of multivariate interval-censored survival data. *Statistics in Medicine*, 21(23):3715–3726, 2002.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *Advances in neural information processing systems*, pages 971–980, 2017.

John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2006.

Ivan Kojadinovic and Jun Yan. Modeling multivariate distributions with continuous margins using the copula R package. *Journal of Statistical Software*, 34(9):1–20, 2010. doi: 10.18637/jss.v034.i09.

Chew Teng Kor, Kuang Fu Cheng, and Yi Hau Chen. A method for analyzing clustered interval-censored data based on Cox model. *Statistics in Medicine*, 32:822–832, 2013.

ML Lakhal-Chaieb. Copula inference under censoring. *Biometrika*, 97(2):505–512, 2010.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

Linxiong Li and Zongwei Pu. Rank estimation of log-linear regression with interval-censored data. *Lifetime data analysis*, 9(1):57–70, 2003.

Faming Liang, Qizhai Li, and Lei Zhou. Bayesian neural networks for selection of drug sensitive genes. *Journal of the American Statistical Association*, 113(523):955–972, 2018.

DY Lin, David Oakes, and Zhiliang Ying. Additive hazards regression with current status data. *Biometrika*, 85(2):289–298, 1998.

Juan Lin and Ximing Wu. A diagnostic test for specification of copulas under censorship. *Econometric Reviews*, pages 1–17, 2020.

George R. Lindfield and John E. T. Penny. *Microcomputers in Numerical Analysis.* Halsted Press, New York, 1989.

George G. Lorentz. *Bernstein Polynomials.* Chelsea Publishing Company, 1986.

Amita K Manatunga and David Oakes. Parametric analysis for matched pair survival data. *Lifetime Data Analysis*, 5(4):371–387, 1999.

Giampiero Marra and Rosalba Radice. Bivariate copula additive models for location, scale and shape. *Computational Statistics & Data Analysis*, 112:99–113, 2017.

Giampiero Marra and Rosalba Radice. Copula link-based additive models for right-censored event time data. *Journal of the American Statistical Association*, pages 1–20, 2019.

Giampiero Marra and Rosalba Radice. *GJRM: Generalised Joint Regression Modelling*, 2020. URL `https://CRAN.R-project.org/package=GJRM`. R package version 0.2-2.

Giampiero Marra, Rosalba Radice, Till Bärnighausen, Simon N Wood, and Mark E McGovern. A simultaneous equation approach to estimating hiv prevalence with nonignorable missing responses. *Journal of the American Statistical Association*, 112(518):484–496, 2017.

Albert W Marshall and Ingram Olkin. Families of multivariate distributions. *Journal of the American Statistical Association*, 83(403):834–841, 1988.

Guido Masarotto and Cristiano Varin. Gaussian copula regression in R. *Journal of Statistical Software*, 77(8):1–26, 2017. doi: 10.18637/jss.v077.i08.

CA McGilchrist and CW Aisbett. Regression with frailty in survival analysis. *Biometrics*, pages 461–466, 1991.

Moyan Mei. A goodness-of-fit test for semi-parametric copula models of right-censored bivariate survival times. Master's thesis, Simon Fraser University, 2016.

Xinlei Mi, Fei Zou, and Ruoqing Zhu. Bagging and deep learning in optimal individualized treatment rules. *Biometrics*, 75(2):674–684, 2019.

Seonwoo Min, Byunghan Lee, and Sungroh Yoon. Deep learning in bioinformatics. *Briefings in Bioinformatics*, 18(5):851–869, 07 2016.

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6): 1236–1246, 2017.

Thomas Nagler and Thibault Vatter. *gamCopula: Generalized Additive Models for Bivariate Conditional Dependence Structures and Vine Copulas*, 2020. URL `https://CRAN.R-project.org/package=gamCopula`. R package version 0.0-7.

Roger B. Nelsen. *An Introduction to Copulas*. Springer-Verlag, New York, 2006.

Daniel Silvestrini Nicole Kraemer. *Bivariate Copula Based Regression Models*, 2014. URL `https://cran.r-project.org/web/packages/CopulaRegression/`. R package version 0.1-5.

David Oakes. A model for association in bivariate survival data. *Journal of the Royal Statistical Society: Series B*, 44(3):414–422, 1982.

David Oakes. Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, 84(406):487–493, 1989.

Ryan Poplin, Avinash V Varadarajan, Katy Blumer, Yun Liu, Michael V McConnell, Greg S Corrado, Lily Peng, and Dale R Webster. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2(3):158, 2018.

Leen Prenen, Roel Braekers, and Luc Duchateau. Extending the archimedean copula methodology to model multivariate survival data grouped in clusters of variable size. *Journal of the Royal Statistical Society: Series B*, 79(2):483–505, 2017a.

Leen Prenen, Roel Braekers, Luc Duchateau, and Ewoud De Troyer. *Sunclarco: Survival Analysis using Copulas*, 2017b. URL `https://CRAN.R-project.org/package=Sunclarco`. R package version 1.0.0.

