NONPARAMETRIC ANALYSIS OF SEMI-COMPETING RISKS DATA

Jing Li

Submitted to the faculty of the University Graduate School

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in the Department of Biostatistics,

Indiana University

April 2020

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Giorgos Bakoyannis, Ph.D., Co-Chair

Ying Zhang, Ph.D., Co-Chair

Doctoral Committee

Sujuan Gao, Ph.D.

November 18, 2019

Yiqing Song, MD, Sc.D.

Chi Zhang, Ph.D.

DEDICATION

I dedicate this work to my advisors, my family and friends, who believed in me, helped me, and stood by me all along this journey.

ACKNOWLEDGMENTS

Throughout my Ph.D. studies I have received tremendous support and assistance from many people. First and foremost, I would like to express my deepest gratitude to my advisors Dr. Ying Zhang and Dr. Giorgos Bakoyannis for their continued support, valuable guidance, and great patience on my dissertation. Their profound understanding and keen insight on research topics was a guiding light on my studies. Over the years, they encouraged me to challenge myself and to improve my research and presentation skills. I am eternally grateful to have both of them as my mentors. Without their guidance, I would not be able to finish this dissertation. Their enthusiasm and commitment to research inspired me greatly to continue conducting quality research in the future. I am also sincerely thankful for Dr. Ying Zhang for always being very supportive, cultivating my professional and career growth, and motivating me through the challenges during my Ph.D. journey.

Dr. Yiqing Song is my graduate research assistant advisor, and he has kindly provided financial support to me the past three years. I am enormously grateful to have the opportunity to work on various epidemiological projects and attend local and national conferences to present our findings. My special thanks go to Dr. Sujuan Gao for being my doctoral committee member, for her constructive suggestions on the manuscript and for the Indianapolis-Ibadan Dementia Project dataset. I am also thankful to Dr. Chi Zhang for being my Ph.D. minor advisor and serving on my doctoral committee.

Getting my Ph.D. has always been a dream of mine. I would not be able to achieve all this without the encouragement from many people. I want to voice my

appreciation to The Department of Biostatistics all the faculty members, staffs, and fellow graduate students for their support over the years. I am overwhelmed with gratitude to Dr. Barry Katz, the Department Chair, and Dr. Ying Zhang, the Education Director for establishing a solid Ph.D. program, good research environment, and providing teaching/research assistant opportunities and travel grants to graduate students. All the training I received at the Indiana University will be beneficial to my future research and career. I also want to extend my heartfelt gratitude to Dr. Wanzhu Tu, Dr. Hao Liu, and Dr. Ziyue Liu for their generous help and guidance. Last but not least, I am very grateful to my family and friends for their sympathetic ears and always being there for me.

Jing Li

NONPARAMETRIC ANALYSIS OF SEMI-COMPETING RISKS DATA

It is generally of interest to explore if the risk of death would be modified by medical conditions (e.g., illness) that have occurred prior. This situation gives rise to semi-competing risks data, which are a mixture of competing risks and progressive state data. This type of data occurs when a non-terminal event can be censored by a well-defined terminal event, but not vice versa.

In the first part of this dissertation, the shared gamma-frailty conditional Markov model (GFCMM) is adopted because it bridges the copula models and illness-death models. Maximum likelihood estimation methodology has been proposed in the literature. However, we found through numerical experiments that the unrestricted model sometimes yields nonparametric biased estimation. Hence a practical guideline is provided for using the GFCMM that includes (i) a score test to assess whether the restricted model, which does not exhibit estimation problems, is reasonable under a proportional hazards assumption, and (ii) a graphical illustration to evaluate whether the unrestricted model yields nonparametric estimation with substantial bias for cases where the test provides a statistical significant result against the restricted model. However, the scientific question of interest that whether the status of non-terminal event alters the risk to terminal event can only be partially addressed based on the aforementioned approach. Therefore in the second part of this dissertation, we adopt a Markov illness-death model, whose transition intensities are essentially equivalent to the marginal hazards defined in GFCMM, but with different interpretations; we de-

velop three nonparametric tests, including a linear test, a Kolmogorov-Smirnov-type test, and a L_2-distance-type test, to directly compare the two transition intensities under consideration. The asymptotic properties of the proposed test statistics are established using empirical process theory. The performance of these tests in finite samples is numerically evaluated through extensive simulation studies. All three tests provide similar power levels with non-crossing curves of cumulative transition intensities, while the linear test is suboptimal when the curves cross. Eventually, the proposed tests successfully address the scientific question of interest. This research is applied to Indianapolis-Ibadan Dementia Project (IIDP) to explore whether dementia occurrence changes mortality risk.

Giorgos Bakoyannis, Ph.D., Co-Chair

Ying Zhang, Ph.D., Co-Chair

TABLE OF CONTENTS

LIST OF TABLES xii

LIST OF FIGURES xiv

CHAPTER 1 Introduction 1

CHAPTER 2 Semi-Competing Risks Data 7

2.1 Introduction 7

2.2 Illness-Death Model 9

2.3 Frailty Models 13

2.3.1 Univariate Frailty Model 14

2.3.2 Shared Frailty Model 16

2.3.3 Gamma Frailty and Shared Gamma Frailty Models 18

2.4 Gamma-Frailty Conditional Markov Model (GFCMM) 20

2.4.1 Model Formulation 20

2.4.2 Likelihood Construction 26

2.5 Markov Illness-Death Model 30

2.5.1 Model Formulation 30

2.5.2 Likelihood Construction 32

2.6 Relationship between GFCMM and Markov Illness-Death Model . 36

2.7 Concluding Remarks 37

CHAPTER 3 Asymptotic Theories 39

3.1 Introduction 39

3.2 Three Inference Procedures with Nuisance Parameters 40

3.2.1 Likelihood Ratio Test 44

3.2.2 Wald Test 45
3.2.3 Rao's Score Test 46
3.3 Empirical Process Theory (EPT) 47
3.3.1 Preliminary Notation and Definitions 47
3.3.2 Empirical Process, Glivenko-Cantelli and Donsker Results 50
3.3.3 Multiplier Central Limit Theorem (MCLT) 60
3.4 Concluding Remarks 62
CHAPTER 4 Nonparametric Maximum Likelihood Estimation (NPMLE) for Semi-Competing Risks Data under GFCMM 64
4.1 Introduction 64
4.2 NPMLE for Semi-Competing Risks Data 66
4.2.1 NPMLE for GFCMM 66
4.2.2 Pitfalls in NPMLE 71
4.2.3 A Practical Guideline for Using GFCMM 78
4.3 Case Study: Indianapolis-Ibadan Dementia Project Data 84
4.4 Discussion 91
CHAPTER 5 Nonparametric Tests for Semi-Competing Risks Data under Markov Illness-Death Model 94
5.1 Introduction 94
5.2 Nonparametric Tests of Transition Intensities 97
5.2.1 Linear Nonparametric Test 100
5.2.2 Kolmogorov-Smirnov-Type Nonparametric Test 107
5.2.3 L_2-Distance-Type Nonparametric Test 112

5.3 Simulation Studies 114

5.4 Case Study: Indianapolis-Ibadan Dementia Project Data 122

5.5 Discussion 125

CHAPTER 6 Summary 128

BIBLIOGRAPHY 131

CURRICULUM VITAE

LIST OF TABLES

4.1 Simulation results of the score test for H_0: $\beta = 0$ 80

5.1 Simulation results of empirical type I error for the linear (Linear), Kolmogorov-Smirnov-type (KS), and L_2-distance-type (L_2) tests under the null hypothesis $H_0 : \alpha_{13}(t) = \alpha_{23}(t), t \in [\tau_1, \tau_2]$ under scenarios I and II based on 2,000 runs. For unweighted test, $K(t) = 1$; for weighted test, $\hat{K}(t) = \frac{\bar{Y}_1(t)\bar{Y}_2(t)}{\bar{Y}_1(t)+\bar{Y}_2(t)}$. Four sets of τ_1 and τ_2 are considered. For the first three scenarios, τ_1 is chosen to be the analytic 5 (or 10) percentile of illness time, and τ_2 was the smaller analytic 90 (or 95) percentile between death with and without illness times; for the last scenario, τ_1 and τ_2 are data dependent and form a non-fixed interval, which are the smallest and largest times such that $Y_1(\tau_2)$, $Y_2(\tau_1)$, and $Y_2(\tau_2)$ are all non-zero for each simulated dataset. 119

5.2 Simulation results of empirical power for the linear (Linear), Kolmogorov-Smirnov-type (KS), and L_2-distance-type (L_2) tests under the null hypothesis $H_0 : \alpha_{13}(t) = \alpha_{23}(t), t \in [\tau_1, \tau_2]$ under scenarios III and IV based on 2,000 runs. For unweighted test, $K = 1$; for weighted test, $\hat{K}(t) = \frac{\bar{Y}_1(t)\bar{Y}_2(t)}{\bar{Y}_1(t)+\bar{Y}_2(t)}$. Four sets of τ_1 and τ_2 are considered. For the first three scenarios, τ_1 is chosen to be the analytic 5 (or 10) percentile of illness time, and τ_2 was the smaller analytic 90 (or 95) percentile between death with and without illness times; for the last scenario, τ_1 and τ_2 are data dependent and form a non-fixed interval, which are the smallest and largest times such that $Y_1(\tau_2)$, $Y_2(\tau_1)$, and $Y_2(\tau_2)$ are all non-zero for each simulated dataset. 120

5.3 Simulation results of empirical power for the linear (Linear), Kolmogorov-Smirnov-type (KS), and L_2-distance-type (L_2) tests under the null hypothesis $H_0 : \alpha_{13}(t) = \alpha_{23}(t), t \in [\tau_1, \tau_2]$ under scenarios V and VI based on 2,000 runs. For unweighted test, $K(t) = 1$; for weighted test, $\hat{K}(t) = \frac{\bar{Y}_1(t)\bar{Y}_2(t)}{\bar{Y}_1(t)+\bar{Y}_2(t)}$. Four sets of τ_1 and τ_2 are considered. For the first three scenarios, τ_1 is chosen to be the analytic 5 (or 10) percentile of illness time, and τ_2 was the smaller analytic 90 (or 95) percentile between death with and without illness times; for the last scenario, τ_1 and τ_2 are data dependent and form a non-fixed interval, which are the smallest and largest times such that $Y_1(\tau_2)$, $Y_2(\tau_1)$, and $Y_2(\tau_2)$ are all non-zero for each simulated dataset. 121

LIST OF FIGURES

2.1 Progressive illness-death model. 11

2.2 Illustration of four data observation scenarios of semi-competing risks data. A dot means both event times T_1 and T_2 are observed. An arrow in the direction of Y_i means T_i is censored for that subject, $i = 1, 2$. 12

2.3 Comparison of (a) competing risks data, (b) semi-competing risks data, and (c) bivariate survival data. A dot means both event times T_1 and T_2 are observed. An arrow in the direction of Y_i means T_i is censored for that subject, $i = 1, 2$. Data are subject to right censoring only. . . 13

2.4 Progressive illness-death model with hazard functions. 21

2.5 Progressive illness-death model with transition intensity functions. . 33

4.1 Profile likelihood for simulated semi-competing risks data under the restricted models. Sample sizes increase from $n = 200$ (left), to $n = 400$ (middle), $n = 800$ (right) for each scenario. 73

4.2 (a): Profile likelihood based on a single data set; (b) and (c): Average of NPMLEs of the conditional cumulative hazards (dashed lines based on 500 repetitions), $\hat{\Lambda}_{01}(\cdot)$ and $\hat{\Lambda}_{02}(\cdot)$, respectively, for the simulated semi-competing risks data under the restricted model with $\lambda_{01} = 2$, $\lambda_{02} = 1$, and $n = 400$. 74

4.3 Profile likelihood for simulated semi-competing risks data under the unrestricted models. Sample sizes increase from $n = 200$ (left), to $n = 400$ (middle), $n = 800$ (right) for each scenario. 75

4.4 (a): Profile likelihood based on a single data set; (b), (c) and (d): Average of NPMLEs of the conditional cumulative hazards (dashed lines based on 500 repetitions), $\hat{\Lambda}_{01}(\cdot)$, $\hat{\Lambda}_{02}(\cdot)$, and $\hat{\Lambda}_{03}(\cdot)$, respectively, for simulated semi-competing risks data under the unrestricted model with $\lambda_{01} = 2$, $\lambda_{02} = 0.5$, $\lambda_{03} = 1.5$, and $n = 400$. 76

4.5 (a): Profile likelihood based on a single data set; (b), (c) and (d): Average of NPMLEs of the conditional hazards (dashed lines based on 500 repetitions), $\hat{\Lambda}_{01}(\cdot)$, $\hat{\Lambda}_{02}(\cdot)$, and $\hat{\Lambda}_{03}(\cdot)$, respectively, for simulated semi-competing risks data under the unrestricted model with $\lambda_{01} = 2$, $\lambda_{02} = 0.5$, $\lambda_{03} = 1$, and $n = 800$. 77

4.6 Histograms of the test statistic $\hat{u}(0)$ under the null hypothesis. 82

4.7 A practical guideline for the use of GFCMM in analysis of semi-competing risks data. 83

4.8 Profile likelihood of θ for dementia data. 87

4.9 Case Study-IIDP: (a) Survival functions for dementia $-$ $\hat{S}_1(\cdot)$; (b) Conditional survival function for death without dementia given alive at age of 75 $-$ $\hat{S}_2(\cdot|75)$, and survival function for death with dementia diagnosed at age of 75 $-$ $\hat{S}_3(\cdot|t_1 = 75)$. 88

4.10 Case Study-IIDP: (a) Naive Kaplan-Meier curve for the non-dementia group $-$ $\widehat{NS}_2(\cdot)$, and survival function for death without dementia $-$ $\hat{S}_2(\cdot)$; (b) Naive Kaplan-Meier curve for the dementia group $-$ $\widehat{NS}_3(\cdot)$, and survival functions for death with dementia diagnosed at age of 75 and 80 $-$ $\hat{S}_3(\cdot|t_1 = 75)$ and $\hat{S}_3(\cdot|t_1 = 80)$. 90

5.1 At-state processes $Y_1(t)$ for the transition from the healthy state and $Y_2(t)$ for the transition from the illness state of a simulated dataset ($n = 200$). 100

5.2 Scenarios I - VI of cumulative transition intensities $A_{13}(t)$ and $A_{23}(t)$ for data simulation. 118

5.3 At-state processes $Y_1(t)$ and $Y_2(t)$ in IIDP data. (a) a full-scale plot; (b) a zoomed-in plot for the grey area in (a). 124

5.4 Cumulative transition intensities $\hat{A}_{13,\tau_1}(t)$ and $\hat{A}_{23,\tau_1}(t)$, $t \in [\tau_1, \tau_2]$, where $\tau_1 = 70$ and $\tau_2 = 104$ in the IIDP data. 125

Chapter 1

Introduction

In biomedical research often times it is of interest to investigate if at any give age, the risk of death would be modified by medical conditions (e.g., illness) that have occurred prior to this age. This situation gives rise to semi-competing risks data. Semi-competing risks data are a mixture of competing risks data and progressive state data. This type of data occurs when a non-terminal event (e.g., onset of illness) can be censored by a well-defined terminal event (e.g., death), but not vice versa. Compared to competing risks data, in which the occurrence of one event censors the rest potential events, semi-competing risks data contain more information which allows to study the transition from the non-terminal event to the terminal event.

There is an increasing interest in semi-competing risks data, especially in biomedical studies, when modeling both time to disease onset and time to mortality is of interest. When semi-competing risks data were first termed by Fine et al. (2001), Clayton copula (Nelsen, 2007) was proposed to study the joint survival distribution and the association between two latent event times in the upper wedge. Since semi-competing risks data are essentially of the same structure as illness-death data within multi-state model framework, the abundant methodology of multi-state modeling (Andersen and Keiding, 2002; Beyersmann et al., 2011; Commenges, 1999; Cook and Lawless, 2018; Hougaard, 1999; Klein et al., 2016; Meira-Machado et al., 2009; Van Den Hout, 2016) can be applied to analyze semi-competing risks data, where Markov illness-death models received much attention.

While the Markov model appears to be widely accepted for illness-death modeling (Frydman, 1995; Touraine et al., 2016), Xu et al. (2010) proposed a shared gamma-frailty conditional Markov model (GFCMM) to analyze semi-competing risks data. It was pointed out that the restricted GFCMM, in which the hazard of death conditional on the latent gamma frailty variable is assumed the same whether or not an individual experiences the illness under study before death, is essentially the Clayton copula model proposed by Fine et al. (2001). This finding bridges the methodology of Markov modeling for illness-death data with copula modeling for semi-competing risks data. Xu et al. (2010) naturally extended the GFCMM to the unrestricted case for more flexibility in modeling semi-competing risks data compared to copula models. Xu et al. (2010) proposed a Newton-Raphson algorithm for non-/semi-parametric maximum likelihood estimation for restricted and unrestricted GFCMM. Bayesian methods for the frailty-based Markov model have also been extensively adopted recently in analyzing semi-competing risks data, see for example, Chapple et al. (2017), Han et al. (2014), and Lee et al. (2016, 2015).

The motivation for this research came from the Indianapolis-Ibadan Dementia Project (IIDP), which is a 20-year National Institute on Aging-funded longitudinal study of dementia and its risk factors in elderly African Americans living in Indianapolis, Indiana and elderly Yoruba residing in Ibadan, Nigeria (Gao et al., 2016; Hendrie et al., 2001, 2017). The observations can be well characterized by semi-competing risks data framework, in which the diagnosis of dementia is considered as a non-terminal event and death as the terminal event. In this research, it is aimed to develop methodology to answer the scientific question of interest that whether the

occurrence of non-terminal event changes the risk to terminal event, which is whether the onset of dementia increases mortality risk specifically for IIDP.

The first part of this dissertation focuses on the nonparametric maximum likelihood estimation (NPMLE) of the hazards under the frequentist framework. The GFCMM is adopted, and an EM algorithm is developed to compute the NPMLE of the hazards for semi-competing risks data. The proposed EM algorithm is more numerically stable than the Newton-Raphson algorithm proposed by Xu et al. (2010) especially for large samples. Through simulation studies, we uncover the issue in NPMLE for the unrestricted GFCMM that the maximizer may occur at the boundary of feasible parameter space, and this results in biased estimates. Thus although the unrestricted GFCMM for semi-competing risks data has merit of flexibility in characterizing the association between the non-terminal and terminal events, it does not always yield consistent likelihood-based inference for model parameters. This numerical issue in NPMLE under the unrestricted GFCMM has not been addressed in any peer-reviewed publications. Hence a more thorough numerical experiment for analyzing semi-computing risks data under both of the restricted and unrestricted GFCMMs to examine the behavior of the NPMLEs is presented.

We alert researchers to be cautious about the unrestricted GFCMM for the nonparametric analysis of semi-competing risks data. Moreover, a practical guideline for the use of GFCMM is provided, which includes (i) a score test to assess whether the restricted model, which does not exhibit estimation problems, is reasonable under a proportional hazards assumption; (ii) a graphical illustration of the profile likelihood to determine whether the unrestricted model yields nonparametric estimation with substantial bias for cases when the score test provides a statistically significant result

against the restricted model. This score test might be the only valid likelihood-based inference procedure for this model given the numerical problem associated with the NPMLE of the unrestricted GFCMM. Finally, the proposed practical guideline is applied to the IIDP data. The resulted p-value from the proposed score test indicates nonsignificant results and suggests the use of restricted model, which implies the occurrence of dementia increases mortality risk as expected. However, if p-value indicates the use of unrestricted model, the scientific question of interest may not be directly answered based on the aforementioned approach due to the numerical issue with NPMLE under the unrestricted model. That is to say, the scientific research question can only be partially addressed based on the first part of this research.

Therefore the second part of this dissertation adopts the Markov illness-death model without engaging frailties, whose transition intensities are essentially equivalent to the defined marginal hazards (unconditional on frailty) for GFCMM but with different interpretations. Furthermore, we focus on the nonparametric testing procedures, which directly assesses the research question that whether the status of non-terminal event alters the risk to terminal event. With three-state illness-death model as a special case, multi-state modeling methodology has been widely used for the analysis of semi-competing risks data as discussed above. Since transition intensity has always been a popular research interest in biomedical applications, much research has been conducted on its nonparametric estimation within multi-state modeling framework (Andersen et al., 2012; Commenges, 2002; Commenges et al., 2004; Datta and Satten, 2001; Frydman, 1995; Frydman et al., 2013; Mau, 1986). However nonparametric testing on transition intensity functions in multi-state models has not received much attention except by Andersen et al. (2012) and Bluhmki et al. (2019).

Three nonparametric tests are developed, which include a linear test, a Kolmogorov-Smirnov-type test, and a L_2-distance-type test, by comparing two transition intensity functions under the Markov illness-death model in the same study sample. The linear test is based on the difference of area under the curve (AUC) between two cumulative transition intensities. The Kolmogorov-Smirnov-type test and a L_2-distance-type test compare the maximum distance and L_2-norm distance between them. The proposed tests are of great importance as they directly address the scientific question of interest. Under mild regularity conditions, the asymptotic distributions of three proposed test statistics under the null hypothesis are established using empirical process theory; the consistency of Kolmogorov-Smirnov-type and L_2-distance-type tests are shown. The performance of these tests in finite samples is numerically evaluated through simulation studies under various scenarios. All three tests provide similar empirical power levels under the scenarios with non-crossing cumulative transition intensity functions, while the Kolmogorov-Smirnov-type and L_2-distance-type tests are more powerful than the linear test for the comparison of crossing curves. The proposed tests can also be adopted for non-Markov processes. In the end, three proposed tests are applied to IIDP dataset. All three p-values provide sufficient evidence to conclude dementia increases the risk of mortality, which is consistent with the findings in the first part of research using GFCMM.

This dissertation is structured as follows. Chapter 2 introduces semi-competing risks data with some literature review. Starting with the well-known illness-death model and frailty models, we then study the GFCMM and the Markov illness-death model, which will be used in Chapter 4 and Chapter 5, respectively. Their model formulation and likelihood construction are specifically studied. The relationship

between these two models is also discussed. Chapter 3 introduces the asymptotic theories that will be utilized in Chapter 4 and Chapter 5. We first review the likelihood ratio test, Wald test, and Rao's score test with nuisance parameters, and their asymptotic distributions under the null hypothesis for parametric models. Then the empirical process theory a powerful tool for non-/semi-parametric inference is introduced. In Chapter 4, findings from the first part of this dissertation are presented. We adopt the GFCMM and focus on the nonparametric estimation of semi-competing risks data. Through simulation studies, the pitfalls in NPMLE with GFCMM is uncovered, and a practical guideline for the use of the GFCMM for nonparametric analysis of semi-competing risks data is provided. Since the research question cannot be fully addressed in Chapter 4, the second part of this dissertation (Chapter 5) adopts the Markov illness-death model and develop three nonparametric tests including a linear test, a Kolmogorov-Smirnov-type test, and a L_2-distance-type test to tackle the research question by directly assessing whether the status of non-terminal event alters the risk to terminal event. Chapter 6 summarizes this research.

Chapter 2

Semi-Competing Risks Data

2.1 Introduction

Semi-competing risks data often occur when there are two events of interest: a non-terminal event (e.g., onset of illness) and a terminal event (e.g., death), where the intermediate non-terminal event can be censored by the terminal event but not vice versa. Such a model is also known as the illness-death model (Cook and Lawless, 2018) under multi-state modeling framework, where Xu et al. (2010) argued that semi-competing risks are merely a half-century old illness-death model. Compared to conventional competing risks data, in which occurrence of one event censors the rest potential events of the study, semi-competing risks data contain more information that allows to study the transition from the non-terminal event to the terminal event. Additionally, semi-competing risks data provide information about the association between the two events in the observable region of their joint distribution, which is the region where the terminal event occurs after the non-terminal event.

There is an increasing interest in semi-competing risks data, especially in biomedical studies when modeling both time to disease progression and time to death is of interest. Fine et al. (2001) appears to be the first to refer illness-death data as semi-competing risks data. They proposed the use of Clayton copula (Nelsen, 2007) to study the joint survival distribution and the association between two latent event times in the upper wedge. After that, Wang (2003) extended the idea to other copula models. Peng and Fine (2007) furthermore proposed time-dependent parametric cop-

ula models for regression modeling under the proportional hazards assumption. Fu et al. (2013) proposed Gaussian copula model under a Bayesian inference framework to study the correlation structure and to predict the overall survival based on the progression free survival. Most recently, copula-based models were also developed for semi-competing risks data under more complex settings, such as those with missing terminal event status (Yu, 2016; Yu and Yiannoutsos, 2015), missing cause of informative terminal event (Zhou et al., 2017), and clustered survival data (Peng et al., 2018).

Since semi-competing risks data are essentially in the same structure as illness-death data within multi-state model framework, the abundant methodology of multi-state modeling based on transition intensities (Andersen and Keiding, 2002; Beyersmann et al., 2011; Commenges, 1999; Cook and Lawless, 2018; Hougaard, 1999; Klein et al., 2016; Meira-Machado et al., 2009; Van Den Hout, 2016) can be applied to analyze semi-competing risks data, where the Markov illness-death model has received much attention. Frydman (1995) developed the nonparametric maximum likelihood methodology for the Markov illness-death model with interval-censored data. Frydman et al. (2013) extended it to the scenario when all transition times are interval-censored. Frydman and Szarek (2009) studied the situation with potential missing status of intermediate transition, and Harezlak et al. (2003) proposed to use the Markov illness-death model for dealing with non-ignorable missingness in longitudinal data specifically for dementia studies.

As the Markov model appears to be widely accepted for illness-death modeling, Xu et al. (2010) proposed a shared gamma-frailty conditional Markov model (GFCMM) to analyze semi-competing risks data. They pointed out that the restricted GFCMM,

in which the hazard of death conditional on the latent gamma frailty variable is assumed to be the same regardless of whether an individual experiences the illness under study before death, is identical to the Clayton copula model in the upper wedge proposed by Fine et al. (2001). This finding bridges the Markov modeling methodology for illness-death data with the Copula modeling methodology for semi-competing risks data. After GFCMM was developed, bayesian methods for the GFCMM have also been extensively adopted recently in analyzing semi-competing risks data, see for example, Chapple et al. (2017), Han et al. (2014), and Lee et al. (2016, 2015).

In this dissertation, illness-death models especially GFCMM and Markov illness-death model are of interest. The remainder of this chapter is structured as follows. Section 2.2 and Section 2.3 introduce illness-death models and frailty models. In Section 2.4 and Section 2.5, the two widely used GFCMM and the Markov illness-death model, which are of interest for this thesis are studied. Their model formulation and likelihood construction are discussed, respectively. The relationship between these two models is presented in Section 2.6. Section 2.7 concludes this chapter with a summary.

2.2 Illness-Death Model

The illness-death multi-state model is widely used for modeling semi-competing risks data, which consists of two living states (e.g., healthy state and diseased state) and an absorbing state (e.g., death). Assume all subjects start with healthy state at time $t = 0$. If the non-terminal event (e.g., onset of illness) occurs, the person transitions from the healthy state to the diseased state. If the terminal event (e.g., death) occurs while the person is either in healthy or diseased state, he or she enters absorbing

state. In this work, the progressive illness-death model is considered (Figure 2.1), since it well describes the practical situation, and it is the main interest in many real data applications. The progressive model means subjects can only move forward to the next state. That is to say, once a subject experiences the non-terminal event and is in the diseased state, it cannot return to the healthy state.

Let T_1 and T_2 denote respectively the non-terminal and terminal event times of an individual under study. If an individual experiences the non-terminal event first, observation for both T_1 and T_2 is possible. However if an individual experiences the terminal event first, T_1 is censored by the terminal event time T_2. In addition, it is assumed that there exists a potential right censoring time C that is independent of and non-informative to the event times (T_1, T_2). In summary, a generic observation of semi-competing risks data can be denoted as $D = (Y_1, Y_2, \delta_1, \delta_2)$, where $Y_2 = T_2 \wedge C, \delta_2 = I(T_2 \leq C), Y_1 = T_1 \wedge Y_2$, and $\delta_1 = I(T_1 \leq Y_2)$. Assume a study consists of n independent and identically distributed (i.i.d) copies of D, $D_1, D_2, \cdots, D_n$ to form the observed data denoted by $\underline{D} = \{D_i : i = 1, 2, \cdots, n\}$.

Figure 2.2 presents four potential scenarios for observed data D, which are

(1) healthy $\rightarrow$ diseased $\rightarrow$ dead: $D = (T_1, T_2, 1, 1)$;

(2) healthy $\rightarrow$ healthy: $D = (C, C, 0, 0)$;

(3) healthy $\rightarrow$ dead: $D = (T_2, T_2, 0, 1)$;

(4) healthy $\rightarrow$ diseased: $D = (T_1, C, 1, 0)$.

A dot means both event times T_1 and T_2 are observed. An arrow in the direction of Y_i means T_i is censored for that subject, $i = 1, 2$. This observation of four data scenarios of semi-competing risks data facilitates the likelihood derivation under illness-death model. In addition, observation of (Y_1, Y_2) is restricted to the upper wedge ($0 <$

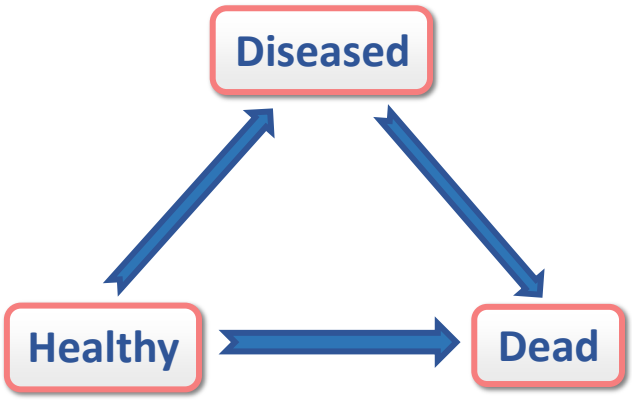


Figure 2.1: Progressive illness-death model.

$Y_1 \leq Y_2$) (shaded area in Figure 2.2) since the terminal event can never happen after the non-terminal event ($0 < T_1 \leq T_2$). The joint density or joint survival functions between T_1 and T_2 can be defined in this upper wedge with a balancing probability at $t_1 = \infty$ (Xu et al., 2010).

Adapted from Jiang et al. (2003), Figure 2.3 compares semi-competing risks data to competing risks data and bivariate survival data, and all are subject to right censoring. As it is just mentioned, semi-competing risks data can only be observed within the upper wedge (shaded area in Figure 2.3 (b)). Competing risks data can only be observed on the diagonal line (Figure 2.3 (a)) and contains the least information among three data types. The reason to this is both events can never happen at the same time; it is only possible to observe the occurrence of one competing event that happened earlier or both competing risks are censored. For bivariate survival data, observations can fill up the whole positive quadrant (shaded area in Figure 2.3 (c)), because there is no restriction on the order of two events like illness-death data. It is possible to observe either one or both events, and both events are also likely to be

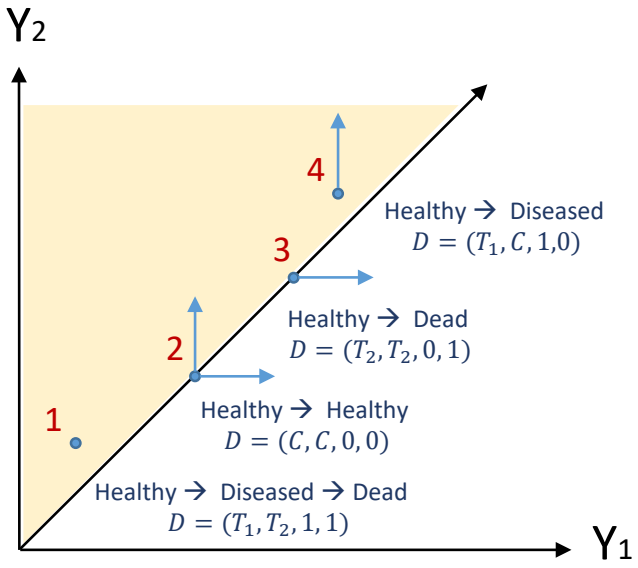
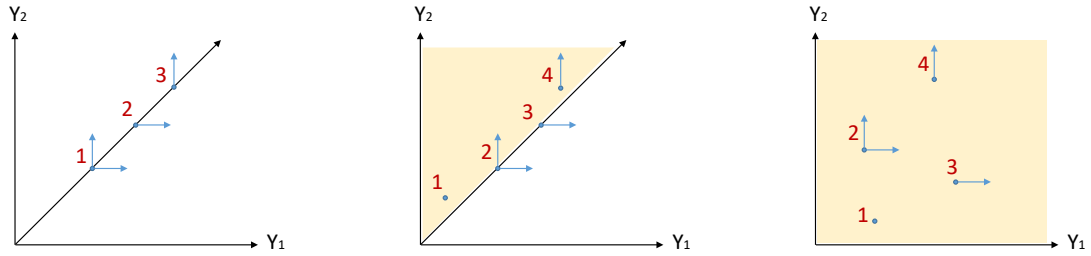


Figure 2.2: Illustration of four data observation scenarios of semi-competing risks data. A dot means both event times T_1 and T_2 are observed. An arrow in the direction of Y_i means T_i is censored for that subject, $i = 1, 2$.

censored. Hence bivariate life history data contain the most information among the three data types.



(a) (b) (c)

Figure 2.3: Comparison of (a) competing risks data, (b) semi-competing risks data, and (c) bivariate survival data. A dot means both event times T_1 and T_2 are observed. An arrow in the direction of Y_i means T_i is censored for that subject, $i = 1, 2$. Data are subject to right censoring only.

2.3 Frailty Models

Frailty term originates from the medical field gerontology, and it indicates an increased risk/susceptibility to morbidity or mortality (Gillick, 2001). If an individual is said to be more frail, he or she is at a higher risk to be exposed to adverse events. Although individuals may appear quite similar to each other with respect to some measurable attributes (e.g., age, race/ethnicity, gender), they are at different risks for morbidity and mortality or respond differently to treatment due to unmeasured attributes (e.g., genotype). Vaupel et al. (1979) introduced the term frailty to statistical literature to account for unobservable covariates and heterogeneity in the analysis of mortality rates of a study population.

Frailty can be modeled as a random effect in survival data analysis, which can be specified at the individual level or at a cluster level. In univariate frailty models, each individual can have his/her own frailty, which allows unobserved individual-specific risk for adverse events. In shared frailty models, frailties are considered at group level (e.g., families) to model the correlations between event times in a group. Family

members tend to be more similar to each other in clinical measures due to family genetics, living environment, and daily life habits. In addition, shared frailties are also useful in modeling multivariate survival data, where a common frailty is shared within each individual among times to different events (e.g., times to diabetes, stroke, and cancer). Shared frailty model is a special case for correlated frailty models.

Although there are more complex frailty models, e.g., nested frailty models, more general correlated frailty models, only two types of frailties models, univariate frailty models and shared frailty models are introduced in this section. There are several excellent books regarding frailty models that can be used for further reference (Duchateau and Janssen, 2007; Hanagal, 2011; Wienke, 2010).

2.3.1 Univariate Frailty Model

In basic survival data models, independent and identically distributed data are usually assumed; however in practice, study population is often heterogeneous, where people are at different levels of risks for adverse events or people respond differently to the treatment. Furthermore, it is often impossible to observe and collect all risk factors to be included in statistical models, and some times we are unable to provide a full list of risk factors. In this situation, frailties as random effects can be incorporated in the model to address the unobserved variability in the study population.

To incorporate the frailty term in the model, the classical way is to adopt a proportional hazards structure conditional on the frailty. That is, the hazard of an individual is multiplicatively related to the time-independent random effect variable (frailty γ), baseline hazard ($\lambda_0(t)$), and the parametric regression part of observed

covariates ($e^{X\beta}$),

$$\lambda(t|X,\gamma) = \gamma\lambda_0(t)e^{X\beta}, \tag{2.1}$$

where $X = (X_1, \cdots, X_k)$ are $n \times k$ dimensional covariates, $\beta = (\beta_1, \cdots, \beta_k)'$ are k dimensional regression parameters, and frailty γ is constant over time.

The frailty model is a generalization of the Cox Proportional Hazards (Cox PH) model. The corresponding survival function given frailty γ is

$$S(t|X,\gamma) = e^{-\int_0^t \lambda(s|X,\gamma)ds} = e^{-\gamma\int_0^t \lambda_0(s|X,\gamma)ds\cdot e^{X\beta}} = e^{-\gamma\Lambda_0(t)\cdot e^{X\beta}},$$

where $\Lambda_0(t) = \int_0^t \lambda_0(s|X,\gamma)ds$ is the cumulative baseline hazard function.

