Copyright Warning & Restrictions

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material.

Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a, user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use" that user may be liable for copyright infringement,

This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve violation of copyright law.

Please Note: The author retains the copyright while the New Jersey Institute of Technology reserves the right to distribute this thesis or dissertation

Printing note: If you do not wish to print this page, then select "Pages from: first page # to: last page #" on the print dialog screen

ABSTRACT

SURVIVAL ANALYSIS USING ARCHIMEDEAN COPULAS

by
Xieyang Jia

This dissertation has three independent parts. The first part studies a variation of the competing risks problem, known as the semi-competing risks problem, in which a terminal event censors a non-terminal event, but not vice versa, in the presence of a censoring event which is independent of these two events. The joint distribution of the two dependent events is formulated under Archimedean copula. An estimator for the association parameter of the copula is proposed, which is shown to be consistent. Simulation shows that the method works well with most common Archimedean copula models.

The second part studies the properties of a special class of frailty models when the frailty is common to several failure times. The model is closely linked to Archimedean copula models. A useful formula for baseline hazard functions for this class of frailty models is established. A new estimator for baseline hazard functions in bivariate frailty models based on dependent censored data with covariates is obtained, and a model checking procedure is presented.

The third part studies the properties of frailty models for bivariate data under fixed left censoring. It turns out that the distribution of observable pairs belongs to a new class of bivariate frailty models. Both the original model for complete data and the new model for observable pairs are members of Archimedean copula family. A new estimation strategy to analyze left-censored data using the corresponding Kendalls distribution is established.

SURVIVAL ANALYSIS USING ARCHIMEDEAN COPULAS

by
Xieyang Jia

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology and
Rutgers, The State University of New Jersey – Newark
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Mathematical Sciences

Department of Mathematical Sciences, NJIT
Department of Mathematics and Computer Science, Rutgers-Newark

May 2018

APPROVAL PAGE

SURVIVAL ANALYSIS USING ARCHIMEDEAN COPULAS

Xieyang Jia

Dr. Antai Wang, Dissertation Advisor — Date
Associate Professor of Mathematical Sciences, NJIT

Dr. Sunil Dhar, Committee Member — Date
Professor of Mathematical Sciences, NJIT

Dr. Ji Meng Loh, Committee Member — Date
Associate Professor of Mathematical Sciences, NJIT

Dr. Sundarraman Subramanian, Committee Member — Date
Associate Professor of Mathematical Sciences, NJIT

Dr. Zhi Wei, Committee Member — Date
Associate Professor of Computer Science, NJIT

BIOGRAPHICAL SKETCH

Author: Xieyang Jia

Degree: Doctor of Philosophy

Date: May 2018

Undergraduate and Graduate Education:

- Doctor of Philosophy in Mathematical Sciences,
  New Jersey Institute of Technology, Newark, NJ, 2018
- Master of Science in Biostatistics,
  New Jersey Institute of Technology, Newark, NJ, 2018
- Bachelor of Science in Public Health Management,
  Fudan University, Shanghai, China, 2009

Major: Mathematical Sciences

Presentations and Publications:

A. Wang and X. Jia, "The analysis of left truncated bivariate data using frailty models", *Scandinavian Journal of Statistics*, 2018. (accepted)

J. Zhong, X. Jia, et al, "Analysis of the health status of infant swimming in Shanghai", *Chinese Journal of Public Health Management*, vol. 3, pp. 361-363, 2012.

J. Zhong, X. Jia, et al, "Analysis of the health status of infant swimming in Shanghai", *Shanghai Journal of Preventive Medicine*, vol. 5, pp. 249-251, 2012.

J. Wang, X. Jia, et al, "Analysis of the status of informed consent in medical research involving human subjects in public hospitals in Shanghai", *Journal of Medical Ethics*, vol. 36, pp. 415-417, 2010.

J. Wang, X. Jia, et al, "Evaluation on the informed consents of medical researches involving human subjects in public hospitals", *Chinese Health Resources*, vol. 13, pp. 116-118, 2010.

X. Jia, J. Wang, et al, "Evaluation on the effectiveness of informed consent in medical researches involving human subjects in public hospitals in Shanghai", *Chinese Health Resources*, vol. 13, pp. 74-75, 2010.

To my grandparents Hongxian and Caizhen:
When I was 8 years old, I took $5 on the table and bought some toy cars. Later on, I overheard you two talking about the lost money. Grandma said: 'Xiaocelao is an honest kid. He would never took the money.' I felt so guilty but never had a chance to confess.
However, I knew I was trusted and loved ever since, and I was determined to earn what I deserve by hard work, instead of taking shortcuts or cheating.

Yuanyuan

ACKNOWLEDGMENT

All praise to God.

Firstly, I would like to offer my special thanks to my dissertation advisor Professor Antai Wang, who guided me and motivated me throughout my five years of PhD study. He is always available when I needed help, willing to give his insightful opinions on my research, as well as sharing his understanding on life and family. His encouragement is the beacon that leaded me out of the dark whenever I'm down.

Secondly, I'm particularly grateful for the assistance and comments given by my committee members, Professor Sunil Dhar, Professor Ji Meng Loh, Professor Sundarraman Subramanian and Professor Zhi Wei.

Thirdly, I would like to express my very great appreciation to NJIT Department of Mathematical Sciences for their support. The positive, cheerful and friendly research environment makes life much easier. The professors are always at hand for help. They deliver enlightening lectures that transform laymen into experts in statistics. The students and colleagues are like family members that share tears and joys.

Finally, I wish to thank my family for their support and encouragement over the years. My parents, Fen and Jingfang put all their efforts on my education, and they always encourage me to eagerly explore the unknown. They want me to succeed more than anyone, but even if I fail, they back me up. My high school Chinese teacher, Fei insisted that she should be mentioned here just for fun. My daughter, Aubrey is truly an angel. She is the melody that comforts me and the string that touches my heart. Her smile is an elixir that expels the darkness and makes life worth fighting for. My always-22-year-old beloved wife, Wenwen quit her job to take care of the family, so that I can focus on my research. I am amazed every time I think about how much she has changed for me. Lao po xin ku la, zhuan qian gei ni mai bao bao.

TABLE OF CONTENTS

Chapter **Page**

1 INTRODUCTION 1

1.1 Survival Analysis Basics 1

1.2 Frailty Models 5

1.3 Archimedean Copula 8

2 A SEMI-COMPETING RISKS PROBLEM 18

2.1 Introduction 18

2.2 Parameter Estimation 20

2.3 Consistency of $\hat{\theta}$ 22

2.4 Marginal Distribution Estimation 25

2.5 Simulation Results 25

2.6 Leukemia Data Example 27

2.7 Conclusion 28

3 A NEW ESTIMATOR OF BASELINE HAZARD FUNCTION 29

3.1 Frailty Model for Clustered Data 29

3.2 A New Estimator of Baseline Hazard Function 36

3.3 A Model Checking Procedure for Frailty Models 41

3.4 Simulation Studies 42

3.5 Discussion 44

4 LEFT CENSORED BIVARIATE DATA ANALYSIS 48

4.1 Properties of Frailty Models for Left Censored Bivariate Data 48

4.2 Parameter Estimation 52

4.3 Simulation Results 53

4.4 Discussion 54

BIBLIOGRAPHY 56

LIST OF TABLES

Table **Page**

2.1 Simulation Results for Clayton Copula 26

2.2 Simulation Results for Gumbel Copula 27

4.1 Estimator Comparison 54

LIST OF FIGURES

Figure		Page
1.1	An example of right-censored data	3
1.2	An example of Kaplan-Meier Estimator	4
1.3	An example of Nelson-Aalen Estimator	4
1.4	Clayton copula with $\tau = 0.2$	10
1.5	Clayton copula with $\tau = 0.6$	11
1.6	Clayton copula with $\tau = 0.9$	11
1.7	Gumbel copula with $\tau = 0$	12
1.8	Gumbel copula with $\tau = 0.5$	13
1.9	Gumbel copula with $\tau = 0.9$	13
1.10	Frank copula with $\tau = -0.85$	14
1.11	Frank copula with $\tau = -0.5$	15
1.12	Frank copula with $\tau = -0.1$	15
1.13	Frank copula with $\tau = 0.1$	16
1.14	Frank copula with $\tau = 0.5$	16
1.15	Frank copula with $\tau = 0.85$	17
2.1	An example of semi-competing risks data	19
2.2	An example of $\hat{S}_X$ vs. S_X	27
2.3	$\hat{S}_X$ of Leukemia data	28
3.1	Model checking procedure	43
3.2	Model checking procedure (wrong model)	45
3.3	Comparison of two estimators	46

CHAPTER 1

INTRODUCTION

Copula models are gaining popularity when modeling dependent random variables. Among them, Archimedean copula is most widely being used. The merit of copula models is that it declares a clear form of the joint survival function with respect to the marginal survival functions. Moreover, the association is captured in a single-parameter generator function that is straightforward to interpret. Oakes(1989)[14] has shown that Archimedean copulas naturally arise from bivariate frailty models, which characterizes the associations among the observable survival data and unobservable latent random variables.

In this chapter, we will discuss some important factors about frailty models and Archimedean copula. This chapter starts from the basic ideas of survival analysis in Section 1.1. Then frailty models and Archimedean copulas are introduced in Section 1.2 and Section 1.3.

1.1 Survival Analysis Basics

Survival analysis studies the expected duration of time until one or more events happen, for example, the time to death of a patient, or time to failure of a machine. In general, let T be the time to event, and we assume T to be an absolute continuous random variable taking on non-negative values. Therefore, T has probability density function $f(t)$ such that

$$P(t_1 \leq T \leq t_2) = \int_{t_1}^{t_2} f(t)\, dt, \quad 0 \leq t_1 \leq t_2,$$

and has cumulative distribution function $F(t)$ defined as

$$F(t) = P(T \leq t) = \int_0^t f(u)\, du, \quad t \geq 0.$$

By the continuity of T, $f(t) = \dfrac{dF(t)}{dt}$. The survival function $S(t)$ is defined as

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(u)\, du, \quad t \geq 0,$$

which measures the probability that the event does not happen by time t. In our examples above, it is the probability that the patient survives beyond time t or the machine does not fail until time t.

The hazard function of T at time t is denoted as $\lambda(t)$ where

$$\lambda(t) = \lim_{h \to 0} \frac{P(t \leq T < t + h | T \geq t)}{h} = \lim_{h \to 0} \frac{P(t \leq T < t + h)}{h\, P(T \geq t)}$$

$$= \frac{1}{P(T \geq t)} \left[\lim_{h \to 0} \frac{P(t \leq T < t + h)}{h} \right] = \frac{f(t)}{S(t)}.$$

The hazard function shows the instantaneous failure rate at time t given that the event has not happened yet at that moment. Note that

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

The cumulative hazard function $\Lambda(t)$ is defined as

$$\Lambda(t) = \int_0^t \lambda(u) du = -\log S(t),$$

or equivalently,

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

The most important and interesting feature of survival analysis is censoring and truncation of data. For example, in many clinical trials, the true time to event is not always observable for each individual because of various reasons, such as lost to follow-up of the participating patients, end of study, competing risks, etc. A more detailed introduction of censoring and truncation can be found on Klein(2003)[9]

Chapter 3. In our proposal, we focus on right censored (see Figure 1.1) and left truncated data. Note that the main difference between censoring and truncation is that censored object is detectable but the value is not known, while the object is not even detectable in the case of truncation due to instrumental limitations.

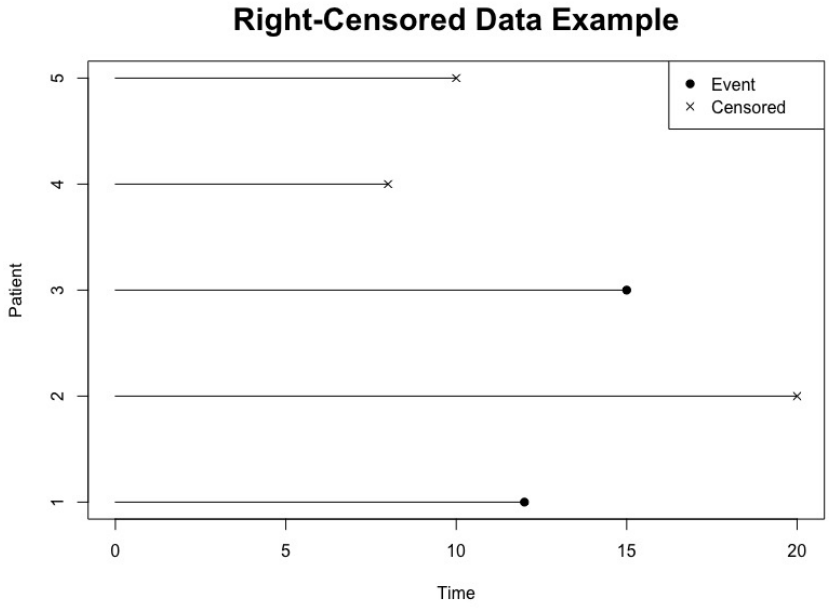

Figure 1.1 An example of right-censored data.

Non-parametric approaches are widely used to estimate the survival function and hazard function of T, such as Kaplan-Meier estimator (see Figure 1.2) and Nelson-Aalen estimator (see Figure 1.3). A detailed explanation of these estimator was addressed on Klein(2003)[9] Chapter 4. These estimators are straightforward in visualization and are easy to apply, but the restrictions are also clear that these non-parametric estimators don't account for covariate effects. Moreover, they are based on an assumption of independent censoring. In other words, the knowledge of a censoring time for an individual provides no further information about this person's likelihood of survival at a future time had the individual continued on the study.

If we also want to include covariates to establish regression models, Cox(1972)[2] introduced proportional hazards model. In this model, the hazard function $\lambda(t)$ is

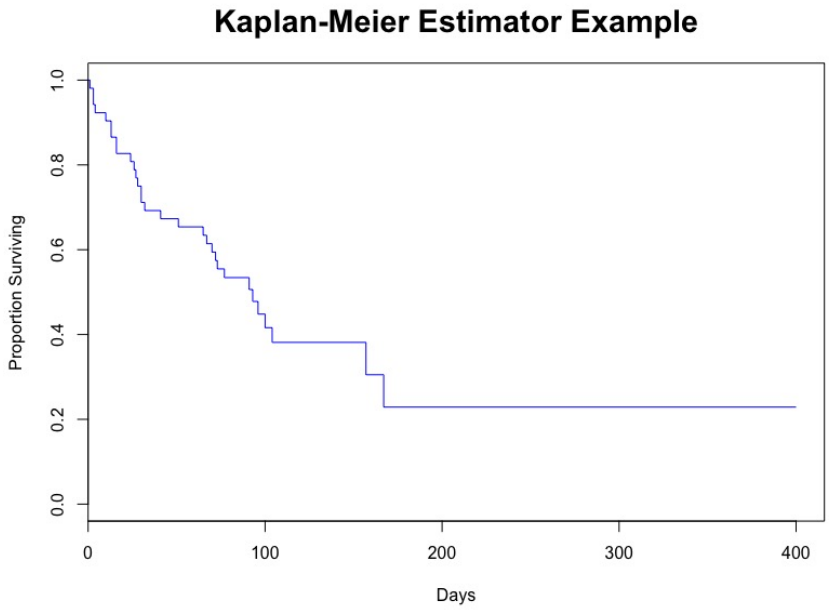

Figure 1.2 An example of Kaplan-Meier Estimator.

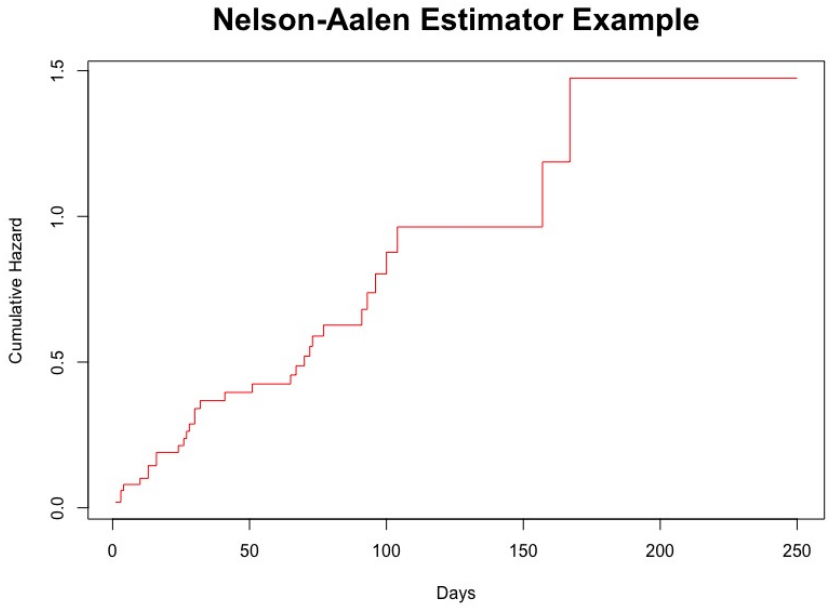

Figure 1.3 An example of Nelson-Aalen Estimator.

defined as

$$\lambda(t, X) = \lambda_0(t) \exp(\beta' X),$$

where X is the observable covariates of interest with associated coefficient β, and $\lambda_0(t)$ is called baseline hazard function, which can be interpreted as the hazard function when all covariates equal to 0. In the proportional hazards model, partial likelihood is used to estimate the unknown parameter β. The partial likelihood is constructed on the conditional probability that a particular subject would fail at t_i given the risk set R_i and the fact that exactly one subject fails at that time, i.e.,

$$PL = \prod_{i=1}^{n} \frac{\lambda(t_i, X_i)}{\sum_{j \in R_i} \lambda(t_i, X_j)}$$

$$= \prod_{i=1}^{n} \frac{\lambda_0(t_i) \exp(\beta' X_i)}{\sum_{j \in R_i} \lambda_0(t_i) \exp(\beta' X_j)} = \prod_{i=1}^{n} \frac{\exp(\beta' X_i)}{\sum_{j \in R_i} \exp(\beta' X_j)}.$$

Note that the risk set R_i is defined as the set of subjects that are alive just before t_i.

