Statistical Inference Using Large Administrative Data on Multiple Event Times, with Application to Cancer Survivorship Research

by

Dongdong Li

M.Sc., University of Calgary, 2010
B.Sc., Nankai University, 2008

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Doctor of Philosophy

in the
Department of Statistics and Actuarial Science
Faculty of Science

SIMON FRASER UNIVERSITY
Fall 2018

Approval

Name:	**Dongdong Li**
Degree:	**Doctor of Philosophy (Statistics)**
Title:	***Statistical Inference Using Large Administrative Data on Multiple Event Times, with Application to Cancer Survivorship Research***
Examining Committee:	**Chair:** Jinko Graham Professor

X. Joan Hu
Senior Supervisor
Professor _____

John J. Spinelli
Co-Supervisor
Adjunct Professor _____

Mary L. McBride
Supervisor
Distinguished Scientist
Cancer Control Research
BC Cancer Agency _____

Lawrence C. McCandless
Internal Examiner
Associate Professor
Faculty of Health Sciences _____

Douglas E. Schaubel
External Examiner
Professor
School of Public Health
University of Michigan _____

Date Defended: December 20, 2018

Ethics Statement

The author, whose name appears on the title page of this work, has obtained, for the research described in this work, either:

a. human research ethics approval from the Simon Fraser University Office of Research Ethics

or

b. advance approval of the animal care protocol from the University Animal Care Committee of Simon Fraser University

or has conducted the research

c. as a co-investigator, collaborator, or research assistant in a research project approved in advance.

A copy of the approval letter has been filed with the Theses Office of the University Library at the time of submission of this thesis or project.

The original application for approval and letter of approval are filed with the relevant offices. Inquiries may be directed to those authorities.

Simon Fraser University Library
Burnaby, British Columbia, Canada

Update Spring 2016

Abstract

Motivated by the breast cancer survivorship research program at BC Cancer Agency, this dissertation develops statistical approaches to analyzing right-censored multivariate event time data.

Following the background and motivation of the research, we introduce the framework of the dissertation, and provide a literature review and a list of the research questions. A description of the motivating study data is then given together with a preliminary analysis before presenting the proposed approaches as follows.

We consider firstly estimation of the joint survivor function of multiple event times when the observations are subject to informative censoring due to a terminating event. We formulate the potential dependence of the multiple event times with the time to the terminating event by the Archimedean copulas. This may account for the informative censoring and, at the same time, allow to adapt the commonly used two-step procedure for estimating the joint distribution of the multiple event times under a copula model. We propose an easy-to-implement pseudo-likelihood based estimation procedure under the model, which reduces computational intensity compared to its MLE counterpart.

A more flexible approach is then proposed to handling informative censoring with particular attention to observations on bivariate event time potentially censored by a terminating event. We formulate the correlation of the bivariate event time with the censoring time by embedding the bivariate event time distribution in a bivariate copula model. This yields the convenience of inference under the conventional copula model. At the same time, the proposed model is more flexible, and thus potentially more appropriate in many practical situations than modeling the event times and the associated censoring time jointly by a single multivariate copula. Adapting the commonly used two-stage estimation procedure under a copula model, we develop an easy-to-implement estimator for the joint survivor function of the two event times. A by-product of the proposed approaches is an estimator for the marginal distribution of a single event time with semicompeting-risks data.

Further, we extend the approach to regression settings to explore covariate effects in either parametric or nonparametric forms. In particular, adjusting for some covariates, we compare two populations based on an event time with observations subject to informative censoring.

We conduct both asymptotic and simulation studies to examine the consistency, efficiency, and robustness of the proposed approaches. The breast cancer program that motivated this research is employed to illustrate the methodological development throughout the dissertation.

Keywords: Copula model; Efficiency and robustness; Informative censoring; Marginal distribution; Multivariate event times; Pseudolikelihood estimation; Regression analysis; Variance estimation

Dedication

To my beloved parents with gratitude, for their endless love, support and encouragement.

Acknowledgements

My deep and sincere gratitude goes to the mentor of my PhD study, Professor Joan Hu, who inspired and guided me through this journey. I feel lucky to have her as my supervisor, and she has always been patient and helpful with my work. I admire her extensive knowledge of statistics, wholehearted commitment to research, and endless patience with students. I really appreciate and will remember her guidance both in research and in life. She has encouraged me in the pursuit and exploration of deeper knowledge and broader thinking, enlightened me with her experience and knowledge, and reminded me of the right direction to go in. This dissertation would not have been possible without her.

I would also like to acknowledge my dissertation committee members. My co-supervisor, Professor John Spinelli, has been supportive throughout my PhD study. He responded to my e-mails promptly, revised my papers, and invited me to his group meeting with other students. I feel very grateful to the PI of the breast cancer study in BC Cancer, Professor Mary McBride, who provided data access and expert advice and comments on my research work. Many thanks to Professor Lawrence McCandless for serving on my committee and Professor Jinko Graham for chairing the defence. Moreover, I would like to thank Professor Douglas Schaubel from the University of Michigan for serving as the external examiner and for travelling here for my dissertation defence.

I would like to express my appreciation to the faculty and staff of the Department of Statistics and Actuarial Science for offering a nice environment in which to learn. The whole department is like a a big family. Special thanks go to my fellow graduate students for their support and friendship. You have made my life brighter and more colorful. Particular mentions go to Yi Xiong, Trevor Thomson, Perry Sang, Zhiyang Zhou, Yuping Yang, Huijing Wang, Chenlu Shi, Tianyu Guan, Terry Tang, Joanna Zhao, Moyan Mei, Charlie Zhou and Bobby Han. I have learnt lots from discussions with them about statistical and life-related topics. The PhD journey has been enjoyable because of them.

Special Acknowledgement: The statistical research was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Canadian Statistical Sciences Institute (CANSSI). The breast cancer research study was funded by the Canadian Institutes of Health Research (TT7-128272). This study was approved by the BC Cancer/University of British Columbia Ethics Board and Simon Fraser University Ethics

Board. BC Cancer (`www.bccancer.bc.ca`) and the BC Ministry of Health approved access to and use of the data; this was facilitated by Population Data BC (`www.popdata.bc.ca`).

Disclaimer: All inferences, opinions, and conclusions drawn in this dissertation are those of the authors, and do not reflect the opinions or policies of the British Columbia Data Steward(s).

Table of Contents

Approval ii

Ethics Statement iii

Abstract iv

Dedication vi

Acknowledgements vii

Table of Contents ix

List of Tables xii

List of Figures xiv

1 Introduction 1

1.1 Background and Motivation . . . 1

1.1.1 General Background: Health Care System in Canada . . . 2

1.1.2 BC Breast Cancer Survivorship (BC-BRCAS) Program . . . 2

1.1.3 Motivation of the Thesis Research . . . 3

1.2 Literature Review . . . 5

1.2.1 Multivariate Event Time . . . 5

1.2.2 Informative Censoring . . . 10

1.3 Notation and Framework . . . 11

1.4 Objectives and Thesis Outline . . . 11

1.4.1 Objectives . . . 11

1.4.2 Thesis Outline . . . 11

2 Data Description and Preliminary Analysis 13

2.1 Introduction . . . 13

2.2 Description of BC-BRCAS Data . . . 13

2.3 Analysis of BC-BRCAS Data (I): Preliminary Data Analyses . . . 15

2.4 Statistical Challenges . . . 16

3 Multiple Event Times in the Presence of Informative Censoring Using Copula - Part One **21**
3.1 Introduction . . . 21
3.2 Notation and Modeling . . . 22
3.2.1 Notation . . . 22
3.2.2 Model Specification . . . 23
3.3 Pseudolikelihood-Based Inference Procedure . . . 23
3.3.1 Likelihood Function Based on the Available Data . . . 24
3.3.2 Pseudo-MLE of Association Parameter . . . 25
3.3.3 Resulting Estimators for Marginal and Joint Survivor Function . . . 26
3.4 Asymptotic Properties . . . 27
3.4.1 Asymptotic Properties for Bivariate Case . . . 27
3.4.2 Asymptotics of a General Pseudo-MLE . . . 33
3.5 Simulation Study . . . 36
3.5.1 Data Generation . . . 36
3.5.2 Simulation Outcomes for $J = 1$ . . . 37
3.5.3 Simulation Outcomes for $J = 2$ . . . 38
3.6 Analysis of BC-BRCAS Data (II) . . . 92
3.6.1 Study Description . . . 92
3.6.2 Correlations amongst Event Times and Death . . . 92
3.7 Discussion . . . 97

4 Multiple Event Times in the Presence of Informative Censoring Using Copula - Part Two: a Flexible Approach **98**
4.1 Modeling . . . 99
4.1.1 Model Specification . . . 99
4.1.2 More on Modeling . . . 100
4.2 Pseudolikelihood-Based Estimation Procedures . . . 101
4.2.1 Estimating Association Parameters with the Observed-Data . . . 101
4.2.2 Resulting Estimators for Marginal and Joint Survivor Function . . . 103
4.3 Simulation Study . . . 104
4.3.1 Simulation Setting and Data Generation . . . 104
4.3.2 Consistency and Efficiency . . . 105
4.3.3 Robustness to Model Misspecification . . . 106
4.3.4 Flexibility on Modeling Correlation Between Event Times . . . 107
4.4 Analysis of BC-BRCAS Data (III) . . . 126
4.4.1 Study Description . . . 126
4.4.2 Estimates of Correlations between Event Times . . . 126
4.5 Discussion . . . 132

5 Regression Analysis of Bivariate Event Time with Observations on Response Subject to Informative Censoring **133**
5.1 Introduction . . . 133
5.2 Notation and Modeling . . . 135
5.2.1 Notation . . . 135
5.2.2 Model Specification . . . 135
5.2.3 More on Modeling . . . 137
5.3 Pseudolikelihood-Based Estimation Procedures . . . 139
5.3.1 Estimating the Parameters with the Observed-Data . . . 139
5.3.2 Resulting Estimators for Marginal and Joint Survivor Function . . . 139
5.3.3 Asymptotic Properties . . . 140
5.4 Simulation . . . 141
5.4.1 Setting and Data Generation . . . 141
5.4.2 Simulation Outcomes . . . 142
5.5 Analysis of BC-BRCAS Data (IV) . . . 161
5.5.1 Study Description . . . 161
5.5.2 Estimates of Correlations between Event Times . . . 161
5.6 Discussion . . . 171

6 Comparison in CVD age Between BC Breast Cancer Cohort and Age–Matched Controls **172**
6.1 Notation and Modeling . . . 172
6.1.1 Notation . . . 172
6.1.2 Model Specification . . . 173
6.1.3 More on Modeling . . . 174
6.2 Pseudolikelihood Based Estimation Procedure . . . 174
6.2.1 Estimating the Association Parameter with the Observed-Data . . . 175
6.3 Analysis of BC-BRCAS Data (V) . . . 176
6.3.1 Study Description . . . 176
6.3.2 Estimates of Conditional Survivor Functions . . . 176
6.4 Discussion . . . 182

7 Final Discussion **183**
7.1 Summary of Contributions . . . 183
7.2 Future Investigations . . . 184

Bibliography **186**

List of Tables

Table 2.1 Descriptive Statistics of BC-BRCAS Data 18
Table 2.2 BC-BRCAS Data Analysis (I). Preliminary Analysis Comparing the Breast Cancer Cohort and Controls 19
Table 2.3 BC-BRCAS Data Analysis (I). Preliminary Analysis with Breast Cancer Cohort 20

Table 3.1 Consistency Study. Estimation of Association Parameter with Simulated Data from Bivariate Clayton Copulas. 40
Table 3.2 Consistency Study. Estimation of Association Parameter with Simulated Data from Bivariate Gumbel Copulas. 41
Table 3.3 Consistency Study. Estimation of Association Parameter with Simulated Data from Bivariate Frank Copulas. 42
Table 3.4 Robustness Study. Estimation of τ with Simulated Data from Bivariate Archimedean Copulas with $\tau = 0.3$. 43
Table 3.5 Robustness Study. Estimation of τ with Simulated Data from Bivariate Archimedean Copulas with $\tau = 0.6$. 44
Table 3.6 Robustness Study. Estimation of τ with Simulated Data from Bivariate Archimedean Copulas with $\tau = 0.8$. 45
Table 3.7 Robustness Study. Estimation of τ with Simulated Data from Bivariate Gaussian Copulas with $\tau = 0.3, 0.6,$ and 0.8. 46
Table 3.8 Consistency Study. Estimation of Association Parameter with Simulated Data from Trivariate Clayton Copulas. 47
Table 3.9 Consistency Study. Estimation of Association Parameter with Simulated Data from Trivariate Gumbel Copulas. 48
Table 3.10 Consistency Study. Estimation of Association Parameter with Simulated Data from Trivariate Frank Copulas. 49
Table 3.11 Robustness Study. Estimation of τ with Simulated Data from Trivariate Archimedean Copulas with $\tau = 0.3$. 50
Table 3.12 Robustness Study. Estimation of τ with Simulated Data from Trivariate Archimedean Copulas with $\tau = 0.6$. 51
Table 3.13 Robustness Study. Estimation of τ with Simulated Data from Trivariate Archimedean Copulas with $\tau = 0.8$. 52

Table 3.14 BC-BRCAS Data Analysis (II). Summary Statistics. 94
Table 3.15 BC-BRCAS Data Analysis (II). Estimates of τ. 95