Due to the model identifiability issue, the mean of frailty distributions are usually assumed to be $E(\gamma) = 1$. The variance of the frailty ($Var(\gamma)$) describes the spread of the distribution and indicates how heterogeneous the study population is. With small $Var(\gamma)$, the values of frailty are close to 1; with large $Var(\gamma)$, frailties are more spread out indicating greater variability across population. It is also important to note this hazard model (2.1) is a conditional frailty model. All the interpretation on regression parameters should be made conditional on the frailty. There is an alternative modeling approach on unconditional or marginal hazard models, which has completely different interpretations on regression parameters. In marginal models, parameters describe the relative risk at the population level, whereas the parameters have a cluster-level interpretation in the conditional frailty model.

It is straightforward to derive the unconditional survival function through

$$S(t|X) = E(S(t|\gamma, X)) = E\left(e^{-\gamma\Lambda_0(t)\cdot e^{X\beta}}\right) = \int_0^\infty e^{-\gamma\cdot(\Lambda_0(t)e^{X\beta})} \cdot f(\gamma)d\gamma$$
$$= \mathcal{L}_\gamma[f(\gamma)](\Lambda_0(t)e^{X\beta}),$$

where $\mathcal{L}$ is Laplace transform and defined by

$$\mathcal{L}_t[f(t)](s) = \int_0^\infty f(t)e^{-st}dt, t \geq 0.$$

Theorem 2.3.1. (Vaupel et al., 1979; Wienke, 2010) Assume a frailty model given by $\lambda(t|\gamma) = \gamma\lambda_0(t)$, the population hazard $\lambda(t) = \frac{f(t)}{S(t)}$ is

$$\lambda(t) = E(\lambda(t|\gamma)|T > t) = \int_0^\infty \lambda(t|\gamma)f(z|T > t)d\gamma = \lambda_0(t)\int_0^\infty \gamma f(\gamma|T > t)d\gamma,$$

where $f(\gamma|T > t)$ represents the frailty density among the survivors of time point t.

This theorem is of great importance since it provides a way to derive marginal models from the conditional hazard models given frailty. Detailed proof can be found in the book by Wienke (2010).

2.3.2 Shared Frailty Model

The above univariate frailty model can be extended to multivariate time-to-event data, which takes into account of the association between event times in the same cluster or group by using the shared frailty. The individual hazard for the shared frailty model is the same as it for the univariate frailty model except one important difference that the frailty is shared by all individuals in the same cluster. It is generally

assumed that conditional on the shared frailty, the event times in the same cluster are independent. Hence for the i^{th} cluster consisting of n_i individuals, the joint survival frailty is

$$S(t_{i,1}, \cdots, t_{i,n_i}|X_i, \gamma_i) = S(t_1|X_{i,1}, \gamma_i) \cdots S(t_{n_i}|X_{i,n_i}, \gamma_i).$$

A few examples are, patients of a family doctor, students from a university, and family members. Shared frailty can also be used on multivariate time-to-event data, where an individual is at risk for developing multiple diseases (e.g., diabetes, stroke, and cancer). The shared frailty is a special case for correlated frailty models; see Wienke (2010) for further reference. Wienke (2010) and Duchateau and Janssen (2007) gave great details on shared frailty models.

It is crucial to note again with great emphasis that the concept of shared frailty is different from the univariate frailty. The shared frailty introduced here only captures the common component shared by individuals in the same group, while it does not take into account of the heterogeneity due to unmeansurable attributes across individuals in the same cluster.

Take bivariate survival data as an example. Let T_1 and T_2 be the two event times from the same individual (e.g., times to diabetes and cancer), which share the same frailty γ. For the i^{th} individual, the conditional bivariate survival function given frailty γ_i and covariates X_i for an individual (its own cluster) is

$$\begin{aligned} S(t_{i1}, t_{i2}|X_i, \gamma_i) &= S(t_{i1}|X_i\gamma_i)S(t_{i2}|X_i, \gamma_i) \\ &= e^{-\gamma_i\left(\Lambda_0(t_{i1})e^{X_i\beta_1} + \Lambda_0(t_{i2})e^{X_i\beta_2}\right)} \end{aligned}$$

where $\Lambda_0(t) = \int_0^t \lambda_0(s)ds$ denotes the cumulative baseline hazard function, and $\beta_i, i = 1, 2$ are regression parameters.

The marginal survival function (or unconditional joint survival function) is derived with averaging conditional survival with respect to the frailty γ_i :

$$\begin{aligned} S(t_{i1}, t_{i2}|X_i) &= E(S(t_{i1}, t_{i2}|X_i, \gamma_i)) \\ &= E\left(e^{-\gamma_i\left(\Lambda_0(t_{i1})e^{X_i\beta_1} + \Lambda_0(t_{i2})e^{X_i\beta_2}\right)}\right) \\ &= \mathcal{L}_{\gamma_i}[f(\gamma_i)]\left(\Lambda_0(t_{i1})e^{X_i\beta_1} + \Lambda_0(t_{i2})e^{X_i\beta_2}\right), \end{aligned}$$

where $\mathcal{L}$ denotes Laplace transform.

2.3.3 Gamma Frailty and Shared Gamma Frailty Models

There are many problems, where gamma, log-normal, Weibull, Poisson, positive stable and power variance function distributions are commonly used for modeling frailties in applications. This work focuses on the discussion of gamma frailty as it is used in the GFCMM. Details for the other frailty distributions can be found in Duchateau and Janssen (2007), Hanagal (2011), and Wienke (2010).

The gamma frailty has been one of the most popular frailties used in statistical applications due to its ease of computation and simple derivation of Laplace transform. The density function for $Gamma(\alpha, \beta)$ distribution is

$$f(\gamma|\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)}\gamma^{\alpha-1}e^{-\beta\gamma},$$

where $\alpha > 0$ is the shape parameter, and $\beta > 0$ is the rate parameter. Its mean is $E(\gamma) = \frac{\alpha}{\beta}$, and its variance is $Var(\gamma) = \frac{\alpha}{\beta^2}$. In order to avoid model non-identifiability issue with frailty models, it is always assumed that $E(\gamma) = 1$, which poses restriction on the scale and rate parameters ($\alpha = \beta$). Here assume $\alpha = \beta = \frac{1}{\theta}$, which makes $Var(\gamma) = \theta$.

With $\gamma \sim Gamma(\theta^{-1}, \theta^{-1})$, for the univariate case, the marginal survival function can be derived, and it is given by

$$S(t) = (1 + \theta(\Lambda_0(t)))^{-\frac{1}{\theta}}. \tag{2.2}$$

For the bivariate case, the survival function can be simplified as follows. For simplicity, the covariates term is omitted.

$$\begin{aligned} S(t_1, t_2) &= \mathcal{L}_\gamma[f(\gamma)]\left(\Lambda_0(t_1) + \Lambda_0(t_2)\right), t_1 > 0, t_2 > 0 \\ &= \frac{\theta^{\frac{1}{\theta}}}{\Gamma(\frac{1}{\theta})} \cdot \mathcal{L}_\gamma[\gamma^{\frac{1}{\theta}-1}]\left(\frac{1}{\theta} + \Lambda_0(t_1) + \Lambda_0(t_2)\right) \\ &= (1 + \theta(\Lambda_0(t_1) + \Lambda_0(t_2)))^{-\frac{1}{\theta}} \\ &= (S_1(t_1)^{-\theta} + S_2(t_2)^{-\theta} - 1)^{-\frac{1}{\theta}} \end{aligned}$$

That is,

$$S(t_1, t_2) = (S_1(t_1)^{-\theta} + S_2(t_2)^{-\theta} - 1)^{-\frac{1}{\theta}}, t_1 > 0, t_2 > 0, \tag{2.3}$$

which coincides with the Clayton copula model for bivariate random variables in the upper wedge.

2.4 Gamma-Frailty Conditional Markov Model (GFCMM)

2.4.1 Model Formulation

Again in this research, the progressive illness-death model is considered. Let T_1 and T_2 denote the non-terminal and terminal event times of a subject under study. Three hazards are defined as follows,

$$\lambda_1(t_1) = \lim_{\Delta \to 0} \frac{P(T_1 \in [t_1, t_1 + \Delta) | T_1 \geq t_1, T_2 \geq t_1)}{\Delta}, \quad t_1 > 0 \tag{2.4}$$

$$\lambda_2(t_2) = \lim_{\Delta \to 0} \frac{P(T_2 \in [t_2, t_2 + \Delta) | T_1 \geq t_2, T_2 \geq t_2)}{\Delta}, \quad t_2 > 0 \tag{2.5}$$

$$\lambda_3(t_2|t_1) = \lim_{\Delta \to 0} \frac{P(T_2 \in [t_2, t_2 + \Delta) | T_1 = t_1, T_2 \geq t_2)}{\Delta}, \quad t_2 \geq t_1 > 0, \tag{2.6}$$

where $\lambda_1(t_1)$ is the hazard of non-terminal event at time t_1, $\lambda_2(t_2)$ the hazard of terminal event at time t_2 without experiencing the non-terminal event, and $\lambda_3(t_2|t_1)$ the hazard of terminal event at time t_2 given the non-terminal event occurs at time t_1 ($t_2 \geq t_1$). The definition of these hazards of event times is along with the framework of survival data analysis. The progressive illness-death model with three defined hazards is depicted in Figure 2.4.

Remark 2.4.1. It is crucial to point out the hazards definitions of $\lambda_1(t_1)$ and $\lambda_2(t_2)$ (2.4)-(2.5) are equivalent to the definitions of cause-specific hazards under competing risks data framework. The **cause-specific hazard** is defined as

$$\lambda_j(t) = \lim_{\Delta \to 0} \frac{P(t \leq T < t + \Delta, J = j | T \geq t)}{\Delta}, \tag{2.7}$$

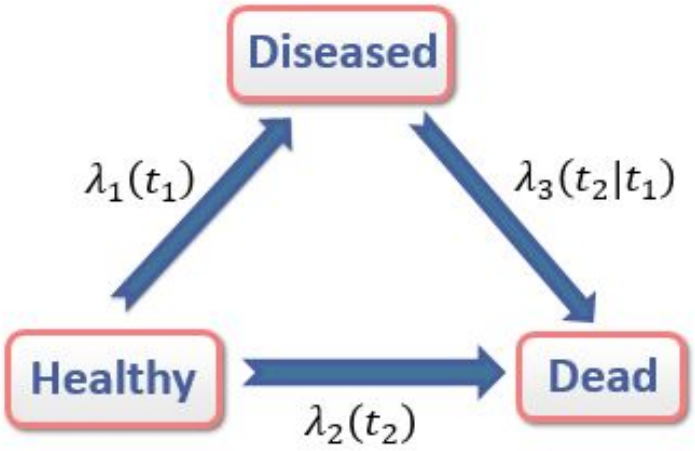


Figure 2.4: Progressive illness-death model with hazard functions.

where T denotes survival time and j represents the type of failure (e.g., cause of mortality), $j = 1, \cdots, k$. This observation is important as it provides insight in underlying mechanism of semi-competing risks data generation, where non-terminal event and terminal event serve as competing risks at the first stage. If the non-terminal event "wins", the subject is still at risk for terminal event, which makes it possible to observe both event times T_1 and T_2. However if the terminal event "wins", it prevents the non-terminal event from happening, that means only T_2 may be observed but never T_1.

Next, to incorporate the dependence of the non-terminal and terminal events in the semi-competing risks setting, Xu et al. (2010) proposed a conditional Markov model, which incorporates a shared frailty variable γ into the illness-death model on

the hazards (2.4)-(2.6), which are conditional hazards

$$\lambda_1(t_1|\gamma) = \gamma\lambda_{01}(t_1), \quad t_1 > 0 \tag{2.8}$$

$$\lambda_2(t_2|\gamma) = \gamma\lambda_{02}(t_2), \quad t_2 > 0 \tag{2.9}$$

$$\lambda_3(t_2|\gamma, t_1) = \gamma\lambda_{03}(t_2), \quad t_2 \geq t_1 > 0. \tag{2.10}$$

Conditional on frailty, this model is Markov because $\lambda_3(t_2|\gamma, t_1) = \gamma\lambda_{03}(t_2), t_2 \geq t_1 > 0$ does not depend on t_1. Detailed explanation on Markovian property can be found in Section 2.5.

Restricted GFCMM

If $\lambda_{02}(t) = \lambda_{03}(t)$, it is referred as a restricted conditional Markov model, where all the dependence structure between non-terminal and terminal event times are captured by the frailty (Xu et al., 2010). This also implies the conditional independence between the non-terminal and terminal event times. Assume that γ is distributed according to $Gamma(\theta^{-1}, \theta^{-1})$, which is a common approach in the multivariate survival analysis literature (Duchateau and Janssen, 2007; Hanagal, 2011; Wienke, 2010) and introduced in Section 2.3. The joint density function of (T_1, T_2) in the upper wedge is

$$f(t_1, t_2) = (\theta + 1)\lambda_{01}(t_1)\lambda_{02}(t_2)[1 + \theta(\Lambda_{01}(t_1) + \Lambda_{02}(t_2))]^{-\frac{1}{\theta}-2}, t_2 \geq t_1 > 0.$$

The corresponding joint survival function of (T_1, T_2) in the upper wedge is

$$\begin{aligned} S(t_1, t_2) &= [1 + \theta\Lambda_{01}(t_1) + \theta\Lambda_{02}(t_2)]^{-\frac{1}{\theta}} \\ &= (S_1^*(t_1)^{-\theta} + S_2^*(t_2)^{-\theta} - 1)^{-\frac{1}{\theta}}, t_2 \geq t_1 > 0, \end{aligned}$$

where $S_1^*(t_1) = (1 + \theta\Lambda_{01}(t_1))^{-\frac{1}{\theta}}$ and $S_2^*(t_2) = (1 + \theta\Lambda_{02}(t_2))^{-\frac{1}{\theta}}$ with $\Lambda_{0i}(t) = \int_0^t \lambda_{0i}(u)du$ for $i = 1, 2$; that is exactly the Clayton copula model proposed by Fine et al. (2001) for analyzing semi-competing risks data (Xu et al., 2010). It indicates that the analysis of semi-competing risks data under the restricted GFCMM is the same as the analysis under the Clayton copula model in the upper wedge. This is also the same as the relationship (2.3) that was derived for bivariate survival data with gamma frailty, except that the constrain on event times $t_2 \geq t_1$.

As it was mentioned earlier and noted by Xu et al. (2010) that in the upper wedge

$$\int_0^\infty \int_{t_1}^\infty f(t_1, t_2)dt_2dt_1 < 1,$$

hence there should exist a probability $f_\infty(t_2)$ such that

$$\int_0^\infty \int_{t_1}^\infty f(t_1, t_2)dt_2dt_1 + \int_0^\infty f_\infty(t_2)dt_2 = 1,$$

which yields

$$f_\infty(t_2) = \int_{t_2}^\infty f(t_1, t_2)dt_1 = \lambda_{02}(t_2)\left(1 + \theta(\Lambda_{01}(t_2) + \Lambda_{02}(t_2))\right)^{-(1/\theta+1)}.$$

With further algebra, the survival function for the non-terminal event can be derived,

$$\begin{aligned} S_1(t) &= \int_t^\infty \int_{t_1}^\infty f(t_1,t_2)dt_2dt_1 + \int_0^\infty f_\infty(t_2)dt_2 \\ &= \int_t^\infty (1+\theta(\Lambda_{01}(t_1)+\Lambda_{02}(t_1)))^{-(1/\theta+1)}\, d\Lambda_{01}(t_1) \\ &\quad + \int_0^\infty (1+\theta(\Lambda_{01}(t_2)+\Lambda_{02}(t_2)))^{-(1/\theta+1)}\, d\Lambda_{02}(t_2), \end{aligned} \tag{2.11}$$

the survival function for the terminal event without occurrence of the non-terminal event,

$$S_2(t) = \frac{\int_t^\infty f_\infty(t_2)dt_2}{\int_0^\infty f_\infty(t_2)dt_2} = \frac{\int_t^\infty (1+\theta(\Lambda_{01}(t_2)+\Lambda_{02}(t_2)))^{-(1/\theta+1)}\, d\Lambda_{02}(t_2)}{\int_0^\infty (1+\theta(\Lambda_{01}(t_2)+\Lambda_{02}(t_2)))^{-(1/\theta+1)}\, d\Lambda_{02}(t_2)}, \tag{2.12}$$

and the survival function for the terminal event with the non-terminal event occurred at time t_1,

$$\begin{aligned} S_3(t|t_1) &= \begin{cases} \frac{\int_t^\infty f(t_1,t_2)dt_2}{\int_{t_1}^\infty f(t_1,t_2)dt_2} & \text{if } t > t_1 \\ 1 & \text{otherwise.} \end{cases} \\ &= \begin{cases} \frac{(1+\theta(\Lambda_{01}(t_1)+\Lambda_{02}(t)))^{-(1/\theta+1)}}{(1+\theta(\Lambda_{01}(t_1)+\Lambda_{02}(t_1)))^{-(1/\theta+1)}} & \text{if } t > t_1 \\ 1 & \text{otherwise.} \end{cases} \end{aligned} \tag{2.13}$$

Unrestricted GFCMM

If $\lambda_{02}(t) \neq \lambda_{03}(t)$, it is unrestricted GFCMM. For the unrestricted GFCMM, the joint density function of (T_1, T_2) in the upper wedge is

$$f(t_1,t_2) = (\theta+1)\lambda_{01}(t_1)\lambda_{03}(t_2)[1+\theta\omega(t_1,t_2)]^{-\frac{1}{\theta}-2}, t_2 \geq t_1 > 0,$$

where $\omega(t_1, t_2) = \Lambda_{01}(t_1) + \Lambda_{02}(t_1) + \Lambda_{03}(t_1, t_2)$ for $t_2 \geq t_1 > 0$ with $\Lambda_{03}(t_1, t_2) = \Lambda_{03}(t_2) - \Lambda_{03}(t_1)$. Then $f_\infty(t_2)$ can be derived, and accordingly $S_1(t)$, $S_2(t)$, and $S_3(t|t_1)$ for the unrestricted GFCMM can be obtained.

Regarding the conditional hazards models (2.8)-(2.10), if the gamma-frailty variable is integrated out using Theorem 2.3.1, the marginal hazards defined by (2.4)-(2.6) are, respectively

$$\lambda_1(t_1) = [1 + \theta\omega(t_1, t_1)]^{-1} \lambda_{01}(t_1), t_1 > 0 \tag{2.14}$$

$$\lambda_2(t_2) = [1 + \theta\omega(t_2, t_2)]^{-1} \lambda_{02}(t_2), t_2 > 0 \tag{2.15}$$

$$\lambda_3(t_2|t_1) = [1 + \theta\omega(t_1, t_2)]^{-1} \lambda_{03}(t_2), t_2 \geq t_1 > 0 \tag{2.16}$$

which were given in Xu et al. (2010).

The marginal hazard $\lambda_1(t_1)$ is derived below as an illustration with the use of Theorem 2.3.1, Laplace transform, and $S(t_1, t_1|\gamma) = e^{-\gamma(\Lambda_{01}(t_1)+\Lambda_{02}(t_1))}$ due to the competing part of semi-competing risks data.

$$\begin{aligned}
\lambda_1(t_1) &= \lambda_0(t_1) \int_0^\infty \gamma f(\gamma|T_1 > t_1, T_2 > t_1) d\gamma \\
&= \lambda_0(t_1) \int_0^\infty \gamma \frac{S(t_1, t_1|\gamma) f(\gamma)}{\int_0^\infty S(t_1, t_1|\gamma) f(\gamma) d\gamma} d\gamma \\
&= [1 + \theta\omega(t_1, t_1)]^{-1} \lambda_{01}(t_1)
\end{aligned}$$

Remark 2.4.2. Marginal hazards (2.15)-(2.16) imply that not only the restricted GFCMM is non-Markov, it also only applies to the scenario that $\lambda_3(t_2|t_1) > \lambda_2(t_2)$ for $t_2 \geq t_1 > 0$ indicating a positive correlation between the non-terminal and terminal events. Therefore the restricted GFCMM will not be applicable to the situation that

the non-terminal event does not impact or decreases the hazard of the terminal event. The unrestricted GFCMM that allows $\lambda_{03}(\cdot)$ to be different from $\lambda_{02}(\cdot)$ will be more flexible to the analysis of semi-competing risks data.

2.4.2 Likelihood Construction

To construct the likelihood for the observed data $\underline{D}$, the likelihood for each of the four data observation scenarios of semi-competing risks data (Figure 2.3) needs to be derived. First take a close look at the definitions of hazards functions $\lambda_1(t_1)$, $\lambda_2(t_2)$, and $\lambda_3(t_2|t_1)$ in (2.4)-(2.6). They reveal the following relationships.

$$\begin{aligned}\lambda_1(t_1) &= \lim_{\Delta\to 0} P(T_1 \in [t_1, t_1+\Delta)|T_1 \geq t_1, T_2 \geq t_1)/\Delta, \quad t_1 > 0 \\ &= \frac{-\frac{\partial S(t_1,t_2)}{\partial t_1}\Big|_{t_2=t_1}}{S(t_1,t_1)} \end{aligned} \tag{2.17}$$

$$\begin{aligned}\lambda_2(t_2) &= \lim_{\Delta\to 0} P(T_2 \in [t_2, t_2+\Delta)|T_1 \geq t_2, T_2 \geq t_2)/\Delta, \quad t_2 > 0 \\ &= \frac{-\frac{\partial S(t_1,t_2)}{\partial t_2}\Big|_{t_1=t_2}}{S(t_2,t_2)} \end{aligned} \tag{2.18}$$

$$\begin{aligned}\lambda_3(t_2|t_1) &= \lim_{\Delta\to 0} P(T_2 \in [t_2, t_2+\Delta)|T_1 = t_1, T_2 \geq t_2)/\Delta) \quad t_2 \geq t_1 > 0 \\ &= \frac{f(t_1,t_2)}{-\frac{\partial S(t_1,t_2)}{\partial t_1}} \qquad (2.19) \\ &= \frac{f_{2|1}(t_2|t_1)}{S_{2|1}(t_2|t_1)}, \qquad (2.20)\end{aligned}$$

where $f_{2|1}(t_2|t_1) = \frac{f(t_1,t_2)}{f_1(t_1)}$ and $S_{2|1}(t_2|t_1) = \frac{-\frac{\partial S(t_1,t_2)}{\partial t_1}}{f(t_1)}$ with $f(t_1) = \int_{t_1}^{\infty} f(t_1,t_2)dt_2$.

It is crucial to derive all the following probabilities in order to construct the likelihood. Considering the conditional hazards (2.8)-(2.9), due to the competing part of the illness-death data, it is obvious that

$$S(t_1,t_1|\gamma) = e^{-\gamma(\Lambda_{01}(t_1)+\Lambda_{02}(t_1))}. \tag{2.21}$$

Based on the conditional hazards (2.8)-(2.10) and (2.17)-(2.20), the following quantities can be derived:

$$\left.\frac{\partial S(t_1,t_2|\gamma)}{\partial t_1}\right|_{t_2=t_1} = -\gamma\lambda_{01}(t_1)e^{-\gamma(\Lambda_{01}(t_1)+\Lambda_{02}(t_1))} \tag{2.22}$$

$$\left.\frac{\partial S(t_1,t_2|\gamma)}{\partial t_2}\right|_{t_1=t_2} = -\gamma\lambda_{02}(t_2)e^{-\gamma(\Lambda_{01}(t_2)+\Lambda_{02}(t_2))} \tag{2.23}$$

$$S_{2|1}(t_2|t_1,\gamma) = e^{-\gamma\Lambda_{03}(t_2)}, t_2 \geq t_1. \tag{2.24}$$

By (2.22) and (2.24), it follows that the density function for $T_1 < \infty$

$$f_1(t_1|\gamma) = \gamma\lambda_{01}(t_1)e^{-\gamma(\Lambda_{01}(t_1)+\Lambda_{02}(t_1)-\Lambda_{03}(t_1))}, t_1 > 0. \tag{2.25}$$

Again data $D = (Y_1, Y_2, \delta_1, \delta_2)$ are observed, where $Y_2 = T_2 \wedge C, \delta_2 = I(T_2 \leq C), Y_1 = T_1 \wedge Y_2$ and $\delta_1 = I(T_1 \leq Y_2)$. The study consists of n i.i.d copies of D, $D_1, D_2, \cdots, D_n$ to form the observed data $\underline{D} = \{D_i : i = 1, 2, \cdots, n\}$. Take the more complicated scenario of "healthy $\rightarrow$ diseased" as an example, the conditional

probability for this case can be derived:

$$
\begin{aligned}
&-\frac{\partial S(t_1,t_2|\gamma)}{\partial t_1}\bigg|\, t_1 = Y_1 \\
=&S_{2|1}(Y_2|Y_1) f_1(Y_1|\gamma) \\
=&e^{-\gamma\Lambda_{03}(Y_2)} \cdot \gamma\lambda_{01}(Y_1) e^{-\gamma(\Lambda_{01}(Y_1)+\Lambda_{02}(Y_1)-\Lambda_{03}(Y_1))} \\
=&\gamma\lambda_{01}(Y_1) e^{-\gamma(\Lambda_{01}(Y_1)+\Lambda_{02}(Y_1)+\Lambda_{03}(Y_1,Y_2))}.
\end{aligned}
$$

Thus the likelihood for "healthy $\rightarrow$ diseased" case conditional on frailty can be written as

$$
L_4(\underline{D}|\underline{\gamma};\ \underline{\Lambda}_0,\theta) = \prod_{i=1}^{n} \left\{\gamma_i\lambda_{01}(Y_{i1}) e^{-\gamma_i(\Lambda_{01}(Y_{i1})+\Lambda_{02}(Y_{i1})+\Lambda_{03}(Y_{i1},Y_{i2}))}\right\}^{\delta_{i1}(1-\delta_{i2})}.
$$

Conditional on frailty, the likelihood for other scenarios can be derived similarly. In summary,

(1) healthy $\rightarrow$ diseased $\rightarrow$ dead ($\delta_{i1}=1, \delta_{i2}=1$):

$$
L_1(\underline{D}|\underline{\gamma};\ \underline{\Lambda}_0,\theta) = \prod_{i=1}^{n} \left\{\gamma_i^2\lambda_{01}(Y_{i1})\lambda_{03}(Y_{i2}) e^{-\gamma_i(\Lambda_{01}(Y_{i1})+\Lambda_{02}(Y_{i1})+\Lambda_{03}(Y_{i1},Y_{i2}))}\right\}^{\delta_{i1}\delta_{i2}} \tag{2.26}
$$

(2) healthy $\rightarrow$ healthy ($\delta_{i1}=0, \delta_{i2}=0, Y_{i1}=Y_{i2}$):

$$
L_2(\underline{D}|\underline{\gamma};\ \underline{\Lambda}_0,\theta) = \prod_{i=1}^{n} \left\{e^{-\gamma_i(\Lambda_{01}(Y_{i1})+\Lambda_{02}(Y_{i1}))}\right\}^{(1-\delta_{i1})(1-\delta_{i2})} \tag{2.27}
$$

(3) healthy $\rightarrow$ dead ($\delta_{i1}=0, \delta_{i2}=1, Y_{i1}=Y_{i2}$):

$$
L_3(\underline{D}|\underline{\gamma};\ \underline{\Lambda}_0,\theta) = \prod_{i=1}^{n} \left\{\gamma_i\lambda_{02}(Y_{i1}) e^{-\gamma_i(\Lambda_{01}(Y_{i1})+\Lambda_{02}(Y_{i2}))}\right\}^{(1-\delta_{i1})\delta_{i2}} \tag{2.28}
$$

(4) healthy $\rightarrow$ diseased ($\delta_{i1} = 1, \delta_{i2} = 0$):

$$L_4(\underline{D}|\underline{\gamma};\ \underline{\Lambda}_0, \theta) = \prod_{i=1}^{n} \left\{ \gamma_i \lambda_{01}(Y_{i1}) e^{-\gamma_i(\Lambda_{01}(Y_{i1}) + \Lambda_{02}(Y_{i1}) + \Lambda_{03}(Y_{i1}, Y_{i2}))} \right\}^{\delta_{i1}(1-\delta_{i2})}. \tag{2.29}$$

Combining the likelihood functions (2.26)-(2.29) and multiplying by the density of gamma frailty, the likelihood for the augmented data $(\underline{D}, \underline{\gamma})$ is

$$\begin{aligned} L(\underline{D}, \underline{\gamma};\ \underline{\Lambda}_0, \theta) = \prod_{i=1}^{n} \Big\{ & \gamma_i^{\delta_{i1}+\delta_{i2}} \lambda_{01}(Y_{i1})^{\delta_{i1}} \lambda_{02}(Y_{i2})^{(1-\delta_{i1})\delta_{i2}} \lambda_{03}(Y_{i2})^{\delta_{i1}\delta_{i2}} \\ & \times e^{-\gamma_i[\Lambda_{01}(Y_{i1}) + \Lambda_{02}(Y_{i1}) + \delta_{i1}\Lambda_{03}(Y_{i1}, Y_{i2})]} \cdot \frac{\theta^{-1/\theta}}{\Gamma(1/\theta)} \gamma_i^{1/\theta - 1} e^{-\gamma_i/\theta} \Big\}. \end{aligned} \tag{2.30}$$

The likelihood for observed data $\underline{D}$ is obtained by integrating out frailty γ using Laplace transform from the above likelihood of complete data:

$$\begin{aligned} L(\underline{D};\ \underline{\Lambda}_0, \theta) = \prod_{i=1}^{n} \Big\{ & \lambda_{01}(Y_{i1})^{\delta_{i1}} \lambda_{02}(Y_{i2})^{(1-\delta_{i1})\delta_{i2}} \lambda_{03}(Y_{i2})^{\delta_{i1}\delta_{i2}} (1+\theta)^{\delta_{i1}\delta_{i2}} \\ & \times (1 + \theta[\Lambda_{01}(Y_{i1}) + \Lambda_{02}(Y_{i1}) + \delta_{i1}\Lambda_{03}(Y_{i1}, Y_{i2})])^{-1/\theta - \delta_{i1} - \delta_{i2}} \Big\}. \end{aligned} \tag{2.31}$$

For the restricted model (i.e., $\lambda_{02}(t) = \lambda_{03}(t)$), the likelihood for the observed data is reduced to:

$$\begin{aligned} L(\underline{D};\ \underline{\Lambda}_0, \theta) = \prod_{i=1}^{n} \Big\{ & \lambda_{01}(Y_{i1})^{\delta_{i1}} \lambda_{02}(Y_{i2})^{\delta_{i2}} (1+\theta)^{\delta_{i1}\delta_{i2}} \cdot \\ & (1 + \theta[\Lambda_{01}(Y_{i1}) + \Lambda_{02}(Y_{i2})])^{-1/\theta - \delta_{i1} - \delta_{i2}} \Big\}. \end{aligned}$$

2.5 Markov Illness-Death Model

2.5.1 Model Formulation

In this section, the progressive Markov illness-death model from the multi-state modeling perspective is studied. To study continuous-time multi-state models, a few concepts including the transition probability and transition intensity need to be established. First a multi-state process is a stochastic process, denoted by $\{X(t), t \in [0, \infty)\}$, which can take values in finite states (e.g., $\{1, ..., K\}$). Transition probabilities or transition intensities are crucial because a multi-state process can be fully described by either of them.

Definition 2.5.1. The **transition probability** from state j at time t_0 to state k at time t is

$$P_{jk}(t_0, t; \mathcal{F}_{t_0-}) = P(X(t) = k | X(t_0) = j; \mathcal{F}_{t_0-}), \tag{2.32}$$

where $\mathcal{F}_{t_0-} = \sigma(\{X(s), s \in [0, t_0)\})$. $\mathcal{F}_{t_0-}$ is a filtration, which represents the trajectory or history of process up to just before time t_0.

Definition 2.5.2. The **transition intensity** at time t between states j and k is

$$\alpha_{jk}(t; \mathcal{F}_{t-}) = \lim_{\Delta \to 0} \frac{P_{jk}(t, t + \Delta; \mathcal{F}_{t-})}{\Delta}, \ j \neq k.$$

Since $\sum_k \alpha_{jk}(t) = 0$, it leads to $\alpha_{jj}(t) = -\sum_{k \neq j} \alpha_{jk}(t)$, when $k = j$.

Remark 2.5.1. The transition intensities can be interpreted as instantaneous rate of transition from state j to state k, which is similar to the hazard definition in the survival data analysis, which is instantaneous rate of occurrence of an event given that person survived up till that time.

It is important to note that the filtration is usually suppressed in both of the notation for transition probability as $P_{jk}(t_0, t)$ and transition intensity as $\alpha_{jk}(t)$, however they do depend on the past history (Meira-Machado et al., 2009). In addition, different assumptions can be made on whether and how the transition probabilities or transition intensities depend on time (Cook and Lawless, 2018; Meira-Machado et al., 2009), which include

1. Time homogeneous models. $\alpha_{jk}(t; \mathcal{F}_{t_0-}) = \alpha_{jk}$, where the transition intensity is constant over time t;
2. Markov models. $P(X(t) = k|X(t_0) = j; \mathcal{F}_{t_0-}) = P(X(t) = k|X(t_0) = j)$, where the transition probability only depends on the history through current state j while not anything happens before time t_0;
3. Semi-Markov models. $\alpha_{jk}(t; \mathcal{F}_{t-}) = \alpha_{jk}(t, t - t_j)$, where the transition intensity depends on the history through current state j and the entry time t_j to current state j (which is the length of stay in state j).

In the literature of multi-state models, the Markov assumption is commonly employed due to its simplicity (Hougaard, 1999; Jackson et al., 2003; Meira-Machado et al., 2009). Details regarding stochastic process and multi-state modeling can be found in many literatures (Andersen and Keiding, 2002; Cook and Lawless, 2018; Hougaard, 1999; Putter et al., 2007; Ross et al., 1996; Van Den Hout, 2016).

Again we are interested in the three-compartment progressive Markov illness-death model. This multi-state process is a stochastic process $\{X(t), t \in [0, \tau]\}$, which takes values on three finite states, where $\mathcal{H} = \{1, 2, 3\}$, and it denotes two transition states, healthy and diseased states, and an absorbing state, death, accordingly. Let $X(0) = 1$, which indicates all individuals start from state 1 (healthy). $\alpha_{jk}(t)$ denotes

the transition intensity from state j to state k at time t, where $j \neq k; j, k = 1, 2, 3$, and assume they are uniformly continuous functions on $[0, \tau]$. The corresponding cumulative transition intensity is denoted as $A_{jk}(t) = \int_0^t \alpha_{jk}(t)dt$, $j, k = 1, 2, 3$, $j \neq k$. To fully characterize this multi-state process, its transition intensity matrix $Q(t)$ can be obtained accordingly, which is

$$Q(t) = \begin{bmatrix} -(\alpha_{12}(t) + \alpha_{13}(t)) & \alpha_{12}(t) & \alpha_{13}(t) \\ 0 & -\alpha_{23}(t) & \alpha_{23}(t) \\ 0 & 0 & 0 \end{bmatrix}. \tag{2.33}$$

Hence this progressive illness-death model can be represented by three transition intensity functions $\alpha_{12}(t)$, $\alpha_{13}(t)$, and $\alpha_{23}(t)$. The corresponding cumulative transition intensities are $A_{12}(t)$, $A_{13}(t)$, and $A_{23}(t)$. With an assumption of the Markovian property, transition intensity from illness to death $\alpha_{23}(t)$ only depends on the status of illness at time t but not the history through illness state before t. This three-state Markov illness-death model without recovery is presented in Figure 2.5

2.5.2 Likelihood Construction

To derive the likelihood for multi-state models, it is important to calculate the transition probabilities of particular sample paths first. For Markov multi-state models, the transition probabilities can be calculated from the intensities by solving the forward Kolmogorov differential equation (Andersen et al., 2012; Ross et al., 1996). To better understand this progressive Markov illness-death model, the transition probability

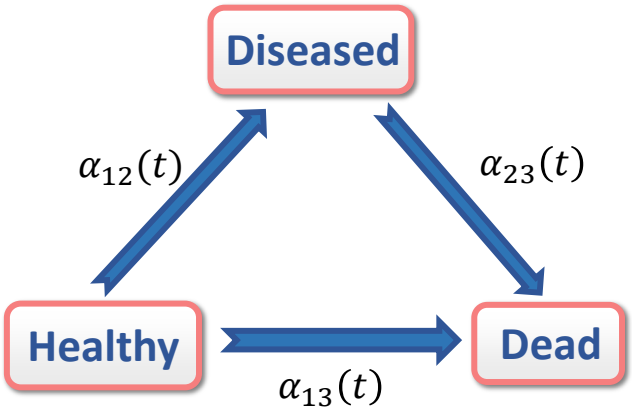


Figure 2.5: Progressive illness-death model with transition intensity functions.

matrix $P(s,t)$ based on transition intensity matrix $Q(t)$ (2.33) is also obtained,

$$P(s,t) = \begin{bmatrix} P_{11}(s,t) & P_{12}(s,t) & 1 - P_{11}(s,t) - P_{12}(s,t) \\ 0 & P_{22}(s,t) & 1 - P_{22}(s,t) \\ 0 & 0 & 1 \end{bmatrix}, \tag{2.34}$$

where $P_{11}(s,t) = e^{-(A_{12}(s,t)+A_{13}(s,t))}$, $P_{22}(s,t) = e^{-A_{23}(s,t)}$,

$P_{12}(s,t) = \int_s^t P_{11}(s,u)\alpha_{12}(u)P_{22}(u,t)du$, with $A_{jk}(s,t) = A_{jk}(t) - A_{jk}(s) = \int_s^t \alpha_{jk}(u)du$.