The beauty of this approach is that we don't have to specify the baseline hazard function, and β corresponds to the increase in the log-hazard. However, the foundation of this model is the independent censoring assumption and the proportional hazard assumption, which can lead to problems if not taken care of. More details could be found on Klein(2003)[9] Chapter 8.

1.2 Frailty Models

When modeling continuous survival data, we are inclined to assume independent censoring, because most canonical approaches such as Cox proportional hazard models or even Kaplan-Meier estimator relies heavily on this critical assumption. However, the analysis of the association between the survival time T and the censoring time C is often an overlooked topic. For example, in an oncology drug study, the progression-free time may have a positive correlation with lost to follow-up, because as time goes

by, more patients are intended to switch to other treatments if the testing drug does not make a big difference. In this case, we're not only interested in the survival time of the patients, but also in the dependence structure, so that we may alter our design due to the accumulating lost to follow-up patients.

The introduction of frailty by Oakes(1989)[14] provided one way to account for such random effects and dependence on the survival model. Generally, frailty W is the common unobserved random effect that modifies multiplicatively the hazard function of T and C. Moreover, when W is given, T and C are conditionally independent, which implies that the common dependence of T and C can be fully explained by frailty W.

In Oakes's frailty model, the conditional marginal survival functions of T and C given W are denoted as

$$Pr(T > t|W = w) = [S_{T_0}(t)]^w$$

and

$$Pr(C > c|W = w) = [S_{C_0}(t)]^w,$$

where $S_{T_0}(t)$ and $S_{C_0}(c)$ are the baseline survival functions of T and C, respectively. Although this set up looks similar to the Cox proportional hazards model, the Cox proportional hazards model won't work with unobservable frailty W.

The unconditional survival function is then

$$S_T(t) = E[Pr(T > t|W)] = E[\{S_{T_0}(t)\}^W].$$

Let the Laplace transform of W be $\psi(s) = E[e^{-sW}]$, we have

$$S_T(t) = E[e^{\log\{[S_{T_0}(t)]^W\}}] = E[e^{W \log S_{T_0}(t)}] = \psi\{-\log S_{T_0}(t)\},$$

and the similar approach shows that $S_C(c) = \psi\{-\log S_{C_0}(c)\}$. If we denote $\psi^{-1}(s)$ the inverse function of $\psi(s)$, we have

$$\psi^{-1}\{S_T(t)\} = -\log S_{T_0}(t)$$

and

$$\psi^{-1}\{S_C(c)\} = -\log S_{C_0}(c).$$

By the assumption that T and C are independent given W, the bivariate survivor function is

$$S(t,c) = E[S(t,c|W)] = E[S(t|W)S(c|W)] = E[S_{T_0}(t)^W S_{C_0}(c)^W]$$

$$= E[e^{\{W[\log S_{T_0}(t) + \log S_{C_0}(c)]\}}] = \psi[-\log S_{T_0}(t) - \log S_{C_0}(c)]$$

$$= \psi\{\psi^{-1}[S_T(t)] + \psi^{-1}[S_C(c)]\}.$$

Therefore, using frailty W to model T and C, their dependence structure naturally follows a bivariate Archimedean copula with copula generator $\psi(s)$, which will be introduced in Section 1.3.

Moreover, Wang(2014)[22] showed that the marginal survival function of T and C is given by

$$S_T(t) = \psi_\theta \left\{ \int_0^t \psi_\theta^{-1\prime}[\pi(u)]\pi(u) d\ln[S_T^\star(u)] \right\}$$

and

$$S_C(c) = \psi_\theta \left\{ \int_0^c \psi_\theta^{-1\prime}[\pi(u)]\pi(u) d\ln[S_C^\star(u)] \right\}.$$

In the formula above, $\pi(u) = P(T > u, C > u)$ for all $u > 0$.

Because of the non-identifiability property of copulas under given dependent censored data (X, δ) as shown in Wang(2012)[21], $S_T^\star$ and $S_C^\star$, defined as the marginal

survival functions of T and C, respectively under the additional assumption that the two variables T and C are independent, can be achieved using independence copula where $\psi(s) = e^{-s}$, and they can be estimated by Kaplan-Meier estimator. These two formula above will be used to derive our new estimator for the baseline hazard function.

1.3 Archimedean Copula

A copula is a multivariate probability distribution where the marginal probability distribution of each variable is uniform. Copula models are popular in survival analysis because of the Sklar's theorem, which claims that we can describe any joint distribution of random variables by the marginal distributions and a copula. Moreover, the copula is unique if the marginal is continuous. In other words, when describing the joint distribution of two correlated random variables, copula separates the marginal distribution from the dependence structure, which is an improved feature comparing with using joint distribution alone.

There are many copula models, and among them Archimedean copula is a special class which is most popular because of its simple settings. Under bivariate setting, denote $C_\theta(U_1, U_2)$ as the copula between two random variables U_1 and U_2 with parameter θ, C_θ is called Archimedean if

$$C_\theta(u_1, u_2) = \psi_\theta[\psi_\theta^{-1}(u_1) + \psi_\theta^{-1}(u_2)],$$

where $\psi^{-1} : [0,1] \times \Theta \to [0, \infty)$ is a continuous, strictly decreasing and convex function such that $\psi_\theta^{-1}(1) = 0$.

In the frailty model, if we choose $U_1 = S_T$ and $U_2 = S_C$, it is clear that the marginal probability distribution of U_1 and U_2 are both uniform. By the formula in the last section,

$$S(t, c) = \psi_\theta\{\psi_\theta^{-1}[S_T(t)] + \psi_\theta^{-1}[S_C(c)]\} = C_\theta[S_T(t), S_C(c)],$$

which explains why frailty models naturally arises from Archimedean Copulas.

To characterize the global association between variables in Archimedean copula, Kendall(1938)[8] introduced τ as a non-parametric rank invariant measure:

$$\tau = 1 + 4 \int_0^1 \frac{\psi_\theta^{-1}(u)}{\psi_\theta^{-1\prime}(u)}\, du,$$

which evaluates the probability of concordance minus the probability of discordance. The association of the random variables is stronger as τ deviates from 0. When τ approaches 1 indicates a positive correlation and -1 a negative correlation.

There are many copula generators ψ we can choose from, and different generators imply different underlying distributions of the frailty because it is the Laplace transform of it. Some examples are given below:

Example 1: Clayton(1978)[1] first introduced the model that when the frailty W follows Gamma distribution with index $(1/\theta, 1)$, the Laplace transform of W is $\psi_\theta(s) = (1+s)^{-\frac{1}{\theta}}$, hence $\psi_\theta^{-1}(s) = s^{-\theta} - 1$. Therefore, the bivariate survival function $S(t, c)$ is

$$S(t,c) = \psi_\theta\{\psi_\theta^{-1}[S_T(t)] + \psi_\theta^{-1}[S_C(c)]\}$$

$$= [S_T(t)^{-\theta} - 1 + S_C(c)^{-\theta} - 1 + 1]^{-\frac{1}{\theta}}$$

$$= [S_T(t)^{-\theta} + S_C(c)^{-\theta} - 1]^{-\frac{1}{\theta}}.$$

Kendall's τ is

$$\tau = 1 + 4 \int_0^1 \frac{\psi_\theta^{-1}(u)}{\psi_\theta^{-1\prime}(u)}\, du = 1 + 4 \int_0^1 \frac{u^{-\theta} - 1}{-\theta u^{-\theta-1}}\, du$$

$$= 1 - \frac{4}{\theta} \int_0^1 (u - u^{\theta+1})\, du = \frac{\theta}{\theta + 2}$$

Figure 1.4 - Figure 1.6 shows the distribution of two random variables under Clayton copulas with different τ levels. As we can see, Clayton copula is heavily concentrated near $(0, 0)$. As τ increases from 0 to 1, a positive correlation between the two random variables is observed.

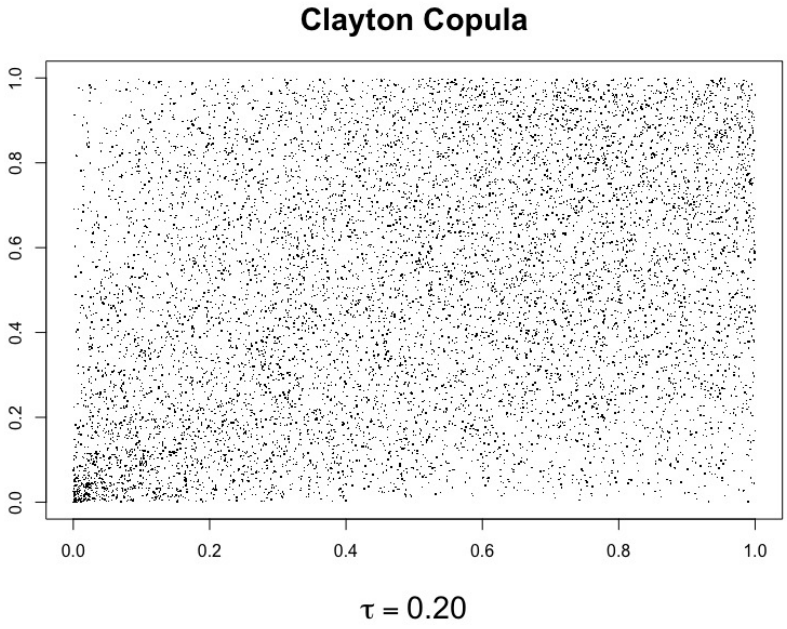

Figure 1.4 Clayton copula with $\tau = 0.2$.

Example 2: Gumbel model assumes that the frailty has a stable distribution. Stable distribution is a family of continuous probability distributions parametrized by location and scale parameters μ and σ, respectively, and two shape parameters θ and β , roughly corresponding to measures of concentration and asymmetry, respectively.

In Gumbel copula, the Laplace transform of the frailty is $\psi_\theta(s) = \exp(-s^{1/\theta})$, with inverse function $\psi_\theta^{-1}(s) = [-\log(s)]^\theta$. The bivariate survival function is

$$S(t, c) = \exp[-\{[-\log S_T(t)]^\theta + [-\log S_C(c)]^\theta\}^{1/\theta}].$$

Under this copula generator, Kendall's τ is

$$\tau = 1 + 4 \int_0^1 \frac{(-\log u)^\theta}{\theta(-\log u)^{\theta-1}(-\frac{1}{u})}\, du$$

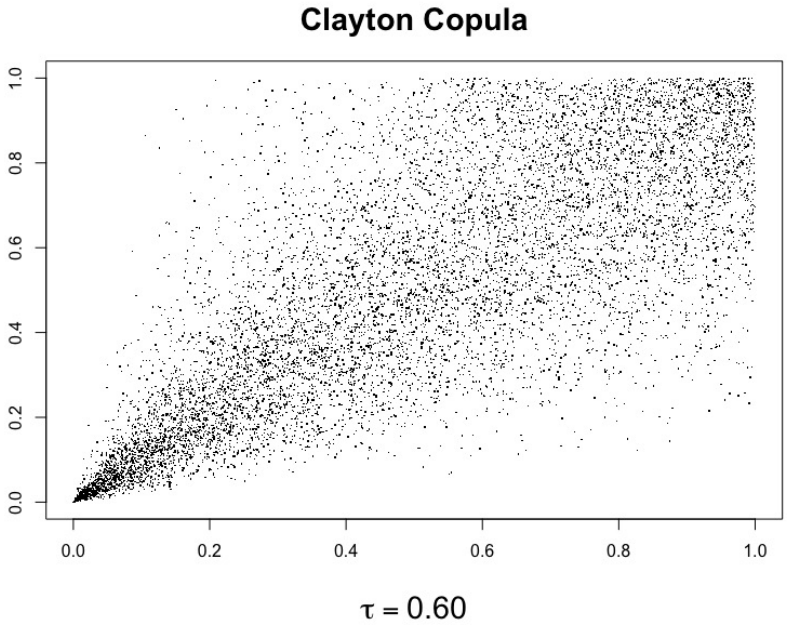

Figure 1.5 Clayton copula with $\tau = 0.6$.

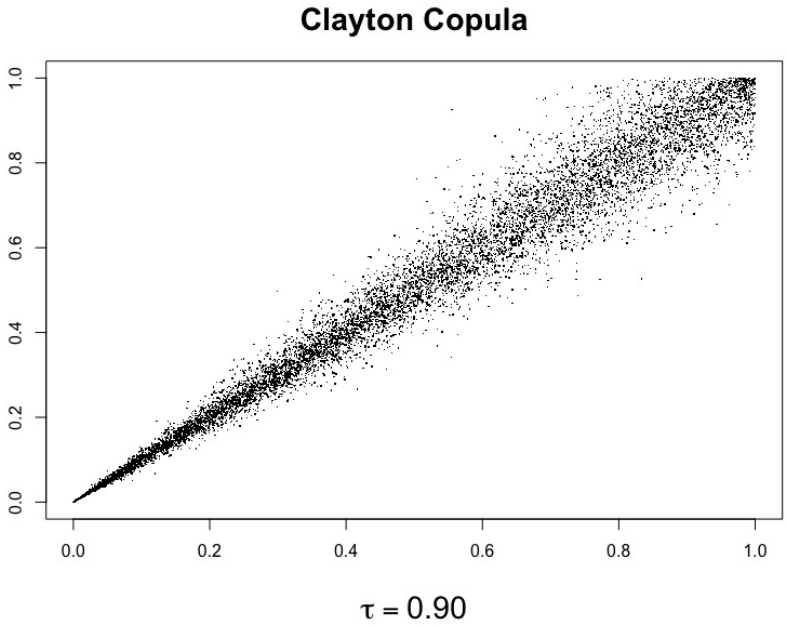

Figure 1.6 Clayton copula with $\tau = 0.9$.

$$= 1 + \frac{4}{\theta} \int_0^1 (u \log u)\, du = \frac{\theta - 1}{\theta}$$

Figures 1.7 - 1.9 shows the distribution of a Gumbel copula. This copula has more probability concentrated in the tails. It is asymmetric, with more weight in the right tail.

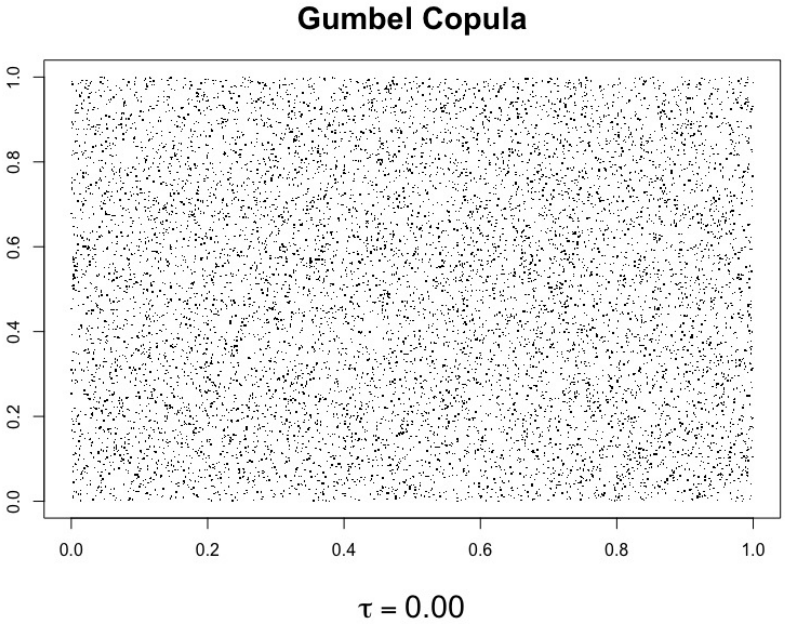

Figure 1.7 Gumbel copula with $\tau = 0$.