Table 4.1 Consistency Study. Estimation of Association Parameters with Simulated Data from Nested Archimedean Copulas 108
Table 4.2 Robustness Study. Estimation of Association Parameters with Simulated Data from Nested Archimedean Copulas - Part 1 109
Table 4.3 Robustness Study. Estimation of Association Parameters with Simulated Data from Nested Archimedean Copulas - Part 2 110
Table 4.4 Robustness Study. Estimation of Association Parameters with Simulated Data from Trivariate Gaussian Copula with $\tau = 0.8$ and$\tau_{12} = 0.8$. 111
Table 4.5 Robustness Study. Estimation of Association Parameters with Simulated Data from Trivariate Gaussian Copula with $\tau = 0.6$ and$\tau_{12} = 0.8$. 111
Table 4.6 Robustness Study. Estimation of Association Parameters with Simulated Data from Trivariate Gaussian Copula with $\tau = 0.8$ and$\tau_{12} = 0.6$. 112
Table 4.7 Estimation of $(\theta_1, \theta_2, \theta_\eta, \theta)$ with Simulated Data From (4.7). 112
Table 4.8 BC-BRCAS Data Analysis (III). Estimates of τ and τ_{12} Using Proposed Approach. 128
Table 4.9 BC-BRCAS Data Analysis (III). Estimates of τ and τ_{12} Using Naïve Estimators 129

Table 5.1 Estimates of Coefficients in Cox Model with Simulated Data 144
Table 5.3 Estimates of (τ, τ_{12}) with Simulated Data 144
Table 5.2 Estimates of Coefficients with Simulated Data 145
Table 5.4 BC-BRCAS Data Analysis (IV). Summary Statistics. 163
Table 5.5 BC-BRCAS Data Analysis (IV). Estimates of Coefficients in Cox Models. 164
Table 5.6 BC-BRCAS Data Analysis (IV). Estimates of τ and τ_{12}. 165

Table 6.1 BC-BRCAS Data Anslysis (V). Summary Statistics. 177

List of Figures

Figure 1.1 Different Time Scale 4

Figure 2.1 Diagram of Breast Cancer Cohort and Controls 14
Figure 2.2 Administrative Data Availability Time Window. 14
Figure 2.3 Administrative Data Window Used in Real Data Analysis (I). . . . 15
Figure 2.4 Illustration of Different Scenarios of Observed Event Times 17

Figure 3.1 Consistency Study. Estimates of Marginal Survivor Functions Using Simulated Data under Bivariate Clayton Copula with $\tau = 0.6$. . . . 53
Figure 3.2 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data from Bivariate Clayton Copula with $\tau = 0.6$. 54
Figure 3.3 Consistency Study. Estimates of Marginal Survivor Functions Using Simulated Data under Bivariate Clayton Copula with $\tau = 0.3$. . . . 55
Figure 3.4 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data from Bivariate Clayton Copula with $\tau = 0.3$. 56
Figure 3.5 Consistency Study. Estimates of Marginal Survivor Functions Using Simulated Data under Bivariate Clayton Copula with $\tau = 0.8$. . . . 57
Figure 3.6 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data from Bivariate Clayton Copula with $\tau = 0.8$. 58
Figure 3.7 Consistency Study. Estimates of Marginal Survivor Functions Using Simulated Data under Bivariate Gumbel Copula with $\tau = 0.6$. . . . 59
Figure 3.8 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data from Bivariate Gumbel Copula with $\tau = 0.6$. 60
Figure 3.9 Consistency Study. Estimates of Marginal Survivor Functions Using Simulated Data under Bivariate Gumbel Copula with $\tau = 0.3$. . . . 61
Figure 3.10 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data from Bivariate Gumbel Copula with $\tau = 0.3$. 62
Figure 3.11 Consistency Study. Estimates of Marginal Survivor Functions Using Simulated Data under Bivariate Gumbel Copula with $\tau = 0.8$. . . . 63
Figure 3.12 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data from Bivariate Gumbel Copula with $\tau = 0.8$. 64

Figure 3.13 Consistency Study. Estimates of Marginal Survivor Functions Using Simulated Data under Bivariate Frank Copula with $\tau = 0.6$. 65
Figure 3.14 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data from Bivariate Frank Copula with $\tau = 0.6$. 66
Figure 3.15 Consistency Study. Estimates of Marginal Survivor Functions Using Simulated Data under Bivariate Frank Copula with $\tau = 0.3$. 67
Figure 3.16 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data from Bivariate Frank Copula with $\tau = 0.3$. 68
Figure 3.17 Consistency Study. Estimates of Marginal Survivor Functions Using Simulated Data under Bivariate Frank Copula with $\tau = 0.8$. 69
Figure 3.18 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data from Bivariate Frank Copula with $\tau = 0.8$. 70
Figure 3.19 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data from Bivariate Gaussian Copula with $\tau = 0.6$. . . . 71
Figure 3.20 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data from Bivariate Gaussian Copula with $\tau = 0.3$. . . . 72
Figure 3.21 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data from Bivariate Gaussian Copula with $\tau = 0.8$. . . . 73
Figure 3.22 Consistency Study. Estimates of Marginal Survivor Functions Using Simulated Data under Trivariate Clayton Copula with $\tau = 0.6$. . . 74
Figure 3.23 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data from Trivariate Clayton Copula with $\tau = 0.6$. . . . 75
Figure 3.24 Consistency Study. Estimates of Marginal Survivor Functions Using Simulated Data under Trivariate Clayton Copula with $\tau = 0.3$. . . 76
Figure 3.25 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data from Trivariate Clayton Copula with $\tau = 0.3$. . . . 77
Figure 3.26 Consistency Study. Estimates of Marginal Survivor Functions Using Simulated Data under Trivariate Clayton Copula with $\tau = 0.8$. . . 78
Figure 3.27 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data from Trivariate Clayton Copula with $\tau = 0.8$. . . . 79
Figure 3.28 Consistency Study. Estimates of Marginal Survivor Functions Using Simulated Data under Trivariate Gumbel Copula with $\tau = 0.6$. . . 80
Figure 3.29 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data from Trivariate Gumbel Copula with $\tau = 0.6$. . . . 81
Figure 3.30 Consistency Study. Estimates of Marginal Survivor Functions Using Simulated Data under Trivariate Gumbel Copula with $\tau = 0.3$. . . 82
Figure 3.31 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data from Trivariate Gumbel Copula with $\tau = 0.3$. . . . 83

Figure 3.32 Consistency Study. Estimates of Marginal Survivor Functions Using Simulated Data under Trivariate Gumbel Copula with $\tau = 0.8$. . . 84
Figure 3.33 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data from Trivariate Gumbel Copula with $\tau = 0.8$. . . . 85
Figure 3.34 Consistency Study. Estimates of Marginal Survivor Functions Using Simulated Data under Trivariate Frank Copula with $\tau = 0.6$. 86
Figure 3.35 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data from Trivariate Frank Copula with $\tau = 0.6$. 87
Figure 3.36 Consistency Study. Estimates of Marginal Survivor Functions Using Simulated Data under Trivariate Frank Copula with $\tau = 0.3$. 88
Figure 3.37 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data from Trivariate Frank Copula with $\tau = 0.3$. 89
Figure 3.38 Consistency Study. Estimates of Marginal Survivor Functions Using Simulated Data under Trivariate Frank Copula with $\tau = 0.8$. 90
Figure 3.39 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data from Trivariate Frank Copula with $\tau = 0.8$. 91
Figure 3.40 BC-BRCAS Data Analysis (II). Estimates of Marginal Survivor Functions. 96

Figure 4.1 Consistency Study. Estimates of Marginal Survivor Functions Using Simulated Data under Nested Clayton Copula with $\tau = 0.4, \tau_{12} = 0.5$. 113
Figure 4.2 Consistency Study. Estimates of Marginal Survivor Functions Using Simulated Data under Nested Clayton Copula with $\tau = 0.3, \tau_{12} = 0.8$. 114
Figure 4.3 Consistency Study. Estimates of Marginal Survivor Functions Using Simulated Data under Nested Frank Copula with $\tau = 0.4, \tau_{12} = 0.5$. 115
Figure 4.4 Consistency Study. Estimates of Marginal Survivor Functions Using Simulated Data under Nested Clayton Copula with $\tau = 0.3, \tau_{12} = 0.8$. 116
Figure 4.5 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data under Nested Clayton Copula with $\tau = 0.4, \tau_{12} = 0.5$ - Part 1. 117
Figure 4.6 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data under Nested Clayton Copula with $\tau = 0.4, \tau_{12} = 0.5$ - Part 2. 118
Figure 4.7 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data under Nested Clayton Copula with $\tau = 0.3, \tau_{12} = 0.8$ - Part 1. 119

Figure 4.8 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data under Nested Clayton Copula with $\tau = 0.3, \tau_{12} = 0.8$ - Part 2. 120
Figure 4.9 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data under Nested Frank Copula with $\tau = 0.4, \tau_{12} = 0.5$ - Part 1. 121
Figure 4.10 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data under Nested Frank Copula with $\tau = 0.4, \tau_{12} = 0.5$ - Part 2. 122
Figure 4.11 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data under Nested Frank Copula with $\tau = 0.3, \tau_{12} = 0.8$ - Part 1. 123
Figure 4.12 Robustness Study. Estimates of Marginal Survivor Functions Using Simulated Data under Nested Frank Copula with $\tau = 0.3, \tau_{12} = 0.8$ - Part 2. 124
Figure 4.13 Marginal Function Estimates with Simulated Data Generated from Model (4.7). Upper Row: S_1; Bottom Row: S_2. Skyblue dashed: naïve. Black solid: true. Orange dotted: Proposed 125
Figure 4.14 BC-BRCAS Data Analysis (III). Estimates of Marginal Survivor Function $S_1(\cdot)$. 130
Figure 4.15 BC-BRCAS Data Analysis (III). Estimates of Marginal Survivor Function $S_2(\cdot)$. 131

Figure 5.1 Estimates of $\tau_{12}(Z)$ for Scenario (II) with Simulated Data 146
Figure 5.2 Estimates of $\tau(Z)$ for Scenario (III) with Simulated Data 147
Figure 5.3 Estimates of $\tau_{12}(Z)$ and $\tau(Z)$ for Scenario (IV) with Simulated Data 148
Figure 5.4 Estimates of Conditional Marginal Survivor Functions for $Z = 0.3$, under Scenario (I). 149
Figure 5.5 Estimates of Conditional Marginal Survivor Functions for $Z = 0.3$, under Scenario (II). 150
Figure 5.6 Estimates of Conditional Marginal Survivor Functions for $Z = 0.3$, under Scenario (III). 151
Figure 5.7 Estimates of Conditional Marginal Survivor Functions for $Z = 0.3$, under Scenario (IV). 152
Figure 5.8 Estimates of Conditional Marginal Survivor Functions for $Z = 0.5$, under Scenario (I). 153
Figure 5.9 Estimates of Conditional Marginal Survivor Functions for $Z = 0.5$, under Scenario (II). 154

Figure 5.10 Estimates of Conditional Marginal Survivor Functions for $Z = 0.5$, under Scenario (III). 155
Figure 5.11 Estimates of Conditional Marginal Survivor Functions for $Z = 0.5$, under Scenario (IV). 156
Figure 5.12 Estimates of Conditional Marginal Survivor Functions for $Z = 0.7$, under Scenario (I). 157
Figure 5.13 Estimates of Conditional Marginal Survivor Functions for $Z = 0.7$, under Scenario (II). 158
Figure 5.14 Estimates of Conditional Marginal Survivor Functions for $Z = 0.7$, under Scenario (III). 159
Figure 5.15 Estimates of Conditional Marginal Survivor Functions for $Z = 0.7$, under Scenario (IV). 160
Figure 5.16 BC-BRCAS Data Analysis (IV). Estimates of Marginal Survivor Function $S_1(\cdot|Z)$. 166
Figure 5.17 BC-BRCAS Data Analysis (IV). Estimates of Marginal Survivor Function $S_1(\cdot|Z)$ when One Covariate is Continuous. 167
Figure 5.18 BC-BRCAS Data Analysis (IV). Estimates of Marginal Survivor Function $S_2(\cdot|Z)$. 168
Figure 5.19 BC-BRCAS Data Analysis (IV). Estimates of Marginal Survivor Function $S_2(\cdot|Z)$ when One Covariate is Continuous. 169
Figure 5.20 BC-BRCAS Data Analysis (IV). Estimates of $\tau_{12}(Z)$ under Scenario (IV). 170
Figure 5.21 BC-BRCAS Data Analysis (IV). Estimates of $\tau(Z)$ for Scenario (IV). 171

Figure 6.1 BC-BRCAS Data Analysis (V). Estimates of Marginal Survivor Function for Cohort and Controls. Era Discrete. 178
Figure 6.2 BC-BRCAS Data Analysis (V). Difference in the Estimates of Marginal Survivor Function Between Cohort and Controls. Era Discrete. . . 179
Figure 6.3 BC-BRCAS Data Analysis (V). Estimates of Marginal Survivor Function for Cohort and Controls. Era Continuous. 180
Figure 6.4 BC-BRCAS Data Analysis (V). Difference in the Estimates of Marginal Survivor Function Between Cohort and Controls. Era Continuous. . 181

Chapter 1

Introduction

The ongoing rise in the annual number of new cancer diagnosis due to a growing and aging population, combined with an improving survival rate for most types of cancer, has meant that a substantial number of people are living with and beyond their cancer diagnosis. These cancer survivors may experience late effects and problems related to the disease and its treatment in their lifetime. For example, breast cancer has constantly been the most common cancer diagnosis in Canadian women over the age of twenty. Although fewer Canadian women are dying from breast cancer today than in the past, breast cancer survivors are at risk from a broad set of late effects. Thus, survivorship research has gained more interest in recent years amongst clinicians and oncologists who want to know the long-term effects and risks to provide quality and targeted treatment and care to patients. It is also of interest to policymakers as well as the survivors themselves. For example, there is evidence that breast cancer survivors, or those who experieced relapse or second cancer (RSC), are at a higher risk of hospitalization related to cardiovascular disease (CVD) compared to their peers (see, for example, Bardia et al. 2012, Hamilton et al. 2015, Park et al. 2017). This is a very important health issue for breast cancer survivors (e.g., Cuzick et al. 1994, Clarke et al. 2005). A recent publication (Mehta et al. 2018) has provided a statement from the American Heart Association (AHA) that breast cancer treatments may increase the risk of heart disease. This is the first scientific statement from the AHA on CVD and breast cancer.