Based on the the transition probability matrix (2.34), it is straightforward to derive the likelihood for all four data scenarios.

(1) Healthy $\rightarrow$ diseased $\rightarrow$ dead:

The individual stays in healthy state till Y_{i1}, transits to diseased state at Y_{i1}, stays at diseased state till Y_{i2}, and then have another transition to death state

at Y_{i2}.

$$L_1(\underline{D}|\underline{A}_0,\theta)=\prod_{i=1}^{n}\left\{e^{-(A_{12}(Y_{i1})+A_{13}(Y_{i1}))}\cdot\alpha_{12}(Y_{i1})\cdot e^{-A_{23}(Y_{i1},Y_{i2})}\cdot\alpha_{23}(Y_{i2})\right\}^{\delta_{i1}\delta_{i2}} \tag{2.35}$$

(2) Healthy $\rightarrow$ healthy:

The individual stays in healthy state till Y_{i1}, where $Y_{i1}=Y_{i2}$.

$$L_2(\underline{D}|\underline{A}_0,\theta)=\prod_{i=1}^{n}\left\{e^{-(A_{12}(Y_{i1})+A_{13}(Y_{i1}))}\right\}^{(1-\delta_{i1})(1-\delta_{i2})} \tag{2.36}$$

(3) Healthy $\rightarrow$ dead:

The individual stays in healthy state till Y_{i1}, and then transition to death at Y_{i1}, where $Y_{i1}=Y_{i2}$.

$$L_3(\underline{D}|\underline{A}_0,\theta)=\prod_{i=1}^{n}\left\{e^{-(A_{12}(Y_{i1})+A_{13}(Y_{i1}))}\cdot\alpha_{13}(Y_{i1})\right\}^{(1-\delta_{i1})\delta_{i2}} \tag{2.37}$$

(4) Healthy $\rightarrow$ diseased:

The individual stays in healthy state till Y_{i1}, and then transition into diseased state at Y_{i1} and stayed in diseased state till Y_{i2}.

$$L_4(\underline{D}|\underline{A}_0,\theta)=\prod_{i=1}^{n}\left\{e^{-(A_{12}(Y_{i1})+A_{13}(Y_{i1}))}\cdot\alpha_{12}(Y_{i1})\cdot e^{-A_{23}(Y_{i1},Y_{i2})}\right\}^{\delta_{i1}(1-\delta_{i2})} \tag{2.38}$$

Combining the likelihood functions (2.35)-(2.38) together, the likelihood for data $\underline{D}$ is

$$L(\underline{D};\ \underline{A}_0) = \prod_{i=1}^{n} \left\{ \alpha_{12}(Y_{i1})^{\delta_{i1}} \alpha_{13}(Y_{i2})^{(1-\delta_{i1})\delta_{i2}} \alpha_{23}(Y_{i2})^{\delta_{i1}\delta_{i2}} \right. \\ \left. \times e^{-[A_{12}(Y_{i1}) + A_{13}(Y_{i1}) + \delta_{i1} A_{23}(Y_{i1}, Y_{i2})]} \right\}. \tag{2.39}$$

Likelihood Function in Counting Process Notation

Suppose there are n individuals in the data, for the i^{th} subject, and define

$$N_{jk,i}(t) = I(u \leq t : X(u-) = j, X(u) = k) \ \ j \neq k; j, k = 1, 2, 3,$$

$$Y_{j,i}(t) = I(X_i(t-) = j),\ j = 1, 2, 3,$$

where $I(\cdot)$ in an indicator function. $\{N_{jk,i}(t),\ j \neq k; j, k = 1, 2, 3\},\ i = 1, \cdots, n$ is the counting process for the i^{th} individual, which is an indicator function for whether there is a direct transition for the i^{th} individual from state j to state k during $[0, t]$, and $N_{jk}(t) = \sum_{i=1}^{n} N_{jk,i}(t)$ count the number of direct transitions from state j to state k during $[0, t]$ in the sample. $\{Y_{j,i}(t),\ j = 1, 2, 3\}$ is a predictable process for i^{th} individual at the j^{th} state, which is an indicator function for whether the i^{th} individual is in state j up till time t, and $Y_j(t) = \sum_{i=1}^{n} Y_{j,i}(t)$ counts the number of individuals in state j right before time t in the sample. In this research, $\{Y_j(t),\ j = 1, 2, 3\}$ is called at-state process for state j. For this three-state illness-death model without recovery, $N_{21}(t) = N_{31}(t) = N_{32}(t) = 0$. Assume the transition intensities $\alpha_{12}(t)$, $\alpha_{13}(t)$, and $\alpha_{23}(t)$ are uniformly continuous. The intensity processes are of multiplicative form $\{\lambda_{jk}(t) = \alpha_{jk}(t) Y_j(t),\ j, k = 1, 2, 3;\ j \neq k\}$.

It is important to note the likelihood derived above is actually the "probability" of trajectory for observed data, which is not the likelihood strictly speaking. Details regarding the derivation of the likelihood in rigorous manner through Jacod's formula and using product integral can be found in (Andersen et al., 2012; Commenges, 2003). With cumulative transition intensities continuous, the likelihood under the counting process notation can be written as

$$\prod_t \prod_j \left\{ \prod_{k \neq j} \left(Y_j(t) dA_{jk}\right)^{dN_{jk}(t)} \left(1 - dA_{j\cdot}(t)\right)^{Y_j(t) - dN_{j\cdot}(t)} \right\},$$

where $\prod$ denotes the product integral, and $N_{j\cdot} = \sum_{k \neq j} N_{jk}$, and $A_{j\cdot} = \sum_{k \neq j} A_{jk}$.

It can also be written as

$$\left\{ \prod_{0 \leq t \leq \tau} \prod_{k \neq j} Y_j(t) dA_{jk}(t)^{dN_{jk}(t)} \right\} e^{-\sum_{k \neq j} \int_0^\tau Y_j(t) dA_{jk}(t)},$$

where $\tau = \sup_t \left\{ \int_0^t \alpha_{jk}(u) du \right\} < \infty$, $j \neq k$, $j, k = 1, 2, 3$.

2.6 Relationship between GFCMM and Markov Illness-Death Model

In Section 2.4, the definition of hazards of event times is along with the framework of survival data analysis. In Section 2.5, the transition intensity between two consecutive states are defined under multi-state modeling framework. The GFCMM and Markov illness-death model may seem like two different models at first sight, however they are actually closely related to each other.

Essentially, the transition intensities defined in Definition 2.5.2 are equivalent to the hazards defined in (2.4)-(2.6). Note, further assumptions on the hazard or

transition intensity functions, such as Markovian property, frailty variable, are not considered here. $\lambda_3(t_2|t_1) = \alpha_{23}(t_2)$ is shown as an example. Simply compare

$$\lambda_3(t_2|t_1) = \lim_{\Delta \to 0} \frac{1}{\Delta} P(T_2 \in [t_2, t_2 + \Delta) | T_1 = t_1, T_2 \geq t_2)$$

to

$$\alpha_{23}(t_2) = \lim_{\Delta \to 0} \frac{1}{\Delta} P(X(t_2 + \Delta) = 3 | X(t_2) = 2; \mathcal{F}_{t_2-}),$$

where $\mathcal{F}_{t_2-}$ indicates the non-terminal event happens at $T_1 = t_1$. Hence although definitions of the hazards in (2.4)-(2.6) focus on event times, and definitions of the transition intensities (Definition 2.5.2) focus on the transient or absorbing states, they are mathematically equivalent. The only difference between these two models is with their interpretations. Hazard is usually interpreted as the instantaneous rate for experiencing an event given that person survived up to that time. Transition intensity is interpreted as instantaneous rate of progression from one state to another state. Therefore, the progressive illness-death model in this research can be investigated from both prospectives of traditional multivariate survival data analysis and multi-state modeling framework.

2.7 Concluding Remarks

This chapter focused on modeling semi-competing risks data. First the three-compartment illness-death model and frailty models were introduced. Specifically, the GFCMM and Markov illness-death model, which are both commonly used for analyzing semi-competing risks data based on the illness-death model, were studied. Here this chapter concludes with a few remarks on these two models.

- The GFCMM restricted model (*i.e.*, $\lambda_{02}(t) = \lambda_{03}(t)$), where all the dependence structure between two event times is captured by the frailty, is equivalent to the well-known Clayton copula model in the upper wedge. This model bridges the illness-death models and the copula models for semi-competing risks data. This model also implies a positive correlation between the non-terminal and terminal event times, which means it can only be applied to the situation where the non-terminal event increases the risk to terminal event.
- The marginal hazards $\lambda_1(t_1)$ and $\lambda_2(t_2)$ defined in (2.4)-(2.5) are cause-specific hazards, which is the competing component of semi-competing risks data. This provides insight for the first stage of semi-competing risks data generation.
- In survival analysis, the focus is on the times to events of interest, while in multi-state models we concentrate on the transitions between states. For this progressive illness-death model, the definitions of marginal hazards (unconditional on frailty) in (2.4)-(2.6) along with survival data analysis framework are essentially equivalent to the transition intensity functions (Definition 2.5.2) in the multi-state modeling framework. Hence it is possible to conduct inference on hazards/transition intensity functions from two perspectives only with different interpretations.

This chapter provided a notion for understanding semi-competing risks data. Before moving on to the nonparametric estimation (Chapter 4) and nonparametric testing (Chapter 5) of semi-competing risks data, in the next chapter, we review the asymptotic properties of three inference procedures, which are likelihood ratio, Wald, and Rao's score tests when nuisance parameters are present and introduce empirical process theory that will be used in future chapters.

Chapter 3

Asymptotic Theories

3.1 Introduction

In statistical literature, asymptotic theories are very important in both theoretical and practical senses. It provides a framework to evaluate properties of statistical procedures, which are not limited to estimators and test statistics, as sample size increases to infinity. Thus it is frequently referred to as large-sample theory.

In this chapter, the asymptotics of maximum likelihood estimation (MLE) are first briefly discussed since the three inference procedures of interest are closely related to it. Regularity conditions for MLE, some related definitions (e.g., influence function, efficient influence function), and its asymptotic properties (e.g., consistency, asymptotic normality and efficiency) are introduced. Nevertheless, in many applications of statistical models, not all parameters are the study of interest. Often, only a subset of model parameters are of interest for inference, and remaining parameters are nuisance parameters. The information bound and efficient influence function with nuisance parameters can be derived in parallel and so are the asymptotic properties of MLE. Then three testing procedures with their asymptotic distributions are reviewed, including likelihood ratio, Wald, and Rao's score tests with nuisance parameters for parametric models (Section 3.2), which helps construct the score test in Chapter 4 for the comparison of the restricted and unrestricted GFCMMs. Next a short introduction to empirical process theory (Section 3.3) is given, which is a powerful tool to study asymptotic theories for non-/semi-parametric models and paves the way for

proving large-sample theories for the nonparametric tests proposed in Chapter 5. The expansion from parametric to non-/semi-parametric inference theories is crucial for having a good understanding of the difference between these models and appreciate the development of empirical process theory.

Much of the material (e.g., definitions, theorems) presented in the chapter is adopted from excellent textbooks (Andersen et al., 2012; Casella and Berger, 2002; Kosorok, 2008; Lehmann and Casella, 2006; Resnick, 2003; Van der Vaart, 2000) and class notes (i.e., Advanced Likelihood Theory, Advanced Survival Analysis).

3.2 Three Inference Procedures with Nuisance Parameters

As the method of maximum likelihood was introduced in 1922 by Fisher (1922), MLE has become one of the most popular and useful methods ever since to conduct statistical inference due to many of its desirable statistical asymptotic properties, for example, consistency, asymptotic normality and efficiency. Built on the usage of MLE, the likelihood ratio test was developed by Wilks (1938), and the Wald test was introduced by Wald (1943). In the meantime, the score test was proposed by Rao (1948). These three classical tests are often referred to as Holy Trinity in statistical literature (Rao, 2005), where numerous research has been conducted.

Generally speaking, the validity of MLE theory requires assumptions on the statistical models or called regularity conditions on the underlying data. Parametric models are considered, where $\mathcal{P}$ is the entire family of distributions, and its members can be identified by a parameter $\underline{\theta}$ through density function $p_{\underline{\theta}}$, and $\Theta \subset R^k$ is the parameter space. That is

$$\mathcal{P} = \left\{ p_{\underline{\theta}} : \underline{\theta} \in \Theta \subset R^k \right\}.$$

The regularity conditions are

(A1) The model is identifiable; that is, $\underline{\theta} \neq \underline{\theta}^*$ implies the densities $p_{\underline{\theta}} \neq p_{\underline{\theta}^*}$ with respect to some dominating measure μ.

(A2) The distributions $p_{\underline{\theta}}$ have common support, $A = \{x : p_{\underline{\theta}}(x) > 0\}$, that is independent of $\underline{\theta}$.

(A3) There is an i.i.d. sample of $\underline{X} = (X_1, \cdots, X_n)$ with probability density $p_{\underline{\theta}}(X)$.

(A4) Θ contains an open neighborhood $\Theta_0 \subset R^k$ of θ_0, where $\underline{\theta}_0$ denotes the true value of θ, for which

  (i) $l(\underline{\theta}; X) = \log p_{\underline{\theta}}(X)$ is twice continuously differentiable in $\underline{\theta}$ for almost everywhere of X;

  (ii) For almost everywhere of X, the third order derivatives of $l(\underline{\theta}; X)$, $l^{(3)}_{jpq}(\underline{\theta}; X)$ exists and satisfy $l^{(3)}_{jpq}(\underline{\theta}; X) \leq M_{jpq}(x)$, $\underline{\theta} \in \Theta_0$, for all $j \geq 1$, $p, q \leq k$ with $E_0(M_{jpq}(X)) < \infty$.

(A5) For the first and second order derivatives of $l(\underline{\theta}; X)$:

  (i) $E_0(\dot{l}_j(\underline{\theta}_0; X)) = 0$ for $j = 1, \cdots, k$

  (ii) $E_0(\dot{l}^2_j(\underline{\theta}_0; X)) < \infty$ for $j = 1, \cdots, k$

  (iii) The information matrix $I(\underline{\theta}_0) = E_0\left(\dot{l}_j(\underline{\theta}_0; X)\dot{l}^T_j(\underline{\theta}_0; X)\right) = -\left(E_0(\ddot{l}_{pq}(X))\right)$ is positive definite.

The above regularity conditions are sufficient to prove the desirable optimality properties of MLE, such as consistency.

Theorem 3.2.1. (Consistency of MLE) Suppose $\underline{X} = (X_1, \cdots, X_n)$ are i.i.d. with density $p_{\underline{\theta}} \in \mathcal{P}$, where $\mathcal{P}$ satisfies A1-A5. Let $\hat{\underline{\theta}}$ denote the MLE of $\underline{\theta}$. $\hat{\underline{\theta}}$ converges in probability to $\underline{\theta}$ or $\hat{\underline{\theta}}$ is weakly consistent denoted by

$$\hat{\underline{\theta}} \xrightarrow{P} \underline{\theta}_0 \text{ as } n \to \infty.$$

Next some very important concepts, the influence function and efficient influence function, are introduced to help establish asymptotic distributions of estimators.

Definition 3.2.1. Consider estimation of $q(\underline{\theta})$, for any estimator $T_n(\underline{X})$, if

$$\sqrt{n}(T_n(\underline{X}) - q(\underline{\theta})) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \varphi_{\underline{\theta}}(X_i) + o_p(1),$$

and $E\left(\varphi_{\underline{\theta}}(X)\right) = 0$, $Var\left(\varphi_{\underline{\theta}}(X)\right) < \infty$, then $\varphi_{\underline{\theta}}$ is called an **influence function**, and $T_n(\underline{X})$ is asymptotic linear with $\varphi_{\underline{\theta}}$, where $o_p(1)$ denotes a quantity going to zero in probability.

Remark 3.2.1. It can be shown by central limit theorem that

$$\sqrt{n}(T_n(\underline{X}) - q(\underline{\theta})) \xrightarrow{D} N\left(0, E\left(\varphi_{\underline{\theta}}(X_1)\varphi_{\underline{\theta}}^T(X_1)\right)\right).$$

However, $T_n(\underline{X})$ may not necessarily be the estimator that achieves the minimum variance among all estimators for $q(\underline{\theta})$, which leads to the introduction of efficient influence function.

Definition 3.2.2. Consider estimation of $q(\underline{\theta})$, the **efficient influence function** for $q(\underline{\theta})$ is

$$\tilde{\varphi}_\theta(X) = \dot{q}^T(\underline{\theta}) I^{-1}(\underline{\theta}) \dot{l}_{\underline{\theta}}(X), \tag{3.1}$$

where $\dot{l}_{\underline{\theta}}(X) = \frac{\partial}{\partial \underline{\theta}} l_{\underline{\theta}}(X)$ is the score function of $\underline{\theta}$, $I(\underline{\theta}) =$ is information for one observation, and $\dot{q}(\underline{\theta}) = \frac{\partial}{\partial \underline{\theta}} q(\underline{\theta})$.

Remark 3.2.2. If an estimator $T_n(\underline{X})$ of $q(\underline{\theta})$ is asymptotic linear with efficient influence function $\tilde{\varphi}_{\underline{\theta}}$,

$$\sqrt{n}(T_n(\underline{X}) - q(\underline{\theta})) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\varphi}_{\underline{\theta}}(X_i) + o_p(1),$$

$T_n(\underline{X})$ is **asymptotic efficient**. It can be shown $E\left(\tilde{\varphi}(\underline{X})\tilde{\varphi}^T(\underline{X})\right) = \dot{q}^T(\underline{\theta})I^{-1}(\underline{\theta})\dot{q}(\underline{\theta})$ and

$$\sqrt{n}(T(\underline{X}) - q(\underline{\theta})) \xrightarrow{D} N(0, \dot{q}^T(\underline{\theta})I^{-1}(\underline{\theta})\dot{q}(\underline{\theta})).$$

$\dot{q}^T(\underline{\theta})I^{-1}(\underline{\theta})\dot{q}(\underline{\theta})$ is the Cramér-Rao lower bound of $q(\underline{\theta})$, which is the lowest bound that $T_n(\underline{X})$ can achieve.

Theorem 3.2.2. (Asymptotic Normality and Efficiency of MLE) Suppose $\underline{X} = (X_1, \cdots, X_n)$ are i.i.d. with density $p_{\underline{\theta}} \in \mathcal{P}$, where $\mathcal{P}$ satisfies A1-A5. Let $\hat{\underline{\theta}}$ denote the MLE of $\underline{\theta}$. Then $\hat{\underline{\theta}}$ is asymptotically normal and efficient estimator for $\underline{\theta}$,

$$\sqrt{n}(\hat{\underline{\theta}} - \underline{\theta}) \xrightarrow{D} N(0, I^{-1}(\underline{\theta})).$$

Remark 3.2.3. Maximum likelihood estimator $\hat{\underline{\theta}}$ is asymptotically linear with efficient influence function $\tilde{\varphi}_{\underline{\theta}}$

$$\sqrt{n}(\hat{\underline{\theta}} - \underline{\theta}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\varphi}_{\underline{\theta}}(X_i) + o_p(1),$$

where $\tilde{\varphi}_{\underline{\theta}}(X) = I^{-1}(\theta)\dot{l}_{\underline{\theta}}(X)$.

The asymptotic normality of maximum likelihood estimator $\hat{\theta}$ is established. However in many cases, researchers are not interested in conducting inference in all pa-

rameters. Maybe only a subset of model parameters is of primary interest, and the remaining parameters are treated as nuisance parameters. Accordingly, the asymptotic properties of MLEs can be derived. Next the asymptotic distributions of likelihood ratio test, Wald test, and score test under the null hypothesis in the situation that when there are nuisance parameters are described.

3.2.1 Likelihood Ratio Test

Suppose the parameter is a k-dimensional vector $\underline{\theta} = (\underline{\theta}_1^T, \underline{\theta}_2^T)^T \in R^m \times R^{k-m}$ $(m < k)$, and the primary interest is only with $\underline{\theta}_1$, while $\underline{\theta}_2$ is the nuisance parameter. For the ease of derivation in this part, the information matrix $I(\underline{\theta})$ is written in block form as follows

$$I(\underline{\theta}) = \begin{pmatrix} I_{11} & I_{12} \\ \\ I_{21} & I_{22} \end{pmatrix},$$

and

$$I^{-1}(\underline{\theta}) = \begin{pmatrix} I_{11\cdot 2}^{-1} & -I_{11\cdot 2} I_{12} I_{22}^{-1} \\ \\ -I_{22\cdot 1} I_{21} I_{11}^{-1} & I_{22\cdot 1}^{-1} \end{pmatrix}$$

where $I_{11\cdot 2} = I_{11} - I_{12} I_{22}^{-1} I_{21}$, $I_{22\cdot 1} = I_{22} - I_{21} I_{11}^{-1} I_{12}$, $I_{11} = E_{\underline{\theta}}(\dot{l}_{\underline{\theta}_1}(X) \dot{l}_{\underline{\theta}_1}^T(X))$, $I_{12} = E_{\underline{\theta}}(\dot{l}_{\underline{\theta}_1}(X) \dot{l}_{\underline{\theta}_2}^T(X))$, $I_{21} = E_{\underline{\theta}}(\dot{l}_{\underline{\theta}_2}(X) \dot{l}_{\underline{\theta}_1}^T(X))$, and $I_{22} = E_{\underline{\theta}}(\dot{l}_{\underline{\theta}_2}(X) \dot{l}_{\underline{\theta}_2}^T(X))$.

Let

$$W = \begin{pmatrix} W_1 \\ \\ W_2 \end{pmatrix} \sim N(0, I(\underline{\theta}_0)),$$

then $W_1 \sim N(0, I_{11})$ and $W_2 \sim N(0, I_{22})$. Let

$$Z = \begin{pmatrix} Z_1 \\ \\ Z_2 \end{pmatrix} = I^{-1}(\underline{\theta}_0)W = \begin{pmatrix} I_{11\cdot 2}^{-1}(W_1 - I_{12}I_{22}^{-1}W_2) \\ \\ I_{22\cdot 1}^{-1}(W_2 - I_{21}I_{11}^{-1}W_1) \end{pmatrix},$$

then it can be shown $Z_1 \sim N(0, I_{11\cdot 2}^{-1})$.

Assume an i.i.d. sample $\underline{X} = (X_1, \cdots, X_n)$ from density distribution $p_{\underline{\theta}}$ and conduct hypothesis testing of $\underline{\theta} = (\underline{\theta}_1^T, \underline{\theta}_2^T)^T$ with the null hypothesis $H_0 : \underline{\theta} \in \Theta_0$ versus the alternative $H_1 : \underline{\theta} \in \Theta$, where $\Theta_0 = \{(\underline{\theta}_1, \underline{\theta}_2) : \underline{\theta}_1 = \underline{\theta}_{10}\}$,
$\Theta = \left\{(\underline{\theta}_1, \underline{\theta}_2) : \underline{\theta}_1 \in R^m, \underline{\theta}_2 \in R^{k-m}\right\}$.

The **likelihood ratio test statistic** for testing $H_0 : \underline{\theta} \in \Theta_0$ versus $H_1 : \underline{\theta} \in \Theta$ is

$$\Lambda_n = -2 \log \frac{\sup_{\Theta_0} L_{\underline{\theta}}(\underline{X})}{\sup_{\Theta} L_{\underline{\theta}}(\underline{X})} = 2(l(\hat{\underline{\theta}}_n) - l(\hat{\underline{\theta}}_n^0)), \tag{3.2}$$

where $\hat{\underline{\theta}}_n^0$ and $\hat{\underline{\theta}}_n$ are consistent estimators of $\underline{\theta}$ under H_1 and H_0.

Theorem 3.2.3. Consider testing $H_0 : \underline{\theta} \in \Theta_0$ versus $H_1 : \underline{\theta} \in \Theta$, A1-A5 are satisfied, and $\underline{\theta}_0 \in \Theta$ is true, then

$$\Lambda_n \xrightarrow{D} Z_1^T I_{11\cdot 2} Z_1 \sim \chi_m^2.$$

3.2.2 Wald Test

The **Wald test statistic** for testing $H_0 : \underline{\theta} \in \Theta_0$ versus $H_1 : \underline{\theta} \in \Theta$ is

$$W_n = \sqrt{n}(\hat{\underline{\theta}}_{n1} - \underline{\theta}_{10})^T \cdot \hat{I}_{11\cdot 2}(\hat{\underline{\theta}}_n) \cdot \sqrt{n}(\hat{\underline{\theta}}_{n1} - \underline{\theta}_{10}), \tag{3.3}$$

where $\hat{I}_{11\cdot 2}(\hat{\underline{\theta}}_n)$ is the empirical version of $I_{11\cdot 2}(\hat{\underline{\theta}}_n)$.

Theorem 3.2.4. Consider testing $H_0 : \underline{\theta} \in \Theta_0$ versus $H_1 : \underline{\theta} \in \Theta$, A1-A5 are satisfied, and $\underline{\theta}_0 \in \Theta$ is true, then

$$W_n \xrightarrow{D} Z_1^T I_{11\cdot 2} Z_1 \sim \chi^2_m.$$

3.2.3 Rao's Score Test

The **score test statistic** is

$$R_n = \frac{1}{\sqrt{n}} \left[\sum_{i=1}^{n} \dot{l}_{\underline{\theta}}(\hat{\underline{\theta}}_n^0 | X_i) \right]^T \cdot I^{-1}(\hat{\underline{\theta}}_n^0) \cdot \frac{1}{\sqrt{n}} \left[\sum_{i=1}^{n} \dot{l}_{\underline{\theta}}(\hat{\underline{\theta}}_n^0 | X_i) \right] \tag{3.4}$$

Theorem 3.2.5. Consider testing $H_0 : \underline{\theta} \in \Theta_0$ versus $H_1 : \underline{\theta} \in \Theta$, A1-A5 are satisfied, and $\underline{\theta}_0 \in \Theta$ is true, then

$$R_n \xrightarrow{D} Z_1^T I_{11\cdot 2} Z_1 \sim \chi^2_m.$$

Remark 3.2.4. The likelihood ratio test, Wald test, and Rao's score test are asymptotically equivalent, if exist. All three tests are consistent.

Remark 3.2.5. Different from the likelihood ratio test and Wald test, Rao's score test only requires the computation of the MLE under the null hypothesis when there are nuisance parameters. This poses advantages for using score test over the other two classical tests in situations where computing MLE is problematic under the full model.

Likelihood technique is a general principle for statistical inference. Maximum likelihood estimator (if exists) grants the attractive asymptotic behavior of likelihood ratio, Wald, and Rao's score tests when there are nuisance parameters. However the likelihood principle for parametric models as described above cannot be easily generalized to non-/semi-parametric settings. In some cases, the parametric component

in semiparametric models are of primary interest, for example, the regression parameters in Cox PH models. For this case, profile or partial likelihood technique can be applied, where the maximum parial likelihood estimator behaves similarly to the maximum likelihood estimators. If the interest lies in the nonparametric components, it is not straightforward to do nonparametric inference. There is some literature on the generalized likelihood ratio test (Fan and Jiang, 2007; Fan et al., 2001), however this research direction is not pursued here. Instead, the empirical process theory, a powerful tool to establish and validate statistical theories when developing new methods for non-/semi-parametric inferences, is introduced.

3.3 Empirical Process Theory (EPT)

Empirical process theory (EPT) is a very powerful tool in establishing and validating statistical theories when developing new methods for non-/semi-parametric inferences. In this thesis, EPT is heavily used when constructing the nonparametric test statistics on cumulative transition intensity functions in multi-state models. Hence some fundamental results in EPT are outlined.

3.3.1 Preliminary Notation and Definitions

EPT is built on probability theory in general sample space. Many notation and definitions used in this section are extracted from Kosorok (2008) and Van Der Vaart and Wellner (1996). First some basic and commonly used concepts, such as metric space, stochastic process, Gaussian process, and weak convergence, are introduced.

In probability theory, it is critical to define the probability space $(\Omega, \mathcal{A}, P)$, where P indicates a probability measure on a sample space Ω with σ-field $\mathcal{A}$. To measure

the relationship between two elements in a sample space quantitatively, it is also very important to define a metric that can be used to numerically measure the distance between the elements.

Definition 3.3.1. A **metric space** is a set $\mathbb{D}$ together with a metric d. The metric (or distance) is a map $d : \mathbb{D} \times \mathbb{D} \to [0, \infty)$, which satisfies

(i) $d(x, y) = d(y, x)$;

(ii) $d(x, z) \leq d(x, y) + d(y, z)$;

(iii) $d(x, y) = 0$ if and only if $x = y$.

A weak version of metric d is called semimetric if it only satisfies (i) and (ii).

Some metrics can be defined in $\mathcal{F}$, a class of measurable functions.

- Uniform or supreme metric:

$$||f||_\infty = \sup_{x \in \Omega} |f(x)| \quad \forall f \in \mathcal{F}$$

The well-known Kolmogorov-Smirnov (KS) statistics measures the distance of two distribution functions in the supreme metric:

$$||F_1 - F_2||_\infty = \sup_{t \in R} |F_1(t) - F_2(t)|$$

- L_r-norm:

$$||f||_{r,P} = \{P(|f|^r)\}^{\frac{1}{r}} = \left\{ \int_\Omega |f(x)|^r dP(x) \right\}^{\frac{1}{r}} \quad \forall f \in \mathcal{F}$$

When $r = 2$, the L_2-norm of a vector is also known as Euclidean norm. When $r = 1$, the L_1-norm can be interested as the "area under the curve" (AUC) when P if chosen to be the probability distribution of x.

Note, the definitions of these metrics are crucial for the construction of the test statistics of nonparametric tests proposed in Chapter 5.

Definition 3.3.2. A **stochastic process** is a collection of measurable random variables $X = \{X_t, t \in T\}$ on the probability space $(\Omega, \mathcal{A}, P)$, indexed by a set T.

The realization of X_t, a function of t, is called a sample path. The sample paths usually lie in the metric space $\ell^\infty(\mathcal{F})$ usually is defined as (in the use of uniform metric),

$$\ell^\infty(\mathcal{F}) = \left\{ \mu : \sup_{f \in \mathcal{F}} |\mu(f)| < \infty \right\},$$

where $\mathcal{F}$ is a class of measurable functions $f : \Omega \to R$. $\ell^\infty(\mathcal{F})$ is a collection of all bounded functionals. The very often used limiting process in $\ell^\infty(\mathcal{F})$ might be a Gaussian process.

Definition 3.3.3. A **Gaussian process** is a stochastic process $\{Z_t, t \in T\}$ where for every finite $T_k \subset T$, $\{Z_t, t \in T_k\}$ is multivariate normal distributed, and all sample paths are uniformly ρ-continuous with respect to some semimetric.

Weak convergence is a form of convergence that is pertinent to sample paths of stochastic processes. Weak convergence is usually denoted by $X_n \rightsquigarrow X$. For example, it might be of interest in showing some empirical process converges weakly

to a Gaussian process. For every $\underline{t} = (t_1, \cdots, t_k)$,

$$\sqrt{n} \begin{pmatrix} F_n(t_1) - F(t_1) \\ \cdots \\ F_n(t_k) - F(t_k) \end{pmatrix} \xrightarrow{D} MVN_k \left(0, \Sigma_F(\underline{t})\right),$$

where $\Sigma_F(\underline{t})$ is the variance-covariance matrix. If $\sqrt{n}(F_n - F)$ is viewed as a stochastic process of a sample, we can write $\sqrt{n}(F_n - F) \rightsquigarrow \mathbb{G}_F(\Sigma_F)$, where $\mathbb{G}_F(\Sigma_F)$ denotes a Gaussian process with mean zero and some variance-covariance matrix Σ_F.

3.3.2 Empirical Process, Glivenko-Cantelli and Donsker Results

Definition 3.3.4. Let $X_1, \cdots, X_n$ be a random sample generated from a probability space $(\Omega, \mathcal{A}, P)$. The **empirical distribution** is the discrete uniform measure on the observations, i.e.,

$$\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i} = \frac{1}{n} \sum_{i=1}^{n} 1 \left[X = X_i\right]$$

Definition 3.3.5. Let $X_1, \cdots, X_n$ be an i.i.d. sample from a probability measure P on an arbitrary sample space $\mathcal{X}$. Let $\mathcal{F}$ be a class of measurable functions $f : \mathcal{X} \to \mathbb{R}$. Define the empirical process as $\{\mathbb{P}_n f, f \in \mathcal{F}\}$, where $\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}$ is the **empirical measure**. More specifically, the **empirical process** is indexed by $\mathcal{F}$

$$\left\{ \mathbb{P}_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i), f \in \mathcal{F} \right\}.$$

The **centralized empirical process** indexed by $\mathcal{F}$ can be defined as

$$\left\{ \mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n f - Pf) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(f(X_i) - E_P f(X_1) \right), f \in \mathcal{F} \right\}.$$

In EPT, there are requirements for the class of functions $\mathcal{F}$ under the study to investigate the asymptotic properties of the estimator in $\mathcal{F}$. Namely, the requirements for $\mathcal{F}$ to be Glivenko-Cantelli or/and Donsker classes are essential.

Definition 3.3.6. A class of measurable functions $f : \Omega \to R$, $\mathcal{F}$, is called **P-Glivenko-Cantelli** if

$$||\mathbb{P}_n - P||_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \xrightarrow{as} 0.$$

Definition 3.3.7. A class of measurable functions $f : \Omega \to R$, $\mathcal{F}$, is called **P-Donsker** if the sequence of centralized empirical processes $\{\mathbb{G}_n f : f \in \mathcal{F}\}$ converges to a Gaussian process in distribution in the space $l^\infty(\mathcal{F})$, i.e.,

$$\mathbb{G}_n(\cdot) \rightsquigarrow \mathbb{G}(\Sigma_P)(\cdot) \quad \text{in} \quad l^\infty(\mathcal{F}),$$

where $\mathbb{G}(\Sigma_P)$ is the Gaussian process with mean zero and variance-covariance matrix Σ_P given by

$$\Sigma_P(i,j) = P(f_i f_j) - (Pf_i)(Pf_j).$$

Glivenko-Cantelli and Donsker classes facilitate an actionable procedure to study the asymptotic behavior of empirical processes. The following empirical distribution function is one of the simplest examples.

Example 3.3.1. Let $X_1, \cdots, X_n$ be i.i.d. random variables with distribution F on the real line, where $F(t) = P\{1[X \leq t]\}$. The **empirical distribution function** is defined as

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1\left[X_i \leq t\right] = \mathbb{P}_n 1_{[X \leq t]}$$

By law of large numbers and multivariate central limit theorem, it immediately results in $||F_n - F||_\infty \xrightarrow{as} 0$ and $\sqrt{n}(F_n - F)$ converges weakly to a tight zero mean Gaussian process. Hence the class $\mathcal{F} = \{1\{x \leq t\}, t \in R\}$ is P-Glivenko-Cantelli and P-Donsker.

Theorems that can verify a much broader class of functions for being Glivenko-Cantelli or Donsker are needed. However it is challenging to prove from the pointwise convergence to uniform convergence. Hence it is of great importance to define the size of class of measurable functions $\mathcal{F}$, since whether or not $\mathcal{F}$ is Glivenko-Cantelli or Donsker class depends on the size of the class.