Daniel Rabinowitz, Anastasios Tsiatis, and Jorge Aragon. Regression with interval-censored data. *Biometrika*, 82(3):501–513, 1995.

Daniel Rabinowitz, Rebecca A Betensky, and Anastasios A Tsiatis. Using conditional logistic regression to fit proportional odds models to interval censored data. *Biometrics*, 56(2): 511–518, 2000.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144. ACM, 2016.

AJ Rossini and AA Tsiatis. A semiparametric proportional odds regression model for the analysis of current status data. *Journal of the American Statistical Association*, 91(434): 713–721, 1996.

Federico Rotolo, Xavier Paoletti, and Stefan Michiels. Surrosurv: an r package for the evaluation of failure time surrogate endpoints in individual patient data meta-analyses of randomized clinical trials. *Computer methods and programs in biomedicine*, 155:189–198, 2018.

Chloé Sarnowski et al. Whole genome sequence analyses of brain imaging measures in the framingham study. *Neurology*, 90(3):e188–e196, 2018.

Ulf Schepsmeier, Jakob Stoeber, Eike Christian Brechmann, Benedikt Graeler, Thomas Nagler, Tobias Erhardt, Carlos Almeida, Aleksey Min, Claudia Czado, Mathias Hofmann, et al. *VineCopula: Statistical Inference of Vine Copulas*, 2018. URL `https://CRAN.R-project.org/package=VineCopula`. R package version 2.1.8.

Martin Schumacher, Norbert Hollander, Guido Schwarzer, Harald Binder, and Willi Sauerbrei. *Prognostic factor studies; in Crowley J, Hoering A (eds): Handbook of Statistics in Clinical Oncology, 3rd Edition*, pages 415–470. Chapman and Hall/CRC, 01 2012. ISBN 978-1-4398-6200-1.

Berthold Schweizer and Abe Sklar. *Probabilistic metric spaces*. Courier Corporation, 2011.

Jobanna M. Seddon, Robyn Reynolds, Yi Yu, and Bernard Rosner. Three new genetic loci are independently related to progression to advanced macular degeneration. *PLoS ONE*, 9(1):1–11, 2014.

Xiaotong Shen. Propotional odds regression and sieve maximum likelihood estimation. *Biometrika*, 85(1):165–177, 1998.

Xiaotong Shen and Wing Hung Wong. Convergence rate of sieve estimates. *The Annals of Statistics*, 22:580–615, 1994.

Joanna H Shih. A goodness-of-fit test for association in a bivariate survival model. *Biometrika*, 85(1):189–200, 1998.

Joanna H. Shih and Thomas A. Louis. Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 51(4):1384–1399, 1995.

Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13, 2011. URL `http://www.jstatsoft.org/v39/i05/`.

Abe Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publications de L'Institut de Statistique de L'Université de Paris*, 8:229–231, 1959.

Xue-Kun Song and Peter X-K Song. *Correlated data analysis: modeling, analytics, and applications.* Springer Science & Business Media, 2007.

Jianguo Sun. *The statistical analysis of interval-censored failure time data.* Springer Science & Business Media, 2007.

Jianguo Sun and Liuquan Sun. Semiparametric linear transformation models for current status data. *Canadian Journal of Statistics*, 33(1):85–96, 2005.

Liuquan Sun, Lianming Wang, and Jianguo Sun. Estimation of the association for bivariate interval-censored failure time data. *Scandinavian Journal of Statistics*, 33(4):637–649, 2006.

Tao Sun and Ying Ding. Copula-based semiparametric regression method for bivariate data under general interval censoring. *Biostatistics*, 2019. doi: 10.1093/biostatistics/kxz032.

Tao Sun and Ying Ding. *CopulaCenR: Copula-Based Regression Models for Bivariate Censored Data*, 2020a. URL `https://CRAN.R-project.org/package=CopulaCenR`. R package version 1.1.2.

Tao Sun and Ying Ding. CopulaCenR: Copula-based regression models for bivariate censored data in R. *The R Journal*, 2020b.

Tao Sun, Yi Liu, Richard J Cook, Wei Chen, and Ying Ding. Copula-based score test for bivariate time-to-event data, with application to a genetic study of amd progression. *Lifetime Data Analysis*, 25(3):546–568, 2019.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Xingwei Tong, Man-Hua Chen, and Jianguo Sun. Regression analysis of multivariate interval-censored failure time data with application to tumorigenicity experiments. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(3):364–374, 2008.

Bruce W. Turnbull. The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society Series B*, 38(3):290–295, 1976.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

Jacques Vanobbergen, Luc Martens, Emmanuel Lesaffre, and Dominique Declerck. The signal-tandmobiel project a longitudinal intervention health promotion study in flanders (belgium): baseline and first year results. *European Journal of Paediatric Dentistry*, 2: 87–96, 2000.

Thibault Vatter and Valérie Chavez-Demoulin. Generalized additive models for conditional dependence structures. *Journal of Multivariate Analysis*, 141:147–167, 2015.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, and Pierre-Antoine Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(Dec):3371–3408, 2010.