1.1 Background and Motivation

It was estimated that in 2017 alone, 206,200 new cases of cancer were diagnosed in Canada, and in 2009, about 810,045 Canadians diagnosed with cancer in the previous ten years were alive. This represented about 2.4 percent of the Canadian population or one out of every forty-two Canadians. Breast cancer accounts for approximately 26% on new cases of cancer and 13% of all cancer deaths in Canadian women. 1 in 8 women are expected to

develop breast cancer during her lifetime and 1 in 31 will die of it. In 2009, an estimated 157,360 women were living with, or surviving from, breast cancer in Canada, This means that there is a large population of breast cancer survivors (`https://www.canada.ca/en/public-health/services/chronic-diseases/cancer/breast-cancer.html`).

1.1.1 General Background: Health Care System in Canada

Canada's publicly funded health care system provides universal coverage for medically necessary health care services provided on the basis of need, rather than the ability to pay. It is available to all eligible residents. The provinces and territories administer and deliver most of Canada's health care services. Each provincial and territorial health insurance plan covers *medically necessary* hospital and doctors' services that are provided on a prepaid basis, without direct charges at the point of service. The provincial and territorial governments fund these services with assistance from federal tax transfers. That is, the provincial administrative databases on hospitalization and physician visits record all of the medically necessary health services provided to the population. Therefore, individualized longitudinal electronic record of health care services are available through administrative databases in each province, and it reflects the care received by each individual.

In British Columbia (BC), a province in Canada, public health insurance is called the Medical Services Plan – or MSP. It covers the cost of medically-necessary insured doctor services. The MSP Registry enrols all eligible BC residents and collects basic demographic information for each individual. All Canadian hospitals (except those in Quebec), including BC, submit their separations records directly to the Canadian Institute of Health information (CIHI: `https://www.cihi.ca`) for inclusion in the Discharge Abstract Database (DAD). The database contains demographic, administrative, and clinical data (e.g. reason for hospitalization) for hospital discharges (inpatient acute, chronic, rehabilitation) and day surgeries. A provincial data set, including various CIHI value-added elements (such as case mix groups, and resource intensity weights) is released on a monthly basis to the respective Ministries of Health.

1.1.2 BC Breast Cancer Survivorship (BC-BRCAS) Program

In BC, as in all provinces, cancer is a reportable disease (Section 9, Health Act, `http://www.bclaws.ca/civix/document/id/complete/statreg/96179_01`). BC Cancer Registry (BCCR), managed by BC Cancer (`www.bccancer.bc.ca`), a provincial cancer care, control and research organization, collects information on all cancers diagnosed for BC residents. The sources of registrations are haematology and pathology reports, death certificates, hospital reports, and admission records at cancer treatment centres. The BC Cancer Registry is estimated to cover at least 95 percent of all cancer cases in BC.

The breast cancer survivorship program in the BC Cancer Research Center is a research program established to study long term health and health care amongst a cohort of breast cancer survivors in BC. The research program has implemented a series of epidemiological, clinical, and health-service studies relating to breast cancer survivorship issues, using existing population-based registries, administrative databases, and record-linkage methodology.

The study subjects included in this dissertation are women diagnosed with breast cancer between 1 January, 1989 and 31 December, 2011 in BC, eighteen years and older and residents of BC, identified from the BCCR. Their relevant demographic, disease, and treatment information as well as death- and RSC-related data up to 31 December, 2014 were extracted from the registry and clinical databases; their records of hospitalizations from 1 January, 1986 to 31 December, 2013 were extracted. In addition, a gender and birth-year matched comparison group (controls) from the MSP registry was identified. Chapter 2 provides more description on the datasets.

The strengths of the breast cancer survivorship research program are that 1) it uses the linked administrative databases and avoids the potential informative loss-to-follow up issues due to non-response, as is often encountered in survey-based studies; 2) the databases collect data over time and thus the records are longitudinal for each individual; therefore, one could examine the time effect on outcomes; 3) the matched comparison group provides the ability to address relevant research questions.

1.1.3 Motivation of the Thesis Research

Many research questions are of interest in monitoring the late effects of breast cancer survivors. It is believed that breast cancer survivors suffer from CVD sooner than their peers in the general population, due to the location of the cancer. Specifically, this can be examined from three aspects, listed as the following three hypothesis, which we will address in this dissertation using data from the BC-BRCAS program.

Hypothesis 1: Breast cancer survivors suffer from CVD at a younger age than their peers in the general population.

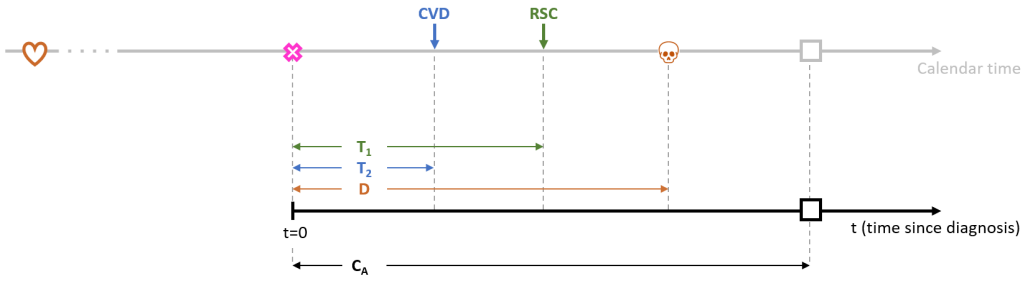
Hypothesis 2: There is a positive association between RSC and CVD.

Hypothesis 3: The treatment on breast cancer is likely associated with higher risk of CVD and RSC.

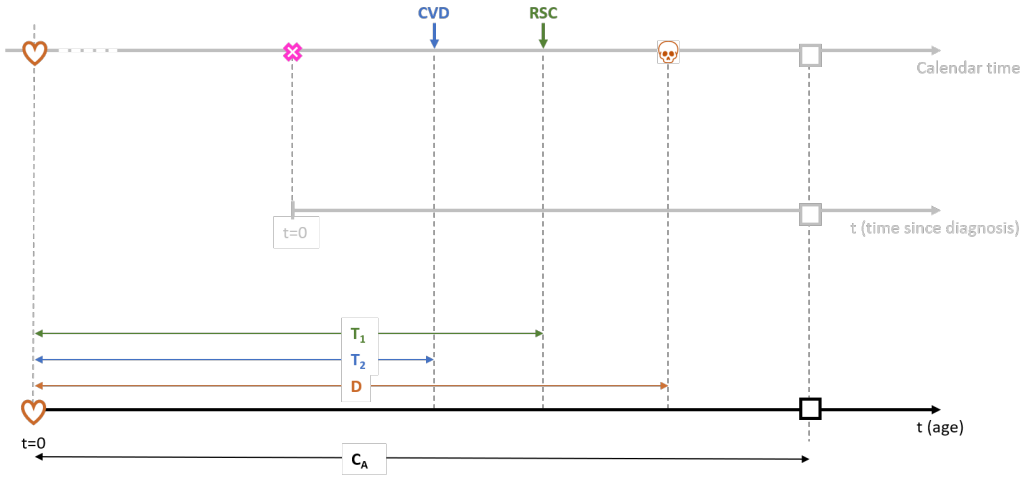
Conventionally, cross-sectional analysis could be used to address the questions above. In Chapter 2, a preliminary cross-sectional analysis is conducted. However, cross-sectional analysis does not account for the effect of time and only uses limited information from the data collected within a short time window. Thus, survival analysis will provide more insight into the process of events.

The data are available as dates in calendar time. In this thesis, two time scales were considered: time since diagnosis time scale, and the age time scale. As shown in figure 1.1, if we denote T_1 as the time at RSC, T_2 as the time at CVD, they are defined as the time gap between diagnosis or birth and the occurence of RSC and CVD, respectively, depending on the time scale used.

(a) (i) calendar time scale



(b) (ii) time since diagnosis scale



(c) (iii) age scale

Figure 1.1: Calendar Time, Time Since Diagnosis, and Age Time Scale.

Further, we denote Z as the covariates, and formulate the statistical problems corresponding to the three hypothesis above as the three following goals:

Goal 1: to estimate T_2's distribution

Goal 2: to estimate (T_1, T_2)'s joint distribution

Goal 3: to estimate the conditional distribution of $T_2|Z$, or $(T_1, T_2)|Z$

Conventional approaches such as the Kaplan–Meier estimator and Cox proportional hazards model in survival analyses have addressed the question of estimating the marginal and conditional distribution of one single event time. To estimate the distribution for bivariate event-time, two commonly used approaches are frailty modeling and copula modeling.

Most of the conventional approaches are based on the assumption that the event-times are subject to noninformative censoring. In many real-life studies, the observation may be terminated earlier than scheduled by the subject's death, and this leads to potential informative censoring, when the goal is to make inference on the whole population. In this study, the observations on the times to CVD and RSC are heavily censored because of either the study follow-up time limit or death. The time to death, denoted as D, is likely correlated to the two event times of interest. It is noted here that inference is to be made for the whole population, rather than for only those who had experienced T_1 and T_2. On one hand, subjects who had experienced T_1, T_2 likely survived longer than those who died before experiencing T_1 or T_2. On the other hand, although not the focus of this dissertation, making inference only on those who had T_1, and T_2 falls within the 'disease-free survival' framework (e.g. Andersen & Keiding 2012). This has also been addressed in this dissertation, through a middle product of the proposed approaches in Chapters 3- 6.

In summary, conventional event time analysis methods, such as the Kaplan–Meier estimator for the survivor function of an event time, are not directly applicable. Furthermore, achieving *goal 2* involves dealing with observations of bivariate event time subject to informative censoring, which adds complexity to the statistical problem. This statistical challenge partially motivated the research presented in this dissertation.

1.2 Literature Review

Given the statistical formulation in the previous section, this dissertation aims to estimate the joint survivor function of multiple event times in presence of informative censoring. This section reviews the relevant topics.

1.2.1 Multivariate Event Time

In literature, many approaches have been developed for modeling bivariate (or multivariate) event time. Existing common frameworks for modeling multivariate failure time data include frailty models and copula models.

Frailty models

The introduction of a common random effect (frailty) is a natural way of modeling the dependence of event times. A frailty model is a multiplicative hazard model consisting of three components: a frailty (random effect), a baseline hazard function (parametric or

nonparametric), and a term modeling the influence of observed covariates (fixed effects). Review of frailty models for multivariate failure time data can be seen in, for example, Liang et al. (1995) and Hougaard (2012).

Definition 1. *Suppose that T is an event time, and that there exists a positive random variable W such that the conditional survivor function of T, given $W = w$, is*

$$Pr(T \geq t|W = w) = S_0(t)^w \tag{1.1}$$

where $S_0(t)$ is the baseline survivor function. Then, W is called latent random effect or frailty, and (1.1) is a frailty model (Hougaard 1984).

The frailty model has the same structure as that of the Cox model, except that the value w of W is not observed, but is a random variable with density function $f_\theta(\cdot)$ with parameter θ. The unconditional survivor function of T is $S(t) = Pr(T \geq t) = \mathcal{L}_\theta\{-\log S_0(t)\}$ where $\mathcal{L}_\theta(u) = \mathrm{E}e^{-uw}$ is the Laplace transform of W, with $\mathrm{E}(\cdot)$ being the expected value. It has been shown in Oakes (1989) that for a bivariate frailty model, which specifies the bivariate event time T_1 and T_2 to be independent conditional on a frailty W, the joint survivor function of (T_1, T_2) is

$$S(t_1, t_2) = \Pr(T_1 \geq t_2, T_2 \geq t_2; \theta) = \mathcal{L}_\theta[\mathcal{L}_\theta(S_1(t_1)) + \mathcal{L}_\theta(S_2(t_2))] \tag{1.2}$$

where $S_j(t_j) = \Pr(T_j \geq t_j)$ is the survivor function of T_j, $j = 1, 2$. Two of the most commonly used frailty models in applications are the gamma frailty model and the positive stable frailty model, where frailty W is assumed to follow gamma distribution and positive stable distribution, respectively.

Copula models

Copula modeling has become an increasingly popular tool in multivariate analysis since the fundamental work of Clayton (1978), which proposes a family of copula models for the analysis of bivariate data. The model specifies the joint distribution of the multiple event times by linking each marginal distribution via a copula function; see, e.g., Joe (1997) and Nelsen (2006) for more discussion and examples. Recent papers on multivariate event times via copulas with various data structures include Zhang et al. (2010), Diao & Cook (2014), and Zhong & Cook (2016). Here we briefly introduce the bivariate copula (Joe 1997, Jaworski et al. 2010).

The concept of copula was introduced by Sklar (1959) to describe in a convenient way the class of distribution functions with given marginals.

Definition 2 (**copula**). *For every $d \geq 2$, a d-dimensional copula (shortly, d-copula) is a d-variate CDF on $[0, 1]^d$ whose univariate marginals are uniformly distributed on $[0, 1]$.*

Thus, each d-copula may be associated with a random variable $U = (U_1, U_2, \ldots, U_d)$ such that $U_j \sim \text{Unif}[0,1]$ for every $j \in \{1, 2, ..., d\}$ and $U \sim \mathcal{C}$. Conversely, any random variable whose components are uniformly distributed on $[0,1]$ is distributed according to some copula. The class of all d-copulas will be denoted by $\mathcal{C}_{[d]}$. Since copulas are multivariate CDFs, they can be characterized in the following equivalent way.