Definition 3.3.8. Let $L_r(P)$ denotes the collection of all functions f satisfies $\|f\|_{r,P} < \infty$. For a class of measurable functions $\mathcal{F}$,

- The ϵ**-ball** of function g in $L_r(P)$ is $\left\{f : \|f - g\|_{r,P} \leq \epsilon, f \in \mathcal{F}\right\}$.
- The **covering number** $N(\epsilon, \mathcal{F}, L_r(P))$ is the minimum number of ϵ-balls needed to cover the set $\mathcal{F}$.
- The **entropy** (without bracketing) is the logarithm of the defined covering number above, $\log(N(\epsilon, \mathcal{F}, L_r(P))$.
- The **uniform covering number** is

$$\sup_P N(\epsilon \|F\|_{r,P}, \mathcal{F}, L_r(P)),$$

where $F(x) = \sup_{f \in \mathcal{F}} |f(x)|$ is an envelop for $\mathcal{F}$.

- The **uniform entropy integral**

$$J(\delta, \mathcal{F}, L_r) = \int_0^\delta \sqrt{\log \sup_P N(\epsilon \|F\|_{r,P}, \mathcal{F}, L_r(P))} d\epsilon$$

is constructed based on the uniform covering number.

Definition 3.3.9. Let $L_r(P)$ denotes the collection of all functions f satisfies $\|f\|_{r,P} < \infty$. For a class of measurable functions $\mathcal{F}$,

- The **ϵ-bracket** of function f in $L_r(P)$ is the set of functions

$$\left\{ f : u_1 \leq f \leq u_2, \|u_1 - u_2\|_{r,P} \leq \epsilon \right\}$$

- The **bracketing number** $N_{[\,]}(\epsilon, \mathcal{F}, \|\cdot\|_{r,P})$ is the minimum number of ϵ-brackets needed to cover the set $\mathcal{F}$.
- The **entropy with bracketing** is the logarithm of the defined bracketing number above $\log(N_{[\,]}(\epsilon, \mathcal{F}, \|\cdot\|_{r,P}))$.
- The **bracketing entropy integray** is

$$J_{[\,]}(\delta, \mathcal{F}, L_r(P)) = \int_0^\delta \sqrt{\log N_{[\,]}(\epsilon, \mathcal{F}, L_r(P))}\, d\epsilon.$$

Remark 3.3.1. It was shown in Lemma 9.18 of Kosorok (2008) that for any norm $\|\cdot\|$ on $\mathcal{F}$, $N(\epsilon, \mathcal{F}, \|\cdot\|) \leq N_{[\,]}(\epsilon, \mathcal{F}, \|\cdot\|)$, for all $\epsilon > 0$, which gives the relationship between the covering number and the bracketing number.

Remark 3.3.2. In Lemma 9.22 of Kosorok (2008), it shows for any norm $\|\cdot\|$ dominated by $\|\cdot\|_\infty$ on $\mathcal{F}$, $\log N_{[\,]}(2\epsilon, \mathcal{F}, \|\cdot\|) \leq \log N(\epsilon, \mathcal{F}, \|\cdot\|)$, which relates the entropy (without bracketing) and entropy with bracketing.

The following theorems provide means to prove Glivenko-Cantelli and Donsker classes with covering number or bracketing number.

Theorem 3.3.1. (**Glivenko-Cantelli Theorem**; Kosorok (2008) Theorem 2.4) Let $\mathcal{F}$ be a class of measurable functions which satisfies

$$\sup_P N(\epsilon \|F\|_{1,P}, \mathcal{F}, L_1(P)) < \infty$$

for every $\epsilon > 0$. If $P\{F\} < \infty$, then $\mathcal{F}$ is P-Glivenko-Cantelli.

Theorem 3.3.2. (**Glivenko-Cantelli Theorem**; Kosorok (2008) Theorem 2.2) Let $\mathcal{F}$ be a class of measurable functions which satisfies

$$N_{[\,]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$$

for every $\epsilon < 0$, then $\mathcal{F}$ is P-Glivenko-Cantelli.

Theorem 3.3.3. (**Donsker Theorem**; Kosorok (2008) Theorem 2.5) Let $\mathcal{F}$ be a class of measurable functions which satisfies

$$J(1, \mathcal{F}, L_2(P)) < \infty.$$

If $P\{F^2\} < \infty$, then $\mathcal{F}$ is P-Donsker.

Theorem 3.3.4. (**Donsker Theorem**; Kosorok (2008) Theorem 2.3) Let $\mathcal{F}$ be a class of measurable functions which satisfies

$$J_{[\,]}(\infty, \mathcal{F}, L_2(P)) < \infty,$$

then $\mathcal{F}$ is P-Donsker.

Remark 3.3.3. Lemma 8.17 in Kosorok (2008) states Donsker classes are automatically Glivenko-Cantelli classes.

Many theorems have been developed to show Glivenko-Cantelli and Donsker classes. Since in Chapter 5 the asymptotic distributions of the test statistics of the proposed nonparametric tests are of interest, next we introduce some theorems that will be of use and especially focus on Donsker results.

Definition 3.3.10. (Kosorok (2008) Section 3.1 Page 35) Consider estimation of the parameter $\varphi(P) \in \mathbb{R}^k$ for the semiparametric model $\mathcal{P}$. For any estimator T_n of $\varphi(P)$, if

$$\sqrt{n}(T_n - \varphi(P)) = \sqrt{n}\,\mathbb{P}_n\breve{\varphi}_P + o_p(1),$$

where $o_p(1)$ denotes a quantity going to zero in probability, then $\breve{\varphi}_P$ is an **influence function** for $\varphi(P)$,and T_n is asymptotically linear.

Remark 3.3.4. Let $\tilde{\varphi}(P)$ denote the efficient influence function, $\sqrt{n}\,\mathbb{P}_n\tilde{\varphi}_P$ converges weakly to a tight zero mean Gaussian process G. Kosorok (2008) page 35-37 states that for any estimator T_n of $\varphi(P)$, $\sqrt{n}(T_n - \varphi(P))$ converges weakly to a convolution of a Gaussian process G and an independent noise process. That is to say, for an inefficient estimator, the noise process is always non-negligible and added to the limiting distribution of efficient estimator.

Remark 3.3.5. Theorem 18.8 of Kosorok (2008) provides a way to establish efficiency of estimator T_n. If T_n is asymptotically linear with influence function $\breve{\varphi}_P$, and it is in a Donsker class, then T_n weakly converges to a tight Gaussian process with mean zero and some covariance, and it is efficient.

Theorem 3.3.5. (Kosorok (2008) Lemma 4.1) For $-\infty < a < b < \infty$, let $\{X(t), t \in [a,b]\}$ be a monotone cadlag or caglad stochastic process with

$$P\left[|X(a)| \vee |X(b)|\right]^2 < \infty,$$

then X is P-Donsker, where cadlag implies right continuous with left limit.

Remark 3.3.6. Theorem 3.3.5 discusses being Donsker in the context of a process instead of classes of function as usual. Kosorok (2008) states for stochastic process $\{X(t) : t \in T\}$, $\sqrt{n}(\mathbb{P}_n - P)X$ converges weakly in $\ell^\infty(T)$ to a tight zero mean Gaussian process if and only if $\mathcal{F} = \{f_t : t \in T\}$ is P-Donsker, where for any $x \in \mathcal{X}$ and $t \in T$, $f_t(x) = x(t)$.

Example 3.3.2. Since counting process $N(t)$ and at-risk process $Y(t)$ both satisfy the conditions of Theorem 3.3.5, they are both P-Donsker in $\ell^\infty([0,\tau])$.

Theorem 3.3.6. (Bakoyannis (2020) Lemma 1) Let $h(t)$ be a fixed and uniformly bounded function on $[0,\tau]$ and $N(t)$ be a counting process with $P[N(\tau)]^2 < \infty$, then the class of functions

$$\mathcal{F}_1 = \left\{ \int_s^t h(u)dN(u) : s \in [0,\tau], t \in [s,t] \right\}$$

is P-Donsker.

Theorem 3.3.7. (Bakoyannis (2020) Lemma 2) Let $h(t)$ be a fixed and uniformly bounded function on $[0,\tau]$ and $Y(t)$ be an at-risk process, and $A(t)$ a continuous

cumulative transition intensity function on $[0, \tau]$, then the class of functions

$$\mathcal{F}_2 = \left\{ \int_s^t h(u)Y(u)dA(u) : s \in [0,\tau], t \in [s,t] \right\}$$

is P-Donsker.

The Donsker preservation results are very useful to build Donsker classes from other Donsker classes.

Theorem 3.3.8. (**Donsker Preservation**; Corollary 9.32 in Kosorok (2008)) Let $\mathcal{F}$ and $\mathcal{G}$ be P-Donsker classes, then

(i) $\mathcal{F} \cup \mathcal{G}$ and $\mathcal{F} + \mathcal{G}$ are P-Donsker;

(ii) If $\|P\|_{\mathcal{F}\cup\mathcal{G}} < \infty$, then the classes of infima and suprema, $\mathcal{F} \wedge \mathcal{G}$ and $\mathcal{F} \vee \mathcal{G}$ are both P-Donsker.

(iii) If $\mathcal{F}$ and $\mathcal{G}$ are both uniformly bounded, $\mathcal{F} \cdot \mathcal{G}$ is P-Donsker.

(iv) If $\bar{R}$ is the union of the ranges of unctions in $\mathcal{F}$ and $\psi : \bar{R} \to \mathbb{R}$ is Lipschitz continuous with $\|\psi(f)\|_{P,2} < \infty$ for at least one $f \in \mathcal{F}$, then $\psi(\mathcal{F})$ is P-Donsker.

(v) If $\|P\|_{\mathcal{F}} < \infty$ and g is a uniformly bounded, measurable function, then $\mathcal{F} \cdot g$ is P-Donsker.

Example 3.3.3. Martingale process is $M(t) = N(t) - \int_0^t Y(t)dA(t)$. By the Theorem 3.3.5, Theorem 3.3.7, and Donsker preservation results (Theorem 3.3.8), the martingale process $M(t)$ is P-Donsker.

Theorem 3.3.9. (Bakoyannis et al. (2019) Lemma 1) Let $h(t)$ be a fixed uniformly bounded function on $[0, \tau]$ and $\phi(t)$ a non-decreasing random function on $[0, \tau]$ that belongs to the P-Donsker class Φ. Then the class of functions

$$\mathcal{F}_3 = \left\{ \int_0^t h(u)d\phi(u) : t \in [0,\tau] \right\}$$

is P-Donsker.

Example 3.3.4. (**Nelson-Aalen estimator**) Nelson-Aalen estimator for cumulative hazard function in the empirical process notation can be written as

$$\hat{A}(t) = \int_0^t \frac{\mathbb{P}_n dN(u)}{\mathbb{P}_n Y(u)}$$

We want to show $\sqrt{n}(\hat{A} - A)$ converges weakly in $D[0, \tau]$ to a tight, mean zero Gaussian process $\mathbb{G}$ with mean 0 and variance-covariance matrix $\Sigma_A = P(\varphi(s)\varphi(t))$, where $\varphi(t) = \int_0^t \frac{dM(u)}{PY(u)}$, and $M(t) = N(t) - \int_0^t Y(u)dA(u)$.

$$\begin{aligned}
\sqrt{n}(\hat{A}(t) - A(t)) &= \sqrt{n}\left(\int_0^t \frac{\mathbb{P}_n dN(u)}{\mathbb{P}_n Y(u)} - \int_0^t \frac{PdN(u)}{PY(u)}\right) \\
&= \sqrt{n}\left(\int_0^t \frac{(\mathbb{P}_n - P)dN(u)}{\mathbb{P}_n Y(u)} + \int_0^t \frac{(\mathbb{P}_n - P)Y(u)}{PY(u)} \cdot \frac{PdN(u)}{\mathbb{P}_n Y(u)}\right) \\
&= B_n(t) + C_n(t) \qquad (3.5)
\end{aligned}$$

It is not hard to verity that

$$\begin{aligned}
B_n(t) &= \sqrt{n}\left(\int_0^t \frac{(\mathbb{P}_n - P)dN(u)}{\mathbb{P}_n Y(u)}\right) \\
&= \sqrt{n}\left(\int_0^t (\mathbb{P}_n - P)dN(u)\left(\frac{1}{PY(u)} + \frac{1}{\mathbb{P}_n Y(u)} - \frac{1}{PY(u)}\right)\right) \\
&= \sqrt{n}\left(\int_0^t \frac{(\mathbb{P}_n - P)dN(u)}{PY(u)}\right) + o_p(1)
\end{aligned}$$

$$
\begin{aligned}
C_n(t) &= \sqrt{n}\left(\int_0^t \frac{(\mathbb{P}_n - P)Y(u)}{PY(u)} \cdot \frac{PdN(u)}{\mathbb{P}_n Y(u)}\right) \\
&= \sqrt{n}\left(\int_0^t \frac{(\mathbb{P}_n - P)Y(u)}{PY(u)} \cdot \left(\frac{\mathbb{P}_n dN(u)}{\mathbb{P}_n Y(u)} + \frac{PdN(u)}{\mathbb{P}_n Y(u)} - \frac{\mathbb{P}_n dN(u)}{\mathbb{P}_n Y(u)}\right)\right) \\
&= \sqrt{n}\left(\int_0^t \frac{(\mathbb{P}_n - P)Y(u)}{PY(u)} dA(u)\right) + o_p(1)
\end{aligned}
$$

Plug $B_n(t)$ and $C_n(t)$ back into equation 3.5, we have

$$
\begin{aligned}
\sqrt{n}(\hat{A}(t) - A(t)) &= \sqrt{n}\left(\int_0^t \frac{(\mathbb{P}_n - P)}{PY(u)} \left(dN(u) - Y(u)dA(u)\right)\right) + o_p(1) \\
&= \sqrt{n}\left(\int_0^t \frac{(\mathbb{P}_n - P)}{PY(u)} dM(u)\right) + o_p(1) \\
&= \sqrt{n}\,\mathbb{P}_n\left(\int_0^t \frac{dM(u)}{PY(u)}\right) + o_p(1).
\end{aligned}
$$

Thus, $\hat{A}(t)$ is asymptotically linear with influence function

$$
\varphi(t) = \int_0^t \frac{dM(u)}{PY(u)}
$$

Since $\varphi(t) = \int_0^t \frac{dM(u)}{PY(u)} = \int_0^t \frac{dN(u) - Y(u)dA(u)}{PY(u)}$, by Theorem 3.3.6 and Theorem 3.3.7 it can be shown that $\{\varphi(t) : t \in T\}$ is P-Donsker. Hence

$$
\sqrt{n}(\hat{A} - A) \rightsquigarrow \mathbb{G}_A(\Sigma_A).
$$

where the variance-covariance functions are

$$
\Sigma_A(s,t) = P(\varphi(s)\varphi(t)) = P\left(\int_0^s \frac{dM(u)}{PY(u)}\right)\left(\int_0^t \frac{dM(u)}{PY(u)}\right),
$$

for any $s, t \in [0, \tau]$. The consistent estimator for $\Sigma_A(s,t)$ is

$$\begin{aligned}\hat{\Sigma}_A(s,t) &= \mathbb{P}_n \left(\int_0^s \frac{d\hat{M}(u)}{\mathbb{P}_n Y(u)} \right) \left(\int_0^t \frac{d\hat{M}(u)}{\mathbb{P}_n Y(u)} \right) \\ &= \mathbb{P}_n \left(\int_0^s \frac{dN(u) - Y(u)d\hat{A}(u)}{\mathbb{P}_n Y(u)} \right) \left(\int_0^t \frac{dN(u) - Y(u)d\hat{A}(u)}{\mathbb{P}_n Y(u)} \right).\end{aligned}$$

The following theorem is useful for establishing weak convergence.

Theorem 3.3.10. (Kosorok (2008) Proposition 7.27) Let $A_n, B_n \in D[a,b]$ be stochastic processes where $An \rightsquigarrow A$ and $B_n \xrightarrow{P} B$ in $D[a,b]$. A is bounded with continuous sample paths. B is fixed, and B_n and B have total variation bounded by $M < \infty$. Hence $\int_s^{\cdot} A_n(t)dB_n(t) \rightsquigarrow \int_s^{\cdot} A(t)dB(t)$.

The following theorem presents a property of a Donsker class in order to conclude convergence in probability to zero.

Theorem 3.3.11. (Kosorok (2008) Lemma 4.2) Let $B_n \in D[a,b]$ and $A_n \in l^\infty([a,b])$ be either cadlag or caglad, and assume $\sup_{t \in (a,b]} |A_n(t)| \xrightarrow{P} 0$, A_n has uniformly bounded total variation, and B_n converges weakly to a tight, mean zero process with sample paths in $D[a,b]$. Then

$$\int_a^b A_n(s)dB_n(s) \xrightarrow{P} 0.$$

3.3.3 Multiplier Central Limit Theorem (MCLT)

The unconditional and conditional multiplier central limit theorems (MCLT) provide techniques to bootstrap empirical process. In this section, they are studied as they are useful to obtain empirical power levels for nonparametric tests proposed in Chapter 5. First the convergence properties of the multiplier processes unconditional on the data are presented.

Theorem 3.3.12. (**Unconditional MCLT**; Van Der Vaart and Wellner (1996) Theorem 2.9.2 and Kosorok (2008) Theorem 10.1) Let $\mathcal{F}$ be a class of measurable functions, and $\xi_1, \cdots, \xi_n$ be i.i.d. random variables with mean zero, variance 1, and $||\xi||_{2,1} = \int_0^\infty \sqrt{P(|\xi| > x)}dx < \infty$, which are independent of the sample data $X_1, \cdots, X_n$. Let $\mathbb{G}'_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i(\delta_{X_i} - P)$, where δ_{X_i} is the probability measure which assigns a mass of 1 to X_i. Then the following are equivalent:

(i) $\mathcal{F}$ is P-Donsker;

(ii) $\mathbb{G}'_n$ converges weakly to a tight process in $l^\infty(\mathcal{F})$;

Remark 3.3.7. Let $Z_i = \delta_{X_i} - P$, by empirical central limit theorem $\frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i$ converges weakly to a tight zero mean Gaussian process in $\ell^\infty(\mathcal{F})$. The unconditional MCLT states that $\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i Z_i$ converges to a tight zero mean Gaussian process if and only if $\mathcal{F}$ is Donsker.

The conditional MCLT prove the convergence properties of the multiplier processes when conditioning on the data. Compared the unconditional MCLT, the conditional MCLT requires slightly stronger conditions. Before the conditinal MCLT is presented, Theorem 3.3.13 is given, since it is a conditional multiplier central limit theorem for i.i.d. data in Euclidean space as a direct consequence of using Lindeberg central limit theorem for easy understanding.

Theorem 3.3.13. (Van Der Vaart and Wellner (1996) Lemma 2.9.5 and Kosorok (2008) Lemma 10.5) Let $Z_1, \cdots, Z_n$ be iid Euclidean random vectors, with $E(Z) = 0$ and $E||Z||^2 < \infty$, independent of the iid sequence of real random variables $\xi_1, \cdots, \xi_n$ with $E(\xi) = 0$ and $E\xi^2 = 1$. Then conditional on $Z_1, Z_2, \cdots$, $\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i Z_i$ converges to $N(0, Cov(Z))$ for almost all sequences $Z_1, Z_2, \cdots$.

The above Theorem 3.3.13 proved the marginal convergence in the conditional MCLT in the Enclidean space. The following theorem proves the convergence in $\ell^\infty(\mathcal{F})$, where stronger conditions need to be satisfied.

Theorem 3.3.14. (**Conditional MCLT**) Let $\mathcal{F}$ be a class of measurable functions, and let $\xi_1, \cdots, \xi_n$ be i.i.d. random variables with mean zero, variance 1, and $||\xi||_{2,1} = \int_0^\infty \sqrt{P(|\xi| > x)}dx < \infty$, independent of the sample data $X_i, \cdots, X_n$. Let $\mathbb{G}'_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i(\delta_{X_i} - P)$, where δ_{X_i} is the probability measure which assigns a mass of 1 to X_i. Then the following are equivalent:

(i) $\mathcal{F}$ is P-Donsker;

(ii) $\sup_{h \in BL_1} |E_\xi h(\mathbb{G}'_n) - Eh(\mathbb{G})| \xrightarrow{P} 0$ in outer probability, and the sequence $\mathbb{G}'_n$ is asymptotically measurable, where E_ξ denotes taking the expectation conditional on the data, $X_i, \cdots, X_n$ and BL_1 is the set of all functions $h : \ell^\infty(\mathcal{F}) \to [0, 1]$ satisfy $|h(z_1) - h(z_2)| \leq \{z_1 - z_2\}_\mathcal{F}$ for every z_1, z_2.

Remark 3.3.8. Let $Z_i = \delta_{X_i} - P$, the conditional MCLT states that conditional on data, $\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i Z_i$ converges to a tight zero mean Gaussian process if and only if $\mathcal{F}$ is Donsker.

3.4 Concluding Remarks

This chapter first reviewed asymptotic theories of MLE and three inference procedures, including likelihood ratio test, Wald test, and score test when there are nuisance parameters in the model. Then the empirical process was also introduced to establish the asymptotic properties of parameters in the non-/semi-parametric models. There are two take-home messages.

- For score test, only the MLE under the null hypothesis instead of the alternative hypothesis needs to be computed when there are nuisance parameters. This is an advantage of score test compared to likelihood ratio test and Wald test especially when the computation of MLE under the full model (alternative hypothesis) is problematic.
- The likelihood techniques in parametric models cannot be easily extended to the settings of non-/semi-parametric models. In this case, the EPT becomes a powerful tool for non-/semi-parametric inference.

The asymptotic theory about the score test will be utilized in the next chapter. The EPT, specifically Donsker class results introduced in this chapter will be heavily used in Chapter 5 for proving the asymptotic properties of the proposed nonparametric tests.

Chapter 4

Nonparametric Maximum Likelihood Estimation (NPMLE) for Semi-Competing Risks Data under GFCMM

4.1 Introduction

Semi-competing risks data have been actively studied in biomedical studies when modeling both time to disease progression and time to death is of interest ever since Fine et al. (2001) first named semi-competing risks data. Fine et al. (2001) proposed the use of Clayton copula (Nelsen, 2007) to study the joint survival distribution and the association between two event times in the upper wedge. After that, their work was either extended to other copula models or more complex data settings (see Section 2.1 for details)

Since semi-competing risks data are essentially in the same structure as illness-death data within multi-state model framework, the abundant methodology of multi-state modeling based on transition intensities (Andersen and Keiding, 2002; Beyersmann et al., 2011; Commenges, 1999; Cook and Lawless, 2018; Hougaard, 1999; Klein et al., 2016; Meira-Machado et al., 2009; Van Den Hout, 2016) can be applied to analyze semi-competing risks data. For example in the recent literature of multi-state modeling, Mandel and Fluss (2009) developed nonparametric estimation methodology for the probability of the non-terminal event under cross-sectional settings. Yu et al. (2010) proposed an estimator for the risk of the non-terminal event and its impact on overall survival using the multiple imputation approach. Zeng et al. (2011) estimated

treatment effects in consideration of treatment switching. Hu and Tsodikov (2014) considered missing non-terminal event status under a joint modeling approach.

While the Markov model appears to be widely accepted for illness-death modeling (Frydman, 1995; Frydman et al., 2013; Frydman and Szarek, 2009; Harezlak et al., 2003), Xu et al. (2010) proposed a shared gamma-frailty conditional Markov model (GFCMM) to analyze semi-competing risks data. They pointed out that the restricted GFCMM, in which the hazard of death conditional on the latent gamma frailty variable is assumed to be the same regardless of whether an individual experiences the illness under study before death, is identical to the Clayton copula model proposed by Fine et al. (2001). This finding bridges the Markov modeling methodology for illness-death data with the Copula modeling methodology for semi-competing risks data. Xu et al. (2010) naturally extended the GFCMM to unrestricted cases that offers more flexibility in modeling semi-competing risks data compared to Copula models. Xu et al. (2010) proposed a nonparametric maximum likelihood estimation (NPMLE) procedure to estimate the parameters of the unrestricted GFCMM. Bayesian methods for the frailty-based Markov model have also been extensively adopted recently in analyzing semi-competing risks data, see for example, Chapple et al. (2017), Han et al. (2014), and Lee et al. (2016, 2015).

In this chapter, an EM-algorithm to compute the NPMLE for the GFCMM (Section 4.2.1) is developed. This EM-algorithm is more numerically stable than the Newton-Raphson algorithm proposed by Xu et al. (2010). Through simulation studies, we uncover the issue in the NPMLE for the unrestricted GFCMM, that the maximizer may occur at the boundary of feasible parameter space resulting in biased estimation (Section 4.2.2). We therefore alert researchers in this field to be cautious

when using the unrestricted GFCMM. Moreover, we provide a practical guideline for using the GFCMM in the analysis of semi-competing risk data that includes (i) the score test to assess if the hypothesis that the restricted model, which does not exhibit estimation problems, holds under a proportional hazards assumption, and (ii) a graphical approach to evaluate whether the unrestricted model yields nonparametric estimation with substantial bias for cases where the test provides a statistical significant result against the restricted model (Section 4.2.3). Finally, this practical guideline is applied to the Indianapolis-Ibadan Dementia Project (IIDP) data as an illustration to explore whether dementia occurrence changes mortality risk (Section 4.3).

4.2 NPMLE for Semi-Competing Risks Data

4.2.1 NPMLE for GFCMM

As it was introduced in Chapter 2, a generic observation of semi-competing risks data can be denoted as $D = (Y_1, Y_2, \delta_1, \delta_2)$, where $Y_2 = T_2 \wedge C, \delta_2 = I(T_2 \leq C), Y_1 = T_1 \wedge Y_2$ and $\delta_1 = I(T_1 \leq Y_2)$, and T_1 and T_2 denote the non-terminal and terminal event times of an individual under study, and C is non-informative administrative censoring time, which is independent from both event times. Assume a study that consists of n i.i.d. copies of D, $D_1, D_2, \cdots, D_n$ to form the observed data denoted by $\underline{D} = \{D_i : i = 1, 2, \cdots, n\}$. Let $\underline{\gamma} = (\gamma_1, \gamma_2, \cdots, \gamma_n)$ be a random sample drawn from $Gamma(\theta^{-1}, \theta^{-1})$, where γ_i is the shared gamma frailty variable associated with the ith individual. Under the conditional Markov assumption proposed by Xu et al. (2010), the likelihood of the augmented data (observed data $\underline{D}$ plus the latent frailty

variables $\underline{\gamma}$) is given by

$$L(\underline{\Lambda}_0, \theta; \underline{D}, \underline{\gamma}) = \prod_{i=1}^{n} \Big\{ \gamma_i^{\delta_{i1}+\delta_{i2}} \lambda_{01}(Y_{i1})^{\delta_{i1}} \lambda_{02}(Y_{i2})^{(1-\delta_{i1})\delta_{i2}} \lambda_{03}(Y_{i2})^{\delta_{i1}\delta_{i2}} \times e^{-\gamma_i[\Lambda_{01}(Y_{i1})+\Lambda_{02}(Y_{i1})+\delta_{i1}\Lambda_{03}(Y_{i1},Y_{i2})]} \cdot \frac{\theta^{-1/\theta}}{\Gamma(1/\theta)} \gamma_i^{1/\theta-1} e^{-\gamma_i/\theta} \Big\}.$$

Details for the derivation of this likelihood are given in Section 2.4.2.

Our interest lies in the nonparametric estimation of the baseline cumulative hazards $\underline{\Lambda}_0 = (\Lambda_{01}, \Lambda_{02}, \Lambda_{03})$, while treating θ as a nuisance parameter. Here two computing techniques, the Newton-Raphson algorithm that Xu et al. (2010) utilized and the EM algorithm that we proposed for GFCMM are presented.

Newton-Raphson Algorithm

By integrating out the gamma-frailty variable $\underline{\gamma}$, Xu et al. (2010) established the likelihood for observed data $\underline{D}$

$$L(\underline{\Lambda}_0, \theta; \underline{D}) = \prod_{i=1}^{n} \Big\{ \lambda_{01}(Y_{i1})^{\delta_{i1}} \lambda_{02}(Y_{i2})^{(1-\delta_{i1})\delta_{i2}} \lambda_{03}(Y_{i2})^{\delta_{i1}\delta_{i2}} (1+\theta)^{\delta_{i1}\delta_{i2}} \times (1+\theta[\Lambda_{01}(Y_{i1}) + \Lambda_{02}(Y_{i1}) + \delta_{i1}\Lambda_{03}(Y_{i1}, Y_{i2})])^{-1/\theta-\delta_{i1}-\delta_{i2}} \Big\}. \quad (4.1)$$

As the convention, letting the NPMLEs of Λ_{0i} $(i = 1, 2, 3)$ to be monotone nondecreasing step functions, where increments or jumps only at corresponding distinct event times, Xu et al. (2010) proposed a Newton-Raphson algorithm to compute the NPMLEs. The score functions and second-order partial derivatives in Hessian matrix need to be derived. However the Hessian matrix in Newton-Raphson algorithm for this particular problem is not sparse, and its size increases linearly with

the sample size. Therefore, it is expected that this algorithm will not be numerically stable especially for large sample size.

EM-Algorithm

Viewing the special structure of the likelihood for the augmented data, an EM-algorithm for computing the NPMLEs is proposed. For the E-step, the expected log conditional likelihood of the augmented data given the observed data, current estimates $\underline{\tilde{\Lambda}}_0(\cdot) = (\tilde{\Lambda}_{01}(\cdot), \tilde{\Lambda}_{02}(\cdot), \tilde{\Lambda}_{03}(\cdot))$, and $\tilde{\theta}$ is

$$\begin{aligned} Q(\underline{\Lambda}_0, \theta | \underline{D}) &= E\left(l(\underline{\Lambda}_0, \theta; \underline{D}, \underline{\gamma}) | \underline{D}, \underline{\tilde{\Lambda}}_0, \tilde{\theta} \right) \\ &= Q_1(\Lambda_{01} | \underline{D}, \underline{\tilde{\Lambda}}_0, \tilde{\theta}) + Q_2(\Lambda_{02} | \underline{D}, \underline{\tilde{\Lambda}}_0, \tilde{\theta}) + Q_3(\Lambda_{03} | \underline{D}, \underline{\tilde{\Lambda}}_0, \tilde{\theta}) + Q_4(\theta | \underline{D}, \underline{\tilde{\Lambda}}_0, \tilde{\theta}) \end{aligned}$$

where

$$Q_1(\Lambda_{01} | \underline{D}, \underline{\tilde{\Lambda}}_0, \tilde{\theta}) = \sum_{i=1}^{n} \left\{ \delta_{i1} \bar{\gamma}_i^{\log} + \delta_{i1} \log \lambda_{01}(Y_{i1}) - \bar{\gamma}_i \Lambda_{01}(Y_{i1}) \right\} \tag{4.2}$$

$$Q_2(\Lambda_{02} | \underline{D}, \underline{\tilde{\Lambda}}_0, \tilde{\theta}) = \sum_{i=1}^{n} \left\{ \delta_{i2} \bar{\gamma}_i^{\log} + (1 - \delta_{i1}) \delta_{i2} \log \lambda_{02}(Y_{i2}) - \bar{\gamma}_i \Lambda_{02}(Y_{i1}) \right\} \tag{4.3}$$

$$Q_3(\Lambda_{03} | \underline{D}, \underline{\tilde{\Lambda}}_0, \tilde{\theta}) = \sum_{i=1}^{n} \left\{ \delta_{i1} \delta_{i2} \log \lambda_{03}(Y_{i2}) - \bar{\gamma}_i \delta_{i1} \left(\Lambda_{03}(Y_{i2}) - \Lambda_{03}(Y_{i1}) \right) \right\} \tag{4.4}$$

$$Q_4(\theta | \underline{D}, \underline{\tilde{\Lambda}}_0, \tilde{\theta}) = \sum_{i=1}^{n} \left\{ -\frac{1}{\theta} \log \theta + \left(\frac{1}{\theta} - 1 \right) \bar{\gamma}_i^{\log} - \frac{\bar{\gamma}_i}{\theta} - \log \Gamma\left(\frac{1}{\theta} \right) \right\} \tag{4.5}$$

with $\bar{\gamma}_i = E(\gamma_i | \underline{D}, \underline{\tilde{\Lambda}}_0, \tilde{\theta})$ and $\bar{\gamma}_i^{\log} = E(\log \gamma_i | \underline{D}, \underline{\tilde{\Lambda}}_0, \tilde{\theta})$.

A straightforward calculation yields that the conditional distribution of γ_i given $\underline{D}$ and the current estimates $(\tilde{\underline{\Lambda}}_0, \tilde{\theta})$ is

$$Gamma\left(\frac{1}{\tilde{\theta}} + \delta_{i1} + \delta_{i2}, \frac{1}{\tilde{\theta}} + \tilde{\Lambda}_{01}(Y_{i1}) + \tilde{\Lambda}_{02}(Y_{i1}) + \delta_{i1}\tilde{\Lambda}_{03}(Y_{i1}, Y_{i2})\right). \tag{4.6}$$

Hence this results in

$$\bar{\gamma_i} = \frac{\frac{1}{\tilde{\theta}} + \delta_{i1} + \delta_{i2}}{\frac{1}{\tilde{\theta}} + \tilde{\Lambda}_{01}(Y_{i1}) + \tilde{\Lambda}_{02}(Y_{i1}) + \delta_{i1}\tilde{\Lambda}_{03}(Y_{i1}, Y_{i2})} \tag{4.7}$$

and

$$\bar{\gamma_i}^{log} = digamma\left(\frac{1}{\tilde{\theta}} + \delta_{i1} + \delta_{i2}\right) - log\left(\frac{1}{\tilde{\theta}} + \tilde{\Lambda}_{01}(Y_{i1}) + \tilde{\Lambda}_{02}(Y_{i1}) + \delta_{i1}\tilde{\Lambda}_{03}(Y_{i1}, Y_{i2})\right), \tag{4.8}$$

for $i = 1, 2, \ldots, n$. Since $\Lambda_{01}, \Lambda_{02}, \Lambda_{03}$, and θ are disjoint in the expression of the "Q-function", the M-step is trivial which is the key consideration for the proposal of EM-algorithm to compute the NPMLEs.

Let $t_{1,1} < ... < t_{1,m}$ denote the m distinct observed non-terminal event times; $t_{2,1} < ... < t_{2,s}$ the s distinct observed terminal event times with no prior non-terminal events observed; and $t_{12,1} < \ldots < t_{12,r}$ the r distinct observed terminal event times with non-terminal event occurred. The NPMLEs of $\Lambda_{01}, \Lambda_{02}$, and Λ_{03} are conventionally defined as the step functions with the jumps only occurred at their corresponding observed even times with sizes $\underline{\lambda}_{01} = (\lambda_{01,1}, \cdots, \lambda_{01,m})$; $\underline{\lambda}_{02} = (\lambda_{02,1}, \cdots, \lambda_{02,s})$; and $\underline{\lambda}_{03} = (\lambda_{03,1}, \cdots, \lambda_{03,r})$, respectively. The M-step leads to the

explicit forms for updating the jump sizes $\underline{\lambda_0} = (\underline{\lambda}_{01}, \underline{\lambda}_{02}, \underline{\lambda}_{03})$ given by

$$\hat{\lambda}_{01,j} = \frac{d_{1,j}}{\sum_{i=1}^{n} \{\bar{\gamma}_i 1[Y_{i1} \geq t_{1,j}]\}} \tag{4.9}$$

$$\hat{\lambda}_{02,j} = \frac{d_{2,j}}{\sum_{i=1}^{n} \{\bar{\gamma}_i 1[Y_{i1} \geq t_{2,j}]\}} \tag{4.10}$$

$$\hat{\lambda}_{03,j} = \frac{d_{3,j}}{\sum_{i=1}^{n} \{\bar{\gamma}_i \delta_{i1} (1[Y_{i2} \geq t_{3,j}] - 1[Y_{i1} \geq t_{3,j}])\}} \tag{4.11}$$

where $\{d_{1,j} : j = 1, \cdots, m\}$, $\{d_{2,j} : j = 1, \cdots, s\}$, and $\{d_{3,j} : j = 1, \cdots, r\}$ are the number of events observed at their corresponding event times.