Example 3: Genest(1987)[6] introduced another important class of frailty models, the Frank models. In Frank model, the copula generator is chosen to be

$$\psi_\theta(s) = -\frac{\log(1 + e^{-s}(e^{-\theta} - 1)}{\theta},$$

and its inverse function is

$$\psi_\theta^{-1}(s) = \log \frac{e^{-\theta} - 1}{e^{-\theta s} - 1}.$$

The bivariate survival function is hence

$$S(t, c) = \psi_\theta\{\psi_\theta^{-1}[S_T(t)] + \psi_\theta^{-1}[S_C(c)]\}$$

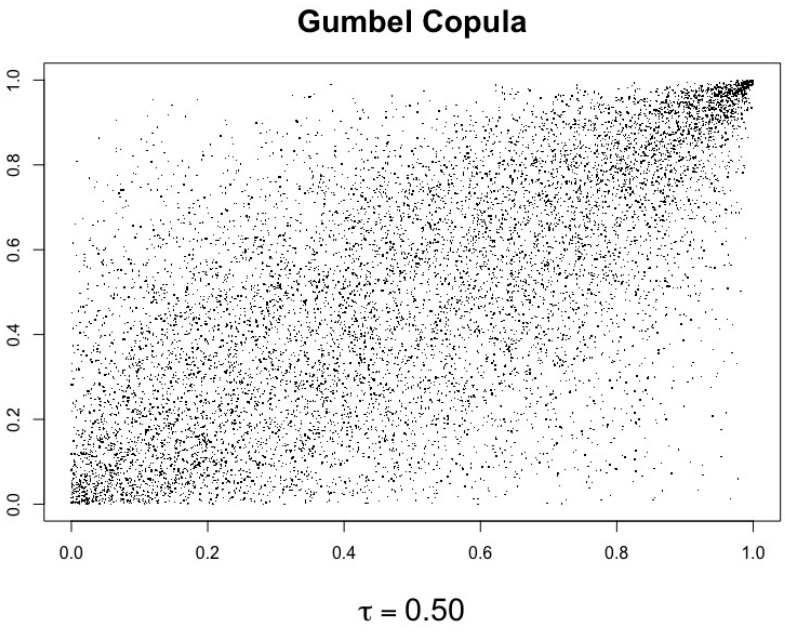

Figure 1.8 Gumbel copula with $\tau = 0.5$.

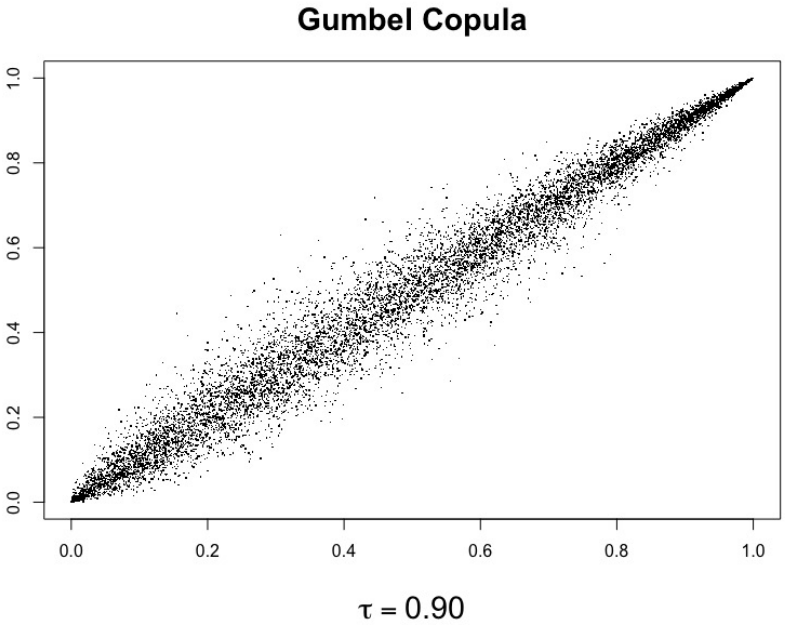

Figure 1.9 Gumbel copula with $\tau = 0.9$.

$$= \psi_\theta \left[\log \frac{e^{-\theta} - 1}{e^{-\theta S_T(t)} - 1} + \log \frac{e^{-\theta} - 1}{e^{-\theta S_C(c)} - 1} \right]$$

$$= \psi_\theta \left\{ \log \frac{(e^{-\theta} - 1)^2}{[e^{-\theta S_T(t)} - 1][e^{-\theta S_C(c)} - 1]} \right\}$$

$$= -\frac{1}{\theta} \log \left\{ 1 + \frac{[e^{-\theta S_T(t)} - 1][e^{-\theta S_C(c)} - 1]}{(e^{-\theta} - 1)^2} (e^{-\theta} - 1) \right\}$$

$$= -\frac{1}{\theta} \log \left\{ 1 + \frac{[e^{-\theta S_T(t)} - 1][e^{-\theta S_C(c)} - 1]}{e^{-\theta} - 1} \right\}.$$

In this model, we use numerical methods to find the value of τ.

Figures 1.10 - 1.15 shows the distribution of a Frank copula. Frank copula is symmetric and has more probability concentrated in the tails like the Gumbel copula. As τ deviates from 0, the association is stronger. Positive τ indicates positive correlation and negative τ suggests negative correlation.

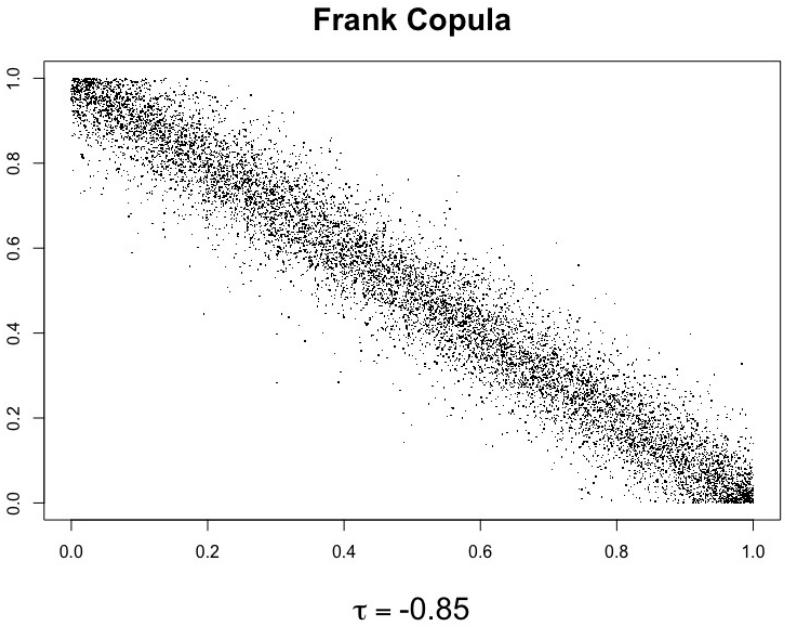

Figure 1.10 Frank copula with $\tau = -0.85$.

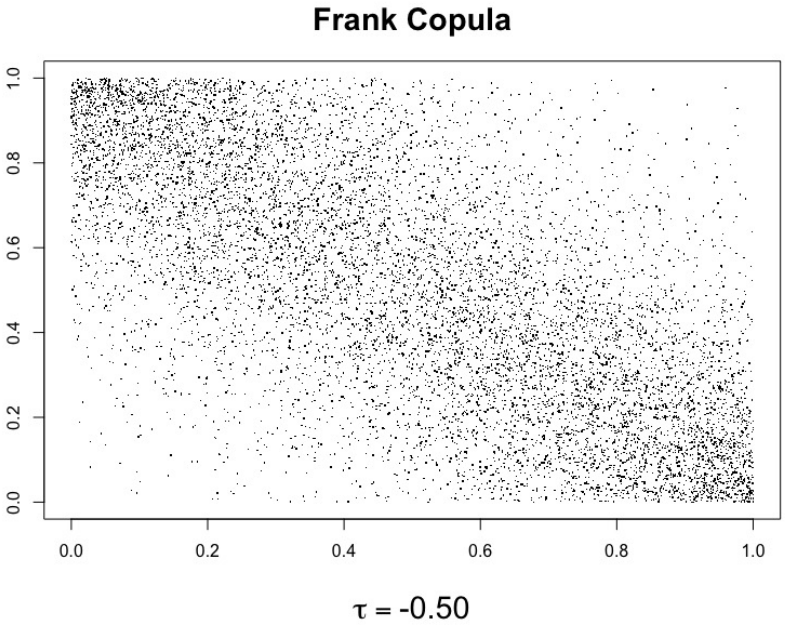

Figure 1.11 Frank copula with $\tau = -0.5$.

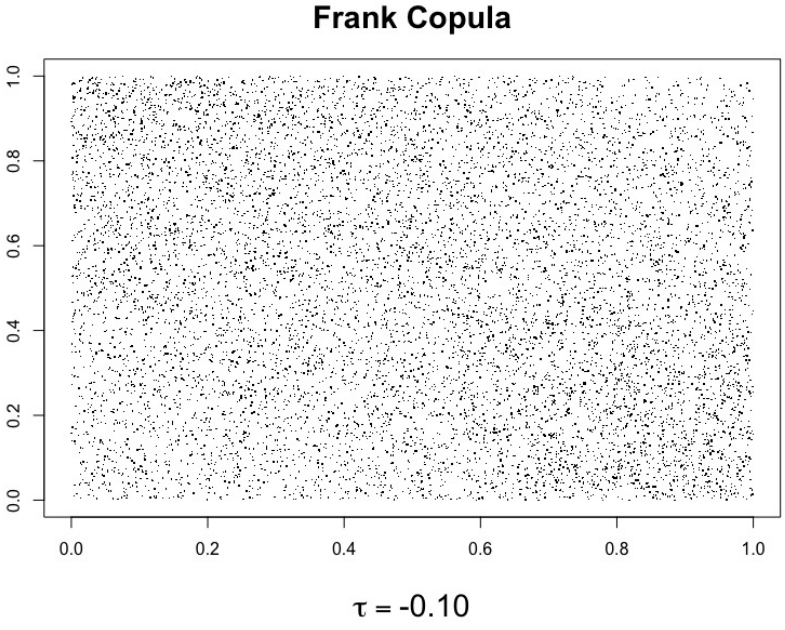

Figure 1.12 Frank copula with $\tau = -0.1$.

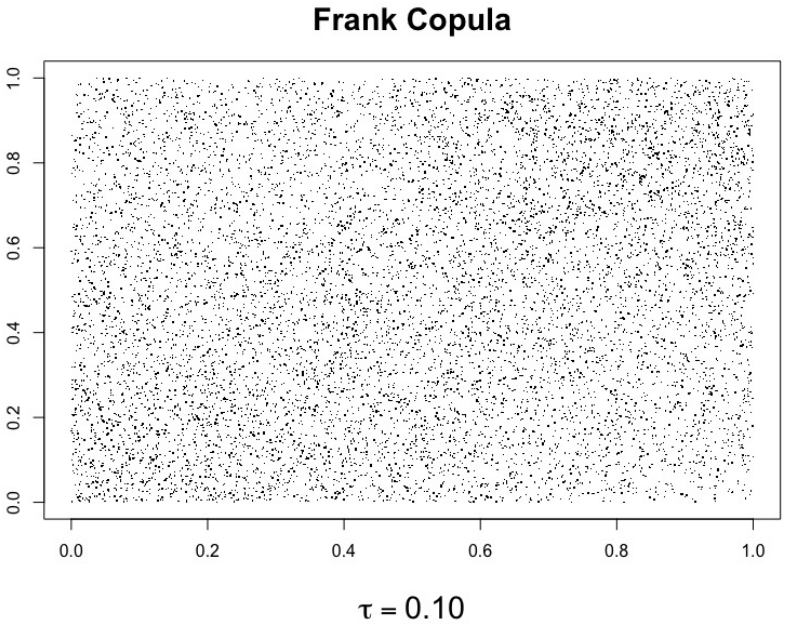

Figure 1.13 Frank copula with $\tau = 0.1$.

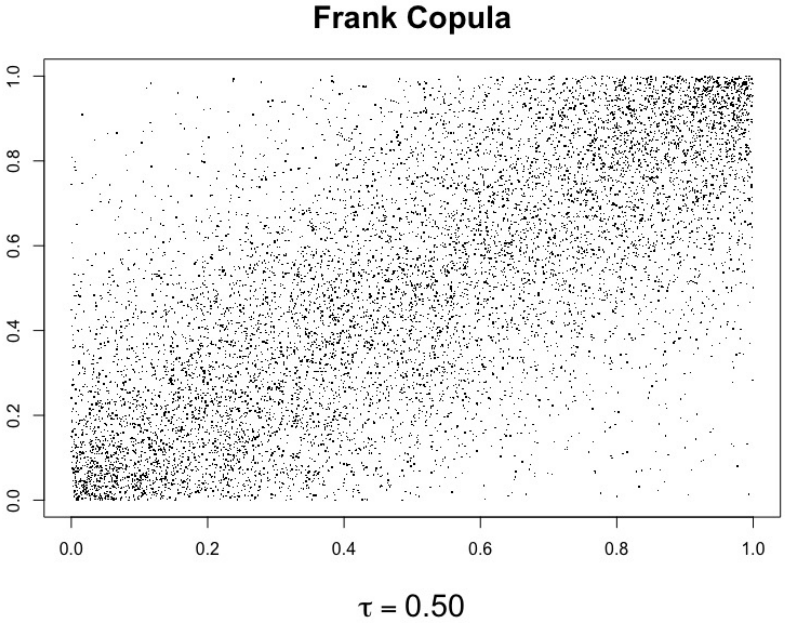

Figure 1.14 Frank copula with $\tau = 0.5$.

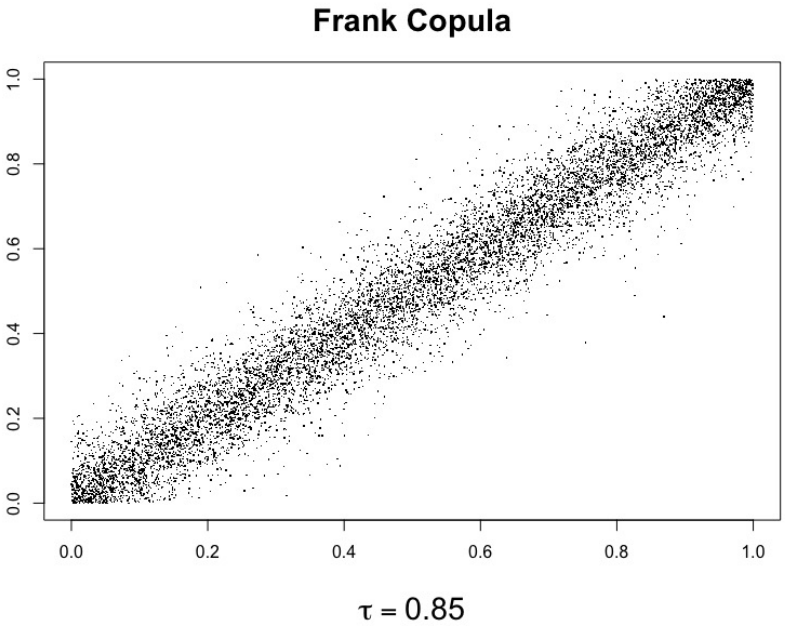

Figure 1.15 Frank copula with $\tau = 0.85$.

CHAPTER 2

A SEMI-COMPETING RISKS PROBLEM

2.1 Introduction

The outcome of clinical trials and medical researches may consist of different kind of events, such as terminal events (i.e. death) and non-terminal events (i.e. relapse, progression of diseases). Traditionally, researchers focus on the behavior of terminal events, such as overall survival probability. But nowadays, due to the more sophisticated nature of diseases, the more complicated progression stages, and the more advanced design techniques, the non-terminal ones carry a lot of practical meanings in the study. Moreover, the underlying dependence structure between these two kinds of events can not be ignored, while in some cases even become important to the decision making.

In contrast to the traditional bivariate competing risks data, in a semi-competing risks data, when a non-terminal event happens, the corresponding terminal event is not censored. An example is that when relapse occurs, death could still be observed, but not vise versa. In fact, death could be caused by either relapse or graft-versus-host diseases (GVHD). In this case, the distribution of relapse and the ability of relapse to predict death may be important.

Moreover, in the presence of a univariate independent censoring (i.e. lost to follow-up) to both events, we face the semi-competing risks problem as introduced in Fine(2001)[4]. In this project, we not only recover the distribution of both terminal and non-terminal events for semi-competing risk data, but also estimate the dependence structure of them.

Let X denote the failure time of the non-terminal event, and Y for the terminal event. As they are very likely to be positively correlated, we impose an Archimedean

copula on their dependency, such that

$$S(x,y) = P(X > x, Y > y) = \psi_\theta^{-1}(\psi_\theta[S_X(x)] + \psi_\theta[S_Y(y)]),$$

where ψ_θ is the copula generator function and S_X and S_Y are marginal survival functions of X and Y respectively. Note that in most common Archimedean copula models, θ is a one-dimensional parameter.