Theorem 1. *A function $\mathcal{C} : [0,1]^d \to [0,1]$ is a copula if, and only if, the following properties hold:*

(C1) for every $j \in \{1, 2, ..., d\}$, $\mathcal{C}(\mathbf{u}) = u_j$ when all the components of $\mathbf{u}$ are equal to 1 with the exception of the j-th one that is equal to u_j;

(C2) $\mathcal{C}$ is isotonic, i.e. $\mathcal{C}(\mathbf{u}) \leq \mathcal{C}(\mathbf{v})$ for all $\mathbf{u}, \mathbf{v} \in [0,1]^d$, $\mathbf{u} \leq \mathbf{v}$;

(C3) $\mathcal{C}$is d-increasing.

Sklar's theorem represents the building block of the modern theory of multivariate CDFs. It can be formulated as follows.

Theorem 2 (Sklar's theorem). *Suppose $X_1, \ldots, X_d$ are random variables with continuous CDFs $F_1, \ldots, F_d$ and joint CDF F. Then there exists a unique copula $\mathcal{C}_{[d]}$, a CDF on $[0,1]^d$ with uniform marginals, such that for all $\mathbf{x} = (x_1, \ldots, x_d)^\top \in R^d$:*

$$F(x_1, \ldots, x_d) = \mathcal{C}_{[d]}(F_1(x_1), \ldots, F_d(x_d)) \tag{1.3}$$

Conversely, given any CDF $F_1, \ldots, F_d$ and copula $\mathcal{C}_{[d]}$, F defined through (1.3) is a d–variate CDF with marginal CDFs $F_1, \ldots, F_d$.
Alternatively, (1.3) can also be rewritten as for $u = (u_1, \ldots, u_d)^\top \in [0,1]^d$,

$$\mathcal{C}_{[d]}(u_1, \ldots u_d) = F(F_1^{-1}(u_1), \ldots, F_d^{-1}(u_d)) \tag{1.4}$$

For survivor functions, we will restate Sklar's theorem as follows. Let $S(t_1, \ldots, t_d)$ be a joint survivor function with marginals $S_1(t_1), \ldots, S_d(t_d)$. Then, there exists a bivariate copula $\bar{\mathcal{C}}$ such that for all $(t_1, \ldots, t_d) \in \mathbb{R}^d$ $S(t_1, \ldots, t_d) = \overline{\mathcal{C}}_{[d]}(S_1(t_1), \ldots, S_d(t_d))$ The survival copula $\overline{\mathcal{C}}_{[d]}$ links the d-variate survivor function to its univariate marginal survivor function analogous to the way in (1.3). There exists a link between survivor copula $\overline{\mathcal{C}}$ and copula $\mathcal{C}$. For example, when $d = 2$, it is given by $\overline{\mathcal{C}}_{[2]}(u_1, u_2) = u_1 + u_2 - 1 + \mathcal{C}_{[2]}(1 - u_1, 1 - u_2)$. Without loss of generality, for the remainder of the dissertation, we use $\mathcal{C}$ to represent the survivor copula.

Example: Archimedean Copula Family

As a preparation for the modeling, inference procedures and numerical studies in the remainder of the dissertation, we now briefly review the Archimedean copulas, following Nelsen (2006).

By Sklar's theorem, every multivariate survivor function can be presented as a copula model, with all the marginal survivor functions linked by a copula function. Archimedean copula models are commonly used in survival analysis. A K-dimensional Archimedean copula has the form

$$\mathcal{A}_{[K]}(w_1, \ldots, w_K; \phi) = \psi^{-1}(\psi(w_1; \phi) + \ldots + \psi(w_K; \phi); \phi), \tag{1.5}$$

where the generator $\psi(\cdot)$ is continuous, strictly decreasing, convex, and with $\psi(1; \phi) = 0$ for $\phi \in \Phi$, a parameter space. Let w^* be $\psi^{-1}(\psi(w_1; \phi) + \ldots + \psi(w_{K-1}; \phi); \phi)$, which is in fact $\mathcal{A}_{[K-1]}(w_1, \ldots, w_{K-1}; \phi)$. We see from (1.5) that

$$\mathcal{A}_{[K]}(w_1, \ldots, w_K; \phi) = \psi^{-1}(\psi(w^*; \phi) + \psi(w_K; \phi); \phi) = \mathcal{A}_{[2]}(w^*, w_K; \phi). \tag{1.6}$$

Moreover,

$$\mathcal{A}_{[K]}(w_1, \ldots, w_K; \phi) = \mathcal{A}_{[K-1]}(w_1, \ldots, w_{k-1}, w_{k+1}, \ldots, w_{K-1}, w_k^*; \phi) \tag{1.7}$$

with $w_k^* = \psi^{-1}(\psi(w_k; \phi) + \psi(w_K; \phi); \phi) = \mathcal{A}_{[2]}(w_k, w_K; \phi)$ for $k = 1, \ldots, K-1$. This feature of Archimedean copulas is convenient for estimation with the models.

Corresponding to the association parameter in a bivariate Archimedean copula, the Kendall rank correlation coefficient (also called Kendall's τ), a widely used measure of correlation between variables, is $\tau = 4 \int_0^1 \int_0^1 \mathcal{A}_{[2]}(w_1, w_2; \phi) \mathcal{A}_{[2]}(dw_1, dw_2; \phi) - 1$. We list below the generators of four commonly used Archimedean copulas and their Kendall's τ, which are employed in the empirical studies reported in numerical studies in later chapters.

1. *Clayton Copula* (Clayton 1978): $\psi(w; \phi) = (w^{-\phi} - 1)/\phi$ for $\phi \in (-1, \infty) \backslash \{0\}$ and Kendall's $\tau = \phi/(\phi + 2)$.

2. *Frank Copula*: $\psi(w; \phi) = -\log\{[\exp(-\phi w) - 1]/[\exp(-\phi) - 1]\}$ for $\phi \in (-\infty, \infty) \backslash \{0\}$ and Kendall's $\tau = 1 + 4[B(\phi) - 1]/\phi$ with $B(\phi) = \phi^{-1} \int_0^\phi t/[\exp(t) - 1] dt$.

3. *Gumbel Copula*: $\psi(w; \phi) = (-\log(w))^\phi$ for $\phi \in [1, \infty)$ and Kendall's $\tau = 1 - \phi^{-1}$.

4. *Joe Copula*: $\psi(w; \phi) = -\log[(1 - (1 - w)^\phi)]$ for $\phi \in [1, \infty)$ and Kendall's $\tau = 1 - 4 \sum_{k=1}^\infty 1/\{k(\phi k + 2)[\phi(k - 1) + 2)]\}$.

Correspondence between copula and frailty modeling

Copulas and frailty models are both widely used to model clustered or multivariate data. Equivalence between Archimedean copula models and shared frailty models, e.g. between the Clayton-Oakes copula model and the shared gamma frailty model, has often been claimed in literature (Oakes 1989, Goethals et al. 2008).

To view the correspondence between the Archimedean copula and (1.2), we take the bivariate Archimedean copula model as an example. The joint distribution of T_1 and T_2 can be expressed as

$$\begin{aligned} & S(t_1, t_2) = \Pr(T_1 \geq t_1, T_2 \geq t_2) = \mathcal{A}_{[2]}(S_1(t_1), S_2(t_2); \phi) \\ = \quad & \psi^{-1}[\psi(S_1(t_1); \phi), S_2(t_2; \phi); \phi] \end{aligned} \tag{1.8}$$

where the generating function $\psi(\cdot)$ only requires $\psi(0) = 1$, $\psi'(u) < 0$ and $\psi''(u) > 0$ for $u \in [0, 1]$. Therefore (1.2) is a special case of (1.8).

Advantages of copula modeling

To summarize, the advantages of copula modeling include the following:

1 By Sklar's theorem, any multivariate distribution can be decomposed into its marginal distributions and a copula function which completely describes the dependence structure.

2 Further, for continuous multivariate distributions, the univariate marginals and the multivariate or dependence structure can be separated, and the multivariate structure can be represented by a copula. This adds convenience in the estimation procedure and make it feasible to apply a two-stage estimation procedure.

3 Thirdly, commonly used dependence measures such as Kendall's τ can be obtained directly from the parameters in copula functions, making the interpretation useful in real data analysis. It also provides a universal measure to compare different copulas and verify robustness.

4 Compared to commonly used frailty models, which can be viewed as a subset of copula models, copula models cover a wider range of model specification.

5 Although beyond the scope of application in this dissertation, asymptotic tail dependence can be used to measure the dependence between extreme values through copula modeling, which is commonly used in fields such as economics and finance.

Likelihood-based Inference Procedures

Likelihood inference is a foundational estimation approach in statistics; the distribution function of the response variable must be fully specified. Under certain conditions of the model, the maximum likelihood estimator (MLE) possesses many properties that make it an appealing choice of estimator. Therefore, the likelihood function has proved to be a powerful tool for inference. On the other hand, obtaining the MLE from the full likelihood

function, can be computationally intensive, especially when there is an unknown function involved as a nuisance parameter.

Likelihood inference has been extended in many ways. For example, various pseudolikelihood functions have been proposed for more complicated models. With some efficiency loss, the pseudo-MLE can be computationally much easier to obtain than its MLE counterpart. Given the structure of copula models reviewed in the previous section, a two-stage estimation procedure is feasible. The Oakes (1994) paper proposes a semiparametric two-stage estimation procedure: in the first stage the unknown marginal distributions are estimated by rescaled empirical distribution functions; in the second stage, the copula dependence parameter is estimated by maximizing the estimated log-likelihood function where the marginal distributions are fixed to the estimates from the first step. Shih & Louis (1995) propose a two-step estimator with right-censored data under copula models, and Lawless & Yilmaz (2011) compare the two-stage pseudolikelihood estimation and the corresponding maximum likelihood estimation in terms of efficiency and robustness.

1.2.2 Informative Censoring

As mentioned, many practical studies of life history processes follow their subjects over a certain time period for particular events. However, the observation may be terminated earlier than scheduled by the subject's death (e.g., Li & Lagakos 1997), and this leads to potentially informative censoring. When a single event is of interest, approaches have been proposed to estimate the distribution of the event time subject to informative censoring in the framework of competing risks (Zheng & Klein 1995) or semicompeting risks (Fine et al. 2001, Wang 2003, Jiang et al. 2005, Cheng & Fine 2012). Li et al. (2007) employed Archimedean copula models for informative censoring in a mixture cure model with a single event time of interest. Bandeen-Roche & Liang (2002) and Ning & Bandeen-Roche (2014) considered a modified conditional hazard ratio measure of association and the associated estimation with bivariate competing-risks data. Cheng et al. (2007) presented nonparametric estimation for the bivariate cause-specific hazard function and the bivariate cumulative incidence function.

This dissertation, however, addresses the problem of modeling multiple event times with observations subject to informative censoring by extending the commonly used copula based dependence modeling approach. Essentially, we treat the informative censoring time as a separate event time, and model the single event time or multiple event times together with the censoring time through a multivariate modeling approach. Furthermore, this dissertation develops likelihood based inference procedures.

1.3 Notation and Framework

We study the joint distribution of multivariate event time in two major ways: unconditional distribution without covariates and conditional distribution in a regression setting. In this section we present the notation that is used throughout the remainder of the dissertation. Specific notation is introduced at the beginning of each chapter.

Let $S(\cdot)$, $f(\cdot)$, $F(\cdot)$ be the survivor function, density function and CDF for random variable(s). Let T be event time, which could be time at RSC, or time at CVD. Let D be the informative censoring time, and in this dissertation we focus on a terminating event such as death. Let $T^* = T \wedge D$ be the so-called 'disease-free' survival time. The introduction of this random variable adds convenience in estimating the marginal survivor function of T, which is of interest in this dissertation. Let C_A be the right-censoring time which is assumed to be noninformative.

The time scale we use in this dissertation includes: a) a time gap between diagnosis of breast cancer and the event, or more succinctly time since diagnosis, in Chapters 3 to 5; it can be reasonably assumed that given a stage at cancer diagnosis, C_A (time since diagnosis to the end of data collection) is independent of T and D; b) age at event, in Chapter 6. Using the age time scale, it can be reasonably assumed that C_A (age at the end of data collection) is independent of T and D.

Let $\boldsymbol{Z}$ be a vector of covariates. $\boldsymbol{Z}$ may include both demographic and clinical factors for the survivor cohort; and it includes only demographic variables for the comparison sample from the general population.

1.4 Objectives and Thesis Outline

1.4.1 Objectives

Our main objective is to estimate the joint survivor function of multiple event times with the right-censored observations subject to informative censoring. From the joint survivor function, we will be able to verify the research *hypothesis 2* directly. The marginal survivor function for each event time can be obtained as a 'by-product' of our approach, which will verify *hypothesis 1* and *hypothesis 3*.

As stated, this dissertation was motivated by the breast cancer survivorship program data, and we will illustrate the proposed approach using this data throughout. The statistical approach, however, can be applied broadly and is not limited to this specific motivating data.

1.4.2 Thesis Outline

The rest of the dissertation is organized as follows. Chapter 2 provides a general description of the data and presents a preliminary analysis using cross-sectional data. Chapter 3 in-

troduces a modeling approach through the multivariate Archimedean copula and proposes the pseudo-likelihood inference approach with an easy-to-implement algorithm. Chapter 4 proposes a more flexible modeling approach which allows the assumption of dependence structure to be different amongst event times. Chapter 5 extends the approach to regression setting, Chapter 6 applies the approach to handle informative censoring with one single event time, and Chapter 7 provides a summary and discussion of future work. All analyses in this thesis are conducted using R (R Core Team 2013).