The M-step update of θ, $\hat{\theta}$ is simply the maximizer of $Q_4(\theta|\underline{D}, \underline{\tilde{\Lambda}}_0, \tilde{\theta})$ that can be calculated using the Newton-Raphosn algorithm described in Zhang and Jamshidian (2003). Let $\alpha = \frac{1}{\theta}$, we have

$$Q_4(\theta|\underline{D}, \underline{\tilde{\Lambda}}_0, \tilde{\theta})/n = \tilde{Q}_4(\theta|\underline{D}, \underline{\tilde{\Lambda}}_0, \tilde{\theta}) = \alpha \log \alpha - \log \Gamma(\alpha) + \alpha(\bar{\gamma}^{\log} - \bar{\gamma}).$$

Let $B = \frac{1}{n} \sum_{i=1}^{n} (\bar{\gamma_i}^{\log} - \bar{\gamma_i})$, we have

$$U(\alpha) = \frac{\partial \tilde{Q}_4(\theta|\underline{D}, \underline{\tilde{\Lambda}}_0, \tilde{\theta})}{\partial \alpha} = (\log \alpha - \frac{\partial \log \Gamma(\alpha)}{\partial \alpha}) + B + (\bar{\gamma}^{\log} - \bar{\gamma})$$

$$U'(\alpha) = \frac{\partial^2 \tilde{Q}_4(\theta|\underline{D}, \underline{\tilde{\Lambda}}_0, \tilde{\theta})}{\partial \alpha^2} = (\frac{1}{\alpha} - \frac{\partial^2 \log \Gamma(\alpha)}{\partial \alpha^2}),$$

where $B^* = B + 1$. The root of $U(\alpha) = 0$ can be updated iteratively through $\alpha^{(k+1)} = \alpha^{(k)} - \frac{U}{U'}$. The initial value can be set as $\alpha^{(0)} = \frac{-3-\sqrt{-9-12B^*}}{12B^*}$ for a fast convergence of root of $U(\alpha)$ (Jamshidian, 2001; Zhang and Jamshidian, 2003).

As outlined, the EM-algorithm has no numerical difficulty in implementation with the starting values using Nelson-Aalen estimates for $\Lambda_{01}, \Lambda_{02}$, Λ_{03}, and any positive value for θ.

4.2.2 Pitfalls in NPMLE

Even though the unrestricted GFCMM for semi-competing risks data has merit in both model construction and numerical computation as described above, it does not, unfortunately, always yield nonparametrically consistent likelihood-based inference for the model parameters based on the numerical experiments. Xu et al. (2010) did not explore the numerical issues in maximum likelihood estimation under the unrestricted GFCMM. Jiang and Haneuse (2015) noted this pitfall in their limited simulation studies. This chapter presents a more thorough numerical experiment for analyzing semi-computing risks data under both of the restricted and unrestricted GFCMM to examine the behavior of the NPMLEs, in order to provide practical guidelines for the possible adoption of the unrestricted GFCMM.

In the numerical experiment, semi-competing risks data with three constant hazards $\lambda_{01}(t) = \lambda_{01}, \lambda_{02}(t) = \lambda_{02}$, and $\lambda_{03}(t) = \lambda_{03}$ with values ranging between 0.5 and 2 were generated. Independent right censoring times according to a mixture distribution of $Unif(3, 5)$ and a degenerated point-mass distribution at 5 representing a possible administrative censoring with 50-50 chance was also generated. The gamma-frailty variable was simulated from *Gamma* distribution with mean 1 and variance $\theta = 1$. For a simulated data set, the gamma-frailty variable θ was varied, and at each given θ value, the EM-algorithm was implemented without M-step for Q_4 to compute

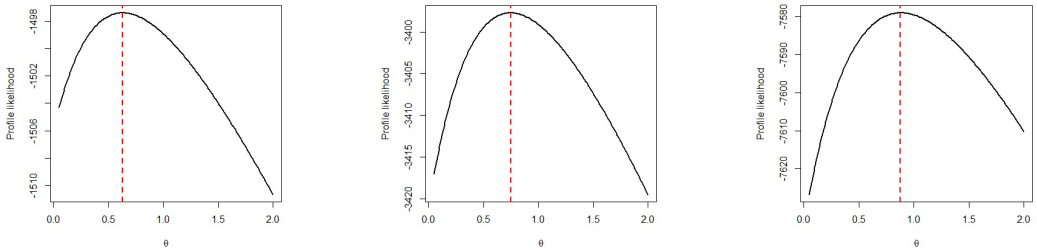
the NPMLEs of $\Lambda_{01}, \Lambda_{02}$, and Λ_{03}, then they were plugged in the log of likelihood for the observed data (4.1) to form the log profile likelihood function of θ.

Figure 4.1 presents the profile likelihood under the restricted model for data generated with $\lambda_{01} = 2$ and $\lambda_{02} = \lambda_{03}$ taking values from 0.5 to 2, and θ ranging from 0 to 2. In the simulation study, sample sizes $n = 200, 400$, and 800 were considered. Figure 4.1 clearly indicates that the profile likelihood peaks at a θ that is inside the feasible region of θ at any scenario considered, which is not surprising because the proposed model is equivalent to Clayton copula model, for which nonparametric inference for semi-competing risks data was well-justified by Fine et al. (2001). The averages of NPMLEs of Λ_{0i} for $i = 1, 2$ for the case of $\lambda_{01}(t) = 2$, $\lambda_{02}(t) = 1$, and $n = 400$ using the proposed EM-algorithm based on 500 repetitions were plotted in Figure 4.2, which clearly demonstrates the consistency of NPMLEs under the restricted GFCMM.

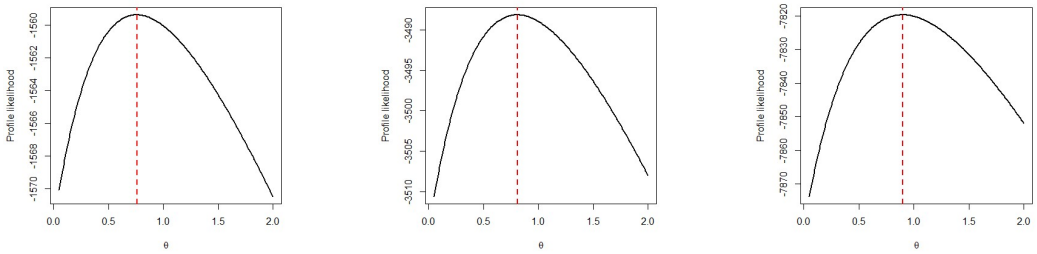
Figure 4.3 depicts the profile likelihood under the unrestricted model for data generated with $\lambda_{01} = 2$ and $\lambda_{02} = 0.5$, while λ_{03} varies from 0.5 and 2. It appears that when there is a large disparity between λ_{02} and λ_{03}, the size of study sample could dramatically affect the behavior of the NPMLEs. For example, when $\lambda_{02} = 0.5$ and $\lambda_{03} = 1.5$, the profile likelihood appears to peak at the boundary of feasible domain, i.e., $\theta = 0$, even for a relatively large sample ($n = 400$), which inevitably results in biased inference based on the NPMLEs as shown in Figure 4.4. But in another example where $\lambda_{02} = 0.5$ and $\lambda_{03} = 1$, the nonparametric maximum likelihood estimation works just fine for $n = 800$ resulting in unbiased NPMLEs as shown in Figure 4.5.

These simulation results reveal a pitfall in the NPMLEs under the unrestricted GFCMM for semi-competing risks data. More precisely, the NPMLEs of the baseline

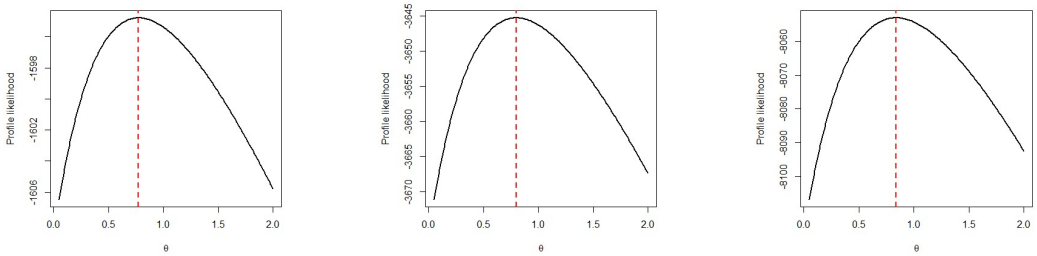
$\lambda_{01}(t) = 2, \lambda_{02}(t) = \lambda_{03}(t) = 2$



$\lambda_{01}(t) = 2, \lambda_{02}(t) = \lambda_{03}(t) = 1.5$



$\lambda_{01}(t) = 2, \lambda_{02}(t) = \lambda_{03}(t) = 1$



$\lambda_{01}(t) = 2, \lambda_{02}(t) = \lambda_{03}(t) = 0.5$

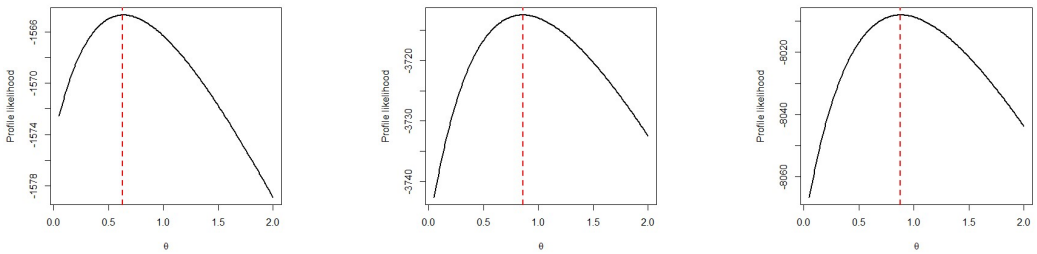
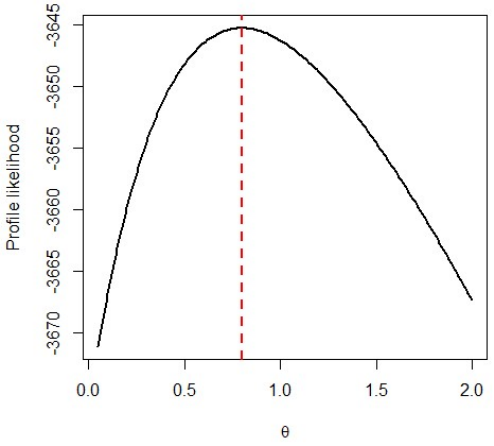
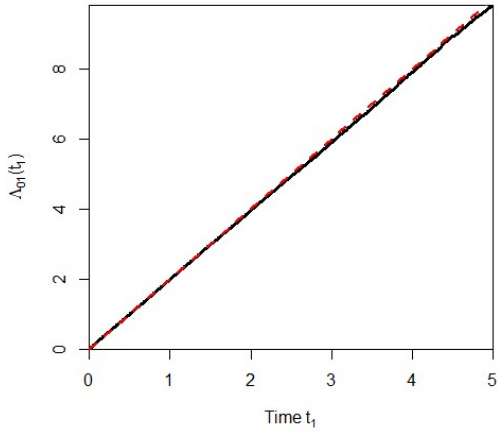


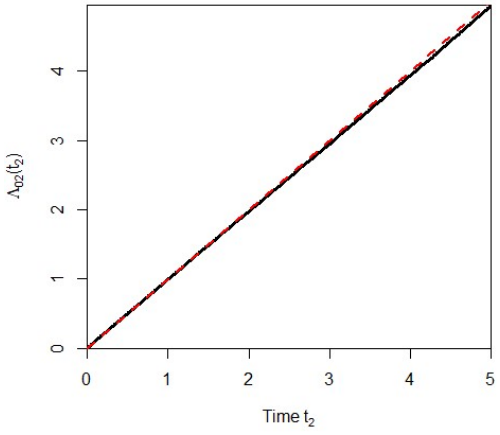
Figure 4.1: Profile likelihood for simulated semi-competing risks data under the restricted models. Sample sizes increase from $n = 200$ (left), to $n = 400$ (middle), $n = 800$ (right) for each scenario.



(a)



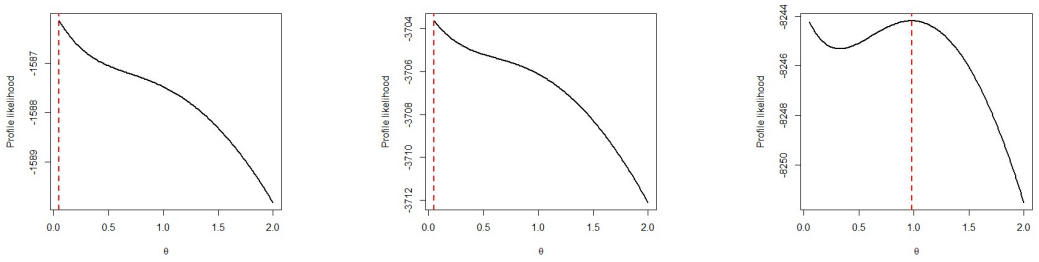
(b)



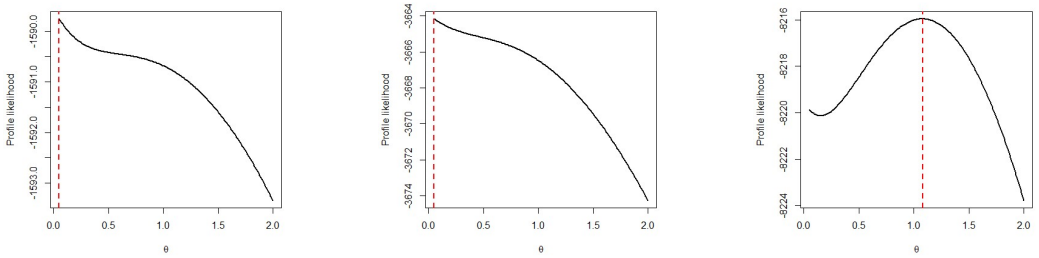
(c)

Figure 4.2: (a): Profile likelihood based on a single data set; (b) and (c): Average of NPMLEs of the conditional cumulative hazards (dashed lines based on 500 repetitions), $\hat{\Lambda}_{01}(\cdot)$ and $\hat{\Lambda}_{02}(\cdot)$, respectively, for the simulated semi-competing risks data under the restricted model with $\lambda_{01} = 2$, $\lambda_{02} = 1$, and $n = 400$.

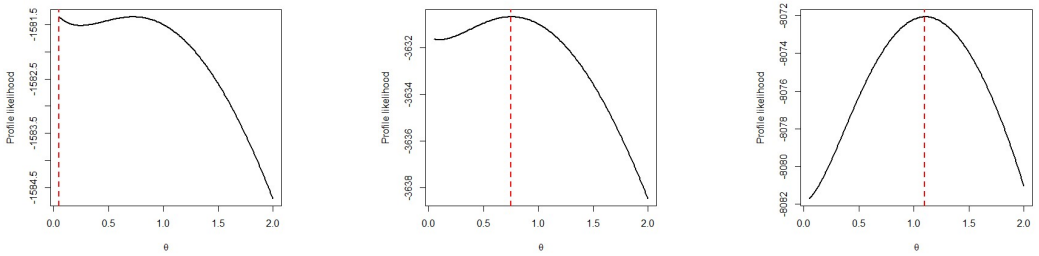
$\lambda_{01}(t) = 2, \lambda_{02}(t) = 0.5, \lambda_{03}(t) = 2$



$\lambda_{01}(t) = 2, \lambda_{02}(t) = 0.5, \lambda_{03}(t) = 1.5$



$\lambda_{01}(t) = 2, \lambda_{02}(t) = 0.5, \lambda_{03}(t) = 1$



$\lambda_{01}(t) = 2, \lambda_{02}(t) = 0.5, \lambda_{03}(t) = 0.5$

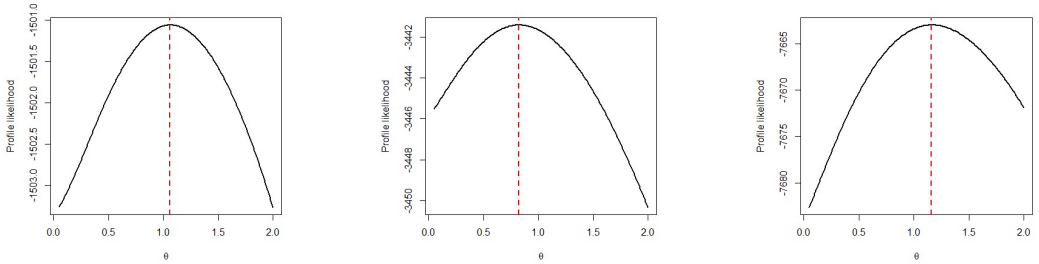


Figure 4.3: Profile likelihood for simulated semi-competing risks data under the unrestricted models. Sample sizes increase from $n = 200$ (left), to $n = 400$ (middle), $n = 800$ (right) for each scenario.

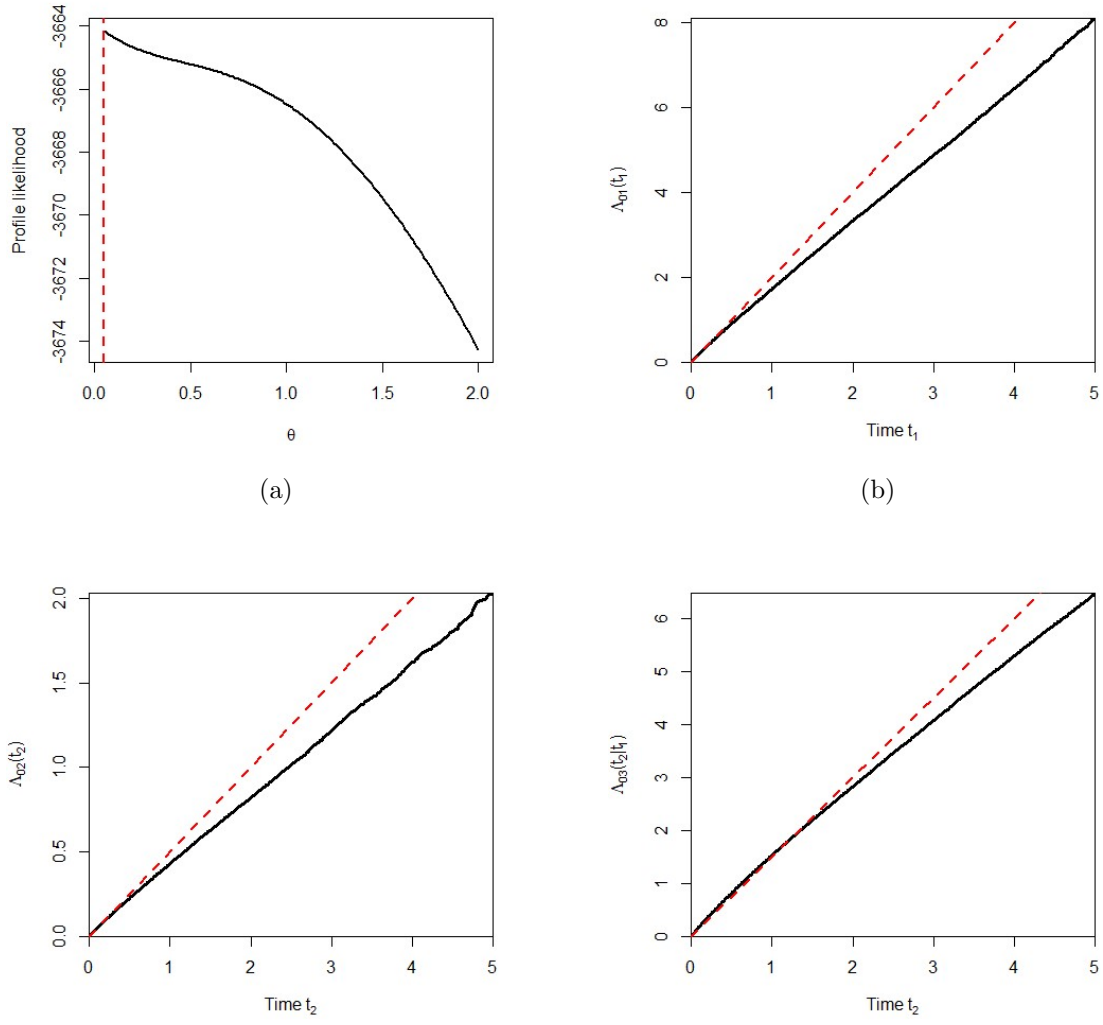


Figure 4.4: (a): Profile likelihood based on a single data set; (b), (c) and (d): Average of NPMLEs of the conditional cumulative hazards (dashed lines based on 500 repetitions), $\hat{\Lambda}_{01}(\cdot)$, $\hat{\Lambda}_{02}(\cdot)$, and $\hat{\Lambda}_{03}(\cdot)$, respectively, for simulated semi-competing risks data under the unrestricted model with $\lambda_{01} = 2$, $\lambda_{02} = 0.5$, $\lambda_{03} = 1.5$, and $n = 400$.

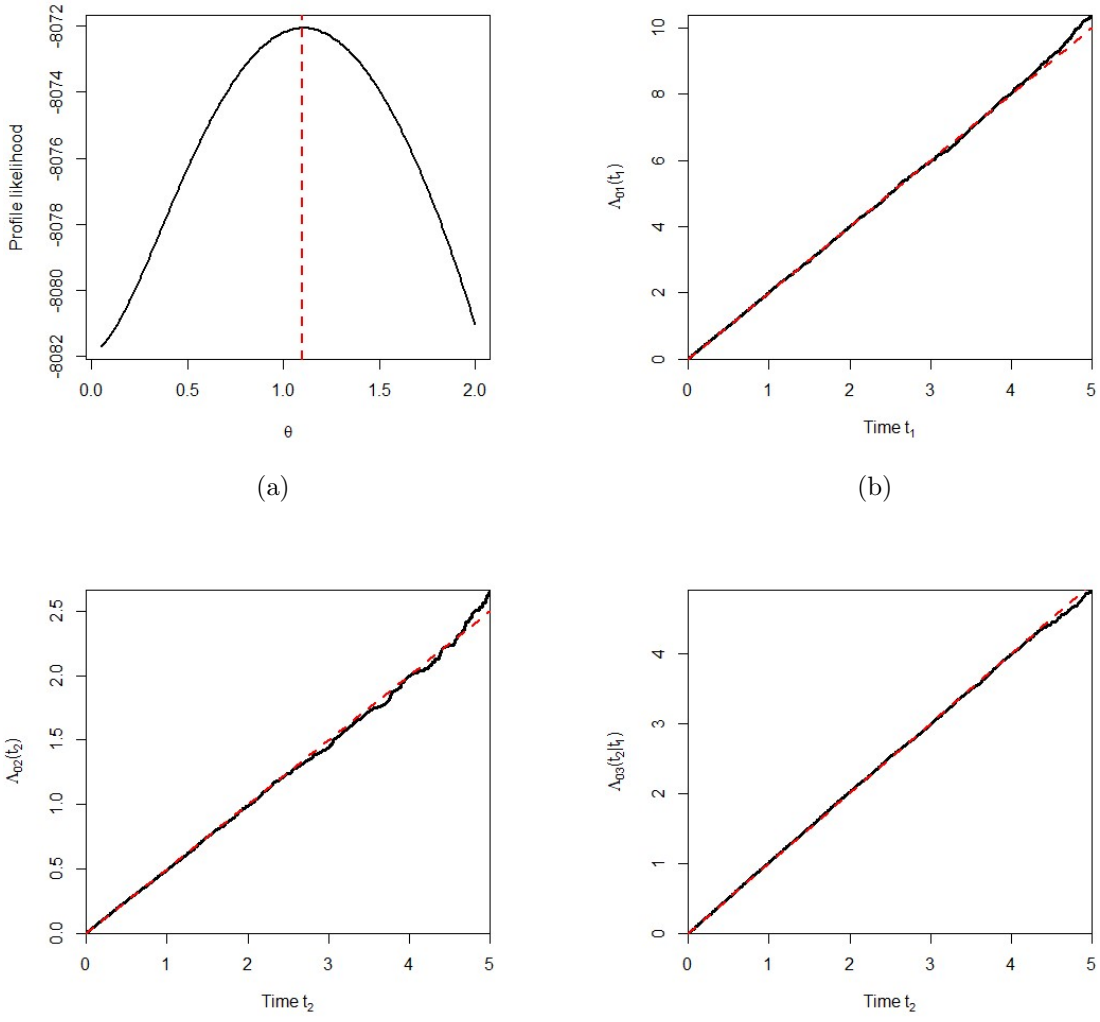


Figure 4.5: (a): Profile likelihood based on a single data set; (b), (c) and (d): Average of NPMLEs of the conditional hazards (dashed lines based on 500 repetitions), $\hat{\Lambda}_{01}(\cdot)$, $\hat{\Lambda}_{02}(\cdot)$, and $\hat{\Lambda}_{03}(\cdot)$, respectively, for simulated semi-competing risks data under the unrestricted model with $\lambda_{01} = 2$, $\lambda_{02} = 0.5$, $\lambda_{03} = 1$, and $n = 800$.

cumulative hazard functions under the unrestricted GFCMM may have substantial bias as the result that the profile likelihood attains its maximum value at $\theta = 0$, the boundary of the feasible domain for the variance of the frailty variable θ. This finding is very important because for MLEs when the regularity conditions are violated, the desirable properties of MLEs are not guaranteed. Although different from the ordinary likelihood, the profile likelihood behaves similarly. In practice, it is not easy to exam if a likelihood is globally concave or not, since it is almost impossible to check all the regularity conditions, especially there are nonparametric components in the model. The profile likelihood is a common exercise to exam the shape of likelihood surface in a dimension-reduced parameter space, particularly when the nuisance parameter estimator is an explicit function of the parameters of interest. This problem exemplifies the practical use of profile likelihood in examining the validity of MLE. In Figure 4.3, it is observed that as sample size increases, it is more likely for the likelihood to be concave. However it may not be possible to say how large the sample size needs to be so the likelihood is concave.

4.2.3 A Practical Guideline for Using GFCMM

The pitfall in the NPMLEs under the unrestricted GFCMM troubles the likelihood-based inference for the comparison between the restricted and unrestricted GFCMMs (testing $H_0 : \lambda_{02}(\cdot) = \lambda_{03}(\cdot)$) because one may obtain biased NPMLEs of $\Lambda_{0i}(\cdot)$ for $i = 1, 2, 3$ under the unrestricted GFCMM. Consequently, likelihood ratio and Wald tests cannot be conducted, since test statistics for both tests require the computation of MLEs under the unrestricted model. However since the NPMLEs under the restricted GFCMM do not exhibit the pitfall for the MLE, one could explore a score test for

$H_0 : \beta = 0$ (i.e., $\lambda_{02}(\cdot) = \lambda_{03}(\cdot)$) assuming the proportional hazards model, $\lambda_{03}(t) = \lambda_{02}(t)e^{\beta}$.

The score function for β under the proportional hazards assumption for unrestricted GFCMM is given by

$$u(\beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \delta_{i1}\delta_{i2} - \frac{(\theta^{-1} + \delta_{i1} + \delta_{i2}) \left(\delta_{i1} \left(\Lambda_{02}(Y_{i2}) - \Lambda_{02}(Y_{i1}) \right) e^{\beta} \right)}{1 + \theta \left[\Lambda_{01}(Y_{i1}) + \Lambda_{02}(Y_{i1}) + \delta_{i1} \left(\Lambda_{02}(Y_{i2}) - \Lambda_{02}(Y_{i1}) \right) e^{\beta} \right]} \right\}.$$

For testing null hypothesis $H_0 : \beta = 0$ (i.e., $\lambda_{02}(\cdot) = \lambda_{03}(\cdot)$), one can calculate the score under H_0,

$$\hat{u}(0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \delta_{i1}\delta_{i2} - \frac{\left(\hat{\theta}^{-1} + \delta_{i1} + \delta_{i2} \right) \left(\delta_{i1} \left(\hat{\Lambda}_{02}(Y_{i2}) - \hat{\Lambda}_{02}(Y_{i1}) \right) \right)}{1 + \hat{\theta} \left[\hat{\Lambda}_{01}(Y_{i1}) + \hat{\Lambda}_{02}(Y_{i1}) + \delta_{i1} \left(\hat{\Lambda}_{02}(Y_{i2}) - \hat{\Lambda}_{02}(Y_{i1}) \right) \right]} \right\}.$$

where $\hat{\Lambda}_{01}(\cdot), \hat{\Lambda}_{02}(\cdot)$, and $\hat{\theta}$ are the MLEs under the restricted GFCMM as outlined above. Therefore the proposed score test is not affected by the pitfall of the NPMLE for the unrestricted model, since it only requires estimates under the restricted model. Therefore it is appropriate to examine the goodness-of-fit of the restricted model under the hypothesis testing by setting the restricted GFCMM model for the null hypothesis. Table 4.1 provides the mean, Monte Carlo standard deviation (MCSD), and the average value of the bootstrap standard error (BSE) of the test statistic $\hat{u}(0)$ for various values of β. It also reports the percentage of rejection (PR) at a two-sided 0.05 significance level for the score test $\hat{u}^2(0)/BSE^2 > 3.84$ based on the simulated data described in the aforementioned unrestricted GFCMM. The numbers in Table 4.1 were calculated based on 1,000 Monte Carlo samples, and the standard error was estimated based on 100 bootstrap samples for each simulated data set. Simulations

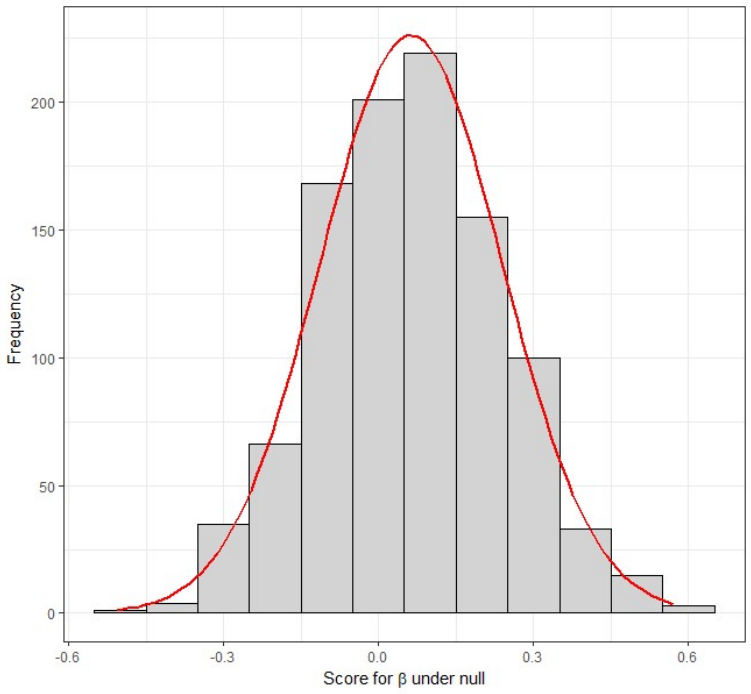
Table 4.1: Simulation results of the score test for H_0: $\beta = 0$

	$n = 200$				$n = 400$			
	Mean	MCSD	BSE	PR	Mean	MCSD	BSE	PR
$\beta = 0$	0.063	0.176	0.182	0.048	0.046	0.180	0.184	0.051
$\beta = \log 2$	0.473	0.220	0.214	0.591	0.616	0.215	0.216	0.814
$\beta = \log 3$	0.764	0.241	0.232	0.920	1.048	0.243	0.235	0.996
$\beta = \log 4$	0.984	0.249	0.243	0.980	1.334	0.249	0.248	1.000

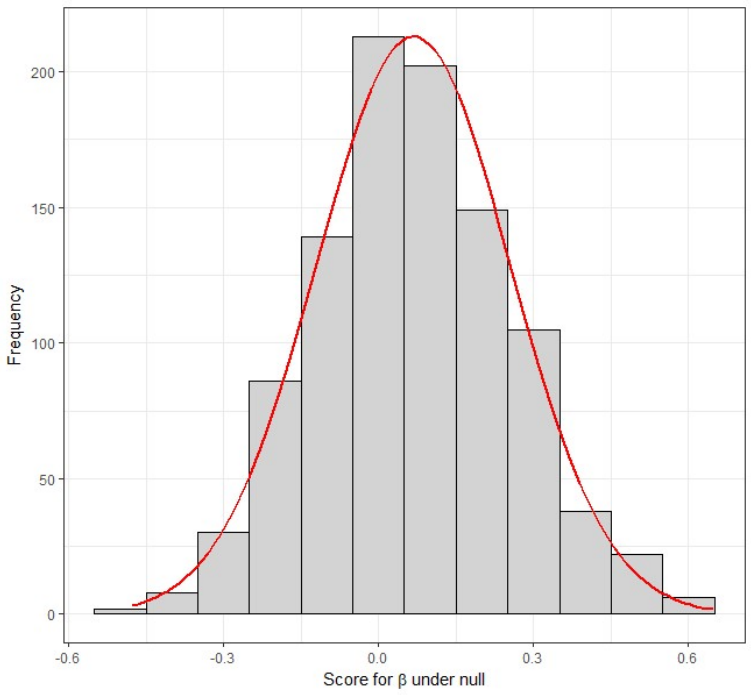
of hypothesis testing were conducted with two sample size scenarios of 200 and 400. True values of β are 0, $\log 2$, $\log 3$, and $\log 4$ corresponding to $\lambda_{03} = 0.5, 1, 1.5$, and 2, respectively, in the simulation setting for the unrestricted GFCMM scenario. Type I error (percentage of rejection under the null $H_0 : \beta = 0$) is well controlled around its nominal value of 0.05. MCSD appears to be fixed when sample size increases, and the average of BSEs is very close to MCSD. The score test appears to be consistent since the power increases as sample size increases and approaches to 1. The histograms of the score test statistic under the null hypothesis resemble well to a normal distribution with a constant variance (Figure 4.6). The simulation results clearly imply that $\hat{u}(0)$ is asymptotically normal with mean 0 (under H_0), which facilitates an adequate procedure for making inference about the difference between $\lambda_{02}(\cdot)$ and $\lambda_{03}(\cdot)$ based on the ordinary normal theory.

Based on what was learnt from the simulation studies, a practical guideline on the use of GFCMM when analyzing semi-competing risks data (Figure 4.7) is suggested. First, one should perform the score test for $H_0 : \beta = 0$ (i.e., $\lambda_{02}(\cdot) = \lambda_{03}(\cdot)$) as the GFCMM under the null hypothesis implies that the occurrence of the non-terminal event would increase the hazards of the terminal event, which is probably a reasonable assumption for illness-death models in most biomedical studies. If the test does not provide statistical evidence to reject the null hypothesis, the restricted GFCMM is

recommended for the analysis. The maximum likelihood estimate of $(\Lambda_{01}(\cdot), \Lambda_{02}(\cdot),$ and $\theta)$ will then be computed using the EM-algorithm proposed in this work, which is used to estimate the survival functions of the non-terminal event, the terminal event without non-terminal event and with non-terminal event occurred at a given time to illustrate how the non-terminal event impacts the terminal event. If the test leads to rejection of $H_0 : \beta = 0$, one should plot the profile likelihood of θ against θ under the unrestricted GFCMM to evaluate whether or not the likelihood is bounded and attains its maximum value in the interior of the parameter space for θ. If it is bounded, the nonparametric maximum likelihood analysis with the unrestricted GFCMM via the proposed EM-algorithm can be readily implemented; otherwise the GFCMM should not be used for nonparametric maximum likelihood analysis. Instead, one would be recommended to consider nonparametric Bayesian approach (Han et al., 2014; Lee et al., 2016, 2015) under the unrestricted GFCMM for the analysis of semi-competing risks data.



(a) $n = 200$



(b) $n = 400$

Figure 4.6: Histograms of the test statistic $\hat{u}(0)$ under the null hypothesis.