In the presence of a censoring time C which is independent of both X and Y, for each individual we can observe $T_2 = \min\{Y, C\}$, $D_2 = \mathbb{1}\{Y < C\}$, $T_3 = \min\{X, T_2\}$ and $D_3 = \mathbb{1}\{X < T_2\}$, where $\mathbb{1}$ is the indicator function. Therefore, the observed data are n independently identically distributed samples denoted by $\{(T_{2i}, D_{2i}, T_{3i}, D_{3i}),\ i = 1, 2, ..., n\}$. Figure 2.1 visualizes different scenarios of semi-competing risks data structure.

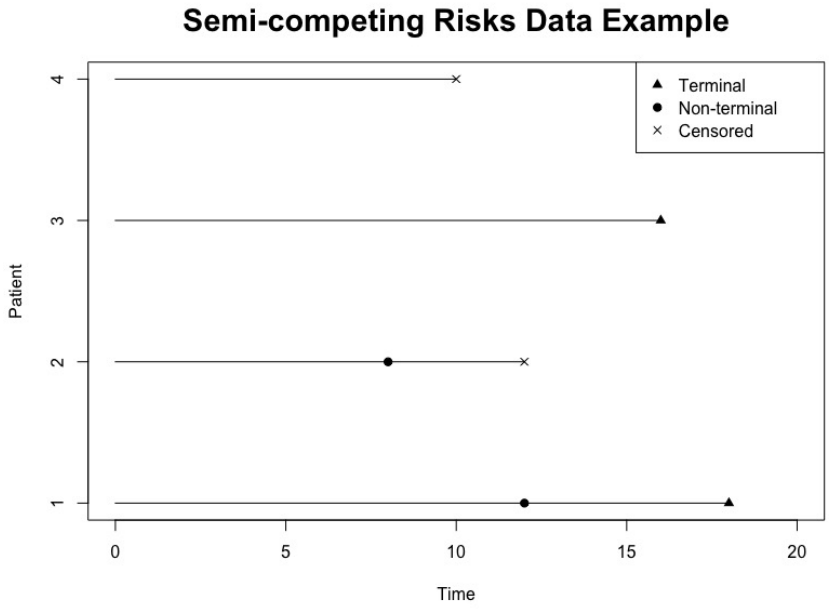

Figure 2.1 An example of semi-competing risks data.

Fine(2001)[4], Lakhal(2008)[10] and other authors have proposed some estimators for marginal distributions and association parameter θ in these situations. However, these approaches have some restrictions that make application infeasible. For

example, Fine(2001)[4] proposed a parameter estimator using the concordance of the data, but only works for Clayton model. Moreover, our assumption of homogeneity of the marginal distribution of X on $X < Y$ and $X > Y$ is plausible. The method we introduce will be straightforward, simple and stable.

This chapter will be organized in the following way. In Section 2.2, we propose our estimator for the copula association parameter. In Section 2.3, we prove the large sample properties of this estimator, and then we recover the marginal distributions in Section 2.4. The following Section 2.5 shows the simulation results, with a real data example in Section 2.6. We end this chapter with some discussions in Section 2.7.

2.2 Parameter Estimation

The copula association parameter θ reveals the relationship between the marginal distribution of the terminal and non-terminal event. To understand the behavior and correlation between these two kinds of events, it is always important to get a solid estimation for it.

Because the existence of X does not censor the occurrence of Y or C, the pair (T_2, D_2) is always observable for each sample, and therefore we can estimate S_Y by the well-established Kaplan-Meier estimator, denoted as $\tilde{S}_Y$. Fleming & Harrington(2005)[5] has shown that it is a uniform consistent estimator of S_Y on $[0, t_0)$ where $t_0 = \max\{T_2\}$. We try to construct another estimator of S_Y, parameterized with θ, denoted $\hat{S}_Y$, so that the association parameter θ can be solved by minimizing the distance between $\tilde{S}_Y$ and $\hat{S}_Y$.

To construct $\hat{S}_Y$, we extend the copula graphical estimator, introduced by Rivest(2001)[16] into semi-competing risks setting. In the original paper, when X and Y follows Archimedean copula and censors each other, denote $Z_i = \min\{X_i, Y_i\}$ and $D_i = \mathbb{1}(Y_i < X_i)$, Rivest suggested that

$$\hat{S}_Y(y) = \psi_\theta^{-1}\left[-\sum_{Z_i \le y, D_i = 1} \psi_\theta[\hat{\pi}(Z_i)] - \psi_\theta[\hat{\pi}(Z_i) - 1/n]\right],$$

where $\hat{\pi}(z) = \sum_{i=1}^{n} \frac{\mathbb{1}\{Z_i \geq z\}}{n}$ is the empirical estimator of $\pi(z) = P(X > z, Y > z)$. The copula graphical estimator is uniformly consistent on $[0, t_0)$ if ψ_θ is correctly specified.

With semi-competing risks data, the presence of C turns Z into a variable that is not always observable, and $\pi(z)$ cannot be estimated using the empirical way. Therefore, the original copula graphical estimator cannot be used directly.

However, $Z = \min\{X, Y\}$ is independently censored by C, thus we can estimate $\pi(z) = P(Z > z)$ by the Kaplan Meier estimator calculated by the observable pair $(T_3, \mathbb{1}\{Z < C\})$, denoted by $\hat{\pi}_2$, and the copula graphical estimator can be modified as

$$\hat{S}_Y(y) = \psi_\theta^{-1}\left[-\sum_{Z_i \leq y, D_i = 1} \psi_\theta[\hat{\pi}_2(Z_i^-)] - \psi_\theta[\hat{\pi}_2(Z_i)]\right].$$

Note that when $C < Z$, both X and Y are censored, hence D is not observable in this case. We have to discard these data when calculating the copula graphical estimator. Fortunately, the independence of C guarantees that this will not affect the consistency of the estimator when the sample size and censoring rate is moderate.

Since $\hat{S}_Y$ and $\tilde{S}_Y$ both consistently estimate S_Y, we can use the minimum discrepancy approach that minimize the Cramér-von Mises distance between these two estimators to find the value of $\hat{\theta}$. In particular,

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{n} \sum_{S} [\hat{S}_Y(Y_i) - \tilde{S}_Y(Y_i)]^2,$$

where the summation is on the set $S = \{Y_i : Y_i = T_{3i}\}$. This is because the step function $\tilde{S}_Y$ jumps on $\{Y_i : Y_i < C_i\}$, where X_i could be less than Y_i. However $\hat{S}_Y$ only jumps on $\{Y_i : Y_i = T_{3i}\}$. Taking the intersection of these two sets gives us set S.

To solve for $\hat{\theta}$, it is equivalent to solve

$$\Psi_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Y_i < \min\{X_i, C_i\}\}[\hat{S}_Y(Y_i) - \tilde{S}_Y(Y_i)]\frac{\partial \hat{S}_Y(Y_i)}{\partial \theta} = 0. \tag{2.1}$$

Therefore, we propose our estimator to be the root $\hat{\theta}$ of equation 2.1, in the sense of a Z-estimator.

2.3 Consistency of $\hat{\theta}$

In this section, we prove the consistency of the estimator $\hat{\theta}$.

Theorem 1: Let X, Y and C be under semi-competing risks setting and $\hat{S}_Y$, $\tilde{S}_Y$ be defined as above. Then

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{n}\sum_{A}[\hat{S}_Y(Y_i) - \tilde{S}_Y(Y_i)]^2$$

is a consistent estimator of θ_0.

Proof: The copula graphical estimator can be expressed in counting process notation as

$$\hat{S}(y) = \psi_\theta^{-1}\left[-\frac{1}{n}\int_0^y \mathbb{1}\{B(u) > 0\}\psi_\theta'\left(\frac{\bar{B}(u)}{n}\right) d\bar{N}(u)\right],$$

where $B(u) = \mathbb{1}\{Z \geq u\}$, $N(u) = \mathbb{1}\{Z \leq u, D = 1\}$, $\bar{B}(u) = \sum_{i=1}^{n} B(u)$ and $\bar{N}(u) = \sum_{i=1}^{n} N(u)$. Rivest(2001)[16] showed that it is a uniformly consistent estimator of

$$S^\star(t) = \psi_\theta^{-1}\left[-\int_0^t \psi_\theta'(\pi(u))\pi(u)\, d\Lambda^{\#}(u)\right],$$

where $\Lambda^{\#}(u)$ is the cumulative crude hazard function. When the copula for the dependency is Archimedean, with generator function ψ_θ, $S^\star = S$.

To begin with, since $\tilde{S}_Y$ and $\hat{S}_Y$ converges in probability to S_Y and $S^\star$ respectively, $\hat{S}_Y(Y_i) - \tilde{S}_Y(Y_i)$ is asymptotically equivalent to $S^\star(Y_i) - S_Y(Y_i)$. When

$\theta = \theta_0$ where θ_0 is the true association parameter, $S^\star = S_Y$ and $S^\star(Y_i) - S_Y(Y_i) = 0$ for all $Y_i > 0$.

Then we prove that $\dfrac{\partial \hat{S}_Y(Y_i)}{\partial \theta}$ is asymptotically equivalent to $\dfrac{\partial S^\star(Y_i)}{\partial \theta}$.

$$\frac{\partial \hat{S}_Y(u)}{\partial \theta} = \frac{\partial}{\partial \theta}\psi_\theta^{-1}\left[-\frac{1}{n}\int_0^t \mathbb{1}\{B(u) > 0\}\psi_\theta'\left(\frac{\bar{B}(u)}{n}\right) d\bar{N}(u)\right]$$

$$= \frac{\partial \psi_\theta^{-1}}{\partial \theta}\left[-\frac{1}{n}\int_0^t \mathbb{1}\{B(u) > 0\}\psi_\theta'\left(\frac{\bar{B}(u)}{n}\right) d\bar{N}(u)\right]$$

$$\left[-\frac{1}{n}\int_0^t \mathbb{1}\{B(u) > 0\}\frac{\partial \psi_\theta'}{\partial \theta}\left(\frac{\bar{B}(u)}{n}\right) d\bar{N}(u)\right]$$

$$= \frac{\partial \psi_\theta^{-1}}{\partial \theta}[\psi_\theta(\hat{S}_Y(u))]\left[-\frac{1}{n}\int_0^t \mathbb{1}\{B(u) > 0\}\frac{\partial \psi_\theta'}{\partial \theta}\left(\frac{\bar{B}(u)}{n}\right) d\bar{N}(u)\right].$$

Under proper assumptions on the smoothness of ψ_θ, by continuous mapping theorem,

$$\frac{\partial \psi_\theta^{-1}}{\partial \theta}[\psi_\theta(\hat{S}_Y(u))] \xrightarrow{P} \frac{\partial \psi_\theta^{-1}}{\partial \theta}[\psi_\theta(S^\star(u))].$$

Moreover, since $\psi_\theta(\hat{S}_Y(u))$ converges to $\psi_\theta(S^\star(u))$ in probability, simply replace the function in the integral results in

$$-\frac{1}{n}\int_0^t \mathbb{1}\{B(u) > 0\}\frac{\partial \psi_\theta'}{\partial \theta}\left(\frac{\bar{B}(u)}{n}\right) d\bar{N}(u) \xrightarrow{P} -\int_0^t \frac{\partial \psi_\theta'}{\partial \theta}(\pi(u))\pi(u)\, d\Lambda^{\#}(u).$$

Combining the two results above,

$$\frac{\partial \psi_\theta^{-1}}{\partial \theta}[\psi_\theta(\hat{S}_Y(u))]\left[-\frac{1}{n}\int_0^t \mathbb{1}\{B(u) > 0\}\frac{\partial \psi_\theta'}{\partial \theta}\left(\frac{\bar{B}(u)}{n}\right) d\bar{N}(u)\right]$$

$$\xrightarrow{P} \frac{\partial \psi_\theta^{-1}}{\partial \theta}[\psi_\theta(S^\star(u))]\left[-\int_0^t \frac{\partial \psi_\theta'}{\partial \theta}(\pi(u))\pi(u)\, d\Lambda^{\#}(u)\right],$$

i.e., $\frac{\partial \hat{S}_Y(Y_i)}{\partial \theta} \xrightarrow{P} \frac{\partial S^\star(Y_i)}{\partial \theta}$. Therefore, $\Psi_n(\theta)$ is asymptotically equivalent to $\Psi_n^{'}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\{Y_i < \min\{X_i, C_i\}\}[S^\star(Y_i) - S_Y(Y_i)]\frac{\partial S^\star(Y_i)}{\partial \theta}$.

Denote $\Psi(\theta)$ the expectation of $\mathbb{1}\{Y_i < \min\{X_i, C_i\}\}[S^\star(Y_i) - S_Y(Y_i)]\frac{\partial S^\star(Y_i)}{\partial \theta}$, notice that $\Psi(\theta_0) = E\left[\mathbb{1}\{Y_i < \min\{X_i, C_i\}\}[S_Y(Y_i) - S_Y(Y_i)]\frac{\partial S^\star(Y_i)}{\partial \theta}\right] = 0$. By law of large numbers, $\sup_\theta \|\Psi_n(\theta) - \Psi(\theta)\| \xrightarrow{P} 0$.

When $|\theta - \theta_0| > \epsilon$ for any fixed $\epsilon > 0$, by Proposition 2 of [16], under most Archimedean copulas, $S^\star(Y_i) - S_Y(Y_i)$ is either always positive (or always negative, depends on the magnitude of θ against θ_0) for all $Y_i > 0$. Moreover,

$$\frac{\partial S^\star(Y_i)}{\partial \theta} = \lim_{\epsilon \to 0} \frac{S^\star_{\theta+\epsilon}(Y_i) - S^\star_\theta(Y_i)}{\epsilon},$$

and since the numerator is stochastically ordered by the same proposition, $\frac{\partial S^\star(Y_i)}{\partial \theta}$ is always positive (or always negative) as well. Finally, $\mathbb{1}\{Y_i < \min\{X_i, C_i\} \geq 0$, which proves that the expectation is always positive (or negative).

Therefore, $\inf_{|\theta-\theta_0|>\epsilon} \|\Psi(\theta)\| > 0 = \|\Psi(\theta_0)\|$. By construction, $\Psi_n(\hat{\theta}) = 0$, and using Theorem 5.9 of [19], $\hat{\theta} \xrightarrow{P} \theta_0$. An example of Clayton copula is given below.

Example (Clayton copula): For any $0 < u < 1$, $v > 0$ and $\theta > 0$,

$$\psi_\theta(u) = u^{-\theta} - 1,$$

$$\psi_\theta^{'}(u) = -\theta u^{-\theta-1},$$

$$\frac{\partial \psi_\theta^{'}(u)}{\partial \theta} = u^{-\theta-1}(\theta \ln u - 1) < 0,$$

$$\psi_\theta^{-1}(v) = (1+v)^{-\frac{1}{\theta}},$$

$$\frac{\partial \psi_\theta^{-1}(v)}{\partial \theta} = \frac{1}{\theta^2}(1+v)^{-\frac{1}{\theta}} \ln(1+v) > 0.$$

Therefore,

$$\frac{\partial S^{\star}(Y_i)}{\partial \theta} = \frac{\partial \psi_{\theta}^{-1}}{\partial \theta}[\psi_{\theta}(S^{\star}(u))] \left[- \int_0^t \frac{\partial \psi_{\theta}^{'}}{\partial \theta}(\pi(u))\pi(u)\, d\Lambda^{\#}(u) \right] > 0.$$

Moreover,

$$\frac{\psi_{\theta_0}^{'}(u)}{\psi_{\theta}^{'}(u)} = \frac{\theta_0}{\theta} u^{\theta - \theta_0}$$

is increasing when $\theta > \theta_0$, which implies that $S^{\star}(Y_i) \leq S_Y(Y_i)$ for any $Y_i > 0$. Therefore, $\Psi(\theta) < 0$. Similarly, $\Psi(\theta) > 0$ for any fixed $\theta < \theta_0$. But either way, $\|\Psi(\theta)\| > 0$,

2.4 Marginal Distribution Estimation

When $\hat{\theta}$ estimating θ_0 consistently, we can retrieve the marginal distribution of X, Y and C. For S_Y and S_C, using Kaplan-Meier estimator calculated by (T_2, D_2) is the most straightforward way. To estimate S_X, we can use the copula graphical estimator with $\psi_{\hat{\theta}}$. That is

$$\hat{S}_X(x) = \psi_{\hat{\theta}}^{-1} \left[- \sum_{Z_i \leq x, D_i = 0} \psi_{\hat{\theta}}[\hat{\pi}_2(Z_i^-)] - \psi_{\hat{\theta}}[\hat{\pi}_2(Z_i)] \right].$$

2.5 Simulation Results

We began by generating 500 pairs of (X, Y) using Clayton copula. Their marginal distributions were chosen to be exponential with parameter 1.5 and 1, because of the nature that the terminal event usually happens after non-terminal event. We chose θ to be 0.5, 2 and 8 so that the corresponding τ is 0.2, 0.5 and 0.8, respectively. We used exponential distribution with parameter 0.5 and 1 for the marginal distribution of the independent censoring time C to see the performance of the estimator under different censoring rate. We repeated this process 1000 times to get the mean square error of $\hat{\tau}$, which is one-to-one mapped to $\hat{\theta}$. We also simulated data when sample size is not quite large. When $n = 200$, the MSE of $\hat{\tau}$ is shown in the parenthesis. In the

end, we generated 200 bootstrap samples to check the performance of the variance. The result is shown on Table 2.1.