Chapter 2

Data Description and Preliminary Analysis

2.1 Introduction

This chapter describes the breast cancer study data which motivated the thesis work and summarizes the CVD-related hospitalization in cross-sectional counts, for both the breast cancer survivor cohort, denoted by $\mathcal{P}_0$, and the comparison group, denoted by $\mathcal{Q}_0$. As what people usually do when dealing with cross-sectional data in epidemiological studies, we conduct a preliminary analysis using Poisson rate regression, to compare CVD events between the breast cancer survivor cohort and the comparison group (to verify *hypothesis 1*), to examine if RSC has any effect on the counts of CVD (to verify *hypothesis 2*), and to quantify the effects of sociodemographic and clinical factors on the outcome (to verify *hypothesis 3*).

2.2 Description of BC-BRCAS Data

This section describes the study subjects, the outcomes of interest and the covariates. We summarize the descriptive statistics in tables and illustrate the data collection mechanism with the information available in timeline figures.

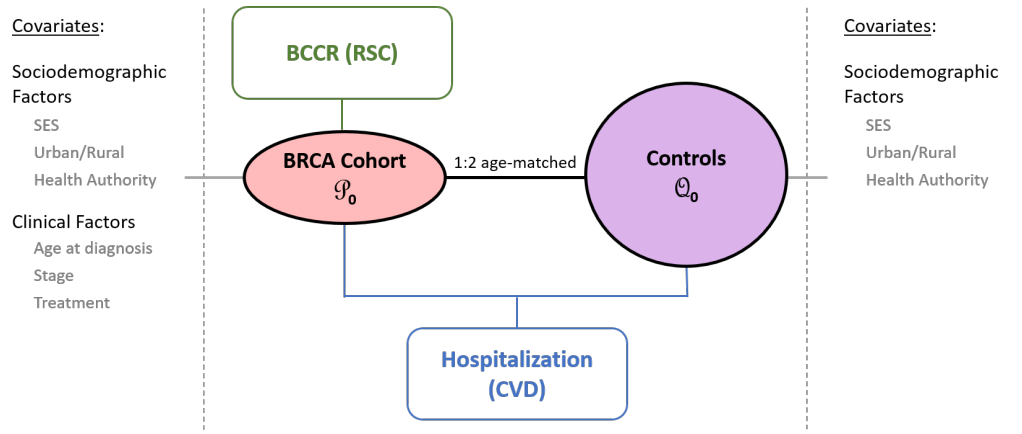
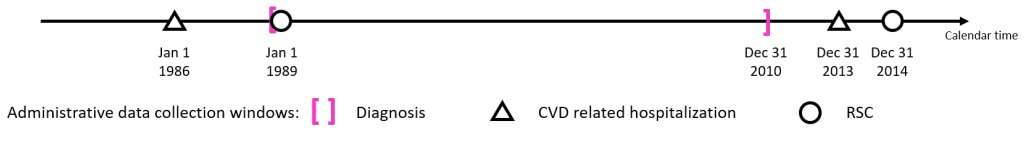


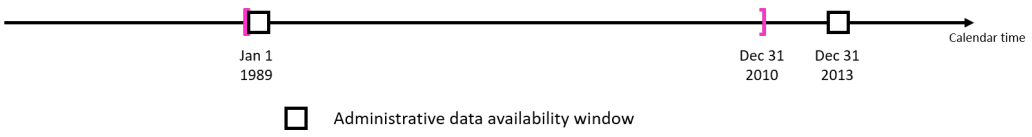
Figure 2.1: Diagram of the Survivor Cohort and the Control Group

As displayed in figure 2.1, the survivor cohort (cases), denoted as $\mathcal{P}_0$, consists of all women diagnosed with breast cancer between 1 January, 1989 and 31 December, 2011 in BC, Canada, who are eighteen years and older and are residents of BC identified from the BC Cancer Registry. Their relevant demographic information, and death- and RSC-related data to 31 December, 2014 were extracted from the registry and clinical databases. The time window of the availability on these administrative databases are shown in figure 2.2(a). There is a one year gap in the end of data collection for RSC and CVD, but there are very few cases. For simplicity, we choose December 31, 2013 as the end of the data extraction, as shown in figure 2.2(b).

A birth-year matched female comparison group (controls) , denoted as $\mathcal{Q}_0$, without diagnosis of breast cancer was selected from the MSP registry. For both cases and controls, their records of CVD-related hospitalizations from 1 January, 1986 to 31 December, 2013 were extracted from the Canadian Institute of Health Information (CIHI) hospital separations database of BC (British Columbia Ministry of Health 2011). Table 2.1 provides the descriptive statistics for $\mathcal{P}_0$ and $\mathcal{Q}_0$ in this study.



(a) Data Availability Window



(b) Data Availability Window: Simplified

Figure 2.2: Breast Cancer Survivor Cohort Data Availability Time Windows

The preliminary cross-sectional analysis presented in this chapter includes the breast cancer survivors who were still alive on 1 January, 2011 (beginning of the cross-sectional time window), and uses CVD-related hospitalizations between 1 January, 2011 and 31 December, 2013 (end of the cross-sectional time window). The time window of the cross-sectional data is illustrated in figure 2.3. We denote the cross-sectional cohorts in this preliminary analysis as $\mathcal{P}_{\text{prelim}}$, and $\mathcal{Q}_{\text{prelim}}$ for cases and controls, respectively. To avoid collinearity in the regression analysis, we chose the following six covariates: *SES* (high versus low), *relapse or second cancer (RSC)* (yes versus no relapse or second cancer at the beginning of the time window), *birth era* (era I: 1900–1927, era II: 1928–1945, era III: 1946–1989, following the definition of generation cohorts that are widely used in North America), *stage of breast cancer* (I, II, III, IV and unknown), and *treatment* (surgery only, chemo only, radiation only, surgery and chemo, surgery and radiation, chemo and radiation, surgery and chemo and radiation).

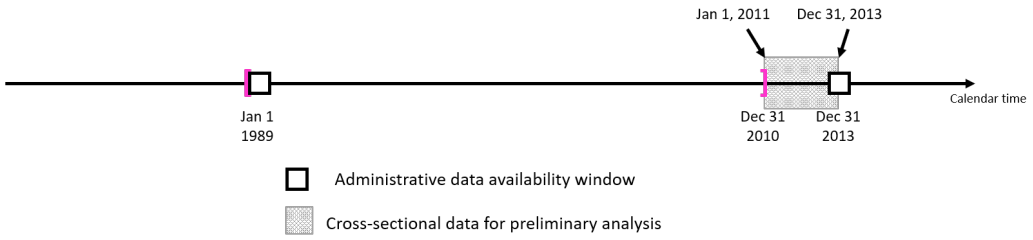


Figure 2.3: Breast Cancer Survivor Cohort Data Collection Windows for $\mathcal{P}_{\text{prelim}}$

Out of the $51,612$ breast cancer survivors, about 81 percent are newly referred patients. For future studies in Chapters 3 and 4 we only included these new patients because events occuring prior to 1989 are unknown for other patients. We call this set of cohorts 'referred cohort', denoted by $\mathcal{P}_{\text{referred}}$. Stage IV survivors were also removed because, by definition, the patients cannot have a relapse.

Furthermore, those with unknown treatment or stage were excluded from $\mathcal{P}_{\text{referred}}$, and we call this set of cleaned-up cohort 'final cohort', denoted by $\mathcal{P}_{\text{final}}$. In Chapters 5 and 6 real data analyses were conducted using $\mathcal{P}_{\text{final}}$. The real data analysis in Chapter 6 uses a case control study which utilized the full set of cohorts $\mathcal{P}_0$ and $\mathcal{Q}_0$.

2.3 Analysis of BC-BRCAS Data (I): Preliminary Data Analyses

In attempt to achieve the goals in Chapter 1, a set of preliminary analyses is conducted using Poisson rate regression on: (i) the count of CVD-related hospitalizations between calendar years 2011 and 2013, and (ii) the event indicator (yes/no) of CVD-related hospitalization. Figure 2.3 illustrates the cross-sectional data that was extracted for analysis. The adjusting

covariates include: *Birth Era*, *SES* (low versus high), *Residential Location: rural versus urban*, *Stage* (stage I as reference), *Treatment* (surgery only as reference), and *Having RSC*. Table 2.2 summarizes the results using $\mathcal{P}_{\text{prelim}} \cup \mathcal{Q}_{\text{prelim}}$. Table 2.3 presents additional results among the breast cancer survivors using $\mathcal{P}_{\text{prelim}}$ only.

It appears on average that cases have significantly higher rates of CVD related hospitalizations as compared to controls. The rate of CVD-related hospitalization is significantly associated with Birth Era. With other covariates fixed, older ages (born in an earlier era) tend to have a higher rate of hospitalization on average. SES and urban/rural residential location do not appear to have any significant effects. In addition, case only analysis indicates having RSC increases the rate of having CVD, while the addition of chemo or radiation treatment seems to decrease the rate of having CVD, compared to those who received surgery only. However, it could be due to the collinearity between Birth Era and Treatment. There may also be a sampling issue due to the data extraction window. For example, the cohort $\mathcal{P}_{\text{prelim}}$ consists only of those who survived until the year 2011, and of those born in era I (1900-1927), 97.6 percent receive surgery only or surgery and radiation. It is likely that those who received any chemotherapy were already deceased by the year 2011. It highlights the importance of using lifetime data analysis to explore the effect of treatment.

2.4 Statistical Challenges

The preliminary analysis presented in the above section utilized only cross-sectional data. However, cross-sectional analyses measure a single outcome for each individual. In addition, cross-sectional analyses, which describe the current situation of the population at the given moment or time period, hardly integrates the dimension of time, thus making the effect of time unmeasurable. Another consequence of failing to consider event-times is that it makes it infeasible to account for informative censoring using cross-sectional analyses. For example, in the breast cancer study described in Chapter 1, death could be a potential informative censoring time. Moreover, event time analyses can deal with time-varying covariates.

As shown in figure 2.2, all subjects were followed throughout the time window, and were censored by either the end of data collection or by death. Following the notation in Chapter 1 to let T_1 be time to the first RSC, T_2 be time to the first CVD admission, and D be time to death, we obtain the event times (T_1, T_2, D). When we use the time scale as time since diagnosis, for example, T_1 is the calendar time gap between time at first RSC and time at diagnosis. As shown in figure 2.4, there are eight different scenarios for the observed event-times on (T_1, T_2, D).

If we can estimate the joint distribution of the event times T_1 and T_2, then we can achieve *goal 1* and *goal 2*. Similarly, if we can estimate the joint distribution of (T_1, T_2) conditional on covariates of interest, then we can achieve *goal 3*. It is not straightforward to address

this statistical problem because both event times T_1 and T_2 are censored by D, which is a potential informative censoring time. Conventional approaches such as the Kaplan–Meier estimator do not apply because of the violation of the noninformative censoring assumption. This leads to the following research presented in Chapters 3 to 6.

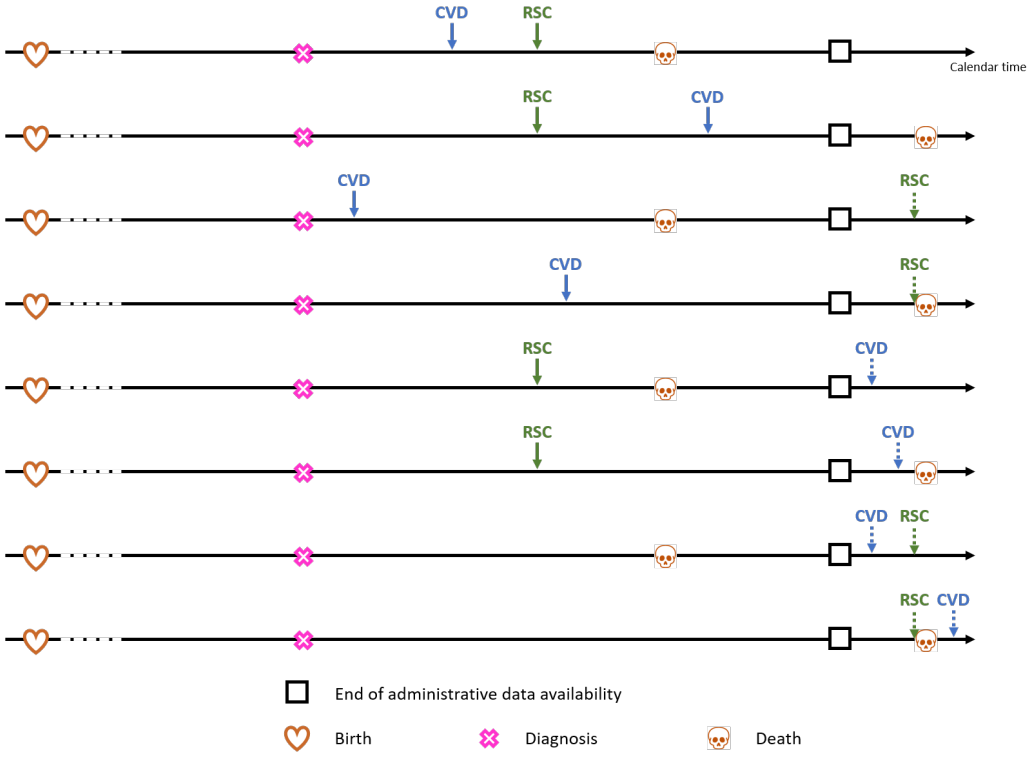


Figure 2.4: Different Scenarios of Observed Event Times. Time Scale: Time Since Diagnosis.

Table 2.1: Charateristics Summary: Breast Cancer Survivor Cohort and the Comparison Group of the Original Cohort ($\mathcal{P}_0$), Cross-sectional Cohort ($\mathcal{P}_{\text{prelim}}$), Referred Cohort ($\mathcal{P}_{\text{referred}}$), and Final Cleaned-up Cohort ($\mathcal{P}_{\text{final}}$), based on the BC-BRCAS Data.