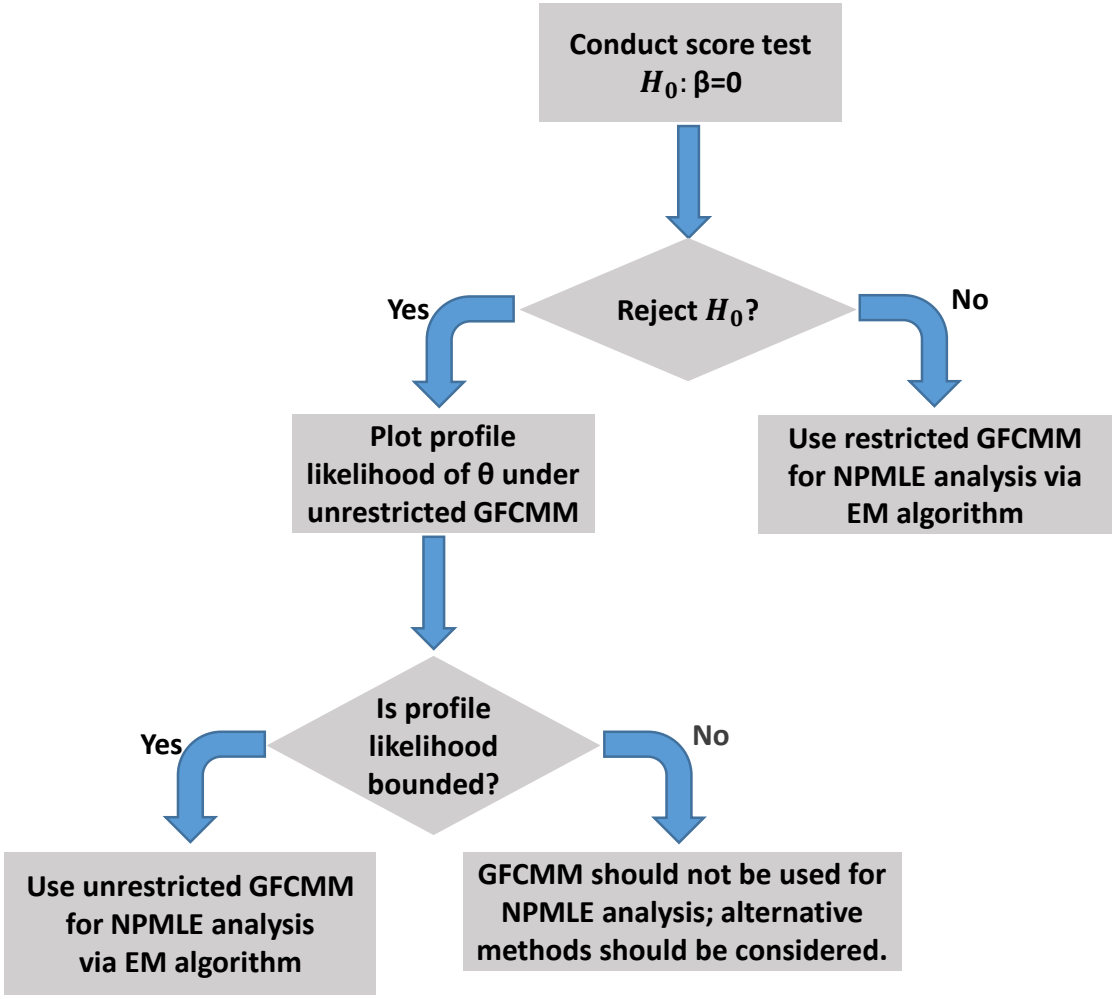


Figure 4.7: A practical guideline for the use of GFCMM in analysis of semi-competing risks data.

4.3 Case Study: Indianapolis-Ibadan Dementia Project Data

It is generally of interest to explore if at any given age, the hazard of death would be modified by medical conditions that have occurred prior to this age. We were particularly interested in addressing this question regarding the medical condition of dementia which is becoming a common problem in the elderly population. Regarding this, data from the Indianapolis-Ibadan Dementia Project (IIDP) were used. It is a 20-year longitudinal study of dementia and its risk factors, in elderly African Americans living in Indianapolis, Indiana and elderly Yoruba residing in Ibadan, Nigeria (Gao et al., 2016; Hendrie et al., 2001, 2017). The study was funded by National Institute on Aging. All participants agreed to participate in follow-up cognitive assessment and clinical evaluations every two or three years, where diagnosis of dementia, mortality, and information on potential risk factors for dementia (e.g., diabetes status with comorbidity conditions) were recorded. This study sample only consisted of African-American (AA) participants of the IIDP, who were aged 65 or older. Recruitment was conducted at two time points with 2,212 recruited in 1992 and additionally 1,893 enrolled in 2001. Participants without dementia at baseline agreed to undergo regular follow-up for cognitive assessments and clinical evaluations. Details on the cohort, diagnosis process and criteria of dementia can be found in Gao et al. (2016), Hendrie et al. (2001), and Hendrie et al. (2017). The dataset consists of 3,973 AA without dementia at baseline, where 1,420 participants were alive without dementia throughout the follow-up period, 78 participants developed dementia but stayed alive by the end of study, 2,232 participants died without developing demen-

tia, and 243 participants developed dementia and died during the course of the study. The average age of the participants was 75.6 years at baseline.

The observations can be characterized as semi-competing risks data, in which the diagnosis of dementia is considered as non-terminal event and death as terminal event. In this research, it was of interest to study the residual life of AA who had survived over 65 years and explore impact of dementia on mortality. The proposed score test resulted in a p-value of 0.17 and thus failed to reject the null hypothesis that the restricted GFCMM holds. The nonsignificant result in this large dataset is to be expected because individuals with dementia were known to have increased mortality, a feature implied by the restricted GFCMM and confirmed by the score test. The profile likelihood of θ (Figure 4.8) was also plotted for both of the restricted and unrestricted GFCMMs, which clearly demonstrated failure of the unrestricted model when applied to the dementia dataset. It is important to note that although the sample size for the dementia dataset is very large (close to 4,000), the profile likelihood of θ still peaks at zero, the boundary of the feasible domain. Following the proposed practical guideline, the restricted GFCMM was used for analysis. Based on the formulas (2.11)-(2.13) from Chapter 2, the survival rates for dementia, death without dementia and death with dementia diagnosed at age t_1, at any given age beyond 65 for AA who had survived over 65 years and had not developed dementia at age 65 can be, respectively, estimated by

$$\begin{aligned}\hat{S}_1(t) &= \int_t^\infty \left(1+\hat{\theta}(\hat{\Lambda}_{01}(t_1)+\hat{\Lambda}_{02}(t_1))\right)^{-(1/\hat{\theta}+1)} d\hat{\Lambda}_{01}(t_1) \\ &\quad + \int_0^\infty \left(1+\hat{\theta}(\hat{\Lambda}_{01}(t_2)+\hat{\Lambda}_{02}(t_2))\right)^{-(1/\hat{\theta}+1)} d\hat{\Lambda}_{02}(t_2), \qquad (4.12)\end{aligned}$$

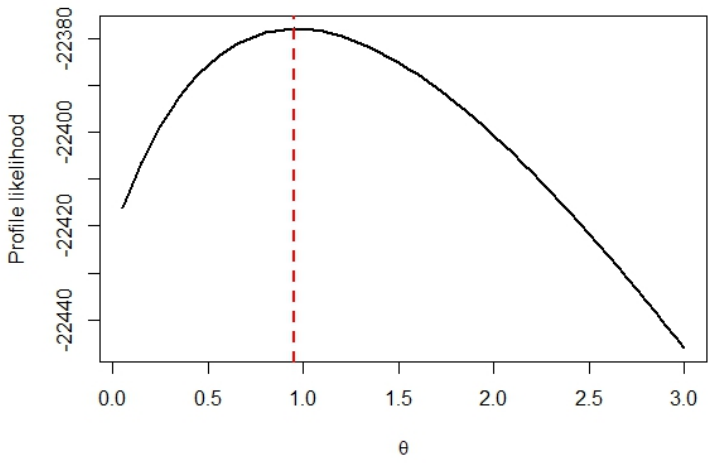
$$\hat{S}_2(t) = \frac{\int_t^\infty \left(1+\hat{\theta}(\hat{\Lambda}_{01}(t_2)+\hat{\Lambda}_{02}(t_2))\right)^{-(1/\hat{\theta}+1)} d\hat{\Lambda}_{02}(t_2)}{\int_0^\infty \left(1+\hat{\theta}(\hat{\Lambda}_{01}(t_2)+\hat{\Lambda}_{02}(t_2))\right)^{-(1/\hat{\theta}+1)} d\hat{\Lambda}_{02}(t_2)}, \tag{4.13}$$

and

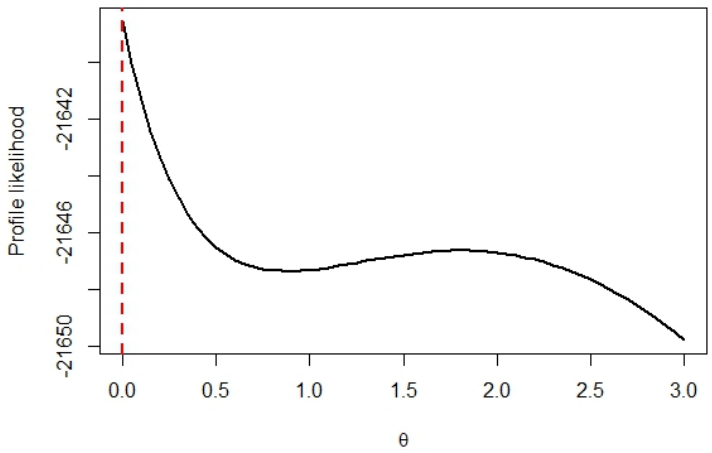
$$\hat{S}_3(t|t_1) = \begin{cases} \frac{\left(1+\hat{\theta}(\hat{\Lambda}_{01}(t_1)+\hat{\Lambda}_{02}(t))\right)^{-(1/\hat{\theta}+1)}}{\left(1+\hat{\theta}(\hat{\Lambda}_{01}(t_1)+\hat{\Lambda}_{02}(t_1))\right)^{-(1/\hat{\theta}+1)}} & \text{if } t > t_1 \\ 1 & \text{otherwise.} \end{cases} \tag{4.14}$$

Figure 4.9 plots the survival functions for dementia $\hat{S}_1(t)$, conditional survival for death without dementia given alive at age of 75 $\hat{S}_2(t|t_1 = 75)$, and survival for death with dementia diagnosed at age of 75 $\hat{S}_3(t|t_1 = 75)$, respectively, based on the IIDP study. From Figure 4.9, it is estimated that 13.9% of AA would develop dementia in their life time, which is consistent with the data. The median residual survival age for AA who were free of dementia given alive at the age of 75 was 86.2, and the median residual survival ages for AA with dementia diagnosed at 75 was 82.9. It should be noted that $\hat{S}_2(t|t_1)$ applies to any AA who had survived over the age of t_1 without developing dementia while $\hat{S}_3(t|t_1)$ applies only to AA who had dementia diagnosed at t_1 and were alive at t_1. Therefore, it is not surprising that $\hat{S}_3(t|t_1 = 75)$ has lower survival rate than $\hat{S}_2(t|t_1 = 75)$ after 75. This is due to the fact that dementia increases the hazard of mortality, which is implied by the restricted GFCMM.

The study of a potential impact of dementia on mortality has been an active research topic in Alzheimer disease research. Using the early data from the IIDP study, Perkins et al. (2002) compared Kaplan-Meier (KM) survival curves of mortality between the AA with and without dementia diagnosed in the study period. James et al. (2014) compared the residual survival between individuals with and without dementia at age 75 to study the impact of dementia on mortality. The com-

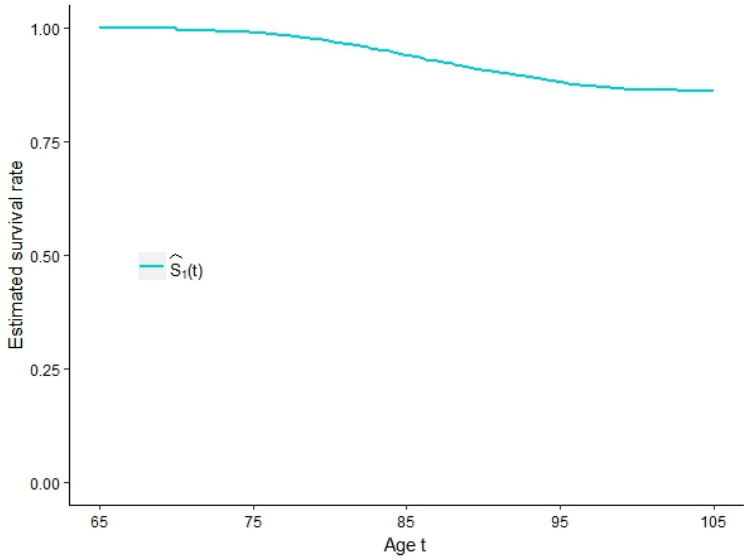


(a) Restricted GFCMM

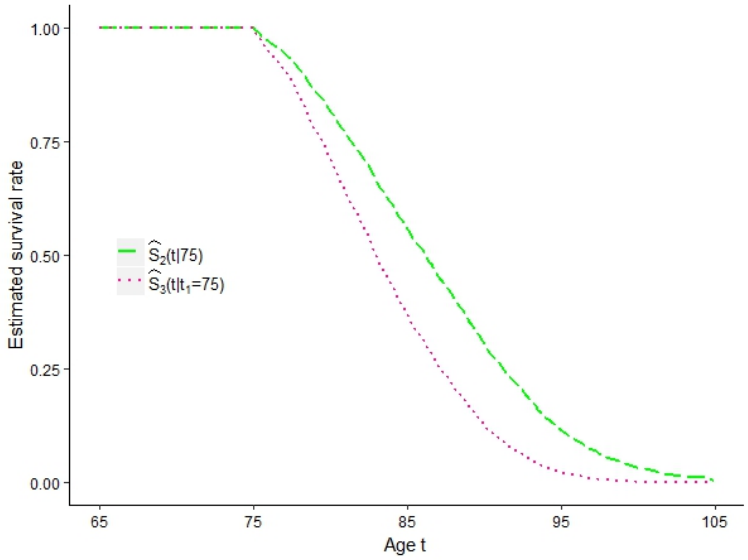


(b) Unrestricted GFCMM

Figure 4.8: Profile likelihood of θ for dementia data.



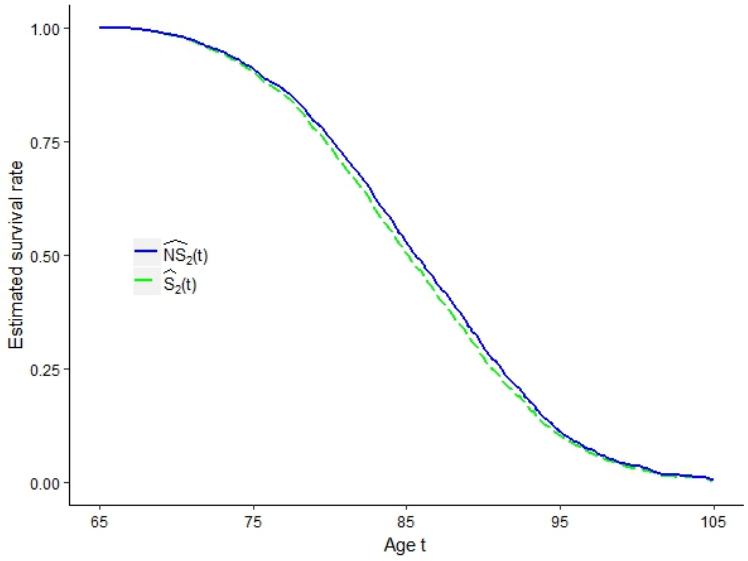
(a)



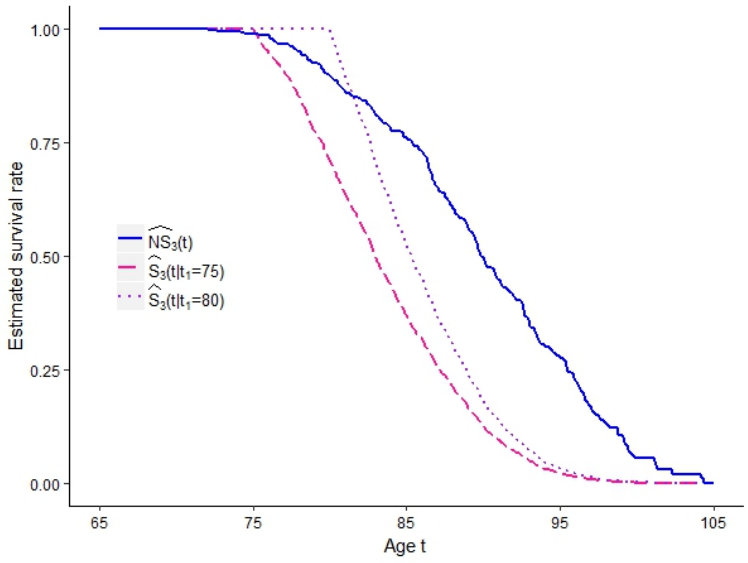
(b)

Figure 4.9: Case Study-IIDP: (a) Survival functions for dementia $-\ \hat{S}_1(\cdot)$; (b) Conditional survival function for death without dementia given alive at age of 75 $-\ \hat{S}_2(\cdot|75)$, and survival function for death with dementia diagnosed at age of 75 $-\ \hat{S}_3(\cdot|t_1 = 75)$.

parison of naive KM-curves (Perkins et al., 2002) is not statistically sound because this approach ignores the timing of dementia, which may lead to misclassification of potential dementia individuals due to censoring. For illustration, Figure 4.10 plots the naive KM-curves against $\hat{S}_2(\cdot)$, $\hat{S}_3(\cdot|t_1 = 75)$, and $\hat{S}_3(\cdot|t_1 = 80)$ using the updated IIDP data. The naive KM-curve slightly overestimates the survival function for those without dementia, which is expected because over 85% of study subjects did not develop dementia during the study period in this study population. However $\hat{S}_3(\cdot|t_1 = 75)$ or $\hat{S}_3(\cdot|t_1 = 80)$ is very different from the naive KM-curve for the dementia group due to the fact that time to dementia diagnosis is also predictive of mortality. Hence the use of a naive KM-curve for individuals with dementia can greatly underestimate the risk of mortality. While the comparison of survival rate between individuals with and without dementia at a fixed time point (James et al., 2014) is statistically reasonable, it does not make good use of the longitudinally observed disease progression data and is not able to assess the impact of the timing of dementia on mortality in a single statistical model. As a take-home message from this case study, we emphasize that semi-competing risks model is readily applicable to prospectively collected data where both time to a disease and time to death are available and when it is of interest to estimate mortality risk for individuals who develop the disease during the course of study.



(a)



(b)

Figure 4.10: Case Study-IIDP: (a) Naive Kaplan-Meier curve for the non-dementia group – $\widehat{NS}_2(\cdot)$, and survival function for death without dementia – $\hat{S}_2(\cdot)$; (b) Naive Kaplan-Meier curve for the dementia group – $\widehat{NS}_3(\cdot)$, and survival functions for death with dementia diagnosed at age of 75 and 80 – $\hat{S}_3(\cdot|t_1 = 75)$ and $\hat{S}_3(\cdot|t_1 = 80)$.

4.4 Discussion

The analysis of semi-competing risks data has been an important topic in both statistical methodology and medical literature in the last couple of decades. Literature review resulted in over 70 peer-reviewed publications in this topic since 2001. The GFCMM proposed by Xu et al. (2010) under the conditional Markov assumption is very appealing to researchers in several aspects, and it provides a more flexible alternative approach to the well-known copula model for semi-competing risks data. This model was widely adopted in the literature for various applications (Chapple et al., 2017; Han et al., 2014; Lee et al., 2016, 2015) since its inception. Xu et al. (2010) proposed a Newton-Raphson algorithm for the maximum likelihood analysis, which may not be numerically stable when the sample size is large. In addition, the fundamental question regarding the validity of non-/semi-parametric likelihood-based inference based on this model under the frequentist inference framework has not been carefully addressed.

To fill in this gap, first an easy-to-implement EM-algorithm was proposed to compute the NPMLEs of the hazard functions for semi-competing risks data under GFCMMs. The validity of NPMLEs under the restricted GFCMM was numerically demonstrated, and a pitfall in the NPMLEs associated with the unrestricted GFCMM was revealed. This fact led to the construction of a score test for the comparison between the restricted and unrestricted GFCMM, which might be the only valid likelihood-based inference procedure given the numerical problem associated with the estimation of the unrestricted GFCMM. The validity of the test was justified through simulation studies. This work adds significantly to the literature of semi-competing

risks data. Not only does it provide an insight of GFCMMs as an alternative analytical model, the proposed score test may also serve as a safeguard to justify the use of Clayton copula model in the practice of semi-competing risks data analysis (Fine et al., 2001; Peng and Fine, 2007; Wang, 2003).

While the validity of the maximum likelihood approach under the GFCMM was explored in the framework of nonparameteric estimation of the hazards functions in this work, a numerically stable EM-algorithm can be also implemented for the case of semiparametric estimation under the proportional hazards model proposed by Xu et al. (2010). A similar numerical pitfall is expected. The score test can be also applied for the comparison between the unrestricted and restricted GFCMM. The take-home message from this work is that the GFCMM, while possessing the merit of nice interpretation and the linkage to the conventional illness-death model in analyzing semi-competing risks model, may not be always yield non-/semi-parametrically consistent estimation based on the proposed inferential procedures. Consequently, one should follow the proposed practical guideline in this thesis for considering the use of this model to analyze semi-competing risks data in real life applications.

However, our scientific question of interest that whether the non-terminal event changes the risk to the terminal event cannot be fully addressed by the GFCMM according to the findings in the chapter. If the restricted GFCMM is used as the final model for the analysis of semi-competing risks data, it can be concluded that the occurrence of intermediate event increases the risk to the terminal event. Unless alternative Bayesian approaches or parametric models are used for analysis, we are unable to come to a conclusion regarding to this question under the unrestricted GFCMM.

Therefore in Chapter 5, it is aimed to develop nonparametric testing procedures to directly address the research question.

Chapter 5

Nonparametric Tests for Semi-Competing Risks Data under Markov Illness-Death Model

5.1 Introduction

In the previous chapters, the GFCMM for semi-competing risks data was studied, and the pitfalls with its NPMLE under the unrestricted model were uncovered. A practical guideline including a semiparametric score test was proposed for the use of the restricted versus the unrestricted version of GFCMM. If the resulted p-value is large, the test fails to provide enough evidence to reject the null hypothesis. Thus the restricted model is possibly appropriate for use in the situation, and it can be concluded that experiencing non-terminal event increases the risk to terminal event. However if the test suggests the use of the unrestricted model, it is not straightforward to answer the scientific question of interest that whether the risk to a terminal event (e.g., mortality) will be altered with the occurrence of the non-terminal event (i.e., dementia) due to the pitfalls in NPMLE with unrestricted GFCMM. Therefore in this chapter we aim to develop nonparametric tests to conduct direct comparison between the risks to terminal event with and without developing the intermediate event.

As it was mentioned in Chapter 2, for the progressive illness-death model, the definitions of marginal hazards (unconditional on frailty) in (2.4)-(2.6) along with survival data analysis framework are essentially equivalent to the transition intensity functions (Definition 2.5.2) in the multi-state modeling framework. Hence conducting inference on hazards functions is equivalent to making inference on transition intensity

functions. Therefore due to the numerical issues in maximum likelihood analysis with GFCMM, in this chapter we adopt the three-state Markov illness-death model without recovery under the multi-state modeling framework, without incorporating frailty terms, for the analysis of semi-competing risks data.

Built on stochastic processes, multi-state models are versatile and of great importance in providing a framework to describe the progression of several time-to-event processes in a single statistical model (Hougaard, 1999). As it was mentioned in previous chapters, much research has been done on multi-state models with illness-death model as a special case (Andersen and Keiding, 2002; Beyersmann et al., 2011; Commenges, 1999; Cook and Lawless, 2018; Hougaard, 1999; Klein et al., 2016; Meira-Machado et al., 2009; Van Den Hout, 2016).

Transition intensity has always been a popular parameter in biomedical and other applications. The nonparametric estimation of cumulative transition intensity has been well studied in literature. The nonparametric Nelson-Aalen estimator of the cumulative transition intensity for multi-state models is described by Andersen et al. (2012) under Markov assumption. Datta and Satten (2001) showed this Nelson-Aalen estimator derived under Markov processes remains uniformly consistent even for non-Markov multi-state models for the marginal, with respect to the prior history $\mathcal{F}_{t-}$, transition intensity. Mau (1986) proposed a nonparametric estimator for cumulative transition intensity when the transition between two transient states is unobservable in the Markov illness-death model. Frydman (1995), Commenges (2002), Commenges et al. (2004), Frydman and Szarek (2009), and Frydman et al. (2013) developed nonparametric estimation methodology for the illness-death model with interval-censored data. Efficient computation tools have been developed and implemented in several

R packages, **p3state.msm** for illness-death model, and **mstate** and **tdc.msm** for multi-state survival models (De Wreede et al., 2010; Meira-Machado et al., 2007; Meira-Machado and Roca-Pardiñas, 2011).

Although much research has been done on the nonparametric estimation of cumulative transition intensity, research on the nonparametric testing has not received enough attention in the illness-death or multi-state models. Andersen et al. (2012) showed the asymptotic properties of the family of weighted rank test statistics, including log-rank and Wilcoxon-type tests using counting process theory to compare a hazard function to a fixed function in one sample or to compare hazard functions across multiple samples. As an example, Andersen et al. (2012) briefly discussed the possibility of testing the equality of mortality risk with and without neuropathy within the insulin-dependent diabetic population under the Markov illness-death model using the log-rank and Wilcoxon-type tests. However there is lack of detailed explanantion for the use of these two tests in a multi-state illness-death model for the comparison of two transition intensities. Bluhmki et al. (2019) developed a Kolmogorov-Smirnov two-sample test on transition intensities under Markovian multi-state models focusing on wild bootstrapping technique of Nelson-Aalen estimators. In addition, the book for multi-state modeling by Cook and Lawless (2018) did not discuss the problem of the nonparametric hypothesis testing. None of the research conducted on nonparametric estimation was extended to the development of nonparametric testing procedures either.

In this chapter, three nonpaprametric tests, including a linear test, a Kolmogorov-Smirnov-type test, and a L_2-distance-type test, are developed to examine whether the status of non-terminal event alters the risk to terminal event under the three-state

Markov illness-death model without recovery. The proposed linear test compares the area under the curve (AUC) and may be more intuitive to understand than the rank-based tests. While it is more powerful when the curves of cumulative transition intensities are not crossing, it suffers from power loss under crossing-curve scenarios. Hence omnibus Kolmogorov-Smirnov-type and L_2-distance-type tests, which are consistent against any fixed alternative hypothesis, are developed to gain power for crossing curves scenarios of cumulative transition intensities. These three tests can be adopted for non-Markov processes (Datta and Satten, 2001).

The remainder of this chapter is organized as follows. In Section 5.2, the Nelson-Aalen estimator for cumulative transition intensities in the Markov illness-death model is presented. Test statistics for the three proposed nonparametric tests with consideration of weight functions are developed. The asymptotic properties of these tests under the null hypothesis are developed with EPT. The consistency of Kolmogorov-Smirnov-type and L_2-distance-type tests is proved. In Section 5.3, simulation studies are conducted, and results are presented for all three proposed tests with weight functions under consideration. In Section 5.4, three nonparametric tests are applied to the Indianapolis-Ibadan Dementia Project (IIDP) data for illustration. In Section 5.5, this chapter concludes with discussion.

5.2 Nonparametric Tests of Transition Intensities

In Section 2.5, the likelihood for the progressive Markov illness-death model was derived. Again suppose there are n individuals in the study, $\{N_{jk,i}(t),\ j \neq k; j,k = 1,2,3\}$ is the counting process for the i^{th} individual, and $N_{jk}(t) = \sum_{i=1}^{n} N_{jk,i}(t)$ count the number of direct transitions from state j to state

k during $[0,t]$ in the sample. $\{Y_{j,i}(t),\ j=1,2,3\}$ is a predictable process for the i^{th} individual for the transition from the j^{th} state, and $Y_j(t)=\sum_{i=1}^{n} Y_{j,i}(t)$ counts the number of individuals in state j up till time t in the sample. In this research, $\{Y_j(t),\ j=1,2,3\}$ is called at-state process for state j. For this three-state illness-death model without recovery, $N_{21}(t)=N_{31}(t)=N_{32}(t)=0$. The likelihood under the counting process notation is

$$\prod_t \prod_j \left\{ \prod_{k\neq j} \left(Y_j(t)dA_{jk}\right)^{dN_{jk}(t)} \left(1-dA_{j\cdot}(t)\right)^{Y_j(t)-dN_{j\cdot}(t)} \right\},$$

where $\prod$ denotes the product integral, $N_{j\cdot}=\sum_{k\neq j} N_{jk}$, $A_{j\cdot}=\sum_{k\neq j} A_{jk}$, and $\tau=\sup_t\left\{\int_0^t \alpha_{jk}(u)du\right\}<\infty$, $j\neq k$, $j,k=1,2,3$. The NPMLE for cumulative transition intensities $A_{12}(t)$, $A_{13}(t)$, and $A_{23}(t)$ can be easily derived with Nelson-Aalen estimators

$$\hat{A}_{jk}(t)=\int_0^t \frac{\mathbb{P}_n dN_{jk}(u)}{\mathbb{P}_n Y_j(u)},\ j,k=1,2,3; j\neq k.$$

We aim to develop nonparametric tests with the goal of directly examining whether the status of non-terminal event alters the risk to the terminal event. This scientific question of interest naturally leads to testing the null hypothesis of the equality between the transition intensities from healthy state to death and from diseased state to death, which is

$$H_0: \alpha_{13}(t)=\alpha_{23}(t),\ t\in[\tau_1,\tau_2], \tag{5.1}$$

for some $\tau_1, \tau_2 \in [0, \tau]$, with $\tau_1 < \tau_2$. This implies that the cumulative transition intensity functions satisfies

$$H_0 : A_{23,\tau_1}(t) - A_{13,\tau_1}(t) = 0,\ t \in [\tau_1, \tau_2], \tag{5.2}$$

where $A_{jk,s}(t) = A_{jk}(t) - A_{jk}(s) = \int_s^t \alpha_{jk}(u)d\mu(u)$. As both $Y_1(t)$ and $Y_2(t)$ will be zero as time goes on, hence it makes sense to set τ_2 is as the upper limit to ensure both $Y_1(\tau_2)$ and $Y_2(\tau_2)$ non-zero for comparison, which is a normal practice in the survival data analysis. The reason for not choosing $\tau_1 = 0$ as the starting point can be well explained by Figure 5.1. Figure 5.1 plots the sample-level at-state processes $Y_1(t)$ and $Y_2(t)$ for estimating the cumulative transition intensities $A_{13}(t)$ and $A_{23}(t)$ from a single simulated dataset ($n = 200$) as illustration. It is straightforward that for the transition from the healthy state, $Y_1(t)$ counts everyone in at time zero, while for the transition from the diseased state, $Y_2(t)$ is zero at time 0, and it only becomes non-zero at or after the first occurrence of illness in the study sample. Hence it will be problematic to compare two cumulative intensities with zero as starting point especially if the first occurrence of illness happens at a much later time. Therefore it is important to restrict the comparison to an interval $[\tau_1, \tau_2]$ which ensures $\inf_{t\in[\tau_1,\tau_2]} PY_i(t) > 0$, $i = 1, 2$.

Under the null hypothesis (5.1), this research aims to develop three nonparametric tests for it. It needs to be noted that although Bakoyannis (2020) developed three nonparametric tests (i.e., a linear test, a Kolmogorov-Smirnov-type test, and a L_2-norm-based test) for comparing a transition probability between two independent samples, these tests are not directly adaptable to test the hypothesis of interest in

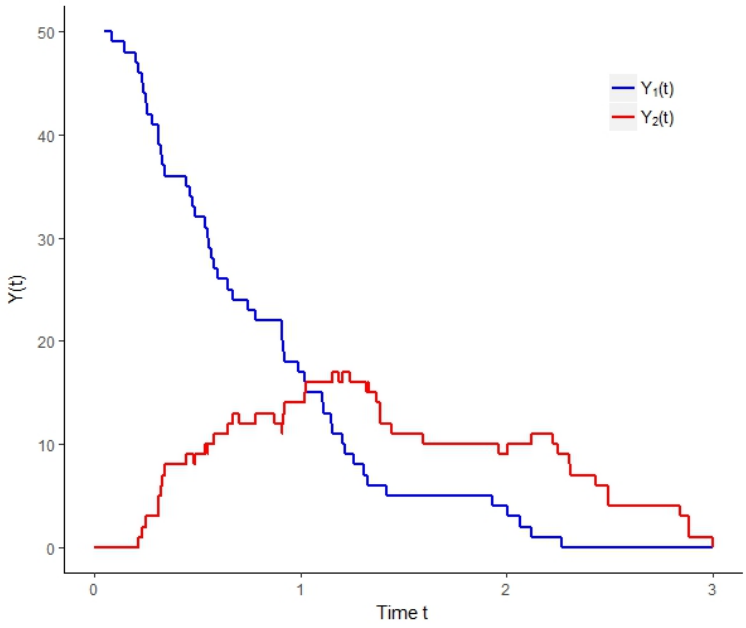


Figure 5.1: At-state processes $Y_1(t)$ for the transition from the healthy state and $Y_2(t)$ for the transition from the illness state of a simulated dataset ($n = 200$).

this thesis. First, these tests by Bakoyannis (2020) are considering transitioning probabilities, while the proposed tests compare transition intensities, which may not be equivalent in terms of inference. Second, tests by Bakoyannis (2020) compare one particular transition probability between two different samples; however, this works aims to test on two transitions in the same study sample.

5.2.1 Linear Nonparametric Test

As it was noted that the null hypothesis on transition intensity rates (5.1) implies that the cumulative transition intensity functions satisfies (5.2) and thus the difference of AUC between two cumulative transition intensities satisfies

$$H_0 : \int_{\tau_1}^{\tau_2} [A_{23,\tau_1}(t) - A_{13,\tau_1}(t)] \, d\mu(t) = 0.$$

This leads to a linear test which examines the difference of AUC between two cumulative transition intensities, $A_{13,\tau_1}(t)$ and $A_{23,\tau_1}(t)$ between times τ_1 and τ_2. In addition, consider the following non-negative weight function

$$\hat{K}(t) = \frac{\bar{Y}_1(t)\bar{Y}_2(t)}{\bar{Y}_1(t) + \bar{Y}_2(t)},$$

where $\bar{Y}_j(t) = \frac{1}{n}\sum_{i=1}^{n} Y_{j,i}(t),\ j = 1, 2$, and $\hat{K}(t)$ converges uniformly to $K(t) = \frac{PY_1(t)PY_2(t)}{PY_1(t)+PY_2(t)}$. Since the classes $\{Y_h(t) : t \in [0, \tau]\}, h = 1, 2$ are P-Donsker and thus also P-Glivenko-Cantelli. This weight function $\hat{K}(t)$ assigns little weight to times when there are few observations counted for $Y_1(t)$ and/or $Y_2(t)$, and it is called weighted test in the simulation. Note, when at least one of $Y_1(t)$ and $Y_2(t)$ becomes zero, $\hat{K}(t)$ assigns zero weight. If $K(t) = 1$, which gives equal weight to all events, it is called unweighted test. Here define the (weighted) estimated pointwise difference between two cumulative transition intensities $A_{13,\tau_1}(t)$ and $A_{23,\tau_1}(t)$ between times τ_1 and τ_2 of one sample

$$\hat{W}(t) = \hat{K}(t)\left(\hat{A}_{23,\tau_1}(t) - \hat{A}_{13,\tau_1}(t)\right).$$

The test statistic for the linear weighted nonparametric test can be constructed as the difference of AUC between two weighted cumulative transition intensity curves

$$T_{LN} = \int_{\tau_1}^{\tau_2} \hat{W}(t)\, d\mu(t). \tag{5.3}$$

Next, make the following assumptions which are needed for proving the asymptotic properties of the test statistic(s):

(B1). The right censoring time C is independent of the counting processes $\{N_{jk}(t), j, k = 1, 2, 3;\ j \neq k;\ t \in [\tau_1, \tau_2]\}$ and noninformative about these transition intensities;

(B2). $\inf_{t\in[\tau_1,\tau_2]} PY_1(t) > 0$ and $\inf_{t\in[\tau_1,\tau_2]} PY_2(t) > 0$;

(B3). The cumulative transition intensities $A_{13}(t)$ and $A_{23}(t)$ are continuous on $[\tau_1, \tau_2]$;

(B4). The weight function $\hat{K}(t)$ converges uniformly to a nonnegative uniformly bounded function $K(t)$ on $[\tau_1, \tau_2]$.

Define the martingale process for transition from state j to k of i^{th} observation as $M_{jk,i}(t) = N_{jk,i}(t) - \int_0^t Y_{jk,i} dA_{jk,i}$. The limiting process of $\hat{W}(t)$ under the null hypothesis is given below.