Table 2.1 Simulation Results for Clayton Copula

τ	0.2		0.5		0.8	
θ_0	0.5		2		8	
Censor	19%	31%	22%	36%	24%	39%
MSE	0.0026	0.0054	0.0017	0.0024	0.0008	0.0010
	(0.0117)	(0.0184)	(0.0056)	(0.0088)	(0.0032)	(0.0085)
Var	0.0019	0.0039	0.0012	0.0016	0.0004	0.0005
VarBS	0.0019	0.0034	0.0014	0.0016	0.0005	0.0007

From Table 2.1, we can see that under moderate censoring rate, the performance of both the parameter estimator and the marginal distribution estimator are extremely good. In fact, as the dependency getting stronger, this approach is producing highly accurate estimation. As censoring rate decreases, we have more data to use in calculating $\hat{\theta}$, therefore the MSE drops as well. The variance and bootstrap variance of the estimator also agrees with one another. However, when we reduce the sample size, MSE under low dependency ($\tau = 0.2$) is not quite appealing. This is not surprising because small sample size, censoring, plus subsetting in estimation can magnify the effect of bad inputs. In addition, the estimator is designed to be used when the dependency is strong. In practice, we suggest that a maximum of 30% censor rate should be hold when sample size is less than 500.

An example of marginal distribution estimation can be found in Figure 2.2. The blue estimation follows quite close to the underlying distribution. In fact, in almost all the cases, the semi-parametric estimator estimates the distribution really well.

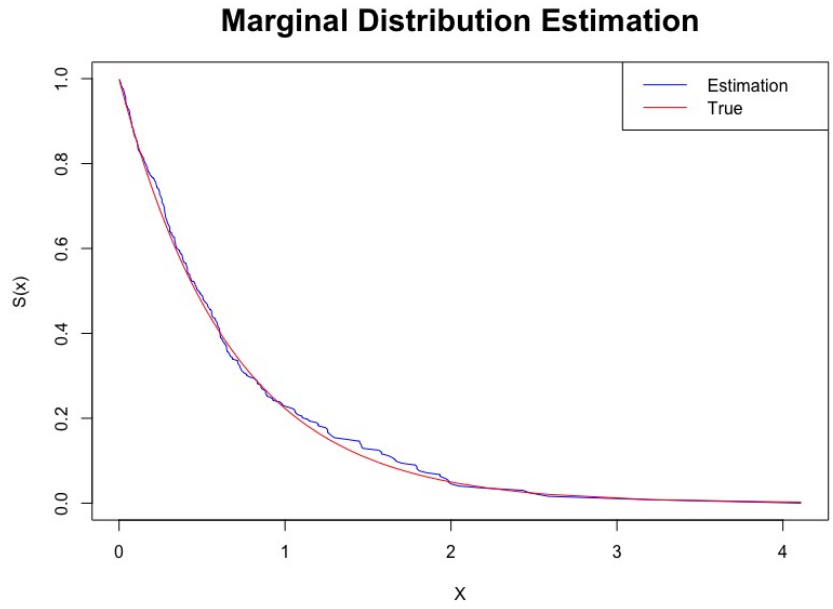

Figure 2.2 An example of $\hat{S}_X$ vs. S_X.

Similar results can be found in Table 2.2, where the copula used is Gumbel with $\psi_\theta(u) = [-\log(u)]^{\frac{1}{\theta}}$.

Table 2.2 Simulation Results for Gumbel Copula

τ	0.2		0.5		0.8	
θ_0	0.8		0.5		0.2	
Censor	19%	31%	22%	36%	24%	39%
MSE	0.0042	0.0046	0.0018	0.0025	0.0011	0.0017
Var	0.0030	0.0032	0.0012	0.0016	0.0002	0.0004
VarBS	0.0028	0.0039	0.0017	0.0013	0.0004	0.0005

2.6 Leukemia Data Example

We analyze the Leukemia data '*bmt*' from *R* package 'KMsurv'. The data records the survival state of 137 patients after taking bone marrow transplant, see Klein(2003)[9].

The patients experience either relapse, GVHD or was alive at the end of the study. Relapse and GVHD are the main causes of death and we are interested in the distribution of relapse, as well as predicting death when relapse happens. Two time points and three indicators were kept, which can be translated into our T_2, T_3, D_2 and D_3.

Using Clayton copula model, our estimate to the association between relapse and death $\hat{\theta} = 6.41$, with corresponding $\tau = 0.76$. This suggests a very strong positive relation between them. We also estimated the marginal distribution $\hat{S}_X$, as shown in Figure 2.3.

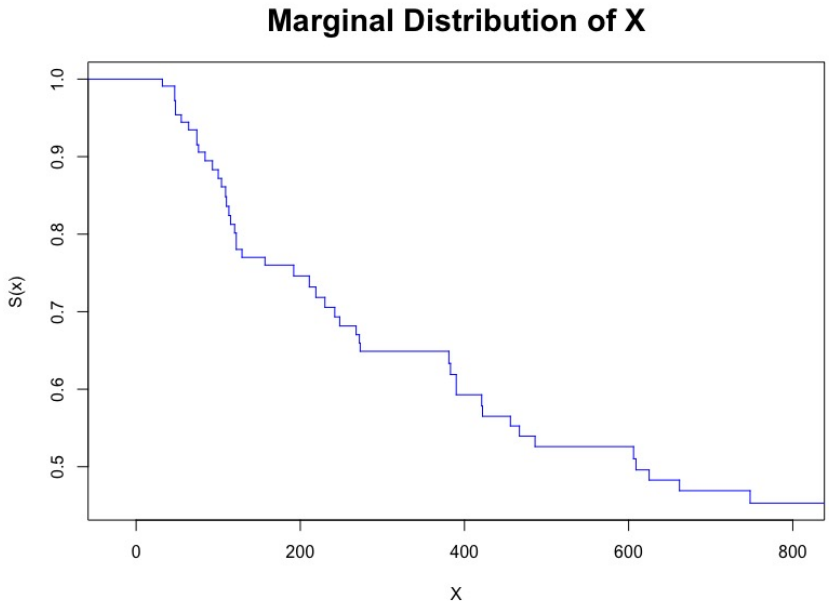

Figure 2.3 $\hat{S}_X$ of Leukemia data.

2.7 Conclusion

Under semi-competing risks data, the estimator we proposed works for most common Archimedean copulas. It performs best when the dependence is strong. A weight term should be considered during estimation process in the future. The asymptotic normality of this estimator can also be proved.

CHAPTER 3

A NEW ESTIMATOR OF BASELINE HAZARD FUNCTION

In clinical trials, we encounter clustered data all the time. For example, most clinical trials assign patients into a treatment group and a control group. The survival time in different groups shall follow different distributions. In such cases, group is a cluster factor. The cluster factor is always observable at the end of the experiment, although in a double blind trial this factor may not be known during the experiment.

Previously, to fit a frailty model to correlated clustered survival data, EM algorithm was applied and Breslow estimator was used to estimate the baseline hazard functions(see Lin(2007)[11]). However, for this class of models, we will show that the baseline hazard function can be estimated using an alternative approach and our estimator is comparable with the Breslow estimator. Furthermore, our estimator can be used as a model checking tool for corresponding frailty distribution based on dependent censored data.

3.1 Frailty Model for Clustered Data

To account for the association between the failure time and the censoring time for clustered data, Manatunga(1999)[13] proposed the following frailty model to fit matched pair survival data (T, C) that

$$\Lambda_T(t|Z_T, Z_C, W) = \Lambda_{T_0}(t)h_T(\beta_T' Z_T)W$$

and

$$\Lambda_C(c|Z_T, Z_C, W) = \Lambda_{C_0}(t)h_C(\beta_C' Z_C)W,$$

where $Z = (Z_T, Z_C)$ is the observable covariate vector denoting cluster, h_T and h_C are known positive convex functions, and W follows some frailty distribution

with the unknown parameter θ. For simplicity, we choose $h_T(u) = h_C(u) = e^u$. Wang(2015)[23] showed that the joint survival function of T and C in the model above satisfies

$$S(t, c|Z) = \psi_\theta\{\psi_\theta^{-1}[S_T(t|Z)] + \psi_\theta^{-1}[S_C(c|Z)]\},$$

which means that this clustered frailty model also naturally arises from Archimedean copula given the cluster covariate.

Now, we extend the distribution of marginal survival functions as shown in Section 1.2 into the clustered setting.

Theorem 1: Assume that the distribution of $(T, C|Z)$ can be modeled by an Archimedean copula with generator ψ_θ such that

$$S(t, c|Z) = \psi_\theta[\psi_\theta^{-1}(S_T(t|Z)) + \psi_\theta^{-1}(S_C(c|Z))].$$

and that the marginal distribution functions of $T|Z$ and $C|Z$ are absolutely continuous. Then we have

$$S_T(t|Z) = \psi_\theta \left\{ \int_0^t \psi_\theta^{-1\prime}[\pi(u|Z)]\pi(u|Z)\, d\ln[S_T^\star(u|Z)] \right\}$$

and

$$S_C(c|Z) = \psi_\theta \left\{ \int_0^c \psi_\theta^{-1\prime}[\pi(u|Z)]\pi(u|Z)\, d\ln[S_C^\star(u|Z)] \right\},$$

respectively for all $t > 0$ and $c > 0$.

In the formula above, $\pi(u|Z) = P(T > u, C > u|Z)$ for all $u > 0$. $S_T^\star(t|Z)$ and $S_C^\star(c|Z)$ are the marginal survival functions of $T|Z$ and $C|Z$, respectively, under the additional assumption that the two variables T and C are conditionally independent given Z.

Under the model assumption, we have that

$$\Lambda_T(t|Z_T, Z_C, W) = \Lambda_{T_0}(t)e^{\beta_T' Z_T}W$$

$$= -\log[S_{T_0}(t)]e^{\beta_T' Z_T}W = -\log S_T(t|Z_T, Z_C, W)$$

where $\Lambda_{T_0}(t) = -\log[S_{T_0}(t)]$ and $S_{T_0}(t)$ is the baseline survival function. If we take the derivative of the equation above with respect to t, we also have

$$\lambda_T(t|Z_T, Z_C, W) = \lambda_{T_0}(t)e^{\beta_T' Z_T}W.$$

The same condition holds for C that

$$\Lambda_C(c|Z_T, Z_C, W) = \Lambda_{C_0}(c)e^{\beta_C' Z_C}W$$

$$= -\log[S_{C_0}(c)]e^{\beta_C' Z_C}W = -\log S_C(c|Z_T, Z_C, W)$$

and

$$\lambda_C(c|Z_T, Z_C, W) = \lambda_{C_0}(c)e^{\beta_C' Z_C}W,$$

where $\Lambda_{C_0}(c) = -\log[S_{C_0}(c)]$.

Similar to what we've shown in Section 1.2,

$$S(t, c|Z) = E[S(t, c|Z)|W] = E[S_T(t|Z, W)S_C(c|Z, W)]$$

$$= E[\exp\{\log S_T(t|Z, W) + \log S_C(c|Z, W)\}]$$

$$= E[\exp\{-\Lambda_{T_0}(t)e^{\beta_T' Z_T}W - \Lambda_{C_0}(c)e^{\beta_C' Z_C}W\}]$$

$$= \psi\{\Lambda_{T_0}(t)e^{\beta_T' Z_T} + \Lambda_{C_0}(c)e^{\beta_C' Z_C}\}$$

Considering the fact that $S(t,0|Z) = S_T(t|Z)$ and $S(0,c|Z) = S_C(c|Z)$, we have

$$S_T(t|Z) = \psi_\theta\{\Lambda_{T_0}(t)e^{\beta_T' Z_T} + \Lambda_{C_0}(0)e^{\beta_C' Z_C}\} = \psi_\theta\{\Lambda_{T_0}(t)e^{\beta_T' Z_T}\}$$

and similarly

$$S_C(c|Z) = \psi_\theta\{\Lambda_{C_0}(c)e^{\beta_C' Z_C}\}$$

Therefore,

$$\psi_\theta^{-1}[S_T(t|Z)] = \Lambda_{T_0}(t)e^{\beta_T' Z_T},$$

and

$$\psi_\theta^{-1}[S_C(c|Z)] = \Lambda_{C_0}(c)e^{\beta_C' Z_C}$$

Comparing the formula above, we have found that

$$\Lambda_{T_0}(t)e^{\beta_T' Z_T} = \int_0^t \psi_\theta^{-1\prime}[\pi(u|Z)]\pi(u|Z)d\ln[S_T^\star(u|Z)]$$

and similarly

$$\Lambda_{C_0}(c)e^{\beta_C' Z_C} = \int_0^c \psi_\theta^{-1\prime}[\pi(u|Z)]\pi(u|Z)d\ln[S_C^\star(u|Z)],$$

and these equations will lead to the main result. Now we propose Theorem 2.

Theorem 2: Assume that the distribution of $(T, C|Z, W)$ can be modeled by a frailty model such that

$$\Lambda_T(t|Z,W) = \Lambda_{T_0}(t)\exp(\beta_T' Z_T)W$$

and

$$\Lambda_C(c|Z,W) = \Lambda_{C_0}(c)\exp(\beta_C' Z_C)W,$$

where W follows some distribution with parameter θ with the Laplace transform $\psi_\theta(s) = E(e^{-sW})$, then the baseline cumulative hazard functions can be expressed as:

$$\Lambda_{T_0}(t) = \int_0^t \frac{\psi_\theta^{-1\prime}[\pi(u|Z)]\pi(u|Z)}{\exp(\beta_T' Z_T)}\, d\ln[S_T^\star(u|Z)]$$

and

$$\Lambda_{C_0}(c) = \int_0^c \frac{\psi_\theta^{-1\prime}[\pi(u|Z)]\pi(u|Z)}{\exp(\beta_C' Z_C)}\, d\ln[S_C^\star(u|Z)]$$

respectively, where $S_T^\star$ and $S_C^\star$ are the marginal survival functions of $T|Z$ and $C|Z$, under the additional assumption that the two variables T and C are in fact independent given Z.

Theorem 2 tells us that for bivariate frailty models with a common frailty, the baseline distributions of failure times are actually not arbitrary. In fact, they are functions of the ψ and the distribution of $(X, \delta|Z) = (\min\{T, C\}, I_{T<C}|Z)$. This means that if we have know the parameters θ, $\beta = (\beta_T, \beta_C)$ and the distribution of $(X, \delta|Z)$, we can determine the baseline distributions of failure times uniquely.

If there are no covariates, i.e., $\beta_T = \beta_C = 0$, the formulas for baseline hazard functions can be simplified to

$$\Lambda_{T_0}(t) = \int_0^t \psi_\theta^{-1\prime}\{\pi(u)\}\pi(u) d\ln[S_T^\star(u)]$$

and

$$\Lambda_{C_0}(c) = \int_0^c \psi_\theta^{-1\prime}\{\pi(u)\}\pi(u) d\ln[S_C^\star(u)],$$

respectively.

We conclude this section by two examples.