		Full Cohort				Cross Sectional Cohort				Referred Cohort		Final Cleaned-up Cohort	
		Cases ($\mathcal{P}_0$)		Controls ($\mathcal{Q}_0$)		Cases ($\mathcal{P}_{\text{prelim}}$)		Controls ($\mathcal{Q}_{\text{prelim}}$)		Case ($\mathcal{P}_{\text{referred}}$)		Cases ($\mathcal{P}_{\text{final}}$)	
		N	%	N	%	N	%	N	%	N	%	N	%
Total		51612		103224		34683		63731		40147		36735	
Known number of deaths by end 2013		19212	37.2	25592	24.8	3635	10.5	3923	6.2	11961	29.8	11100	30.2
Birth Era	1 (1900 - 1927)	13311	25.8	26622	25.8	4502	13.0	8024	12.6	7767	19.3	7180	19.5
	2 (1928 - 1945)	18807	36.4	37614	36.4	13934	40.2	26917	42.2	15287	38.1	14166	38.6
	3 (1946 - 1989)	19494	37.8	38988	37.8	16247	46.8	28790	45.2	17093	42.6	15389	41.9
Sociodemographic Factors													
SES	4,5-Highest	17031	33.0	29399	28.5	13043	37.6	21814	34.2	13919	34.7	12651	34.4
	1,2,3-Lowest	24814	48.1	46410	45	17744	51.2	31934	50.1	19575	48.8	17693	48.2
	Unknown	9767	18.9	27415	26.6	3896	11.2	9983	15.7	6653	16.6	6391	17.4
Urban/Rural of Residence	Rural	4480	8.7	7989	7.7	3541	10.2	6224	9.8	3344	8.3	2986	8.1
	Urban	27912	54.1	50353	48.8	22641	65.3	39808	62.5	23542	58.6	20911	56.9
	Unknown	19220	37.2	44882	43.5	8501	24.5	17699	27.8	13261	33.0	12838	34.9
Health Authority	Interior	9704	18.8	17333	16.8	6380	18.4	10526	16.5	6551	16.3	5933	16.2
	Fraser	15908	30.8	28121	27.2	10805	31.2	18022	28.3	13081	32.6	11966	32.6
	Coastal	12629	24.5	24166	23.4	8662	25.0	14523	22.8	9980	24.9	9120	24.8
	Island	10387	20.1	17645	17.1	6797	19.6	10820	17.0	9035	22.5	8337	22.7
	Northern	2647	5.1	5321	5.2	1805	5.2	3215	5.0	1399	3.5	1295	3.5
	Unknown	337	0.7	10638	10.3	234	0.7	519	0.8	101	0.0	84	0.2
Clinical Factors (for cases)													
Age at diagnosis (continuous)	Mean(std)	61.5 (14.1)	-	-	-	58.9 (12.8)	-	-	-	59.6 (13.3)		50.3 (13.1)	-
Age at diagnosis (group)	< 40	3238	6.3	-	-	2236	6.4	-	-	2801		2617	7.1
	≥ 40	48374	93.7	-	-	32447	93.6	-	-	37346		34118	92.9
Stage at diagnosis	I	19661	38.1	-	-	15830	45.6	-	-	19063		17832	48.5
	II	16982	32.9	-	-	11677	33.7	-	-	16276		15200	41.4
	III	4668	9.0	-	-	2900	8.4	-	-	4486		3703	10.1
	IV	1932	3.7	-	-	513	1.5	-	-	-		-	-
	Unknown	8369	16.2	-	-	3763	10.8	-	-	322		-	-
Treatment Received	Surgery only	9518	18.4	-	-	5610	16.2	-	-	8347		8240	22.4
	Chemo only	12	0.0	-	-	≤ 5	0.0	-	-	-		-	-
	Radiation only	41	0.1	-	-	13	0.0	-	-	-		-	-
	Surgery+Chemo	3322	6.4	-	-	2347	6.8	-	-	2883		2866	7.8
	Surgery+Radiation	14754	28.6	-	-	10982	31.7	-	-	14560		14470	39.4
	Chemo+Radiation	68	0.1	-	-	39	0.1	-	-	-		-	-
	Surgery+Chemo+Radiation	11376	22.0	-	-	8837	25.5	-	-	11210		11159	30.4
	Unknown	12521	24.3	-	-	6851	19.8	-	-	3147		-	-
Having Relapse or Second Cancer (RSC)	No	37292	72.3	-	-	28410	81.9	-	-	28384		25710	70.0
	Yes	14320	27.7	-	-	6273	18.1	-	-	11763		11025	30.0
Referral Status*	1	381	0.7	-	-	261	0.8	-	-	-		-	-
	2	42069	81.5	-	-	30323	87.4	-	-	40147		36735	100.0
	3	938	1.8	-	-	237	0.7	-	-	-		-	-
	4	89	0.2	-	-	22	0.1	-	-	-		-	-
	Unknown	8135	15.8	-	-	3840	11.1	-	-	-		-	-

*: 1=follow-up, 2=new patient, 3= recurrence, 4= residual disease, Blank = not referred

Description:

New patient = Patient is referred shortly after diagnosis of breast cancer. May be referred pre or post-operatively.

Follow-up= Patient is referred for follow-up of breast cancer that was initially treated elsewhere. No active disease is present.

Recurrence= Patient is referred with active disease after a period of remission for breast cancer that was initially treated elsewhere.

Residual disease = Patient is referred with active disease after an interval of time has elapsed between diagnosis and first visit to the Agency. Disease has never been in complete remission.

Table 2.2: BC-BRCAS Data Analysis (I). Preliminary Cross-sectional Analysis Using Poisson Rate Regression for Case ($\mathcal{P}_0$) Control ($\mathcal{Q}_0$) Comparison.

		Case - Control		Case		Control	
		Yes/No	Count	Yes/No	Count	Yes/No	Count
		$\hat{\beta}$ (se)	$\hat{\beta}$ (se)	$\hat{\beta}$ (se)	$\hat{\beta}$ (se)	$\hat{\beta}$ (se)	$\hat{\beta}$ (se)
Intercept		-5.12 (0.319)	-5.08 (0.357)	-2.14 (0.317)	-2.08 (0.397)	-4.66 (1.012)	-4.99 (1.239)
case control		3.00 (0.076)	2.99 (0.087)				
SES	Low (1-3)	0.11 (0.047)	0.15 (0.054)	0.11 (0.052)	0.16 (0.062)	0.04 (0.175)	0.11 (0.189)
	High (4-5)	-	-	-	-	-	-
	Unknown	0.00 (0.079)	0.00 (0.091)	0.03 (0.088)	0.02 (0.104)	-0.25 (0.307)	-0.14 (0.326)
Urban	Rural	-0.02 (0.074)	-0.05 (0.086)	-0.06 (0.082)	-0.12 (0.101)	0.28 (0.248)	0.51 (0.245)
	Urban	-	-	-	-	-	-
	Unknown	-0.04 (0.057)	0.02 (0.065)	-0.03 (0.063)	0.03 (0.074)	-0.13 (0.217)	-0.13 (0.235)
Era	I	-	-	-	-	-	-
	II	-0.47 (0.051)	-0.43 (0.059)	-0.45 (0.057)	-0.41 (0.068)	-0.66 (0.176)	-0.65 (0.187)
	III	-1.73 (0.066)	-1.72 (0.077)	-1.65 (0.072)	-1.63 (0.087)	-2.91 (0.337)	-3.05 (0.380)
log(days)		-0.02 (0.042)	0.03 (0.051)	-0.02 (0.046)	0.02 (0.058)	-0.04 (0.147)	0.06 (0.180)
Dispersion parameter		0.850	1.739	0.952	2.087	1.015	1.801

log-days as offset		Case - Control		Case		Control	
		Yes/No	Count	Yes/No	Count	Yes/No	Number
		$\hat{\beta}$ (se)	$\hat{\beta}$ (se)	$\hat{\beta}$ (se)	$\hat{\beta}$ (se)	$\hat{\beta}$ (se)	$\hat{\beta}$ (se)
Intercept		-12.10 (0.099)	-11.73 (0.109)	-9.12 (0.067)	-8.755 (0.079)	-11.82 (0.209)	-11.46 (0.219)
case control		3.01 (0.084)	3.00 (0.092)				
SES	Low (1-3)	0.12 (0.052)	0.17 (0.057)	0.13 (0.057)	0.17 (0.066)	0.05 (0.188)	0.11 (0.196)
	High (4-5)	-	-	-	-	-	-
	Unknown	0.02 (0.088)	0.02 (0.096)	0.05 (0.096)	0.04 (0.111)	-0.19 (0.328)	-0.10 (0.337)
Urban	Rural	-0.01 (0.082)	-0.04 (0.091)	-0.05 (0.091)	-0.11 (0.108)	0.28 (0.266)	0.51 (0.254)
	Urban	-	-	-	-	-	-
	Unknown	-0.04 (0.063)	0.02 (0.068)	-0.03 (0.069)	0.03 (0.079)	-0.11 (0.233)	-0.12 (0.243)
Era	I	-	-	-	-	-	-
	II	-0.60 (0.055)	-0.56 (0.061)	-0.58 (0.061)	-0.54 (0.071)	-0.77 (0.186)	-0.74 (0.191)
	III	-1.88 (0.072)	-1.85 (0.080)	-1.80 (0.079)	-1.77 (0.091)	-3.03 (0.360)	-3.15 (0.392)
Dispersion parameter		1.045	1.944	1.153	2.381	1.171	1.928

Table 2.3: BC-BRCAS Data Analysis (I). Preliminary Cross-sectional Analysis Using Poisson Rate Regression for Survivor ($\mathcal{P}_0$) Only Study.

		log(days) as covariate		log(days) as offset	
		Yes/No	Count	Yes/No	Count
		$\hat{\beta}$ (se)	$\hat{\beta}$ (se)	$\hat{\beta}$ (se)	$\hat{\beta}$ (se)
Intercept		-2.31 (0.404)	-2.33 (0.504)	-9.67 (0.260)	
Era	**I**	-	-	-	-
	II	-0.55 (0.061)	-0.51 (0.073)	-0.66 (0.065)	-0.61 (0.077)
	III	-1.72 (0.086)	-1.69 (0.103)	-1.84 (0.093)	-1.81 (0.108)
Sociodemographic Factors					
SES	**4,5-Highest**	-	-	-	-
	1,2,3-Lowest	0.10 (0.054)	0.15 (0.065)	0.12 (0.059)	0.16 (0.069)
	Unknown	0.02 (0.089)	0.01 (0.106)	0.01 (0.097)	0.00 (0.112)
Urban/Rural of Residence	**Rural**	0.06 (0.091)	0.02 (0.111)	0.06 (0.099)	0.03 (0.118)
	Urban	-	-	-	-
	Unknown	0.04 (0.064)	0.11 (0.075)	0.04 (0.070)	0.12 (0.080)
Region of Residence	**Interior**	0.09 (0.080)	0.13 (0.096)	0.10 (0.087)	0.14 (0.102)
	Fraser	0.16 (0.066)	0.24 (0.079)	0.17 (0.072)	0.25 (0.084)
	Coastal	-	-	-	-
	Island	0.01 (0.073)	-0.06 (0.090)	0.03 (0.079)	-0.04 (0.095)
	Northern	0.23 (0.141)	0.37 (0.160)	0.24 (0.153)	0.37 (0.170)
	Unknown	-12.14 (132.705)	-12.47 (189.565)	-11.86 (133.163)	-12.21 (186.810)
Clinical Factors					
Age at diagnosis	< 40	-	-	-	-
	≥ 40	0.55 (0.219)	0.66 (0.272)	0.54 (0.239)	0.65 (0.289)
Stage at diagnosis	**I**	-	-	-	-
	II	0.21 (0.055)	0.23 (0.066)	0.24 (0.060)	0.26 (0.070)
	III	0.28 (0.099)	0.27 (0.119)	0.36 (0.108)	0.34 (0.127)
Treatment Received	**Surgery only**	-	-	-	-
	Chemo only	-0.09 (0.118)	-0.01 (0.138)	-0.11 (0.129)	-0.03 (0.146)
	Radiation only	0.00 (0.059)	0.03 (0.072)	-0.01 (0.065)	0.02 (0.076)
	Chemo + Radiation	-0.10 (0.084)	-0.06 (0.100)	-0.13 (0.091)	-0.09 (0.106)
Having Relapse or Second Cancer (RSC)	**No**	-	-	-	-
	Yes	0.15 (0.058)	0.08 (0.071)	0.28 (0.062)	0.20 (0.074)
log(days)		-0.07 (0.048)	-0.04 (0.060)	-	-
dispersion parameter		0.932	1.105	2.060	2.325

Chapter 3

Multiple Event Times in the Presence of Informative Censoring Using Copula - Part One

This chapter is concerned with the joint survivor function of multiple event times when the observations are subject to informative censoring caused by a terminating event. We formulated the correlation of the multiple event times together with the time to the terminating event by an Archimedean copula. This accounts for the informative censoring and adapts the commonly used two-step procedure for estimating the joint distribution of the multiple event times under a copula model. We propose an easy-to-implement pseudolikelihood based estimation procedure under the model. A by-product of the approach is a new estimator for the marginal distribution of a single event time with semicompeting-risks data. We derived asymptotic properties and conducted simulation studies to examine the consistency, efficiency, and robustness of the proposed approach. Data from the breast cancer project is employed to motivate and illustrate the method.