Theorem 5.2.1. Assume conditions $B1-B4$ are satisfied. Under the null hypothesis,

$$\sqrt{n}\,\hat{W}(\cdot) \rightsquigarrow \mathbb{G}_W,$$

where $\mathbb{G}$ is a tight zero-mean Gaussian process with variance-covariance functions

$$\Sigma_W(s,t) = P\left(K(s)\varphi_1(s)\right)\left(K(t)\varphi_1(t)\right).$$

and

$$\varphi_i(t) = \int_{\tau_1}^{t} \frac{dM_{23,i}(u)}{PY_2(u)} - \int_{\tau_1}^{t} \frac{dM_{13,i}(u)}{PY_1(u)}.$$

A consistent estimator of $\Sigma_W(s,t)$ is

$$\hat{\Sigma}_W(s,t) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{K}(s)\hat{\varphi}_i(s)\right)\left(\hat{K}(t)\hat{\varphi}_i(t)\right),$$

where

$$\hat{\varphi}_i(t) = \int_{\tau_1}^{t} \frac{dN_{23,i}(u) - Y_{2,i}(t)d\hat{A}_{23,i}(u)}{\mathbb{P}_n Y_2(u)} - \int_{\tau_1}^{t} \frac{dN_{13,i}(u) - Y_{1,i}(t)d\hat{A}_{13,i}(u)}{\mathbb{P}_n Y_1(u)}.$$

Proof. First we want to show the asymptotic linearity of the estimator $\int_{\tau_1}^{t} \hat{\alpha}_{23}(t)d\mu(t) - \int_{\tau_1}^{t} \hat{\alpha}_{13}(t)d\mu(t)$. Recall in Example 3.3.4, for one-sample Nelson-Aalen estimator $\hat{A}(t) = \int_0^t \frac{\mathbb{P}_n dN(u)}{\mathbb{P}_n Y(u)}$, it is proved to be asymptotically linear with its influence functions,

$$\sqrt{n}(\hat{A}(t) - A(t)) = \sqrt{n}\,\mathbb{P}_n \left(\int_0^t \frac{dM(u)}{PY(u)} \right) + o_p(1).$$

Similarly, it can be shown $\sqrt{n}\left(\int_{\tau_1}^{t} \hat{\alpha}_{23}(t)d\mu(t) - \int_{\tau_1}^{t} \alpha_{23}(t)d\mu(t)\right) = \sqrt{n}\,\mathbb{P}_n \left(\int_{\tau_1}^{t} \frac{dM_{23}(u)}{PY_2(u)}\right) + o_p(1)$ and $\sqrt{n}\left(\int_{\tau_1}^{t} \hat{\alpha}_{13}(t)d\mu(t) - \int_{\tau_1}^{t} \alpha_{13}(t)d\mu(t)\right) = \sqrt{n}\,\mathbb{P}_n \left(\int_{\tau_1}^{t} \frac{dM_{13}(u)}{PY_1(u)}\right) + o_p(1)$. Hence under the null hypothesis,

$$\sqrt{n}\,\left(\hat{A}_{23,\tau_1}(t) - \hat{A}_{13,\tau_1}(t)\right) = \sqrt{n}\,\mathbb{P}_n \varphi(t) + o_p(1)$$

where

$$\varphi(t) = \int_{\tau_1}^{t} \frac{dM_{23}(u)}{PY_2(u)} - \int_{\tau_1}^{t} \frac{dM_{13}(u)}{PY_1(u)}.$$

In Example 3.3.4, it was shown that the class of functions $\left\{\int_0^t \frac{dM(u)}{PY(u)} : t \in [0,\tau]\right\}$ is P-Donsker. Similarly it can be shown that the classes $\left\{\int_{\tau_1}^t \frac{dM_{23}(u)}{PY_2(u)} : t \in [\tau_1,\tau_2]\right\}$ and $\left\{\int_{\tau_1}^t \frac{dM_{13}(u)}{PY_1(u)} : t \in [\tau_1,\tau_2]\right\}$ are both P-Donsker. By the Donsker preservation property

(Theorem 3.3.8),

$$\mathcal{G} = \left\{ \varphi(t) = \int_{\tau_1}^{t} \frac{dM_{23}(u)}{PY_2(u)} - \int_{\tau_1}^{t} \frac{dM_{13}(u)}{PY_1(u)} : t \in [\tau_1, \tau_2] \right\} \tag{5.4}$$

formed by a finite sum of Donsker classes is still P-Donsker.

Hence

$$\sqrt{n}\ \left(\hat{A}_{23,\tau_1}(t) - \hat{A}_{13,\tau_1}(t) \right) \rightsquigarrow \mathbb{G}_A \text{ in } D[\tau_1, \tau_2],$$

where $\mathbb{G}_A$ is tight Gaussian process with zero mean and variance-covariance function $P(\varphi(s)\varphi(t))$.

Next, by condition B4 on weight function $\hat{K}(t)$ and the fact that

$$\sqrt{n}\ \left(\hat{A}_{23,\tau_1}(t) - \hat{A}_{13,\tau_1}(t) \right) = O_p(1),$$

as a result of the weak convergence stated before, it follows that

$$\sup_{t \in [\tau_1, \tau_2]} \left| \left(\hat{K}(t) - K(t) \right) \cdot \sqrt{n} \left(\hat{A}_{23,\tau_1}(t) - \hat{A}_{13,\tau_1}(t) \right) \right| = o_p(1) O_p(1) = o_p(1).$$

Hence also due to the asymptotic linearity of $\hat{A}_{23,\tau_1}(t) - \hat{A}_{13,\tau_1}(t)$,

$$\begin{aligned}
\sqrt{n}\, \hat{W}(t) =& \sqrt{n}\ \hat{K}(t) \left(\hat{A}_{23,\tau_1}(t) - \hat{A}_{13,\tau_1}(t) \right) \\
=& \sqrt{n}\ K(t) \left(\hat{A}_{23,\tau_1}(t) - \hat{A}_{13,\tau_1}(t) \right) + o_p(1) \\
=& \sqrt{n}\ \mathbb{P}_n K(t) \varphi(t) + o_p(1).
\end{aligned}$$

Since $\mathcal{G}$ is P-Donsker, and weight function $K(t)$ is fixed and uniformly bounded on $[\tau_1, \tau_2]$ due to condition B4, it can be shown that $\{K(t)\varphi(t) : t \in [\tau_1, \tau_2]\}$ is Donsker and thus

$$\sqrt{n}\,\hat{W}(\cdot) \rightsquigarrow \mathbb{G}_W \text{ in } D[\tau_1, \tau_2],$$

where $\mathbb{G}_W$ is a tight zero-mean Gaussian process with variance-covariance function

$$\Sigma_W(s,t) = P\left(K(s)\varphi_1(s)\right)\left(K(t)\varphi_1(t)\right).$$

$\Sigma_W(s,t)$ can be consistently (in probability) estimated by

$$\hat{\Sigma}_W(s,t) = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{K}(s)\hat{\varphi}_i(s)\right)\left(\hat{K}(t)\hat{\varphi}_i(t)\right),$$

where

$$\begin{aligned}\hat{\varphi}_i(t) &= \int_{\tau_1}^{t}\frac{d\hat{M}_{23,i}(u)}{\mathbb{P}_n Y_2(u)} - \int_{\tau_1}^{t}\frac{d\hat{M}_{13,i}(u)}{\mathbb{P}_n Y_1(u)} \\ &= \int_{\tau_1}^{t}\frac{dN_{23,i}(u) - Y_{2,i}(t)d\hat{A}_{23,i}(u)}{\mathbb{P}_n Y_2(u)} - \int_{\tau_1}^{t}\frac{dN_{13,i}(u) - Y_{1,i}(t)d\hat{A}_{13,i}(u)}{\mathbb{P}_n Y_1(u)}.\end{aligned}$$

□

Next, the asymptotic distribution of the test statistic T_{LN} under the null hypothesis is proved.

Theorem 5.2.2. Assume conditions $B1-B4$ are satisfied. Under the null hypothesis,

$$\frac{T_{LN}}{\hat{\sigma}/\sqrt{n}} \xrightarrow{D} N(0,1), \tag{5.5}$$

where

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\int_{\tau_1}^{\tau_2}\hat{K}(t)\hat{\varphi}_i(t)d\mu(t)\right)^2. \tag{5.6}$$

Proof. Since $\hat{A}_{23,\tau_1}(t)-\hat{A}_{13,\tau_1}(t)$ is asymptotically linear with influence functions $\varphi_i(t)$,

$$\begin{aligned}\sqrt{n}\,T_{LN} &= \sqrt{n}\int_{\tau_1}^{\tau_2}\hat{W}(t)d\mu(t)\\ &= \sqrt{n}\int_{\tau_1}^{\tau_2}\hat{K}(t)\left(\hat{A}_{23}(\tau_1,t)-\hat{A}_{13}(\tau_1,t)\right)d\mu(t)\\ &= \sqrt{n}\,\mathbb{P}_n\int_{\tau_1}^{\tau_2}K(t)\varphi(t)\;d\mu(t)+o_p(1).\end{aligned}$$

Since T_{LN} is asymptotically linear with influence function $\int_{\tau_1}^{\tau_2}K(t)\varphi(t)\;d\mu(t)$, its asymptotic distribution is

$$\sqrt{n}\,T_{LN}\xrightarrow{D}N(0,\sigma^2),$$

where

$$\sigma^2 = P\left(\int_{\tau_1}^{\tau_2}K(t)\varphi(t)\;d\mu(t)\right)^2,$$

and it can be consistently (in probability) estimated by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\int_{\tau_1}^{\tau_2}\hat{K}(t)\hat{\varphi}_i(t)\;d\mu(t)\right)^2,$$

where

$$\hat{\varphi}_i(t) = \int_{\tau_1}^{t}\frac{dN_{23,i}(u)-Y_{2,i}(t)d\hat{A}_{23,i}(u)}{\mathbb{P}_nY_2(u)} - \int_{\tau_1}^{t}\frac{dN_{13,i}(u)-Y_{1,i}(t)d\hat{A}_{13,i}(u)}{\mathbb{P}_nY_1(u)}.$$

□

Remark 5.2.1. The linear nonparametric test is not consistent when the curves of cumulative transition intensity functions cross each other, which can be observed in the simulation (Section 5.3). This can be well explained since the difference of AUC can be positive or negative for certain sections resulting in a potential zero sum.

5.2.2 Kolmogorov-Smirnov-Type Nonparametric Test

Although the linear test for comparing AUC between two cumulative transition intensity functions is intuitive to understand, it may not be powerful under crossing curves scenarios. Hence two other omnibus tests, Kolmogorov-Smirnov-type test and L_2-distance-type test, which are expected more powerful than the linear test when comparing crossing curves, are also developed. The asymptotic distribution of Kolmogorov-Smirnov-type test is studied in this part. The null hypothesis (5.1) can also be expressed as

$$H_0 : \sup_{t\in[\tau_1,\tau_2]} |A_{23,\tau_1}(t) - A_{13,\tau_1}(t)| = 0.$$

This leads to the Kolmogorov-Smirnov-type test, which compares the maximum distance between the two cumulative transition intensity functions within the interval $[\tau_1, \tau_2]$. With consideration of weight functions, the test statistic is

$$T_{KS} = \sup_{t\in[\tau_1,\tau_2]} \left|\hat{W}(t)\right|$$

Theorem 5.2.3. Assume conditions $B1-B4$ are satisfied. Under the null hypothesis,

$$\sqrt{n}\, T_{KS} \xrightarrow{D} \sup_{t\in[\tau_1,\tau_2]} |\mathbb{G}_W| .$$

Proof. Based on Theorem 5.2.1 and by continuous mapping theorem it follows. □

Theorem 5.2.4. For every fixed alternative hypothesis, the Komogorov-Smirnov-type test is consistent.

Proof. Due to the uniform convergence of $\hat{K}(t)$ by condition B4 and the uniform convergence of Nelson-Aalen estimator,

$$\hat{K}(t) \xrightarrow{P} K(t) \text{ uniformly over } [\tau_1, \tau_2]$$

and

$$\hat{A}_{23,\tau_1}(t) - \hat{A}_{13,\tau_1}(t) \xrightarrow{P} A_{23,\tau_1}(t) - A_{13,\tau_1}(t),$$

hence the test statistic

$$T_{KS} \xrightarrow{P} \sup_{t\in[\tau_1,\tau_2]} |K(t)\,(A_{23,\tau_1}(t) - A_{13,\tau_1}(t))| ,$$

by continuous mapping theorem.

By Lemma 14.15 in Van der Vaart (2000), which states that if T_n is a sequence of statistics such that $T_n \xrightarrow{P} \mu(\theta)$ uniformly for every θ, then the test which rejects the $H_0 : \theta = 0$ for large values of T_n is consistent against every θ when $\mu(\theta) > \mu(0)$, therefore the Kolmogorov-Smirnov-type test is consistent. □

Theorem 5.2.5. Assume conditions $B1-B4$ are satisfied. Under the null hypothesis and conditional on the observed data,

$$\sqrt{n}\,\mathbb{P}_n\hat{K}(\cdot)\hat{\varphi}(\cdot)\xi \rightsquigarrow \mathbb{G}_W \text{ in } D[\tau_1,\tau_2],$$

where ξ is standard-normal distributed and $\mathbb{G}_W$ is a tight Gaussian process with mean zero and variance-covariance function

$$\Sigma_W(s,t) = P\left(K(s)\varphi_1(s)\right)\left(K(t)\varphi_1(t)\right).$$

Proof. As it has been shown $\{K(t)\varphi(t) : t \in [\tau_1,\tau_2]\}$ is a Donsker class in Theorem 5.2.1, and by the conditional multiplier central limit theorem (Theorem 3.3.12), we have

$$\sqrt{n}\,\mathbb{P}_n K(t)\varphi(t)\xi \rightsquigarrow \mathbb{G}_W$$

when conditioning on the observed data, where ξ is $N(0,1)$ distributed.

It remains to prove $\sup_{t\in[\tau_1,\tau_2]}\left|\sqrt{n}\,\mathbb{P}_n\left(\hat{K}(t)\hat{\varphi}(t) - K(t)\varphi(t)\right)\xi\right| = o_p(1)$ holds unconditionally on the observed data. It is not hard to see that

$$\sup_{t\in[\tau_1,\tau_2]}\left|\sqrt{n}\,\mathbb{P}_n\left(\hat{K}(t)\hat{\varphi}(t) - K(t)\varphi(t)\right)\xi\right| \le S_1 + S_2$$

holds due to triangle inequality theorem, where

$$S_1 = \sup_{t\in[\tau_1,\tau_2]}\left|\sqrt{n}\;\hat{K}(t)\mathbb{P}_n\left[\left(\hat{\varphi}(t) - \varphi(t)\right)\xi\right]\right|$$

and

$$S_2 = \sup_{t\in[\tau_1,\tau_2]} \left|(\hat{K}(t) - K(t)) \cdot \sqrt{n}\, \mathbb{P}_n\varphi(t)\xi\right|.$$

It can be easily shown that $S_2 = o_p(1)$ due to condition B4 $\sup_{t\in[\tau_1,\tau_2]} \left|\hat{K}(t) - K(t)\right| = o_p(1)$, and Theorem 3.3.12 $\sqrt{n}\, \mathbb{P}_n\varphi(t)\xi = O_p(1)$.

To prove $S_1 = o_p(1)$ it requires some work. Due to condition B4 and triangle inequality, it follows that

$$\begin{aligned} S_1 &\leq \sup_{t\in[\tau_1,\tau_2]} \left|\sqrt{n}\, \mathbb{P}_n \left[(\hat{\varphi}(t) - \varphi(t))\, \xi\right]\right| \cdot [o_p(1) + O(1)] \\ &\leq (S_{11} + S_{12}) \cdot O_p(1) \end{aligned}$$

where

$$S_{11} = \sup_{t\in[\tau_1,\tau_2]} \left|\sqrt{n}\, \mathbb{P}_n\xi \left(\int_{\tau_1}^{t} \frac{d\hat{M}_{23}(u)}{\mathbb{P}_n Y_2(u)} - \int_{\tau_1}^{t} \frac{dM_{23}(u)}{PY_2(u)}\right)\right|$$

and

$$S_{12} = \sup_{t\in[\tau_1,\tau_2]} \left|\sqrt{n}\, \mathbb{P}_n\xi \left(\int_{\tau_1}^{t} \frac{d\hat{M}_{13}(u)}{\mathbb{P}_n Y_1(u)} - \int_{\tau_1}^{t} \frac{dM_{13}(u)}{PY_1(u)}\right)\right|.$$

Next $S_{11} = o_p(1)$ is proved by showing $S_{11} \leq S_{111} + S_{112}$, where

$$S_{111} = \sup_{t\in[\tau_1,\tau_2]} \left|\sqrt{n}\mathbb{P}_n\xi \cdot \int_{\tau_1}^{t} \left(\frac{1}{\mathbb{P}_n Y_2(u)} - \frac{1}{PY_2(u)}\right) d\hat{M}_{23}(u)\right| = o_p(1)$$

and

$$S_{112} = \sup_{t\in[\tau_1,\tau_2]} \left|\sqrt{n}\mathbb{P}_n\xi \cdot \int_{\tau_1}^{t} \frac{Y_2(u)}{PY_2(u)} d\left[\hat{A}_{23}(t) - A_{23}(t)\right]\right| = o_p(1).$$

$$
\begin{aligned}
S_{111} &= \sup_{t\in[\tau_1,\tau_2]} \left| \int_{\tau_1}^{t} \sqrt{n} \left(\frac{1}{\mathbb{P}_n Y_2(u)} - \frac{1}{PY_2(u)} \right) \mathbb{P}_n \xi [dN_{23}(u) - Y_2(u) d\hat{A}_{23}(u)] \right| \\
&\leq \sup_{t\in[\tau_1,\tau_2]} \left| \int_{\tau_1}^{t} \sqrt{n} \left(\frac{1}{\mathbb{P}_n Y_2(u)} - \frac{1}{PY_2(u)} \right) d[\mathbb{P}_n \xi N_{23}(u)] \right| + \\
&\quad \sup_{t\in[\tau_1,\tau_2]} \left| \int_{\tau_1}^{t} \sqrt{n} \left(\frac{1}{\mathbb{P}_n Y_2(u)} - \frac{1}{PY_2(u)} \right) \mathbb{P}_n \xi Y_2(u) d\hat{A}_{23}(u) \right|
\end{aligned}
$$

By functional delta method, $\sqrt{n}\left(\frac{1}{\mathbb{P}_n Y_2(u)} - \frac{1}{PY_2(u)}\right)$ converges weakly to a tight process, say $\mathbb{G}_2$, and $\mathbb{P}_n \xi N_{23}(u) \xrightarrow{as^*} P\xi N_{23}(u)$ by P-Glivenko-Cantelli property. Hence by Theorem 3.3.10,

$$
\int_{\tau_1}^{t} \sqrt{n} \left(\frac{1}{\mathbb{P}_n Y_2(u)} - \frac{1}{PY_2(u)} \right) d[\mathbb{P}_n \xi N_{23}(u)] \rightsquigarrow \int_{\tau_1}^{t} \mathbb{G}_2(u) dP N_{23}(u) \cdot P\xi = 0.
$$

Similarly, by Theorem 3.3.10 $\int_{\tau_1}^{t} \sqrt{n}\left(\frac{1}{\mathbb{P}_n Y_2(u)} - \frac{1}{PY_2(u)}\right) d\hat{A}_{23}(u) \rightsquigarrow \int_{\tau_1}^{t} \mathbb{G}_2(u) dA_{23}(u)$ and $\mathbb{P}_n \xi = o_p(1)$, it can be shown that

$$
\int_{\tau_1}^{t} \sqrt{n} \left(\frac{1}{\mathbb{P}_n Y_2(u)} - \frac{1}{PY_2(u)} \right) \mathbb{P}_n \xi Y_2(u) d\hat{A}_{23}(u) = o_p(1).
$$

Thus $S_{111} = o_p(1)$.

Since $\mathbb{P}_n \xi = o_p(1)$ and $\sqrt{n}[\hat{A}_{23}(u) - A_{23}(u)]$ converges weakly to a tight, mean zero process, by Theorem 3.3.11 it can be shown that $S_{112} = o_p(1)$.

As $S_{11} = o_p(1)$ has been proved, similarly $S_{12} = o_p(1)$ can be shown, hence $S_1 = o_p(1)$.

Since $S_1 = o_p(1)$ and $S_2 = o_p(1)$, we have

$$
\sup_{t\in[\tau_1,\tau_2]} \left| \sqrt{n}\, \mathbb{P}_n \left(\hat{K}(t)\hat{\varphi}(t) - K(t)\varphi(t) \right) \xi \right| = o_p(1).
$$

Therefore,

$$\sqrt{n}\,\mathbb{P}_n\hat{K}(t)\hat{\varphi}(t)\xi \rightsquigarrow \mathbb{G}_W.$$

□

Remark 5.2.2. By Theorem 5.2.1 and Theorem 5.2.5, $\hat{W}(t)$ and $\mathbb{P}_n\hat{K}(t)\hat{\varphi}(t)\xi$ (conditional on data) have the same asymptotic distribution $\mathbb{G}_W$. Hence by continuous mapping theorem, $T_{KS} = \sup_{t\in[\tau_1,\tau_2]}\left|\hat{W}(t)\right|$ and $\sup_{t\in[\tau_1,\tau_2]}\left|\mathbb{P}_n\hat{K}(t)\hat{\varphi}(t)\xi\right|$ (conditional on data) have the same asymptotic distribution $\sup_{t\in[\tau_1,\tau_2]}|\,\mathbb{G}_W|$.

It is a daunting job to analytically calculate the probability of the asymptotic distribution of this omnibus Kolmogorov-Smirnov-type test statistic under the null hypothesis. Remark 5.2.2 provides an alternative and practical way of calculating the significance level for this test, which can be obtained through the following steps.

1. Simulate R samples of independent standard normal variables ξ_i with size n from N(0,1), that is $\{\xi_{ir}, i = 1, \cdots, n\}$ for $r = 1, \cdots, R$;
2. Calculate $\sup_{t\in[\tau_1,\tau_2]}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\hat{K}(t)\hat{\varphi}_i(t)\xi_{ir}\right)\right|$ for each of $r = 1, \cdots, R$ and compare them to the calculated test statistic T_{KS} based on the observed data;
3. The significance level is the proportion of $\sup_{t\in[\tau_1,\tau_2]}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\hat{K}(t)\hat{\varphi}_i(t)\xi_{ir}\right)\right|$ values greater than or equal to the calculated T_{KS} value.

5.2.3 L_2-Distance-Type Nonparametric Test

Besides Kolmogorov-Smirnov-type test, another omnibus test based on the L_2-distance is developed. This test might be more powerful against some alternative hypotheses. The test statistic is

$$T_{L_2} = \left\{\int_{\tau_1}^{\tau_2}\hat{W}(t)^2\,d\mu(t)\right\}^{1/2}.$$

Theorem 5.2.6. Assume conditions $B1-B4$ are satisfied. Under the null hypothesis,

$$\sqrt{n}\, T_{L_2} \xrightarrow{D} \left\{ \int_{\tau_1}^{\tau_2} \mathbb{G}_W(t)^2 \, d\mu(t) \right\}^{1/2}. \tag{5.7}$$

Proof. Similar to the proof of Theorem 5.2.3, it can be proved by Theorem 5.2.5 and continuous mapping theorem. □

Theorem 5.2.7. For every fixed alternative hypothesis, the L_2-distance-type test is consistent.

Proof. The proof is similar to Theorem 5.2.4. Following Lemma 14.15 in Van der Vaart (2000), it can be shown

$$T_{L_2} \xrightarrow{P} \left\{ \int_{\tau_1}^{\tau_2} \left[K(t) \left(A_{23,\tau_1}(t) - A_{13,\tau_1}(t) \right) \right]^2 \, d\mu(t) \right\}^{1/2}.$$

Hence the L_2-distance-type test is consistent. □

To get the significance level from the L_2-distance-type test, the procedure is similar to Kolmogorov-Smirnov-type test as it can be shown that $\sqrt{n}\, T_{L_2}$ and $\left\{ \sqrt{n} \int_{\tau_1}^{\tau_2} \left(\mathbb{P}_n \hat{K}(t) \hat{\varphi}(t) \xi \right)^2 d\mu(t) \right\}^{1/2}$ (conditionally of the observed data) both converge in distribution to $\left\{ \int_{\tau_1}^{\tau_2} \mathbb{G}_W(t)^2 \, d\mu(t) \right\}^{1/2}$. Instead of calculating $\sup_{t \in [\tau_1, \tau_2]} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\hat{K}(t) \hat{\varphi}_i(t) \xi_{ir} \right) \right|$,

$$\left\{ \int_{\tau_1}^{\tau_2} \left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\hat{K}(t) \hat{\varphi}_i(t) \xi_{ir} \right) \right)^2 d\mu(t) \right\}^{1/2}$$

is computed for the L_2-distance-type test.

5.3 Simulation Studies

Again a progressive Markov illness-death model, which consists of three states, the transient states 1 (healthy state), 2 (diseased states), and the absorbing state 3 (death) is adopted to evaluate the hypothesis of $H_0 : \alpha_{13}(t) = \alpha_{23}(t),\ t \in [\tau_1, \tau_2]$. Suppose all subjects are at healthy state when study starts, i.e. $X(0) = 1$. In this simulation study, 6 simulation scenarios were considered: scenarios I and II are simulated under the null hypothesis, and the other four scenarios are under the alternative hypothesis; scenarios III and IV are with non-crossing curves of cumulative transition intensities; scenarios V and VI are with crossing curves. Details regarding parameter choice for the cumulative transition intensities $A_{12}(t), A_{13}(t)$, and $A_{23}(t)$ for the 6 scenarios are as follows.

- Scenario I: $A_{12}(t) = t, A_{13}(t) = A_{23}(t) = t^2$
- Scenario II: $A_{12}(t) = t, A_{13}(t) = A_{23}(t) = 1.5t^2$
- Scenario III: $A_{12}(t) = t, A_{13}(t) = t^2; A_{23}(t) = 1.2t^2$
- Scenario IV: $A_{12}(t) = t, A_{13}(t) = t^2, A_{23}(t) = 1.5t^2$
- Scenario V: $A_{12}(t) = 0.8t^{\frac{4}{7}}, A_{13}(t) = 0.5t^{1.75}, A_{23}(t) = 0.8t^{\frac{4}{7}}$
- Scenario VI: $A_{12}(t) = t, A_{13}(t) = 1.5t^{0.5}, A_{23}(t) = t^{1.2}$

The weight function $\hat{K}(t) = \frac{\bar{Y}_1(t)\bar{Y}_2(t)}{\bar{Y}_1(t)+\bar{Y}_2(t)}$ is considered for the weighted version of three proposed tests. The unweighted tests are also considered in simulation studies. Four different sets of τ_1 and τ_2 values are considered to examine how sensitive three testing procedures are to the constrained interval in terms of power performance. For the first three intervals $[\tau_1, \tau_2]$, τ_1 was chosen to be the analytic 5 (or 10) percentile of illness time with $Y_2(\tau_1) > 0$. τ_2 was similarly chosen to be the smaller value of

analytic 90 (or 95) percentile between death with and without illness times with $Y_h(\tau_2) > 0$, $h = 1, 2$. For the last set of τ_1 and τ_2, they are data dependent, which are the smallest and largest times to ensure $\inf_{t\in[\tau_1,\tau_2]} Y_i(t) > 0$, $i = 1, 2$, which form a non-fixed interval. Independent right censoring times that follow the mixture distribution of $Unif(3, 5)$ and a point-mass distribution at 5 with 50-50 chance, were generated. Samples sizes of 200, 400 and 800 were considered. All empirical rejection rates are based on 2,000 Monte Carlo samples at the two-sided significance level of 0.05. The type I error rate and power for the linear test is calculated based on the asymptotic Normal distribution with closed-form standard error of test statistic. The percentage of rejection for L_2-distance-type and Kolmogorov-Smirnov-type tests were computed based on $R = 1,000$ independent simulations of $\xi_{i1} \sim N(0, 1)$ and $\xi_{i2} \sim N(0, 1)$, $i = 1, \cdots, n$, separately for each simulated dataset. This choice of R is typically used in the literature (Bakoyannis, 2020).

Table 5.1 presents the empirical type I error rates for scenario I and II under the null hypothesis. For the weighted tests, empirical type I error rates are well-controlled around the nominal level of 0.05 for all cases considered regardless of the choices of $[\tau_1, \tau_2]$. For the unweighted tests, the empirical type I errors for all three tests are around 0.05 in most cases for the first three sets of fixed intervals $[\tau_1, \tau_2]$ except for Kolmogorov-Smirnov-type test under the first and second choices of $[\tau_1, \tau_2]$ with type I error under scenario I: 0.067, 0.067 and scenario II: 0.074, 0.073. There is slight inflation on type I error particularly for Kolmogorov-Smirnov-type test when τ_1 and τ_2 are data-dependent and chosen to ensure, $Y_2(\tau_1) > 0$, $Y_1(\tau_2) > 0$, and $Y_2(\tau_2)$, where the type I errors are all greater than 0.10. Therefore it requires more caution on the choice of τ_1 and τ_2 values to ensure the validity of the unweighted tests. The inflated

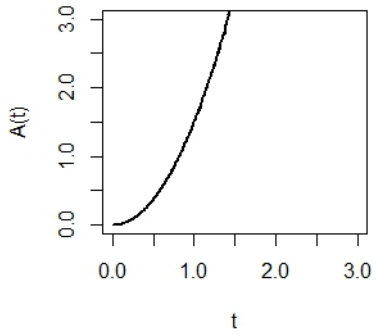
type I error might be contributed by the more pronounced difference at tail area due to chance whereas few observations are under study. However with weight function $\hat{K}(t)$ that assigns less weight to the time where there are few observations at risk, the weighted test is robust to different choice of $[\tau_1, \tau_2]$ in terms of controlling type I error.

Table 5.2 presents the empirical power for non-crossing cures scenarios III and IV. Since the difference between two curves is larger for scenario IV compared to III, the power levels for scenario IV are always greater than that for scenario III. For both unweighted and weighted tests, the empirical power increases as the sample size increases and approaches to 1, which indicates numerically the consistency of all three tests for this non-crossing curves scenario. All three tests seem to present similar performance under the same situation. The power levels for unweighted and weighted tests seem to depend on the choice of $[\tau_1, \tau_2]$. For weighted linear and L_2-distance-type tests, the power is higher when $[\tau_1, \tau_2]$ is chosen such that at-state processes satisfy $Y_2(\tau_1) > 0$, $Y_1(\tau_2) > 0$, and $Y_2(\tau_2)$. This might be due to the difference between the two curves is more pronounced at the tail area, the power is hence higher when τ_2 value is chosen based on at-state processes, as it is usually larger than the other three cases considered. Similar observation can be made on the unweighted version of these two tests, however this may be due to the slightly inflated type I error rates.

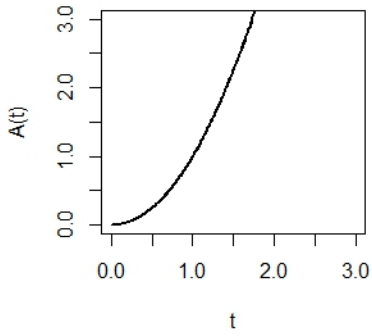
Table 5.3 presents the empirical power for crossing curves scenarios V and VI. Scenario V presents an early crossing-curve case, while scenario VI presents a late crossing-curve case. The weighted tests are more powerful compared to the unweighted tests for the late crossing scenario VI. It is also observed that the power increases as sample size increases for Kolmogorov-Smirnov-type and L_2-distance-type

tests but not for the linear test, which implies the inconsistency of linear test under certain crossing curves situations. This makes sense since when two curves cross each other, the difference of AUC will be positive for certain section, while negative for the rest, which may possibly result in a zero sum. Both unweighted and weighted tests are sensitive to the choice of $[\tau_1, \tau_2]$ which makes sense for crossing curves scenarios. The Kolmogorov-Smirnov-type and L_2-distance-type tests are more powerful for crossing curves of cumulative transition intensities compared to the linear test.

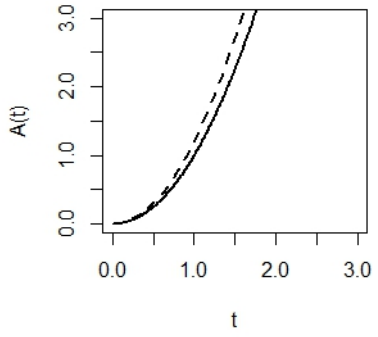
Based on the simulation studies results, the weighted tests with weight $\hat{K}(t) = \frac{\bar{Y}_1(t)\bar{Y}_2(t)}{\bar{Y}_1(t)+\bar{Y}_2(t)}$ are recommended for use in practice compared to the unweighted tests due to its well-controlled type I error across all scenarios regarding τ_1, τ_2. If $\hat{K}(t) = \frac{\bar{Y}_1(t)\bar{Y}_2(t)}{\bar{Y}_1(t)+\bar{Y}_2(t)}$ is used, it is not necessary to specify $[\tau_1, \tau_2]$ as $\hat{K}(t)$ assigns zero weight when $Y_1(t) = 0$ and/or $Y_2(t) = 0$. For the non-crossing curves scenarios, linear, Kolmogorov-Smirnov-type, and L_2-distance-type tests are all consistent and perform similarly in terms of empirical power. For the crossing curves scenarios, the linear test may not be consistent under certain situations, and the Kolmogorov-Smirnov-type and L_2-norm tests are more powerful than the linear test.



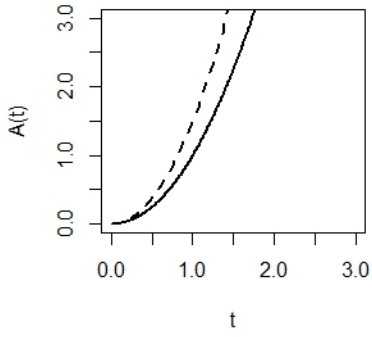
(a) I



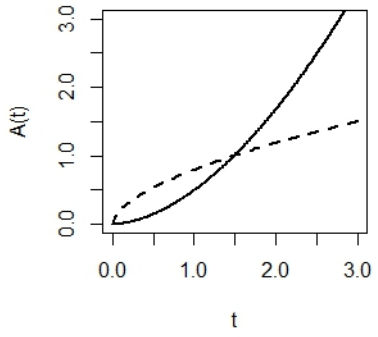
(b) II



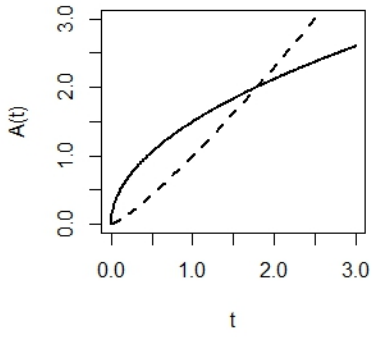
(c) III



(d) IV



(e) V



(f) VI

Figure 5.2: Scenarios I - VI of cumulative transition intensities $A_{13}(t)$ and $A_{23}(t)$ for data simulation.

Table 5.1: Simulation results of empirical type I error for the linear (Linear), Kolmogorov-Smirnov-type (KS), and L_2-distance-type (L_2) tests under the null hypothesis $H_0 : \alpha_{13}(t) = \alpha_{23}(t), t \in [\tau_1, \tau_2]$ under scenarios I and II based on 2,000 runs. For unweighted test, $K(t) = 1$; for weighted test, $\hat{K}(t) = \frac{\bar{Y}_1(t)\bar{Y}_2(t)}{\bar{Y}_1(t)+\bar{Y}_2(t)}$. Four sets of τ_1 and τ_2 are considered. For the first three scenarios, τ_1 is chosen to be the analytic 5 (or 10) percentile of illness time, and τ_2 was the smaller analytic 90 (or 95) percentile between death with and without illness times; for the last scenario, τ_1 and τ_2 are data dependent and form a non-fixed interval, which are the smallest and largest times such that $Y_1(\tau_2)$, $Y_2(\tau_1)$, and $Y_2(\tau_2)$ are all non-zero for each simulated dataset.