Example 1: Suppose that (T, C) follows the Clayton copula model with association parameter θ. T and C follow the same marginal distributions as $\exp(\lambda)$

so that the joint survivor function of (T, C) is:

$$S(t,c) = [S_T(t)^{-\theta} + S_C(c)^{-\theta} - 1]^{-\frac{1}{\theta}} = (e^{\theta\lambda t} + e^{\theta\lambda c} - 1)^{-\frac{1}{\theta}}.$$

For simplicity, we don't consider the covariate Z, i.e., $\beta_T = \beta_C = 0$. By the non-identifiability property, we note that

$$\pi(u) = S(u,u) = S^\star(u,u) = S_T^\star(u) S_C^\star(u).$$

If we further assume $S_T^\star(u) = S_C^\star(u)$, it follows that

$$[S_T^\star(u)]^2 = (e^{\theta\lambda u} + e^{\theta\lambda u} - 1)^{-\frac{1}{\theta}} = (2e^{\theta\lambda u} - 1)^{-\frac{1}{\theta}}.$$

Therefore,

$$[S_T^\star(u)] = [S_C^\star(u)] = (2e^{\theta\lambda u} - 1)^{-\frac{1}{2\theta}}$$

Then we use the formulas in Theorem 2 to estimate the baseline hazard. In a Clayton model, $\psi_\theta^{-1}(s) = s^{-\theta} - 1$, therefore,

$$\psi_\theta^{-1\prime}[\pi(u)]\pi(u) = -\theta\pi(u)^{-\theta-1}\pi(u) = -\theta\pi(u)^{-\theta},$$

where $\pi(u) = S_T^\star(u) S_C^\star(u) = (2e^{\theta\lambda u} - 1)^{-\frac{1}{\theta}}$ and

$$d\ln[S_T^\star(u)] = d\ln(2e^{\theta\lambda u} - 1)^{-\frac{1}{2\theta}} = -\frac{1}{2\theta} d\ln(2e^{\theta\lambda u} - 1)$$

$$= -\frac{1}{2\theta(2e^{\theta\lambda u} - 1)} d(2e^{\theta\lambda u} - 1).$$

Finally,

$$\Lambda_{T_0}(t) = \int_0^t -\theta(2e^{\theta\lambda u} - 1)^{(-\frac{1}{\theta})(-\theta)} \left[-\frac{1}{2\theta(2e^{\theta\lambda u} - 1)}\right] d(2e^{\theta\lambda u} - 1)$$

$$= \frac{1}{2}\int_0^t d(2e^{\theta\lambda u} - 1) = \frac{1}{2}(2e^{\theta\lambda u} - 1)\Big|_0^t = e^{\theta\lambda t} - 1.$$

Similarly, $\Lambda_{C_0}(c) = e^{\theta\lambda c} - 1$. And the corresponding bivariate frailty model is

$$\Lambda_T(t|W) = \Lambda_{T_0}(t)W = (e^{\theta\lambda t} - 1)W,$$

$$\Lambda_C(c|W) = \Lambda_{C_0}(c)W = (e^{\theta\lambda t} - 1)W.$$

Example 2: Suppose that (T, C) follows the Gumbel copula model with association parameter θ. T and C follow the same marginal distributions as $\exp(\lambda)$ so that the joint survivor function of (T, C) can be written as:

$$S(t, c) = \exp[-\{[-\log S_T(t)]^\theta + [-\log S_C(c)]^\theta\}^{1/\theta}]$$

$$= \exp\{-[(\lambda t)^\theta + (\lambda c)^\theta]^{1/\theta}\}.$$

Similar to Example 1,

$$\pi(u) = [S_T^\star(u)]^2 = \exp\{-[(\lambda u)^\theta + (\lambda u)^\theta]^{1/\theta}\} = \exp\{-2^{1/\theta}\lambda u\},$$

and

$$[S_T^\star(u)] = [S_C^\star(u)] = \exp\{-2^{1/\theta-1}\lambda u\}.$$

As in Gumbel model, $\psi_\theta^{-1}(s) = (-\log s)^\theta$, therefore,

$$\psi_\theta^{-1\prime}[\pi(u)]\pi(u) = \theta[-\log \pi(u)]^{\theta-1}\frac{1}{-\pi(u)}\pi(u)$$

$$= -\theta[-\log \pi(u)]^{\theta-1} = -\theta(2^{1/\theta}\lambda u)^{\theta-1}$$

$$= -2^{1-1/\theta}\theta\lambda^{\theta-1}u^{\theta-1},$$

while

$$d\ln[S_T^\star(u)] = d(-2^{1/\theta-1}\lambda u) = -2^{1/\theta-1}\lambda du.$$

Therefore,

$$\Lambda_{T_0}(t) = \int_0^t -2^{1-1/\theta}\theta\lambda^{\theta-1}u^{\theta-1}(-2^{1/\theta-1}\lambda)du$$

$$= \lambda^\theta \int_0^t \theta u^{\theta-1}du = \lambda^\theta(u^\theta)\Big|_0^t = (\lambda t)^\theta$$

and $\Lambda_{C_0}(c) = (\lambda c)^\theta$. Finally, the corresponding bivariate frailty model is

$$\Lambda_T(t|W) = \Lambda_{T_0}(t)W = (\lambda t)^\theta W$$

and

$$\Lambda_C(c|W) = \Lambda_{C_0}(c)W = (\lambda c)^\theta W$$

3.2 A New Estimator of Baseline Hazard Function

Let our observed dependent censored data set be (X_i, δ_i, Z_i), where $X_i = \min\{T_i, C_i\}$ and $\delta_i = I_{T_i<C_i}$. Using Theorem 2 we can construct an alternative estimator of baseline cumulative hazard function of Tas

$$\hat{\Lambda}_{T_0}(t) = \sum_{X_i<t} \frac{-\psi_{\hat{\theta}}^{-1\prime}[\hat{\pi}(X_i|Z)]\hat{\pi}(X_i|Z)}{\exp(\hat{\beta}_T' Z_T)} \times \frac{\Delta\bar{N}(X_i|Z)}{\bar{Y}(X_i|Z)}$$

$$= \sum_{X_i<t} \frac{-\psi_{\hat{\theta}}^{-1\prime}[\hat{\pi}(X_i|Z)]\Delta\bar{N}(X_i|Z)}{n_Z \exp(\hat{\beta}_T' Z_T)}.$$

In this formula, $\hat{\theta}$ and $\hat{\beta}_T$ is the estimation of θ and β_T by EM algorithm(see Dempster(1977)[3]). Let $N_i(t|Z) = I_{X_i<t,\delta_i=1}|Z$ and $\bar{N}(t|Z) = \sum_i N_i(t|Z)$, so that $\Delta\bar{N}(t|Z)$ is the number of events at time t. Similarly, we define $Y_i(t|Z) = I_{X_i\geq t}|Z$ so that $\bar{Y}(t|Z) = \sum_i Y_i(t|Z)$ is the number of people at risk at time t. $n_Z = \sum_i(I_{Z_i=Z})$ is the total number of people in group i.

Using counting process, our estimator can be written as:

$$\hat{\Lambda}_{T_0}(t) = \int_0^t \frac{-\psi_{\hat{\theta}}^{-1\prime}\{\hat{\pi}(u|Z)\}}{n_Z \exp(\hat{\beta}_T' Z_T)}\, d\bar{N}(u|Z)$$

$$= \int_0^t \frac{-\psi_\theta^{-1\prime}\{\pi(u|Z)\}}{n_Z \exp(\beta_T' Z_T)}\, d\bar{N}(u|Z)$$

$$+ \int_0^t \left(\frac{\psi_\theta^{-1\prime}\{\pi(u|Z)\}}{n_Z \exp(\beta_T' Z_T)} - \frac{\psi_{\hat{\theta}}^{-1\prime}\{\hat{\pi}(u|Z)\}}{n_Z \exp(\hat{\beta}_T' Z_T)} \right) d\bar{N}(u|Z).$$

Define

$$M_i(t|Z) = N_i(t|Z) - \int_0^t Y_i(t|Z)\, d\Lambda^\star(t|Z)$$

and

$$\bar{M}(t|Z) = \bar{N}(t|Z) - \int_0^t \bar{Y}(t|Z)\, d\Lambda^\star(t|Z),$$

by Theorem 1.3.1 on Fleming(2005)[5], M_i and $\bar{M}$ are martingales with respect to the σ fields

$$F_t^i = \sigma\{I_{X_i \le u, \delta_i = 1}, I_{X_i \le u, \delta_i = 0}, 0 \le u \le t|Z\}$$

and $F_t = \vee_{i=1}^n F_t^i$ respectively, where

$$\Lambda^\star(t|Z) = \int_0^t \frac{f^\star(u|Z)}{S^\star(u|Z)}\, du$$

is the cumulative hazard function of $T|Z$ under the assumption of $T|Z$ and $C|Z$ are independent. Therefore we have

$$\hat{\Lambda}_{T_0}(t) = \int_0^t \frac{-\psi_\theta^{-1\prime}\{\pi(u|Z)\}}{n_Z \exp(\beta_T' Z_T)}\, d\bar{M}(u|Z) + \int_0^t \frac{-\psi_\theta^{-1\prime}\{\pi(u|Z)\}}{n_Z \exp(\beta_T' Z_T)} \bar{Y}(u|Z)\, d\Lambda^\star(u|Z)$$

$$+ \int_0^t \left(\frac{\psi_\theta^{-1\prime}\{\pi(u|Z)\}}{n_Z \exp(\beta_T' Z_T)} - \frac{\psi_{\hat{\theta}}^{-1\prime}\{\hat{\pi}(u|Z)\}}{n_Z \exp(\hat{\beta}_T' Z_T)} \right) d\bar{M}(u|Z)$$

$$+ \int_0^t \left(\frac{\psi_\theta^{-1\prime}\{\pi(u|Z)\}}{n_Z \exp(\beta_T' Z_T)} - \frac{\psi_{\hat{\theta}}^{-1\prime}\{\hat{\pi}(u|Z)\}}{n_Z \exp(\hat{\beta}_T' Z_T)} \right) \bar{Y}(u|Z)\, d\Lambda^\star(u|Z).$$

After some simplification we have:

$$\hat{\Lambda}_{T_0}(t) - \Lambda_{T_0}(t) = \int_0^t \frac{-\psi_\theta^{-1\prime}\{\pi(u|Z)\}}{n_Z \exp(\beta_T' Z_T)}\, d\bar{M}(u|Z)$$

$$+ \int_0^t \frac{-\psi_\theta^{-1\prime}\{\pi(u|Z)\}}{\exp(\beta_T' Z_T)}(\hat{\pi}(u|Z) - \pi(u|Z))\, d\Lambda^\star(u|Z)$$

$$+ \frac{1}{n_Z}\int_0^t \left(\frac{\psi_\theta^{-1\prime}\{\pi(u|Z)\}}{\exp(\beta_T' Z_T)} - \frac{\psi_{\hat{\theta}}^{-1\prime}\{\hat{\pi}(u|Z)\}}{\exp(\hat{\beta}_T' Z_T)}\right) d\bar{M}(u|Z)$$

$$+ \int_0^t \left(\frac{\psi_\theta^{-1\prime}\{\pi(u|Z)\}}{\exp(\beta_T' Z_T)} - \frac{\psi_{\hat{\theta}}^{-1\prime}\{\hat{\pi}(u|Z)\}}{\exp(\hat{\beta}_T' Z_T)}\right) \hat{\pi}\, d\Lambda^\star(u|Z).$$

Using Lengart's inequality and similar arguments to prove Theorem 3.4.2 in Fleming(2005)[5], we can show that the first term and the third term go to zero in probability when $n_Z \to \infty$. Using the Glivenko-Cantelli Theorem, it is easy to show that other two terms go to zero uniformly in probability under the boundedness assumptions of the first and second derivatives of ψ^{-1}. Therefore we have proved the uniform consistency of our estimator.

To derive large sample results for our estimator, we have

$$\sqrt{n_Z}\left(\hat{\Lambda}_{T_0}(t) - \Lambda_{T_0}(t)\right) = \int_0^t \frac{-\psi_\theta^{-1\prime}\{\pi(u|Z)\}}{\sqrt{n_Z}\exp(\beta_T' Z_T)}\, d\bar{M}(u|Z)$$

$$+\sqrt{n_Z}\int_0^t \frac{-\psi_\theta^{-1\prime}\{\pi(u|Z)\}}{\exp(\beta_T' Z_T)}(\hat{\pi}(u|Z) - \pi(u|Z))\, d\Lambda^\star(u|Z)$$

$$+ \frac{1}{\sqrt{n_Z}}\int_0^t \left(\frac{\psi_\theta^{-1\prime}\{\pi(u|Z)\}}{\exp(\beta_T' Z_T)} - \frac{\psi_{\hat{\theta}}^{-1\prime}\{\hat{\pi}(u|Z)\}}{\exp(\hat{\beta}_T' Z_T)}\right) d\bar{M}(u|Z)$$

$$+\sqrt{n_Z}\int_0^t \left(\frac{\psi_\theta^{-1\prime}\{\pi(u|Z)\}}{\exp(\beta_T' Z_T)} - \frac{\psi_{\hat{\theta}}^{-1\prime}\{\hat{\pi}(u|Z)\}}{\exp(\hat{\beta}_T' Z_T)}\right) \hat{\pi}(u|Z)\, d\Lambda^\star(u|Z).$$

The third term converges uniformly to zero in probability because the corresponding predictive variation process of the third term has a compensator

$$\int_0^t \left(\frac{\psi_\theta^{-1\prime}\{\pi(u|Z)\}}{\exp(\beta_T' Z_T)} - \frac{\psi_{\hat{\theta}}^{-1\prime}\{\hat{\pi}(u|Z)\}}{\exp(\hat{\beta}_T' Z_T)} \right)^2 \frac{\bar{Y}(u|Z)}{n_Z} \, d\Lambda^\star(u|Z)$$

which converges to zero in probability. The first term, the second term and the fourth term all converge to Gaussian processes. Using the Taylor expansion, we have

$$X_{n_Z}(u) = \sqrt{n_Z} \left\{ \frac{\psi_\theta^{-1\prime}\{\pi(u|Z)\}}{\exp(\beta_T' Z_T)} - \frac{\psi_{\hat{\theta}}^{-1\prime}\{\hat{\pi}(u|Z)\}}{\exp(\hat{\beta}_T' Z_T)} \right\}$$

$$\approx \exp(-\beta_T' Z_T) \{ \frac{\partial \psi_\theta^{-1\prime}\{\pi(u|Z)\}}{\partial \pi} \sqrt{n_Z} (\hat{\pi}(u|Z) - \pi(u|Z))$$

$$- \left\{ \frac{\partial \psi_\theta^{-1\prime}\{\pi(u|Z)\}}{\partial \Theta} \right\}^T \sqrt{n_Z}(\hat{\Theta} - \Theta) + \psi_\theta^{-1\prime}\{\pi(u|Z)\} Z^T \sqrt{n_Z}(\hat{\beta}_T - \beta_T) \}$$

which converges weakly to a mean zero Gaussian process $X(u)$ on $D[0, t_0)$. Define the limiting covariance of X_{n_Z} as:

$$\mathrm{cov}(X_{n_Z}(s), X_{n_Z}(t)) = V_0(s,t).$$

Define the covariance between X_{n_Z} and $\bar{M}$ as

$$\mathrm{cov}(X_{n_Z}(s), \bar{M}(t)) = \sqrt{n_Z} V_1(s,t).$$

Let $Y_{n_Z}(u) = \sqrt{n_Z}(\hat{\pi}(u|Z) - \pi(u|Z))$. The covariance between Y_{n_Z} and $\bar{M}$ is

$$\mathrm{cov}(Y_{n_Z}(s), \bar{M}(t)) = -\sqrt{n_Z} \pi(s) \Lambda^\star(s \wedge t)$$

as has been shown in Rivest(2001)[16]. Define the covariance between X_{n_Z} and Y_{n_Z} as

$$\mathrm{cov}(X_{n_Z}(s), Y_{n_Z}(t)) = V_2(s,t).$$

Define $A(u) = \psi_\theta^{-1\prime}(\pi(u))/\exp(2\beta Z)$. The asymptotic variance of

$$\sqrt{n_Z}\left(\hat{\Lambda}_{T_0}(t) - \Lambda_{T_0}(t)\right)$$

is $I_1 + I_2 + I_3 + C_1 + C_2 + C_3$, where

$$I_1 = \int_0^t A^2(u)\pi(u)d\Lambda^\star(u),$$

$$I_2 = 2\int_0^t \int_0^s A(u)A(s)[\pi(u) - \pi(u)\pi(s)]d\Lambda^\star(u)d\Lambda^\star(s),$$

$$I_3 = \int_0^t \int_0^t V_0(u,s)\pi(u)\pi(s)d\Lambda^\star(u)d\Lambda^\star(s),$$

$$C_1 = -A(t)\int_0^t A(s)\pi(s)\Lambda^\star(s)d\Lambda^\star(s) + \int_0^t \int_0^t \pi(s)\Lambda^\star(s\wedge u)A(s)dA(u)d\Lambda^\star(s),$$

$$C_2 = A(t)\int_0^t V_1(s,t)\pi(s)d\Lambda^\star(s) - \int_0^t \int_0^t V_1(s,u)\pi(s)dA(u)d\Lambda^\star(s)$$

and

$$C_3 = \int_0^t \int_0^t V_2(u,s)A(u)\pi(s)d\Lambda^\star(u)d\Lambda^\star(s).$$

In above expression, $s \wedge u$ represents the minimum value of s and t. In summary, we have proved:

Theorem 3: Let $t_0 > 0$, be such that $\pi(t_0) > 0$. Assume that the distribution of $(T, C)|Z, W$ can be modeled by a frailty model such that

$$\Lambda_T(t|Z, W) = \Lambda_{T_0}(t)\exp(\beta_T' Z_T)W$$

and

$$\Lambda_C(c|Z, W) = \Lambda_{C_0}(c)\exp(\beta_C' Z_C)W,$$

where W follows some parametric distribution with the Laplace transform $\psi(s) = E[\exp(-sW)]$. Suppose that the first two derivatives of $\psi^{-1}(s)$ with respect to s and θ are bounded for $s \in (t_0, 1)$ and the parameter estimates of unknown parameters θ, β_T and β_C are all asymptotically normal, the process $\sqrt{n_Z}\left(\hat{\Lambda}_{T_0}(t) - \Lambda_{T_0}(t)\right)$ converges weakly on $D[0, t_0)$ to a mean zero Gaussian process with variance function $v(t) = \sum_{i=1}^{3} I_i + \sum_{i=1}^{3} C_i$.