3.1 Introduction

In many studies with human subjects, the researchers are usually not confident in formulating the distributions of interest into parametric models. The problem becomes more involved when multiple event times are of interest and the observations on them are subject to informative censoring. In the breast cancer study, for example, the occurrence of death is likely correlated with both the time to CVD and the time to RSC. The death time censors the observation on either of the two event times but not vice versa: the censoring is thus potentially correlated with the event times. Motivated by the breast cancer study, this chapter focuses on the informative censoring with observations on multiple event times due to a terminating event that is correlated with the event times. Leaving the marginal

distribution of the terminating event unspecified, we model the correlation of the multiple event times together with the terminating event time via an Archimedean copula model. It allows us to adapt naturally the commonly-used two-step estimation procedure with a copula model and, in the meantime, account for the informative censoring.

We motivate the proposed model and illustrate the associated estimation procedure using a study at BC Cancer with more recent data. The methodology, however, has broader applicability. The rest of this chapter is organized as follows. Section 3.2 presents the model after introducing the notation and framework. Section 3.3 proposes a pseudolikelihood-based semiparametric procedure to estimate the model parameters. We then derive the asymptotic properties of the resulting estimators, and in particular the maximum pseudo-likelihood estimator (pseudo-MLE) for the model parameter that measures the association between the event times. Section 3.5 reports a simulation study that evaluated the finite-sample performance of the estimation procedure, and section 3.6 presents an analysis of the real data from the breast cancer study from the proposed approach. Section 3.7 provides a summary of this chapter.

3.2 Notation and Modeling

3.2.1 Notation

Let T_j with survivor function $S_j(\cdot)$ for $j = 1, \ldots, J$ be the J (≥ 1) event times of interest in a study. Denote their joint survivor function by $S(\underline{t}) = Pr(T_1 \geq t_1, \ldots, T_J \geq t_J)$ with $\underline{t} = (t_1, \ldots, t_J)$. Suppose the study observations on the event times T_j are subject to right-censoring where the censoring time C is either the time to a terminating event D with survivor function $S_D(\cdot)$ or the study's administrative follow-up time C_A, whichever comes sooner. That is, C is the minimum of D and C_A, denoted as $C = D \wedge C_A$. The observations on T_j may be censored by D but not vice versa; this structure is referred to as semicompeting-risks data (e.g., Fine et al. 2001).

We aim to estimate the joint survivor function $S(\underline{t})$ given the study's right-censored multivariate event times when T_j for $j = 1, \ldots, J$ are potentially correlated among each other and with D. Adopting the conventional notation, let Δ_D be the indicator $I\{D \leq C_A\}$, and $U_j = T_j \wedge C$ with $\Delta_j = I\{T_j \leq C\}$ for $j = 1, \ldots, J$. Suppose that the study data are n independent realizations of $\{(U_1, \Delta_1), \ldots, (U_J, \Delta_J), (C, \Delta_D)\}$, denoted by

$$\text{Observed-Data} = \bigcup_{i=1}^{n} \Big\{ \{(u_{ji}, \delta_{ji}) : j = 1, \ldots, J\} \bigcup \{(c_i, \delta_{Di})\} \Big\}. \tag{3.1}$$

This is the union of the J semicompeting-risks data sets on T_j together with D:

$$\text{Observed-Data}_j = \big\{(u_{ji}, \delta_{ji}, c_i, \delta_{Di}) : i = 1, \ldots, n\big\}. \tag{3.2}$$

We perform inference on the distributions of the event times T_j over the intervals $[0, v_j^\star]$ with predetermined $v_j^\star$. In practice, $v_j^\star$ are usually chosen to be slightly smaller than anticipated $\max_i\{u_{ji}\}$ for $j = 1, \ldots, J$.

3.2.2 Model Specification

Denote a k-dimensional Archimedean copula function by $\mathcal{C}_{[k]}(v_1, \ldots, v_k; \theta)$ for integer $k \geq 1$ with its generator function $\psi(\cdot; \theta)$. We assume that the administrative censoring time C_A is independent of both $(T_1, \ldots, T_J)$ and D, and the joint survivor function of the multiple event times with D follows the $(J+1)$-dimensional Archimedean copula model (e.g., Nelsen 2006):

$$Pr(T_1 \geq t_1, \ldots, T_J \geq t_J, D \geq d) = \mathcal{C}_{[J+1]}(S_1(t_1), \ldots, S_J(t_J), S_D(d); \theta). \tag{3.3}$$

The association parameter θ characterizes the correlation of the $J+1$ event times $T_1, \ldots, T_J$ and the terminating event time D.

Following (1.6), the Archimedean copula model in (3.3) can be presented as

$$Pr(T_1 \geq t_1, \ldots, T_J \geq t_J, D \geq d) = \mathcal{C}_{[2]}(S(\underline{t}; \theta), S_D(d); \theta), \tag{3.4}$$

where the joint survivor function of $T_1, \ldots, T_J$ is

$$S(\underline{t}; \theta) = \mathcal{C}_{[J]}(S_1(t_1), \ldots, S_J(t_J); \theta). \tag{3.5}$$

Further, by (1.7), we see that, for $j = 1, \ldots, J$, (3.3) is

$$\mathcal{C}_{[K]}(S_1(t_1), \ldots, S_{j-1}(t_{j-1}), S_{j+1}(t_{j+1}), \ldots, S_J(t_J), S_{jD}(t_j, d; \theta); \theta)$$

with the joint survivor function of T_j and D:

$$S_{jD}(t_j, d; \theta) = Pr(T_j \geq t_j, D \geq d) = \mathcal{C}_{[2]}(S_j(t_j), S_D(d); \theta). \tag{3.6}$$

Estimating the marginal survivor functions of the event times T_j $(j = 1, \ldots, J)$ with the semicompeting-risks data (denoted Observed-Data$_j$ in (3.2)), is of interest in many situations. It can now be viewed as the special case of $J = 1$ in the estimation presented in section 3.3.

3.3 Pseudolikelihood-Based Inference Procedure

This section presents the likelihood function of the parameters in the model (3.3) based on the available data. It then proposes a procedure for estimating the joint survivor function of the multiple event times. Asymptotic properties of the estimators are also provided.

3.3.1 Likelihood Function Based on the Available Data

Denote $\sum_{j=1}^{J}\delta_{ji}=\delta_{\cdot i}$, and $\underline{u}_i=(u_{1i},\ldots,u_{Ji})$. Let $\dot{h}(r)$ be the derivative for a function $h(r)$, and $h^{(a_1,\ldots,a_k)}(r_1,\ldots,r_k)$ be $\partial h^{(a_1+\ldots+a_k)}(r_1,\ldots,r_k)/\partial r_1^{a_1}\ldots\partial r_k^{a_k}$ for a function $h(r_1,\ldots,r_k)$ with well-defined partial derivatives. The likelihood function with the available data under the copula model (3.3) is

$$\begin{aligned}
& L(\theta;S_1(\cdot),\ldots,S_J(\cdot),S_D(\cdot)|\text{Observed-Data})\\
= & \prod_{i=1}^{n}\Big\{(-1)^{\delta_{\cdot i}+\delta_{Di}}\mathcal{C}_{[J+1]}^{(\delta_{1i},\ldots,\delta_{Ji},\delta_{Di})}(S_1(u_{1i}),\ldots,S_J(u_{Ji}),S_D(c_i);\theta)\Big[\prod_{j=1}^{J}\dot{S}_j(u_{ji})^{\delta_{ji}}\Big]\dot{S}_D(c_i)^{\delta_{Di}}\Big\}\\
= & \prod_{i=1}^{n}\Big\{(-1)^{\delta_{\cdot i}+\delta_{Di}}\frac{\partial^{\delta_{\cdot i}}\mathcal{C}_{[2]}^{(0,\delta_{Di})}(S(\underline{u}_i;\theta),S_D(c_i);\theta)}{\partial S_1(u_1)^{\delta_{1i}}\ldots\partial S_J(u_J)^{\delta_{Ji}}}\Big[\prod_{j=1}^{J}\dot{S}_j(u_{ji})^{\delta_{ji}}\Big]\dot{S}_D(c_i)^{\delta_{Di}}\Big\}. \qquad (3.7)
\end{aligned}$$

Here $I(A)$ is the indicator of set A. When $J=2$,

$$\begin{aligned}
& \frac{\partial^{\delta_{\cdot i}}\mathcal{C}_{[2]}^{(0,\delta_{Di})}(S(\underline{u}_i;\theta),\tilde{S}_D(c_i);\theta)}{\partial S_1(u_1)^{\delta_{1i}}\ldots\partial S_J(u_J)^{\delta_{Ji}}}\\
= & \begin{cases}
\mathcal{C}_{[2]}^{(0,\delta_{Di})}(S(\underline{u}_i;\theta),\tilde{S}_D(c_i);\theta), & \delta_{1i}=\delta_{2i}=0\\
\mathcal{C}_{[2]}^{(1,\delta_{Di})}(S(\underline{u}_i;\theta),\tilde{S}_D(c_i);\theta)\mathcal{C}_{[2]}^{(\delta_{1i},\delta_{2i})}(S_1(u_{1i}),S_2(u_{2i});\theta), & \delta_{1i}\neq\delta_{2i}\\
\mathcal{C}_{[2]}^{(2,\delta_{Di})}(S(\underline{u}_i;\theta),\tilde{S}_D(c_i);\theta)\mathcal{C}_{[2]}^{(0,1)}(S_1(u_{1i}),S_2(u_{2i});\theta)\mathcal{C}_{[2]}^{(1,0)}(S_1(u_{1i}),S_2(u_{2i});\theta) & \\
+\mathcal{C}_{[2]}^{(1,\delta_{Di})}(S(\underline{u}_i;\theta),\tilde{S}_D(c_i);\theta)\mathcal{C}_{[2]}^{(1,1)}(S_1(u_{1i}),S_2(u_{2i});\theta), & \delta_{1i}=\delta_{2i}=1.
\end{cases}
\end{aligned}$$

It is not easy to obtain the MLE of θ by maximizing (3.7) with respect to θ jointly with the unspecified survivor functions $S_j(\cdot)$ and $S_D(\cdot)$. Note that the right-censored observations on D are due to noninformative censoring. Thus there is a readily available consistent estimator for $S_D(\cdot)$, e.g., the Kaplan–Meier estimator, denoted as $\tilde{S}_D(\cdot)$. Following the idea of the pseudolikelihood estimation procedure under a copula model (e.g., Lawless & Yilmaz 2011), we may consider (3.7) with $S_D(\cdot)$ substituted by its estimate $\tilde{S}_D(\cdot)$:

$$\begin{aligned}
& L(\theta;S_1(\cdot),\ldots,S_J(\cdot)|\tilde{S}_D(\cdot);\text{Observed-Data})\\
= & \prod_{i=1}^{n}\Big\{(-1)^{\delta_{\cdot i}+\delta_{Di}}\frac{\partial^{\delta_{\cdot i}}\mathcal{C}_{[2]}^{(0,\delta_{Di})}(S(\underline{u}_i;\theta),\tilde{S}_D(c_i);\theta)}{\partial S_1(u_1)^{\delta_{1i}}\ldots\partial S_J(u_J)^{\delta_{Ji}}}\Big[\prod_{j=1}^{J}\dot{S}_j(u_{ji})^{\delta_{ji}}\Big]\Big\}, \qquad (3.8)
\end{aligned}$$

and maximize it with respect to θ together with $S_j(\cdot)$ to obtain a pseudo-MLE of θ. The resulting estimator for θ, with the trade-off of some efficiency loss, can be easier to implement than its MLE counterpart.

However, since (3.8) involves the unspecified survivor functions $S_j(\cdot)$ of T_j, that pseudo-MLE of θ is still rather hard to compute. It is especially so when the number of multiple

event times J is larger than 1. This consideration motivated the following estimation procedure.

3.3.2 Pseudo-MLE of Association Parameter

When the marginal survivor functions $S_j(\cdot)$ for $j = 1, \ldots, J$ and $S_D(\cdot)$ are known, the likelihood function (3.7) is proportional to

$$\begin{aligned} & L(\theta | S_1(\cdot), \ldots, S_J(\cdot), S_D(\cdot); \text{Observed-Data}) \\ = & \prod_{i=1}^{n} \Big\{ (-1)^{\delta_{\cdot i} + \delta_{Di}} \mathcal{C}_{[J+1]}^{(\delta_{1i}, \ldots, \delta_{Ji}, \delta_{Di})} (S_1(u_{1i}), \ldots, S_J(u_{Ji}), S_D(c_i); \theta) \Big\}. \end{aligned} \tag{3.9}$$

The MLE of θ from (3.9) is easy to calculate. However, the marginal survivor functions are unknown in many practical situations, and thus the MLE of θ is not evaluable.

Under model (3.6) induced from model (3.3) specified in section 3.2, the marginal survivor function $S_j(\cdot)$ can be expressed as a function of the marginal survivor function of $T_j^* = T_j \wedge D$, denoted by $S_j^*(\cdot)$, and the marginal survivor function of D: for $j = 1, \ldots, J$,

$$S_j(t) = g(S_j^*(t), S_D(t); \theta) = \psi^{-1}\{\psi(S_j^*(t); \theta) - \psi(S_D(t); \theta); \theta\}, \tag{3.10}$$

where $\psi(\cdot; \theta)$ is the generator of the Archimedean copula chosen in model (3.3).

Note that the observations on $T_j^* = T_j \wedge D$ and D are subject to noninformative censoring due to the administrative follow-up time C_A. Well-established survival approaches such as the Kaplan–Meier estimator and the Nelson–Aalen estimator (e.g., Andersen et al. 1993) can be used to consistently estimate their survivor functions $S_j^*(\cdot)$ and $S_D(\cdot)$. According to (3.10), $\tilde{S}_j(t; \theta) = g(\tilde{S}_j^*(t), \tilde{S}_D(t); \theta)$ is then a consistent estimator of $S_j(\cdot)$ with fixed θ.