			Unweighted test			Weighted test		
Scenario	$[\tau_1, \tau_2]$	n	Linear	KS	L_2	Linear	KS	L_2
I	5% - 90% [0.05, 2.30]	200	0.053	0.067	0.056	0.054	0.046	0.043
		400	0.054	0.061	0.056	0.052	0.048	0.050
		800	0.046	0.054	0.049	0.049	0.049	0.049
	10% - 90% [0.11, 2.30]	200	0.053	0.067	0.052	0.048	0.049	0.046
		400	0.053	0.062	0.057	0.052	0.052	0.049
		800	0.044	0.052	0.049	0.047	0.047	0.049
	10% - 80% [0.11, 1.61]	200	0.049	0.053	0.050	0.049	0.049	0.046
		400	0.056	0.056	0.060	0.050	0.052	0.050
		800	0.047	0.049	0.044	0.049	0.047	0.048
	Non-fixed interval	200	0.064	0.106	0.068	0.051	0.046	0.047
		400	0.056	0.110	0.073	0.051	0.050	0.054
		800	0.065	0.124	0.086	0.046	0.047	0.049
II	5% - 90% [0.05, 1.52]	200	0.049	0.074	0.051	0.053	0.055	0.050
		400	0.049	0.059	0.050	0.054	0.057	0.053
		800	0.045	0.052	0.048	0.048	0.050	0.050
	10% - 90% [0.11, 1.52]	200	0.050	0.073	0.051	0.051	0.052	0.050
		400	0.048	0.057	0.051	0.051	0.056	0.055
		800	0.048	0.052	0.046	0.049	0.051	0.048
	10% - 80% [0.11, 1.27]	200	0.054	0.056	0.045	0.054	0.053	0.053
		400	0.052	0.056	0.051	0.054	0.056	0.056
		800	0.052	0.050	0.046	0.050	0.051	0.050
	Non-fixed interval	200	0.066	0.147	0.090	0.051	0.052	0.053
		400	0.068	0.150	0.099	0.051	0.055	0.052
		800	0.078	0.170	0.106	0.048	0.049	0.052

Table 5.2: Simulation results of empirical power for the linear (Linear), Kolmogorov-Smirnov-type (KS), and L_2-distance-type (L_2) tests under the null hypothesis $H_0 : \alpha_{13}(t) = \alpha_{23}(t), t \in [\tau_1, \tau_2]$ under scenarios III and IV based on 2,000 runs. For unweighted test, $K = 1$; for weighted test, $\hat{K}(t) = \frac{\bar{Y}_1(t)\bar{Y}_2(t)}{\bar{Y}_1(t)+\bar{Y}_2(t)}$. Four sets of τ_1 and τ_2 are considered. For the first three scenarios, τ_1 is chosen to be the analytic 5 (or 10) percentile of illness time, and τ_2 was the smaller analytic 90 (or 95) percentile between death with and without illness times; for the last scenario, τ_1 and τ_2 are data dependent and form a non-fixed interval, which are the smallest and largest times such that $Y_1(\tau_2)$, $Y_2(\tau_1)$, and $Y_2(\tau_2)$ are all non-zero for each simulated dataset.

			Unweighted test			Weighted test		
Scenario	$[\tau_1, \tau_2]$	n	Linear	KS	L_2	Linear	KS	L_2
III	5% - 90% [0.05, 1.39]	200	0.184	0.209	0.221	0.139	0.125	0.150
		400	0.337	0.325	0.353	0.258	0.224	0.267
		800	0.604	0.545	0.627	0.487	0.437	0.510
	10% - 90% [0.11, 1.39]	200	0.192	0.213	0.212	0.153	0.132	0.160
		400	0.342	0.323	0.351	0.276	0.234	0.277
		800	0.608	0.543	0.622	0.504	0.447	0.522
	10% - 80% [0.11, 1.16]	200	0.145	0.187	0.177	0.120	0.132	0.136
		400	0.275	0.301	0.311	0.224	0.229	0.240
		800	0.499	0.554	0.552	0.426	0.439	0.460
	Non-fixed interval	200	0.288	0.303	0.273	0.183	0.126	0.174
		400	0.438	0.370	0.385	0.338	0.224	0.300
		800	0.628	0.441	0.501	0.591	0.432	0.554
IV	5% - 90% [0.05, 1.24]	200	0.650	0.663	0.688	0.547	0.518	0.580
		400	0.916	0.917	0.935	0.846	0.836	0.879
		800	0.998	0.996	0.999	0.993	0.992	0.995
	10% - 90% [0.11, 1.24]	200	0.660	0.665	0.694	0.577	0.526	0.595
		400	0.927	0.919	0.942	0.860	0.848	0.891
		800	0.998	0.996	1	0.997	0.992	0.996
	10% - 80% [0.11, 1.04]	200	0.539	0.618	0.593	0.455	0.501	0.516
		400	0.832	0.892	0.888	0.767	0.820	0.818
		800	0.992	0.995	0.995	0.975	0.987	0.989
	Non-fixed interval	200	0.799	0.623	0.723	0.684	0.521	0.651
		400	0.954	0.767	0.887	0.932	0.838	0.926
		800	0.997	0.874	0.975	0.998	0.991	0.998

Table 5.3: Simulation results of empirical power for the linear (Linear), Kolmogorov-Smirnov-type (KS), and L_2-distance-type (L_2) tests under the null hypothesis $H_0 : \alpha_{13}(t) = \alpha_{23}(t), t \in [\tau_1, \tau_2]$ under scenarios V and VI based on 2,000 runs. For unweighted test, $K(t) = 1$; for weighted test, $\hat{K}(t) = \frac{\bar{Y}_1(t)\bar{Y}_2(t)}{\bar{Y}_1(t)+\bar{Y}_2(t)}$. Four sets of τ_1 and τ_2 are considered. For the first three scenarios, τ_1 is chosen to be the analytic 5 (or 10) percentile of illness time, and τ_2 was the smaller analytic 90 (or 95) percentile between death with and without illness times; for the last scenario, τ_1 and τ_2 are data dependent and form a non-fixed interval, which are the smallest and largest times such that $Y_1(\tau_2)$, $Y_2(\tau_1)$, and $Y_2(\tau_2)$ are all non-zero for each simulated dataset.

			Unweighted test			Weighted test		
Scenario	$[\tau_1, \tau_2]$	n	Linear	KS	L_2	Linear	KS	L_2
V	5% - 90% [0.01, 2.39]	200	0.096	0.804	0.754	0.124	0.764	0.766
		400	0.099	0.987	0.998	0.240	0.991	0.998
		800	0.108	1	1	0.507	1	1
	10% - 90% [0.01, 2.39]	200	0.140	0.860	0.811	0.069	0.690	0.743
		400	0.202	0.995	0.998	0.113	0.989	0.999
		800	0.301	1	1	0.242	1	1
	10% - 80% [0.03, 1.95]	200	0.066	0.519	0.461	0.164	0.688	0.591
		400	0.085	0.884	0.951	0.345	0.988	0.989
		800	0.128	1	1	0.660	1	1
	Non-fixed interval	200	0.616	0.988	0.980	0.048	0.708	0.834
		400	0.936	1	1	0.068	0.991	0.997
		800	0.997	1	1	0.141	1	1
VI	5% - 90% [0.05, 2.00]	200	0.066	0.229	0.144	0.227	0.527	0.418
		400	0.076	0.321	0.259	0.332	0.772	0.674
		800	0.087	0.555	0.633	0.538	0.963	0.932
	10% - 90% [0.11, 2.00]	200	0.067	0.294	0.168	0.085	0.308	0.231
		400	0.055	0.456	0.270	0.107	0.500	0.387
		800	0.063	0.698	0.574	0.146	0.776	0.725
	10% - 80% [0.11, 1.15]	200	0.191	0.098	0.153	0.244	0.305	0.261
		400	0.282	0.141	0.242	0.364	0.493	0.403
		800	0.476	0.307	0.468	0.591	0.766	0.676
	Non-fixed interval	200	0.188	0.493	0.454	0.435	0.830	0.777
		400	0.258	0.708	0.716	0.637	0.960	0.948
		800	0.454	0.882	0.933	0.866	0.992	0.989

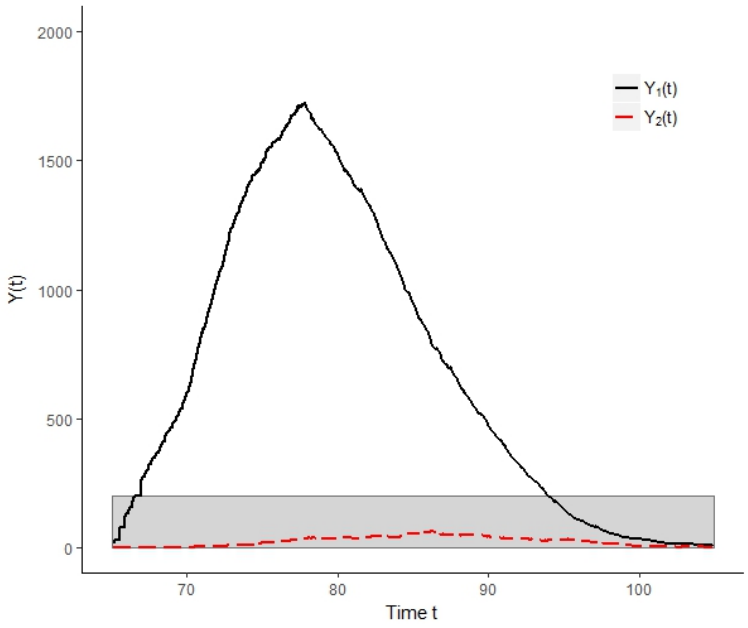
5.4 Case Study: Indianapolis-Ibadan Dementia Project Data

As in Chapter 4, we were interested in exploring whether the risk of death would be modified by dementia status in the elderly population while using the GFCMM. Here all three proposed nonparametric tests, the linear test, the Kolmogorov-Smirnov-type test, and the L_2-distance-type test are applied to the dataset from the Indianapolis-Ibadan Dementia Project (IIDP) to directly test the null hypothesis $H_0 : \alpha_{13}(t) = \alpha_{23}(t),\ t \in [\tau_1, \tau_2]$. One thing important to note is that only elder adults over 65 years old at the first visit were enrolled in the study. Details regarding IIDP data can be found in Section 4.3.

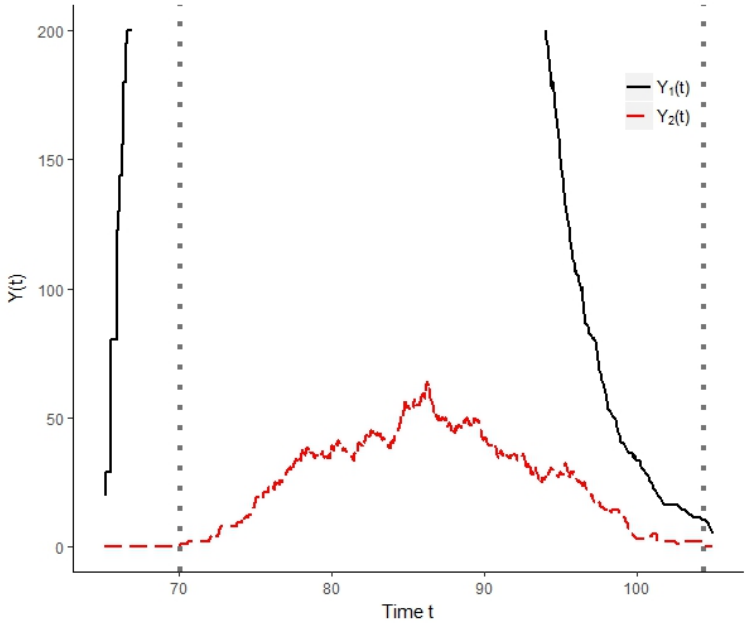
Based on Figure 5.3, the smallest time τ_1 is 70 years old to ensure $Y_2(\tau_1) > 0$, while the largest time τ_2 can attain is 104 years old to have $Y_i(\tau_2) > 0$, $i = 1, 2$. Hence the estimated cumulative transition intensities from the healthy state to death and from dementia to death are compared within the age range of 70 and 104 (Figure 5.4). Figure 5.4 shows the cumulative transition intensity from dementia to death is always much higher than that from the healthy state to death, and the two curves do not cross in the studied age range. The weight function $\hat{K}(t) = \frac{\bar{Y}_1(t)\bar{Y}_2(t)}{\bar{Y}_1(t)+\bar{Y}_2(t)}$ is considered. The p-values from the weighted linear, Kolmogorov-Smirnov-type, and L_2-distance-type tests are all < 0.001. This provides sufficient evidence to reject H_0 and conclude that dementia increases the risk to death.

Recall in Chapter 4, the proposed practical guideline including a score test was applied to the IIDP dataset. The resulted p-value of 0.17 implies the use of restricted version of GFCMM, and hence it was concluded that the occurrence of dementia increases the risk to mortality due to the nature of the restricted model. The results

of the proposed nonparametric tests in this chapter also suggest that mortality risk increases after having dementia, and this is consistent with the findings from Chapter 4 using GFCMM. Hence there is more empirical evidence to conclude that dementia does increase the risk to death.



(a)



(b)

Figure 5.3: At-state processes $Y_1(t)$ and $Y_2(t)$ in IIDP data. (a) a full-scale plot; (b) a zoomed-in plot for the grey area in (a).

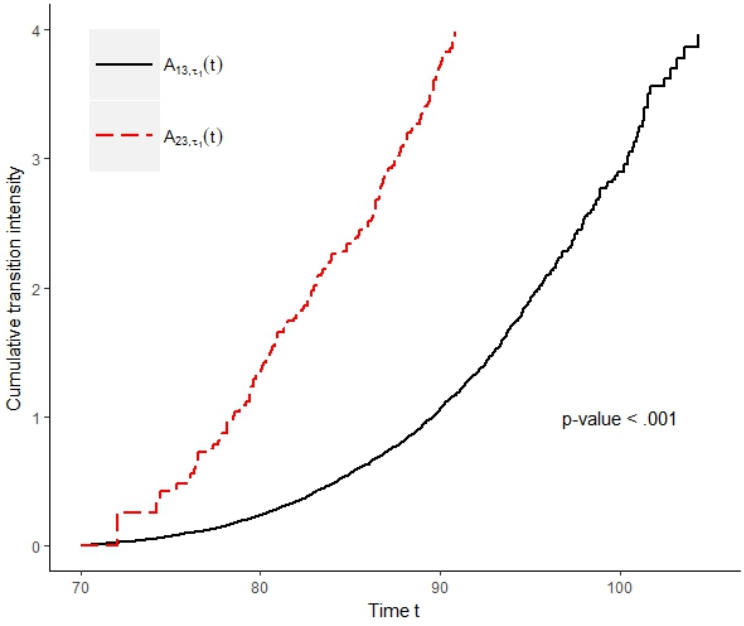


Figure 5.4: Cumulative transition intensities $\hat{A}_{13,\tau_1}(t)$ and $\hat{A}_{23,\tau_1}(t)$, $t \in [\tau_1, \tau_2]$, where $\tau_1 = 70$ and $\tau_2 = 104$ in the IIDP data.

5.5 Discussion

In this chapter, three nonparametric tests for transition intensity functions, including a linear test, a Kolmogorov-Smirnov-type test, and a L_2-distance-type test were developed to directly assess whether the occurrence of non-terminal event changes the risk to terminal event under a progressive Markov illness-death model. The EPT was used to show the asymptotic distributions of three test statistics under the null hypothesis. The Kolmogorov-Smirnov-type and L_2-distance-type tests were shown to be consistent. Simulation studies were conducted to illustrate that type I error rates for all three tests are well-controlled with weight function $\hat{K}(t) = \frac{\bar{Y}_1(t)\bar{Y}_2(t)}{\bar{Y}_1(t)+\bar{Y}_2(t)}$; therefore weighted tests are recommended for use in practice. The power performance of the three tests is also numerically evaluated under various scenarios, including different crossing and non-crossing curves of cumulative transition intensities. All three tests

seem to perform similarly under non-crossing curves scenarios. The Kolmogorov-Smirnov-type and L_2-distance-type tests appear to be more powerful compared to the linear test for crossing curves scenarios. Both weighted and unweighted tests are sensitive to the choice of $[\tau_1, \tau_2]$ values when evaluating their empirical power. All three weighted tests were applied to the IIDP dataset, which was analyzed under GFCMM in Chapter 4. The resulted p-values from the proposed nonparametric tests are very small indicating the occurrence of dementia increases the risk to death. These results are consistent with the findings in Chapter 4.

Multi-state models have been widely used to study the disease progression with multiple states. Much research has been done on the nonparametric estimation for the multi-state models. However nonparametric testing for multi-state models, especially for transition intensity functions have not received enough attention. Utilizaing the Markov illness-death model, Andersen et al. (2012) briefed about the potential of application of two-sample log-rank and Wilcoxon-type tests to the comparison of two transition intensities without further justification or guidance, and Bluhmki et al. (2019) developed a Kolmogorov-Smirnov test to compare a particular transition intensity function between two samples, which may not be directly applied to our case. The contribution of this research is three nonparametric tests and their asymptotic properties for comparing two transition intensities in a Markov illness-death model within the same study sample were developed. The proposed tests hence directly answered the scientific question of interest whether developing an intermediate event changes the risk to the terminal event.

The proposed tests have some strengths. First, these tests are nonparametric, which did not make any parametric or proportional hazards assumptions. Second,

the theoretical properties of the proposed test statistics are established using the EPT. Third, the proposed tests do not only apply to Markov illness-death models, but can also be adopted for non-Markov models, since the Nelson-Aalen estimators remain consistent NPMLE for cumulative transition intensities under non-Markov models (Datta and Satten, 2001). Similar to Bakoyannis (2020), if interest lies in the marginal transition intensities (unconditional on the past history) instead, i.e., $\alpha_{jk}(t) = \lim_{\Delta \to 0} \frac{P_{jk}(t,t+\Delta)}{\Delta}$, $j \neq k$, where $P_{jk}(t_0, t) = P(X(t) = k | X(t_0) = j)$ is the marginal transition probability from state j at time t_0 to state k at time t ignoring the filtration $\mathcal{F}_{t-}$, the past history up to time t, the proved theorems on asymptotics of three test statistics under the null hypothesis are still valid to use.

Chapter 6

Summary

Semi-competing risks data became an important research topic in survival data analysis and biomedical applications for the last couple of decades since it was first introducted by Fine et al. (2001). Semi-competing risks data occur when an individual is subject to a non-terminal event and a terminal event, where the non-terminal event can be ceonsored by the terminal event, but not vice versa. The copula-based models and illness-death models under multi-state modeling framework have been widely used for the analysis of semi-competing risks data. Ever since Xu et al. (2010) proposed a GFCMM, which bridges the copula models and illness-death models, it has been adopted for various applications in biomedical studies.

In this research, we were ultimately interested in addressing the scientific question of interest that whether the occurrence of non-terminal event alters the risk of terminal event. In the first part of this dissertation, we adopted the GFCMM and developed an easy-to-implement EM algorithm to compute the NPMLE of the hazards for semi-competing risks data (Section 4.2.1). Through extensive simulation studies, we uncovered the pitfalls with its NPMLE under unrestricted GFCMM, which does not always yield consistent likelihood-based inference for model parameters (Section 4.2.2); moreover, a practical guideline was provided for using the GFCMM, which includes a score test for the comparison between the restricted and unrestricted GFCMM, in the nonparametric analysis of semi-competing risk data (Section 4.2.3). However, only if the score test suggests the use of restricted model, one can conclude the occurrence of non-terminal event increases the risk to terminal event. Due to the

numerical issues with NPMLE for the unrestricted GFCMM, the research question cannot be fully addressed by the aforementioned approach for this situation.

In the second part of this dissertation, the Markov illness-death model, where the transition intensities are essentially equivalent to the marginal hazards (unconditional on frailty) defined for GFCMM was adopted, and three non-parametric tests under Markov illness-death models were developed to directly answer the scientific question of interest by comparing two transition intensity functions. These three nonparametric tests include a linear test (Section 5.2.1), a Kolmogorov-Smirnov-type test (Section 5.2.2), and a L_2-distance-type test (Section 5.2.3), where the asymptotic properties of the proposed tests are established by empirical process theory. The performance of these tests were numerically compared under various scenarios through extensive simulation (Section 5.3).

This research is of importance for the following reasons. First, it addressed the fundamental question regarding the validity of nonparametric likelihood-based inference of GFCMM under frequentist inference framework, which had not been carefully explored in any peer-reviewed publications. Second, this research did not stop at only pointing out the pitfalls in NPMLE with the unrestricted GFCMM, it took a step further and provided a guideline for the nonpaparametric analysis using GFCMM. Third, the proposed score test in the guideline might be the only valid likelihood-based inference procedure for this model given the numerical problem associated with NPMLE of the unrestricted GFCMM. Last but not least, although the research question was not fully answered in the first part of research, the research question was tackled by developing three nonparametric tests in the second part under the Markov illness-death model, where the asymptotic properties were established with empirical

process theory. Furthermore, these tests can also be adopted for non-Markov models, constructed by marginal transition intensities (unconditional on the past history).

BIBLIOGRAPHY

Andersen, P. K., O. Borgan, R. D. Gill, and N. Keiding (2012). *Statistical models based on counting processes*. Springer Science & Business Media.

Andersen, P. K. and N. Keiding (2002). Multi-state models for event history analysis. *Statistical Methods in Medical Research 11*(2), 91–115.

Bakoyannis, G. (2020). Nonparametric tests for transition probabilities in nonhomogeneous markov processes. *Journal of Nonparametric Statistics 32*(1), 131–156.

Bakoyannis, G., Y. Zhang, and C. T. Yiannoutsos (2019). Nonparametric inference for markov processes with missing absorbing state. *Statistica Sinica 29*(4), 2083–2104.

Beyersmann, J., A. Allignol, and M. Schumacher (2011). *Competing risks and multistate models with R*. Springer Science & Business Media.

Bluhmki, T., D. Dobler, J. Beyersmann, and M. Pauly (2019). The wild bootstrap for multivariate nelson–aalen estimators. *Lifetime Data Analysis 25*(1), 97–127.

Casella, G. and R. L. Berger (2002). *Statistical inference*, Volume 2. Duxbury Pacific Grove, CA.

Chapple, A. G., M. Vannucci, P. F. Thall, and S. Lin (2017). Bayesian variable selection for a semi-competing risks model with three hazard functions. *Computational Statistics & Data Analysis 112*, 170–185.

Commenges, D. (1999). Multi-state models in epidemiology. *Lifetime Data Analysis 5*(4), 315–327.

Commenges, D. (2002). Inference for multi-state models from interval-censored data. *Statistical Methods in Medical Research 11*(2), 167–182.

Commenges, D. (2003). Likelihood for interval-censored observations from multi-state models. *SORT-Statistics and Operations Research Transactions 27*(1), 1–12.

Commenges, D., P. Joly, L. Letenneur, and J.-F. Dartigues (2004). Incidence and mortality of alzheimer's disease or dementia using an illness-death model. *Statistics in Medicine 23*(2), 199–210.

Cook, R. J. and J. F. Lawless (2018). *Multistate models for the analysis of life history data*. Chapman and Hall/CRC.

Datta, S. and G. A. Satten (2001). Validity of the aalen–johansen estimators of stage occupation probabilities and nelson–aalen estimators of integrated transition hazards for non-markov models. *Statistics & Probability Letters 55*(4), 403–411.

De Wreede, L. C., M. Fiocco, and H. Putter (2010). The mstate package for estimation and prediction in non-and semi-parametric multi-state and competing risks models. *Computer Methods and Programs in Biomedicine 99*(3), 261–274.

Duchateau, L. and P. Janssen (2007). *The frailty model*. Springer Science & Business Media.

Fan, J. and J. Jiang (2007). Nonparametric inference with generalized likelihood ratio tests. *Test 16*(3), 409–444.

Fan, J., C. Zhang, J. Zhang, et al. (2001). Generalized likelihood ratio statistics and wilks phenomenon. *The Annals of Statistics 29*(1), 153–193.

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

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character 222*(594-604), 309–368.

Frydman, H. (1995). Nonparametric estimation of a markov 'illness-death' process from interval-censored observations, with application to diabetes survival data. *Biometrika 82*(4), 773–789.

Frydman, H., T. Gerds, R. Grøn, and N. Keiding (2013). Nonparametric estimation in an 'illness-death' model when all transition times are interval censored. *Biometrical Journal 55*(6), 823–843.

Frydman, H. and M. Szarek (2009). Nonparametric estimation in a markov "illness–death" process from interval censored observations with missing intermediate transition status. *Biometrics 65*(1), 143–151.

Fu, H., Y. Wang, J. Liu, P. M. Kulkarni, and A. S. Melemed (2013). Joint modeling of progression-free survival and overall survival by a bayesian normal induced copula estimation model. *Statistics in Medicine 32*(2), 240–254.

Gao, S., A. Ogunniyi, K. S. Hall, O. Baiyewu, F. W. Unverzagt, K. A. Lane, J. R. Murrell, O. Gureje, A. M. Hake, and H. C. Hendrie (2016). Dementia incidence declined in african-americans but not in yoruba. *Alzheimer's & Dementia 12*(3), 244–251.

Gillick, M. (2001). Guest editorial: pinning down frailty. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences 56*(3), M134–M135.

Han, B., M. Yu, J. J. Dignam, and P. J. Rathouz (2014). Bayesian approach for flexible modeling of semicompeting risks data. *Statistics in Medicine 33*(29), 5111–5125.

Hanagal, D. D. (2011). *Modeling survival data using frailty models*. Chapman and Hall/CRC.

Harezlak, J., S. Gao, and S. L. Hui (2003). An illness–death stochastic model in the analysis of longitudinal dementia data. *Statistics in Medicine 22*(9), 1465–1475.

Hendrie, H. C., A. Ogunniyi, K. S. Hall, O. Baiyewu, F. W. Unverzagt, O. Gureje, S. Gao, R. M. Evans, A. Ogunseyinde, A. Adeyinka, et al. (2001). Incidence of dementia and alzheimer disease in 2 communities: Yoruba residing in ibadan, nigeria, and african americans residing in indianapolis, indiana. *JAMA 285*(6), 739–747.

Hendrie, H. C., M. Zheng, W. Li, K. Lane, R. Ambuehl, C. Purnell, F. W. Unverzagt, A. Torke, A. Balasubramanyam, C. M. Callahan, et al. (2017). Glucose level decline precedes dementia in elderly african americans with diabetes. *Alzheimer's & Dementia 13*(2), 111–118.

Hougaard, P. (1999). Multi-state models: a review. *Lifetime Data Analysis 5*(3), 239–264.

Hu, C. and A. Tsodikov (2014). Joint modeling approach for semicompeting risks data with missing nonterminal event status. *Lifetime Data Analysis 20*(4), 563–583.

Jackson, C. H., L. D. Sharples, S. G. Thompson, S. W. Duffy, and E. Couto (2003). Multistate markov models for disease progression with classification error. *Journal of the Royal Statistical Society: Series D (The Statistician) 52*(2), 193–209.

James, B. D., S. E. Leurgans, L. E. Hebert, P. A. Scherr, K. Yaffe, and D. A. Bennett (2014). Contribution of alzheimer disease to mortality in the united states. *Neurology 82*(12), 1045–1050.

Jamshidian, M. (2001). A note on parameter and standard error estimation in adaptive robust regression. *Journal of Statistical Computation and Simulation 71*(1), 11–27.

Jiang, F. and S. Haneuse (2015). Simulation of semicompeting risk survival data and estimation based on multistate frailty model. *Harvard University Biostatistics Working Paper Series*.

Jiang, H., R. Chappell, and J. P. Fine (2003). Estimating the distribution of nonterminal event time in the presence of mortality or informative dropout. *Controlled clinical trials 24*(2), 135–146.

Klein, J. P., H. C. Van Houwelingen, J. G. Ibrahim, and T. H. Scheike (2016). *Handbook of survival analysis*. CRC Press.

Kosorok, M. R. (2008). *Introduction to empirical processes and semiparametric inference*. Springer.

Lee, K. H., F. Dominici, D. Schrag, and S. Haneuse (2016). Hierarchical models for semicompeting risks data with application to quality of end-of-life care for

pancreatic cancer. *Journal of the American Statistical Association 111*(515), 1075–1095.

Lee, K. H., S. Haneuse, D. Schrag, and F. Dominici (2015). Bayesian semiparametric analysis of semicompeting risks data: investigating hospital readmission after a pancreatic cancer diagnosis. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 64*(2), 253–273.

Lehmann, E. L. and G. Casella (2006). *Theory of point estimation*. Springer Science & Business Media.

Mandel, M. and R. Fluss (2009). Nonparametric estimation of the probability of illness in the illness-death model under cross-sectional sampling. *Biometrika 96*(4), 861–872.

Mau, J. (1986). Nonparametric estimation of the integrated intensity of an unobservable transition in a markov illness-death process. *Stochastic Processes and Their Applications 21*(2), 275–289.

Meira-Machado, L., C. Cadarso-Suárez, and J. de Uña-Álvarez (2007). tdc. msm: an r library for the analysis of multi-state survival data. *Computer Methods and Programs in Biomedicine 86*(2), 131–140.

Meira-Machado, L., J. de Uña-Álvarez, C. Cadarso-Suarez, and P. K. Andersen (2009). Multi-state models for the analysis of time-to-event data. *Statistical Methods in Medical Research 18*(2), 195–222.

Meira-Machado, L. and J. Roca-Pardiñas (2011). p3state. msm: Analyzing survival data from an illness-death model. *Journal of Statistical Software 38*(3), 1–18.

Nelsen, R. B. (2007). *An introduction to copulas*. Springer Science & Business Media.

Peng, L. and J. P. Fine (2007). Regression modeling of semicompeting risks data. *Biometrics 63*(1), 96–108.

Peng, M., L. Xiang, and S. Wang (2018). Semiparametric regression analysis of clustered survival data with semi-competing risks. *Computational Statistics & Data Analysis 124*, 53–70.

Perkins, A. J., S. L. Hui, A. Ogunniyi, O. Gureje, O. Baiyewu, F. W. Unverzagt, S. Gao, K. S. Hall, B. S. Musick, and H. C. Hendrie (2002). Risk of mortality for dementia in a developing country: the yoruba in nigeria. *International Journal of Geriatric Psychiatry 17*(6), 566–573.

Putter, H., M. Fiocco, and R. B. Geskus (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine 26*(11), 2389–2430.

Rao, C. (2005). Score test: historical review and recent developments. In *Advances in Ranking and Selection, Multiple Comparisons, and Reliability*, pp. 3–20. Springer.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, Volume 44, pp. 50–57. Cambridge University Press.

Resnick, S. (2003). *A probability path*. Birkhauser Verlag AG.

Ross, S. M., J. J. Kelly, R. J. Sullivan, W. J. Perry, D. Mercer, R. M. Davis, T. D. Washburn, E. V. Sager, J. B. Boyce, and V. L. Bristow (1996). *Stochastic processes*, Volume 2. Wiley New York.

Touraine, C., C. Helmer, and P. Joly (2016). Predictions in an illness-death model. *Statistical Methods in Medical Research 25*(4), 1452–1470.

Van Den Hout, A. (2016). *Multi-state survival models for interval-censored data*. Chapman and Hall/CRC.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

Van Der Vaart, A. W. and J. A. Wellner (1996). Weak convergence and empirical processes with applications to statistics. In *Weak convergence and empirical processes*, pp. 16–28. Springer.

Vaupel, J. W., K. G. Manton, and E. Stallard (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography 16*(3), 439–454.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society 54*(3), 426–482.

Wang, W. (2003). Estimating the association parameter for copula models under dependent censoring. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 65*(1), 257–273.

Wienke, A. (2010). *Frailty models in survival analysis*. Chapman and Hall/CRC.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics 9*(1), 60–62.

Xu, J., J. D. Kalbfleisch, and B. Tai (2010). Statistical analysis of illness–death processes and semicompeting risks data. *Biometrics 66*(3), 716–725.

Yu, B., J. S. Saczynski, and L. Launer (2010). Multiple imputation for estimating the risk of developing dementia and its impact on survival. *Biometrical Journal 52*(5), 616–627.

Yu, M. (2016). Improving estimation efficiency for semi-competing risks data with partially observed terminal event. *Journal of Nonparametric Statistics 28*(4), 860–874.

Yu, M. and C. T. Yiannoutsos (2015). Marginal and conditional distribution estimation from double-sampled semi-competing risks data. *Scandinavian Journal of Statistics 42*(1), 87–103.

Zeng, D., Q. Chen, M.-H. Chen, J. G. Ibrahim, and A. R. Groups (2011). Estimating treatment effects with treatment switching via semicompeting risks models: an application to a colorectal cancer study. *Biometrika 99*(1), 167–184.

Zhang, Y. and M. Jamshidian (2003). The gamma-frailty poisson model for the nonparametric estimation of panel count data. *Biometrics 59*(4), 1099–1106.

Zhou, R., H. Zhu, M. Bondy, and J. Ning (2017). Analyzing semi-competing risks data with missing cause of informative terminal event. *Statistics in Medicine 36*(5), 738–753.

CURRICULUM VITAE

Jing Li

EDUCATION

- Ph.D. in Biostatistics, Indiana University Richard M. Fairbanks School of Public Health, Indianapolis, IN 2020 (minor in Bioinformatics)
- M.S. in Statistics, The Ohio State University, Columbus, OH 2012
- B.S. in Mathematics and Statistics, University of Minnesota, Morris, MN 2010

WORKING EXPERIENCE

- Summer Biostatistics Intern, Merck, Rahway, NJ

  May 2018 - Aug. 2018
- Graduate Research Assistant, Department of Epidemiology, Indiana University Richard M. Fairbanks School of Public Health, Indianapolis, IN

  Jan. 2017 - Nov. 2019
- Graduate Teaching Assistant, Department of Biostatistics, Indiana University Richard M. Fairbanks School of Public Health, Indianapolis, IN

  Sep. 2016 - Dec. 2016
- Research Associate - Biostatistician (full-time), Department of Epidemiology and Biostatistics, Indiana University School of Public Health, Bloomington, IN

  May 2013 - Jul. 2015
- Graduate Teaching Assistant, Department of Statistics, The Ohio State University, Columbus, OH

  Aug. 2011 - May. 2013

PUBLICATIONS

- **Li, J.,** Li, X., Gathirua-Mwangi, W., & Song, Y. (2020). Prevalence and trends in dietary supplement use among US adults with diabetes: the National Health and Nutrition Examination Surveys, 1999–2014. *BMJ Open Diabetes Research and Care*, 8(1).
- Kong, X., Liu, J., **Li, J.,** Kwong, W., Koh, M., Sukijthamapan, P., Guo, J., Sun, J., & Song, Y. (2020). The effects of probiotics and oxytocin nasal spray on neuro-social behaviors of Autism Spectrum Disorders (ASD) children: study protocol for a randomized, double-blinded, placebo-controlled, parallel-group clinical trial. *Pilot and Feasibility Studies*, 6(1), 20.
- **Li, J.,** Zhang, Y., Myers, L., & Bravata, D. (2019). Power calculation in stepped-wedge cluster randomized trial with reduced intervention sustainability effect. *Journal of Biopharmaceutical Statistics*, 29(4), 663-674.
- Dawson, S., McCormick, B., & **Li, J.** (2018). A Network Analysis of Youth with Physical Disabilities Attending a Therapeutic Camp, *Therapeutic Recreation Journal*, 52(2), 154-169.
- Park, J. S., Xun, P., **Li, J.**, Morris, S. J., Jacobs, D. R., Liu, K., & He, K. (2016). Longitudinal association between toenail zinc levels and incidence of diabetes: The CARDIA trace element study. *Scientific Reports*, 6, 23155.
- Macy, J. T., **Li, J.**, Xun, P., Presson, C. C., & Chassin, L. (2015). Dual trajectories of cigarette smoking and smokeless tobacco use from adolescence to midlife among males in a Midwestern US community sample. *Nicotine & Tobacco Research*, 18(2), 186-195.

- Piatt, J. A., Nagata, S., Zahl, M., **Li, J.,** & Rosenbluth, J. P. (2016). Problematic secondary health conditions among adults with spinal cord injury and its impact on social participation and daily life. *The Journal of Spinal Cord Medicine*, 39(6), 693-698.
- Chen, Z., Zhang, G., & **Li, J.** (2015). Goodness-of-fit test for meta-analysis. *Scientific Reports*, 5, 16983.
- Chen, Z., Yang, W., Liu, Q., Yang, J. Y., **Li, J.**, & Yang, M. Q. (2014). A new statistical approach to combining p-values using gamma distribution and its application to genome-wide association study. *BMC Bioinformatics*, 15(17), S3.
- Chen, Z., Ng, H. K. T., **Li, J.**, Liu, Q., & Huang, H. (2017). Detecting associated single-nucleotide polymorphisms on the X chromosome in case control genome-wide association studies. *Statistical Methods in Medical Research*, 26(2), 567-582.