In practice, $v(t)$ is hard to estimate and bootstrap estimators will be applied to estimate corresponding variances. It is worth mentioning that our estimator presented above is an estimator of cumulative hazard function given the covariate Z (we used $\hat{\Lambda}_{T_0}(t)$ instead of $\hat{\Lambda}_{T_0}(t|Z)$ because the baseline cumulative functions of T are the same for different covariate values). Notice the fact that for each $Z = z_j$ (here we assume Z is a discrete covariate), we have an estimator of the baseline cumulative hazard function of T. An overall estimator of the baseline cumulative hazard function for T can thus be given by the weighted average of $\hat{\Lambda}_{T_0}(t|Z)$:

$$\hat{\Lambda}_{\text{Overall}}(t) = \sum_j \hat{\Lambda}_{T_0}(t|Z = z_j)\hat{\mathrm{P}}(Z = z_j).$$

3.3 A Model Checking Procedure for Frailty Models

Under our frailty model assumption, the baseline hazard functions are independent of covariate values based on Theorem 2. This fact motivates us to establish a model checking procedure for our frailty model assumption when the covariate Z takes finite values. For simplicity, we assume that the covariate $Z_T = Z_C = Z$ is a binary variable and for different Z values, we have independent estimators of corresponding baseline hazard functions (we denote them by $\hat{\Lambda}_{T_0}(t|Z = Z_i)$ for $i = 1, 2$ respectively) which should be the same asymptotically because:

$$\hat{\Lambda}_{T_0}(t|Z = Z_1) - \hat{\Lambda}_{T_0}(t|Z = Z_2)$$

$$= \left(\hat{\Lambda}_{T_0}(t|Z = Z_1) - \Lambda_{T_0}(t)\right) - \left(\hat{\Lambda}_{T_0}(t|Z = Z_2) - \Lambda_{T_0}(t)\right) \to 0$$

almost surely when $n \to \infty$. If we plot two estimators $\hat{\Lambda}_{T_0}(t|Z = Z_1)$ and $\hat{\Lambda}_{T_0}(t|Z = Z_2)$ against T respectively (or $\hat{\Lambda}_{C_0}(c|Z = Z_1)$ and $\hat{\Lambda}_{C_0}(c|Z = Z_2)$ against censoring time C respectively), they should look similar graphically under the correct model assumption. A test may be established based on the asymptotic properties proved in Theorem 3, however, as the analytic form of the variance formulas is not available, a bootstrap procedure has to be applied to perform such a test based on the difference between baseline cumulative functions corresponding to different covariate values.

3.4 Simulation Studies

In this section, we conduct simulation studies to compare our estimator with the Breslow estimator. We generate dependent censored data $(T, C)|Z$ from Clayton copula for $Z = 0$ and $Z = 1$ respectively. The Kendall's τ is chosen to be on four levels: $0.2, 0.4, 0.6$ and 0.8 so that the association parameter θ is $0.5, 1.33, 3$ and 8. The sample sizes corresponding to each covariate Z is chosen to be 500. The baseline hazard functions are assumed to be constant 1. Then we calculate the baseline cumulative hazard functions $\hat{\Lambda_{T_0}}$ and$\tilde{\Lambda_{T_0}}$ using our estimator and Breslow estimator, respectively.

First, we could apply our graphical model checking procedure to see if the assumed frailty distribution fits the data. As we can see from Figure 3.1, the red line($Z = 0$) and the blue line($Z = 1$) are very close, which supports our theory that our estimator is independent of the covariate. In other words, under different covariate levels, our estimators share a common distribution.

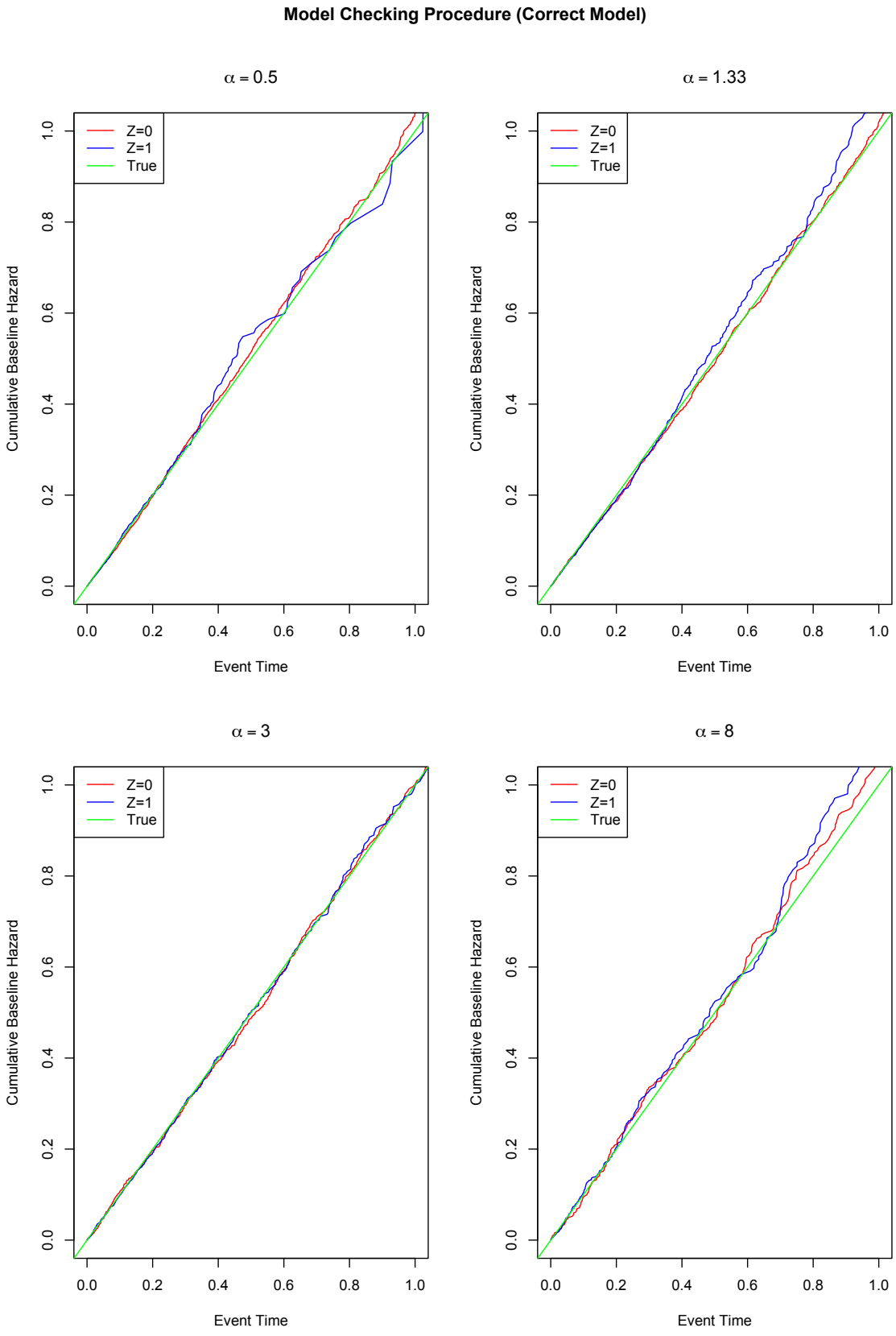

Figure 3.1 Model checking procedure.

More specifically, in our simulation, we used Clayton model to generate data, which implies that the frailty follows a Gamma distribution. When estimating the baseline cumulative hazard function, we used the Clayton copula generator in calculating our estimator, which is graphically consistent with the assumption. However, if we use the Gumbel copula generator for our estimator, the result, as shown in Figure 3.2, is clear that the two estimators are not consistent.

We ran 100 replications of the simulations above and compare the MSE of our estimator with Breslow estimator, as shown in Figure 3.3, we find that the two estimators are comparable in terms of mean square error. Although Breslow estimator slightly reduces the MSE, our estimator has some properties over the Breslow estimator that are very useful in practice.

For one thing, our estimator provides a model checking tool for the underlying frailty distribution. Sometimes we would like to know the distribution of these latent effects, so as to have a better understanding of our data. Then our approach provides a way into such concerns.

For another, as a semi-perimetric estimator, our estimator gives an explicit form of the baseline cumulative hazard function. The Breslow estimator uses EM algorithm to solve for the baseline cumulative hazard functions numerically, but does not have an analytical form as ours.

3.5 Discussion

In this project, we have established a formula for the baseline cumulative hazard functions in bivariate frailty models described in Oakes(1989)[14] and Manatunga (1999)[13]. We propose a new estimator of the baseline hazard functions based on our formula. From our simulation studies, we can see that our estimator is comparable with the Breslow estimator for this type of models. A clear advantage of our estimator is that it can be used to check the frailty model assumption or perform the frailty

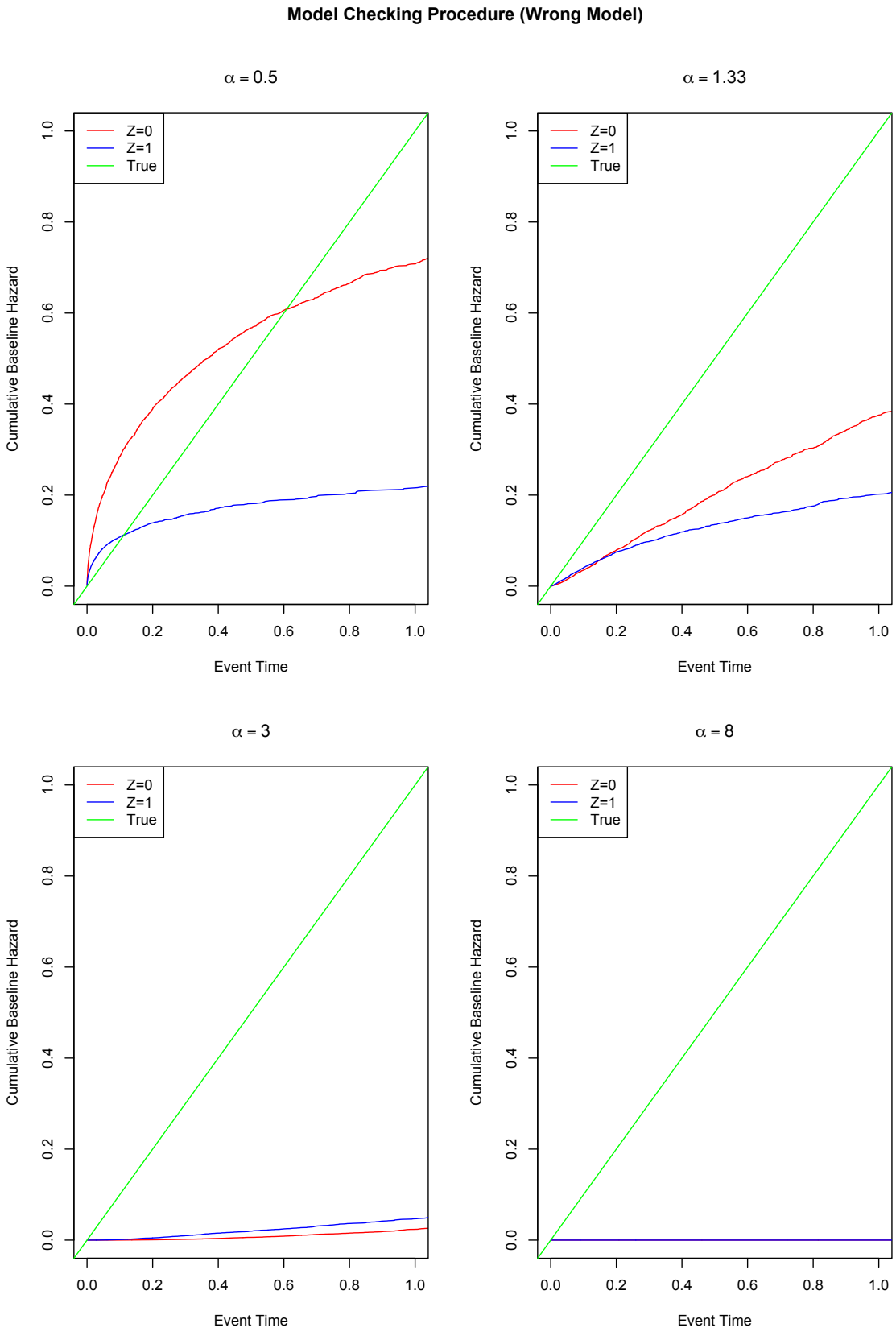

Figure 3.2 Model checking procedure (wrong model).

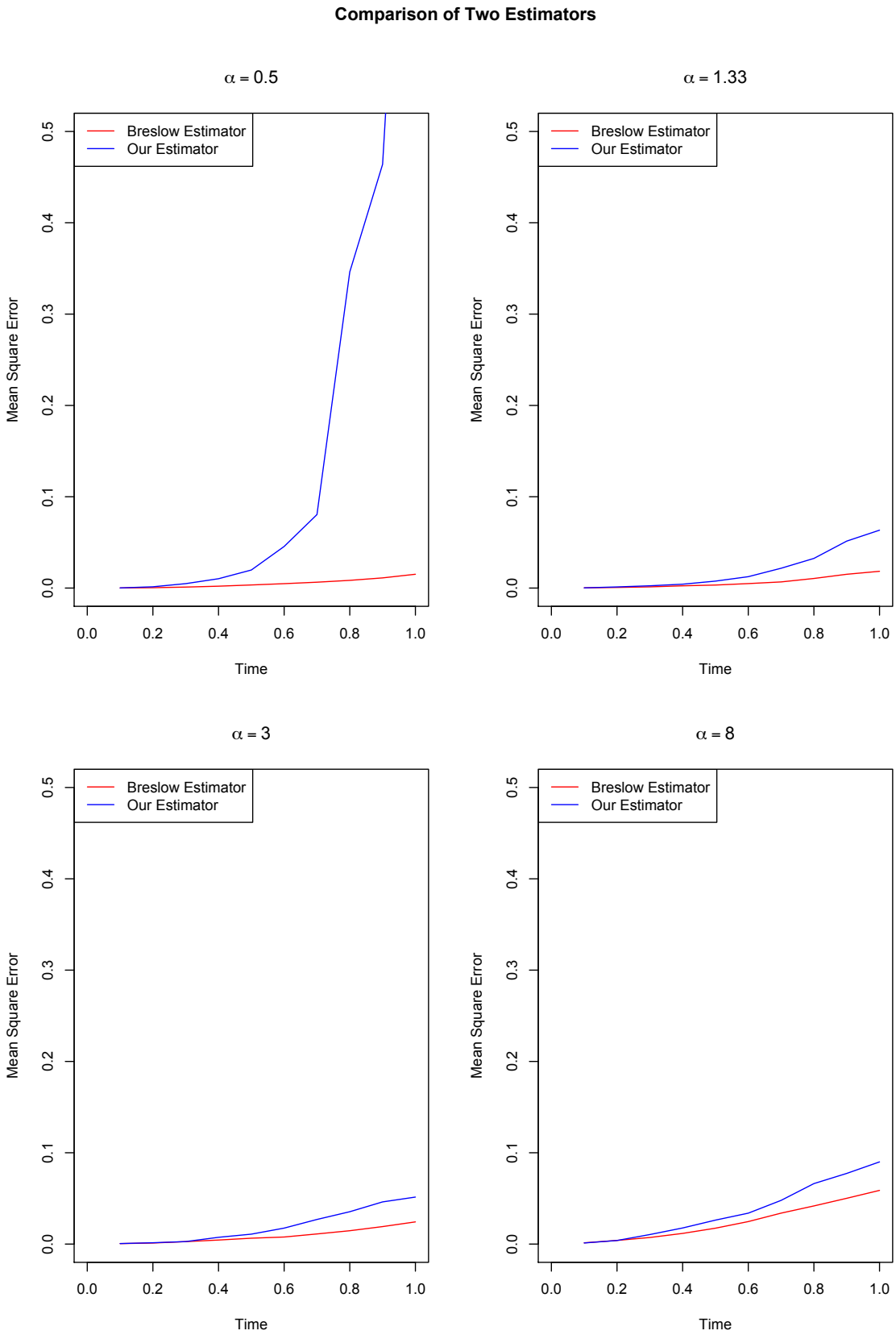

Figure 3.3 Comparison of two estimators

model selection. Because our estimator can be applied for groups of patients with different covariate values, subgroup analysis can be conducted using our proposed approach.