Antai Wang. Goodness-of-fit tests for archimedean copula models. *Statistica Sinica*, pages 441–453, 2010.

Lianming Wang, Jianguo Sun, and Xingwei Tong. Efficient estimation for the proportional hazards model with bivariate current status data. *Lifetime Data Analysis*, 14:134–153, 2008.

Lianming Wang, Jianguo Sun, and Xingwei Tong. Regression analysis of case ii interval-censored failure time data with the additive hazards model. *Statistica Sinica*, 20(4):1709, 2010.

Naichen Wang, Lianming Wang, and Christopher S McMahan. Regression analysis of bivariate current status data under the gamma-frailty proportional hazards model using the em algorithm. *Computational Statistics & Data Analysis*, 83:140–150, 2015.

Weijing Wang and A Adam Ding. On assessing the association for bivariate current status data. *Biometrika*, 87(4):879–893, 2000.

Weijing Wang and Martin T Wells. Model selection and semiparametric inference for bivariate failure-time data. *Journal of the American Statistical Association*, 95(449):62–72, 2000.

L. J. Wei, Danyu Lin, and Lisa Weissfeld. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84(408):1065–1073, 1989.

Lee-Jen Wei. The accelerated failure time model: a useful alternative to the cox regression model in survival analysis. *Statistics in medicine*, 11(14-15):1871–1879, 1992.

Jon A. Wellner and Ying Zhang. Two likelihood-based semiparametric estimation methods for panel count data with covariates. *The Annals of Statistics*, 35:2106–2142, 2007.

Chi-Chung Wen and Yi-Hau Chen. A frailty model approach for regression analysis of bivariate interval-censored survival data. *Statistica Sinica*, 23:383–408, 2013.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, pages 1–25, 1982.

Andreas Wienke. *Frailty models in survival analysis*. Chapman and Hall/CRC, 2010.

Jun Yan. Enjoy the joy of copulas: With a package copula. *Journal of Statistical Software*, 21(4):1–21, 2007. doi: 10.18637/jss.v021.i04.

Qi Yan, Ying Ding, Yi Liu, Tao Sun, Lars G Fritsche, Traci Clemons, Rinki Ratnapriya, Michael L Klein, Richard J Cook, Yu Liu, Ruzong Fan, Lai Wei, Gonalo R Abecasis, Anand Swaroop, Emily Y Chew, AREDS2 Research Group, Daniel E Weeks, and Wei Chen. Genome-wide analysis of disease progression in Age-related Macular Degeneration. *Human Molecular Genetics*, 27(5):929–940, 2018.

Yildiz E Yilmaz and Jerald F Lawless. Likelihood ratio procedures and tests of fit in parametric and semiparametric copula models with censored data. *Lifetime data analysis*, 17 (3):386–408, 2011.

Safoora Yousefi, Fatemeh Amrollahi, Mohamed Amgad, Chengliang Dong, Joshua E Lewis, Congzheng Song, David A Gutman, Sameer H Halani, Jose Enrique Velazquez Vega, Daniel J Brat, et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Scientific reports*, 7(1):11707, 2017.

Donglin Zeng, Jianwen Cai, and Yu Shen. Semiparametric additive risks model for interval-censored data. *Statistica Sinica*, 16(1):287–302, 2006.

Donglin Zeng, Lu Mao, and Danyu Lin. Maximum likelihood estimation for semiparametric transformation models with interval-censored data. *Biometrika*, 103(2):253–271, 2016.

Donglin Zeng, Fei Gao, and Danyu Lin. Maximum likelihood estimation for semiparametric regression models with multivariate interval-censored data. *Biometrika*, 104(3):505–525, 2017.

Shulin Zhang, Ostap Okhrin, Qian M Zhou, and Peter X-K Song. Goodness-of-fit test for specification of semiparametric copula dependence models. *Journal of econometrics*, 193 (1):215–233, 2016.

Zhigang Zhang and Yichuan Zhao. Empirical likelihood for linear transformation models with interval-censored failure time data. *Journal of Multivariate Analysis*, 116:398–409, 2013.

Zhigang Zhang, Liuquan Sun, Xingqiu Zhao, and Jianguo Sun. Regression analysis of interval-censored failure time data with linear transformation models. *Canadian Journal of Statistics*, 33(1):61–70, 2005.

Qian M Zhou, Peter X-K Song, and Mary E Thompson. Information ratio test for model misspecification in quasi-likelihood inference. *Journal of the American Statistical Association*, 107(497):205–213, 2012.

Qingning Zhou, Tao Hu, and Jianguo Sun. A sieve semiparametric maximum likelihood approach for regression analysis of bivariate interval-censored failure time data. *Journal of the American Statistical Association*, 112(518):664–672, 2017.

Liang Zhu, Xingwei Tong, and Jianguo Sun. A transformation approach for the analysis of interval-censored failure time data. *Lifetime data analysis*, 14(2):167–178, 2008.