Although the estimator is proposed based on dependent censored data, the method can certainly be applied to multivariate failure time data if we assume that (T_1, T_2) follows our bivariate frailty model. In fact, dependent censored data contains less information than bivariate failure time data because both T_1 and T_2 are available in the latter case and it is easier for us to estimate the parameters in our model accordingly.

CHAPTER 4

LEFT CENSORED BIVARIATE DATA ANALYSIS

In this project, we study the properties of frailty models for bivariate data under fixed left censoring. It turns out that the distribution of observable pairs belongs to a new class of bivariate frailty models. Both the original model for complete data and the new model for observable pairs are members of Archimedean copula family. We propose a new estimation strategy to analyze left censored data using the corresponding Kendall's distribution. A general goodness-of-fit test procedure is then established for original models based on left censored data. Our strategies are generalization of the methodologies proposed in Wang(2007)[20], Romdhani(2011)[17] and Genest(2006)[**?**]. We demonstrate our new strategies using simulations and an illustrative example.

4.1 Properties of Frailty Models for Left Censored Bivariate Data

In this section, we assume that $(T_{11}, T_{21}), \ldots, (T_{1n}, T_{2n})$ are independent and identically distributed pairs which can be modeled by a bivariate frailty model such that:

$$\Lambda_{T_1}(t_1|W) = \Lambda_{T_{10}}(t_1)W$$

and

$$\Lambda_{T_2}(t_2|W) = \Lambda_{T_{20}}(t_2)W,$$

where W is the frailty whose distribution can be specified with unknown parameter θ. Denote the Laplace transform of W by $\psi(s) = E[\exp(-sW)]$ and the density function of W by $G_\theta(W)$. λ_{T_1}, λ_{T_2} and $\lambda_{T_{10}}$, $\lambda_{T_{20}}$ are defined as the hazard and baseline hazard functions for T_1 and T_2 respectively. The baseline cumulative hazards $\Lambda_{T_{10}}$ and $\Lambda_{T_{20}}$

satisfies

$$\Lambda_{T_{10}}(t_1) = \int_0^{t_1} \lambda_{T_{10}}(u)du < \infty \quad \text{and} \quad \Lambda_{T_{20}}(t_2) = \int_0^{t_2} \lambda_{T_{20}}(u)du < \infty$$

for all $t_1 \in [0, \infty)$ and $t_2 \in [0, \infty)$. Similar to Chapter 2,

$$S(t_1, t_2) = \psi[\psi^{-1}\{S_{T_1}(t_1)\} + \psi^{-1}\{S_{T_2}(t_2)\}],$$

where ψ^{-1} is the inverse function of ψ. Therefore (T_1, T_2) follows an Archimedean copula model with generator $\psi(s)$.

Suppose (T_1, T_2) is subject to fixed left censoring/truncation at (L_1, L_2), then the joint survival function of (T_1, T_2) given $T_1 > L_1$ and $T_2 > L_2$ is:

$$S(t_1, t_2|T_1 > L_1, T_2 > L_2) = \frac{\Pr(T_1 > t_1, T_2 > t_2)}{\Pr(T_1 > L_1, T_2 > L_2)}$$

$$= \frac{\psi[-\log(S_{T_{10}}(t_1)) - \log(S_{T_{20}}(t_2))]}{\psi[-\log(S_{T_{10}}(L_1)) - \log(S_{T_{20}}(L_2))]}$$

$$= \frac{\psi[-\log(S_{T_{10}}(t_1)/S_{T_{10}}(L_1)) - \log(S_{T_{20}}(t_2)/S_{T_{20}}(L_2)) - \log(S_{T_{10}}(L_1)) - \log(S_{T_{20}}(L_2))]}{\psi[-\log(S_{T_{10}}(L_1)) - \log(S_{T_{20}}(L_2))]}$$

$$= \frac{\psi(s + L)}{\psi(L)} = \psi^\star(s)$$

where

$$L = -\log(S_{T_{10}}(L_1)) - \log(S_{T_{20}}(L_2)) = \psi^{-1}(S(L_1, L_2))$$

is independent of t_1 and t_2, and

$$s = -\log(S_{T_{10}}(t_1)/S_{T_{10}}(L_1)) - \log(S_{T_{20}}(t_2)/S_{T_{20}}(L_2))$$

(see Manatunga(1996)[12]).

Based on above derivations, if we let $t_2 = L_2$, we have

$$S(t_1|T_1 > L_1, T_2 > L_2) = S(t_1, L_2|T_1 > L_1, T_2 > L_2) = \psi^\star(-\log(S_{T_{10}}(t_1)/S_{T_{10}}(L_1))),$$

therefore,

$$(\psi^\star)^{-1}[S(t_1|T_1 > L_1, T_2 > L_2)] = -\log(S_{T_{10}}(t_1)/S_{T_{10}}(L_1)).$$

Similarly, we can show that

$$(\psi^\star)^{-1}[S(t_2|T_1 > L_1, T_2 > L_2)] = -\log(S_{T_{20}}(t_2)/S_{T_{20}}(L_1)).$$

Combining above results, we can conclude that

$$S(t_1, t_2|T_1 > L_1, T_2 > L_2) = \psi^\star\{(\psi^\star)^{-1}[S(t_1|T_1 > L_1, T_2 > L_2)]+(\psi^\star)^{-1}[S(t_2|T_1 > L_1, T_2 > L_2)]\}.$$

Therefore, the conditional distribution of (T_1, T_2) given $T_1 > L_1$ and $T_2 > L_2$ still follows an Archimedean copula model with the copula generator $\psi^\star(s) = \psi(s + L)/\psi(L)$. It turns out that $\psi^\star(s)$ is the Laplace transform of the frailty W_1 that follows the distribution with density function:

$$f_{W_1}(w_1) = \frac{\exp(-Lw_1)dF(w_1)}{\psi(L)}$$

for $w_1 \in (0, \infty)$ where F is the distribution function of W_1. In summary, we have reached a similar conclusion as described in Manatunga(1996)[12].

Theorem 1: Suppose that (T_1, T_2) follows a bivariate frailty model such that:

$$\Lambda_{T1}(t_1|W) = \Lambda_{T_{10}}(t_1)W$$

and

$$\Lambda_{T2}(t_2|W) = \Lambda_{T_{20}}(t_2)W$$

where W is the frailty whose distribution can be specified with unknown parameter θ. Denote the Laplace transform of W by $\psi(s) = E[\exp(-sW)]$. Assume that (T_1, T_2) is subject to fixed left censoring/truncation with the censoring vector (L_1, L_2), then

$(T_1, T_2|T_1 > L_1, T_2 > L_2)$ follows an Archimedean copula model with generator

$$\psi^\star(s) = \frac{\psi(s+L)}{\psi(L)},$$

where

$$L = -\log(S_{T_{10}}(L_1)) - \log(S_{T_{20}}(L_2)) = \psi^{-1}(S(L_1, L_2))$$

and

$$s = -\log(S_{T_{10}}(t_1)/S_{T_{10}}(L_1)) - \log(S_{T_{20}}(t_2)/S_{T_{20}}(L_2))$$

Now we use an examples to illustrate this theorem.

Example 1: When the frailty W follows a Gamma distribution, $\psi(s) = (1+s)^{-1/\theta}$. Therefore,

$$\psi^\star(s) = \frac{\psi(s+L)}{\psi(L)} = (1 + s/(1+L))^{-1/\theta}.$$

The corresponding survival function for $S(t_1, t_2|T_1 > L_1, T_2 > L_2)$ is:

$$S(t_1, t_2|T_1 > L_1, T_2 > L_2) = \psi^\star(\{(\psi^\star)^{-1}[S(t_1|T_1 > L_1, T_2 > L_2)] + (\psi^\star)^{-1}[S(t_2|T_1 > L_1, T_2 > L_2)]\}$$

$$= \left\{ \frac{1}{S(t_1|T_1 > L_1, T_2 > L_2)^{-\theta} + S(t_2|T_1 > L_1, T_2 > L_2)^{-\theta} - 1} \right\}^{1/\theta}$$

which has the same form as the original Clayton copula. This result basically shows that if the original data follows the Clayton copula with parameter θ, the uncensored/untruncated data also follows the Clayton model with the same parameter value θ. This is the invariance property of the Clayton copula under left censoring/truncation shown in Oakes(2005)[15].

4.2 Parameter Estimation

For left censored bivariate data, because the uncensored pairs still follow the Archimedean copula models, we can directly apply the existing strategies to fit the Archimedean copula model based on completely observable pairs. However, their estimation procedure tends to be quite complicated and the performance of their estimators is quite unstable based on our simulation studies. In this section, we propose an alternative estimation approach based on the new frailty distribution derived in the previous section. We have:

Theorem 2: Suppose that (T_1, T_2) are defined as the previous section, then the random variables

$$V = S(T_1, T_2 | T_1 > L_1, T_2 > L_2)$$

and

$$U = \psi^{\star -1}(S_1(T_1 | T_1 > L_1, T_2 > L_2))/\psi^{\star -1}(S(T_1, T_2 | T_1 > L_1, T_2 > L_2))$$

are independently distributed with the Kendall distribution and Uniform$(0, 1)$ distribution respectively. Moreover, the Kendall distribution function of V can be written as:

$$K^{\star}(v) = v - \psi^{\star -1}(v)/\psi^{\star -1\prime}(v) = v - [\psi^{-1}(vv^{\star}) - \psi^{-1}(v^{\star})]/(\psi^{-1\prime}(vv^{\star})v^{\star})$$

for $v \in (0, 1)$ where $v^{\star} = S(L_1, L_2)$.

Based on Theorem 2, The log-likelihood function of $S(T_1, T_2 | T_1 > L_1, T_2 > L_2)$ can be written as:

$$l = \sum_{i:T_{1i} > L_1, T_{2i} > L_2} \log(k(V_i))$$

$$= \sum_{i:T_{1i} > L_1, T_{2i} > L_2} \left[\log(\psi^{-1}(V_i V^{\star}) - \psi^{-1}(V^{\star})) + \log(\psi^{-1\prime\prime}(V_i V^{\star})) - 2\log(-\psi^{-1\prime}(V_i V^{\star}))\right].$$

To estimate the unknown parameter θ in frailty distribution, we first replace $V^\star$ and V_i's by corresponding empirical estimates

$$\hat{V}^\star = \frac{\#\{T_{1i} > L_1, T_{2i} > L_2\}}{n} \quad \text{and} \quad \hat{V}_i = \frac{\#\{T_{1j} > \max\{T_{1i}, L_1\}, T_{2j} > \max\{T_{2i}, L_2\}\}}{n}$$

to establish our estimating equation:

$$\frac{1}{n}\frac{\partial l}{\partial \theta}(\hat{\theta}) = \frac{1}{n}\sum_{i:T_{1i}>L_1, T_{2i}>L_2} \frac{\partial l_i}{\partial \theta}(\hat{V}^\star, \hat{V}_i, \hat{\theta}) = 0.$$

Using the Taylor expansion, we have

$$\frac{1}{n}\frac{\partial l}{\partial \theta}(\hat{\theta}) \approx \frac{1}{n}\frac{\partial l}{\partial \theta}(\theta) + \frac{1}{n}\frac{\partial^2 l}{\partial \theta^2}(\theta)(\hat{\theta} - \theta).$$

It then follows from above equation that

$$n^{1/2}(\hat{\theta} - \theta) \approx n^{1/2}\frac{\{(1/n)\partial l/\partial\theta(\theta)\}}{\{(1/n)\partial^2 l/\partial\theta^2(\theta)\}}.$$

We can actually show

Theorem 3: Under necessary regularity conditions, our parameter estimator $\hat{\theta}_n$ is consistent and $n^{1/2}(\hat{\theta}_n - \theta)$ is asymptotically normal with zero mean and variance σ^2.

4.3 Simulation Results

In this section, we conduct simulation studies to demonstrate our estimation and test procedures. We generate bivariate data (T_1, T_2) with standard exponential marginal distributions from the Hougaard model corresponding to different dependence levels (measured by Kendall's τ values: $\tau = 0.2, 0.4, 0.6$ and 0.8). (T_1, T_2) is also subject to fixed left censoring with the detection limits $L_1 = L_2 = 0.1$ (i.e., we can observe (T_1, T_2) only if $T_1 > L_1, T_2 > L_2$). Then we apply two estimation strategies: our proposed strategy based on the Kendall distribution function and the strategy

Table 4.1 Estimator Comparison

	Sample Size $n = 100$		**Sample Size** $n = 200$	
τ	$\hat{\theta}$	$\tilde{\theta}$	$\hat{\theta}$	$\tilde{\theta}$
0.2	0.8687	0.9149	0.9132	0.9464
	(0.0204)	(0.0258)	(0.0226)	(0.0276)
0.4	0.6260	0.6535	0.6799	0.6651
	(0.0109)	(0.0169)	(0.0068)	(0.0334)
0.6	0.3894	0.3456	0.4017	0.3738
	(0.0034)	(0.0258)	(0.0016)	(0.0195)
0.8	0.2008	0.0896	0.1971	0.1602
	(0.0104)	(0.0206)	(0.0005)	(0.0081)

proposed by Genest(1995)[7] and Shih(1995)[18] to fit assumed frailty models (they are all Archimedean copula models. The results are presented in Table 4.1. We compare our estimator$\hat{\theta}$ and the traditional estimator $\tilde{\theta}$. For each τ, the first row is the estimate mean and the second row represents the mean square error. It turns out that our estimator has smaller bias and MSE.

4.4 Discussion

Oakes(2005)[15] has shown that the Clayton model is invariant under under left truncation. In this paper, we have shown that the same fact holds for Archimedean copula models. The main difference between the frailty model for original data and the frailty model for uncensored data lies in the corresponding frailty distributions (i.e., corresponding copula generators). Based on above fact, we propose a new parameter estimator when the bivariate data is under fixed left truncation or censoring. From our simulation study results, we can see that our estimator is less biased and more efficient than the popular estimator proposed by Shih(1995)[18] under the Hougaard

model assumption. Under the Clayton model assumption, the performances of two estimators are similar(in a simulation study not presented in this paper).

BIBLIOGRAPHY

[1] D. 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, Apr. 1978.

[2] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[4] J.P. Fine, H. Jiang, and R. Chappell. On semi-competing risks data. *Biometrika*, 88(4):907–919, Dec. 2001.

[5] T.R. Fleming and D.P. Harrington. *Counting Processes and Survival Analysis*. Hoboken, NJ, Wiley, 2005.

[6] C. Genest. Frank's family of bivariate distributions. *Biometrika*, 74(3):549–555, 1987.

[7] C. Genest, K. Ghoudi, and LP. Rivest. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552, Sep. 1995.

[8] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81–93, Jun. 1938.

[9] J. P. Klein and M. L. Moeschberger. *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, 2003.

[10] L. Lakhal, L. P. Rivest, and B. Abdous. Estimating survival and association in a semicompeting risks model. *Biometrics*, 64(1):180–188, Mar. 2008.

[11] D.Y. Lin. On the breslow estimator. *Lifetime Data Analysis*, 13(4):471–480, 2007.

[12] A. K. Manatunga and D. Oakes. A measure of association for bivariate frailty distributions. *Journal of Multivariate Analysis*, 56(1):60–74, Jan. 1996.

[13] A. K. Manatunga and D. Oakes. Parametric analysis for matched pair survival data. *Lifetime Data Analysis*, 5(4):371–387, Dec. 1999.

[14] D. Oakes. Bivariate survival models induced by frailty. *Journal of the American Statistical Association*, 84(406):487–493, Jun. 1989.

[15] D. Oakes. On the preservation of copula structure under truncation. *Canadian Journal of Statistics*, 33(3):465–468, Sep. 2005.

[16] LP. Rivest and M. T. Wells. A martingale approach to the copula-graphic estimator for the survival function under dependent censoring. *Journal of Multivariate Analysis*, 79(1):138–155, Oct. 2001.

[17] H. Romdhani and L. Lakhal-Chaieb. On the association between variables with lower detection limits. *Statistics in Medicine*, 30(26):3137–3148, Nov. 2011.

[18] J. H. Shih and T. A. Louis. Inferences on the association parameter in copula models for bivariate survival data. *Biometrics*, 51(4):1384–1399, Dec. 1995.

[19] A.W. van der Vaart. *Asymptotic Statistics*. New York, NY, Cambridge, 2007.

[20] A. Wang. The analysis of bivariate truncated data using the clayton copula model. *The International Journal of Biostatistics*, 3(1), 2007.

[21] A. Wang. On the nonidentifiability property of archimedean copula models under dependent censoring. *Statistics and Probability Letters*, 82(3):621–625, Mar. 2012.

[22] A. Wang. Properties of the marginal survival functions for dependent censored data under an assumed archimedean copula. *Journal of Multivariate Analysis*, 129(3):57–68, Aug. 2014.

[23] A. Wang et al. The identifiability of dependent competing risks models induced by bivariate frailty models. *Scandinavian Journal of Statistics*, 42(2):427–437, Jun. 2015.