The discussion above leads to the following estimation procedure. Plugging in (3.9) the consistent estimator for the unspecified survivor functions $S_j(\cdot)$ and $S_D(\cdot)$, we maximize the resulting pseudolikelihood function of θ or, equivalently, its log-transformation with respect to θ, to derive a pseudo-MLE of θ:

$$\hat{\theta}_n = \text{argmax}_\theta L(\theta | \tilde{S}_1(\cdot; \theta), \ldots, \tilde{S}_J(\cdot; \theta), \tilde{S}_D(\cdot); \text{Observed-Data}). \tag{3.11}$$

This pseudo-MLE procedure is computationally easy to implement. We present below an iterative algorithm to calculate $\hat{\theta}_n$.

ALGORITHM. Using the Kaplan–Meier estimates $\tilde{S}_j^*(\cdot)$ and $\tilde{S}_D(\cdot)$ together with the current estimate $\theta^{(k-1)}$ and $S_j^{(k-1)}(\cdot)$ for $j = 1, \ldots, J$ and with $k \geq 1$,

Step 1. obtain the updated estimate for θ as

$$\theta^{(k)} = \text{argmax}_\theta L(\theta | S_1^{(k-1)}(\cdot), \ldots, S_J^{(k-1)}(\cdot), \tilde{S}_D(\cdot); \text{Observed-Data});$$

Step 2. obtain the updated estimates for $S_j(\cdot)$ as $S_j^{(k)}(t) = \tilde{S}_j(t;\theta^{(k)}) = g(\tilde{S}_j^*(t), \tilde{S}_D(t); \theta^{(k)})$ for $j = 1, \ldots, J$.

Repeat steps 1 and 2 until the sequence $\{\theta^{(k)} : k = 0, 1, \ldots\}$ converges. The limit is $\hat{\theta}_n$ defined in (3.11). The initial estimate $\theta^{(0)}$ is in fact not needed. The Kaplan–Meier estimates of $S_j(\cdot)$ may be used as the initial estimates $S_j^{(0)}(\cdot)$ for $j = 1, \ldots, J$.

The following proposition establishes the consistency and asymptotic normality of the resulting estimator.

Proposition 1. *Under the regularity conditions (RC1)–(RC4) presented in section 3.4, and provided $\tilde{S}_j^*(t)$ and $\tilde{S}_D(t)$ satisfy condition (AC1) in section 3.4, as $n \to \infty$, $\hat{\theta}_n \xrightarrow{a.s.} \theta$ and $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, AV(\theta))$, where the asymptotic variance is*

$$AV(\theta) = V_B(\theta)^{-1} V_A(\theta) V_B(\theta)^{-1} \tag{3.12}$$

with $V_B(\theta)$ and $V_A(\theta)$ the limits of

$$-\frac{1}{n}\sum_{i=1}^{n} \partial^2 \log L(\theta | \tilde{S}_1(\cdot;\theta), \ldots, \tilde{S}_J(\cdot;\theta), \tilde{S}_D(\cdot); \textit{Observed-Data})/\partial\theta^2 \tag{3.13}$$

and

$$\frac{1}{n} Var\Big\{ \sum_{i=1}^{n} \partial \log L(\theta | \tilde{S}_1(\cdot;\theta), \ldots, \tilde{S}_J(\cdot;\theta), \tilde{S}_D(\cdot); \textit{Observed-Data})/\partial\theta \Big\}, \tag{3.14}$$

respectively, and $\tilde{S}_j(t;\theta) = g(\tilde{S}_j^*(t), \tilde{S}_D(t); \theta)$.

In section 3.4 we outline a proof of this proposition after presenting the regularity conditions (RC1)–(RC4) and additional assumptions (AC1)–(AC2). One may estimate the variance of $\hat{\theta}_n$ by a bootstrap approach (e.g., Lawless & Yilmaz 2011). A natural and practical variance estimator evaluates (3.12) at $\theta = \hat{\theta}_n$ and uses (3.13) and (3.14) to replace their limits. The resulting variance estimator is often referred to as Huber's robust sandwich estimator (Huber 1967).

Note that when $S_j(\cdot)$ for $j = 1, \ldots, J$ are known and used to estimate θ, $\hat{\theta}_n$ is an MLE, and $V_A(\theta)$ and $V_B(\theta)$ in (3.13) and (3.14) are the same as the corresponding inverse Fisher information matrix. See section 3.5 for our empirical comparison of the robust variance estimator with the estimator based on the Fisher information inverse, a consistent estimator for the variance of the MLE.

3.3.3 Resulting Estimators for Marginal and Joint Survivor Function

Plugging $\hat{\theta}_n$, $\tilde{S}_j^*(t)$, and $\tilde{S}_D(t)$ from the above section in (3.10) gives a natural estimator for the marginal survivor function $S_j(\cdot)$:

$$\hat{S}_{jn}(t) = g(\tilde{S}_j^*(t), \tilde{S}_D(t); \hat{\theta}_n). \tag{3.15}$$

Proposition 2. *Under the regularity conditions (RC1)–(RC4) presented in section 3.4 and provided $\tilde{S}_j^*(\cdot)$ and $\tilde{S}_D(\cdot)$ satisfy condition (AC1) in section 3.4, as $n \to \infty$, $\hat{S}_{jn}(t) \xrightarrow{a.s.} S_j(t)$ uniformly and $\sqrt{n}(\hat{S}_{jn}(t) - S_j(t)) \xrightarrow{w} \mathcal{G}_j(t)$ with $t \in [0, v_j^\star]$, where $\mathcal{G}_j(t)$ is a Gaussian process with mean zero and variance function $\sigma_j^2(t)$ as defined in, for example, Andersen et al. (1993).*

An outline of a proof for the proposition is given in section 3.4. When the sample size is large and the censoring rate is not too high, we may choose to ignore the variation of the Kaplan–Meier estimates $\tilde{S}_j^*(\cdot)$ and $\tilde{S}_D(\cdot)$. It then yields an approximate confidence band (CB) for $S_j(\cdot)$ based on (3.10) with $\hat{\theta}_n$ plugged in, and using the proposed variance estimator of $\hat{\theta}_n$ in the above section.

Using the idea underlying two-stage estimation procedures with a copula model (e.g., Oakes 1994), we estimate the joint survivor function $S(\underline{t})$ of $(T_1, \ldots, T_J)$ based on (3.5):

$$\hat{S}_n(\underline{t}) = \mathcal{C}_{[J]}(\hat{S}_{1n}(t_1), \ldots, \hat{S}_{Jn}(t_J); \hat{\theta}_n). \tag{3.16}$$

The following proposition establishes the consistency and asymptotic normality/weak convergence of the resulting estimator.

Proposition 3. *Under the regularity conditions (RC1)–(RC4) presented in section 3.4 and provided $\tilde{S}_j^*(t)$ and $\tilde{S}_D(t)$ satisfy condition (AC1) in Section 3.4, as $n \to \infty$, $\hat{S}_n(\underline{t}) \xrightarrow{a.s.} S(\underline{t})$ uniformly and $\sqrt{n}(\hat{S}_n(\underline{t}) - S(\underline{t})) \xrightarrow{w} \mathcal{G}(\underline{t})$ with $\underline{t} \in [0, v_1^\star] \times \ldots \times [0, v_J^\star]$, where $\mathcal{G}(\underline{t})$ is a Gaussian field with mean zero and variance function $\sigma^2(\underline{t})$.*

We outline a proof for this proposition in section 3.4.

3.4 Asymptotic Properties

This section consists of two subsections. The first one outlines the derivation of asymptotic properties for bivariate cases when $J = 2$, as is the case for the motivating breast cancer example. The second subsection gives the asymptotic derivation for a general case.

3.4.1 Asymptotic Properties for Bivariate Case

This section outlines a derivation of the consistency and the asymptotic normality of the pseudo-MLE obtained through (3.11) when $J = 2$. Define the following regularity conditions:

> *(RC1)* Suppose θ is in an open interval Θ in the real line, $\mathcal{C}^{ab}(r_1, r_2; \theta)$, $\mathcal{C}^{abc}(r_1, r_2; \theta)$ exist and are continuous and uniformly bounded by some constant M for $a, b, c \in$ (0,1,2,3), $0 \le r_1 \le 1$, $0 \le r_2 \le 1$, $0 \le v_1 \le 1$, $0 \le v_2 \le 1$.

(RC2) For each $\theta \in \Theta$, $0 < E_\theta\left[\frac{\mathcal{C}^{abc}(r_1,r_2;\theta)}{\mathcal{C}^{ab}(r_1,r_2;\theta)}\right] \leq \infty$ for $a,b,c \in$ (0,1), $0 \leq r_1 \leq 1$, $0 \leq r_2 \leq 1$, $0 \leq v_1 \leq 1$, $0 \leq v_2 \leq 1$.

(RC3) Under an Archimedean copula, $g^{ab}(v_1, v_2; \theta)$ and $g^{abc}(v_1, v_2; \theta)$ exist and are continuous and uniformly bounded by some constant M for $a,b,c \in$ (0,1,2,3), $0 \leq r_1 \leq 1$, $0 \leq r_2 \leq 1$, $0 \leq v_1 \leq 1$, $0 \leq v_2 \leq 1$.

(RC4) Under an Archimedean copula, for each $\theta \in \Theta$, $0 < E_\theta\left[\frac{\mathcal{C}^{abc}(r_1,r_2;\theta)g^{001}(v_1,v_2;\theta)}{\mathcal{C}^{ab}(r_1,r_2;\theta)}\right] \leq \infty$ for $a,b,c \in$ (0,1), $0 \leq r_1 \leq 1$, $0 \leq r_2 \leq 1$, $0 \leq v_1 \leq 1$, $0 \leq v_2 \leq 1$.

Given data $Data = \{X_1, \cdots, X_n\}$, where $X_1, \cdots X_n$ are n i.i.d. observations, and denote S_1 and S_2 as two unknown survival functions, we write the likelihood function as $L(\alpha; S_1, S_2, Data)$, with α being the unknown parameter of interest. If the estimator functions $\hat{S}_1$ and $\hat{S}_2$ of S_1, S_2 satisfy the following two conditions:

C1 $\hat{S}_1$ and $\hat{S}_2$ converge uniformly to S_1 and S_2, respectively.

C2 $\sqrt{n}(\hat{S}_1 - S_1) \xrightarrow{w} \mathcal{G}_1$, and $\sqrt{n}(\hat{S}_2 - S_2) \xrightarrow{w} \mathcal{G}_2$, where $\mathcal{G}_1$ and $\mathcal{G}_2$ are two mean zero Gaussian processes with limiting covariance $cov(G_j(s_1), G_j(s_2)) = \sigma_j^2(s_1 \wedge s_2)$ for $j = 1, 2$, with σ_j^2 defined as in Andersen et al. (1993). For simplification, we denote $\sigma_1^2(\cdot)$ and $\sigma_2^2(\cdot)$ as the limiting variance function for $\sqrt{n}(\hat{S}_1 - S_1)$ and $\sqrt{n}(\hat{S}_2 - S_2)$, respectively.

Then we have the following lemma:

Lemma 1. *Under regularity condition (R1), the pseudo-MLE* $\hat{\alpha}^* \triangleq \arg\max_{\alpha\in\Theta} L(\alpha; \hat{S}_1, \hat{S}_2, Data)$ *satisfies*

(a) $\hat{\alpha}^* \xrightarrow{a.s.} \alpha$, *as* $n \to \infty$

(b) $\sqrt{n}(\hat{\alpha}^* - \alpha) \xrightarrow{d} N(0, V^*)$, *as* $n \to \infty$, *where* V^* *is the limiting variance.*

Proof. If S_1 and S_2 are known, then the 'pseudo'-MLE $\hat{\alpha}^*$ will be the regular MLE, denoted as $\hat{\alpha}$, and the above statements will hold following standard arguments, (Serfling 1980, see, for example). A sketch of proof is provided as follows: Define $Q(\alpha; S_1, S_2) = \partial \log f(\alpha; X, S_1, S_2)/\partial\alpha$ as the score function; and let

$$A_n(\alpha) = \frac{1}{n}\sum_{i=1}^{n} Q(\alpha; X_i, S_1, S_2) \triangleq \frac{1}{n}\sum_{i=1}^{n} a_i(\alpha),$$

where $a_i(\alpha)$'s are i.i.d because of the i.i.d. observations. In addition, define

$$B_n(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial Q(\alpha; X_i, S_1, S_2)}{\partial\alpha} \triangleq \frac{1}{n}\sum_{i=1}^{n} b_i(\alpha),$$

and

$$C_n = \frac{1}{n}\sum_{i=1}^{n} H(X_i).$$

Now

$$\frac{1}{n}\frac{\partial \log L(\alpha; S_1, S_2, Data)}{\partial \alpha} = A_n + B_n(\alpha - \alpha_0) + \frac{1}{2}C_n(\alpha - \alpha_0)^2\xi^2.$$

Since $E(Q(\alpha)) = 0$ and $Var(Q(\alpha)) = -E(\frac{\partial Q(\alpha; S_1, S_2)}{\partial \alpha}) \triangleq FI(\alpha)$ exists by regularity condition (1). We have $A_n \xrightarrow{a.s.} 0$, thus $\partial \log L(\alpha; S_1, S_2, Data)/\partial\theta \xrightarrow{a.s.} 0$ by Strong Law of Large Numbers (SLLN). Besides, by Central Limit Theorem, we have $\sqrt{n}\frac{\partial \log L(\alpha; S_1, S_2, Data)}{\partial \alpha} \xrightarrow{d} N(0, FI(\alpha))$. Since b_i's are i.i.d and $Var(\frac{\partial Q(\alpha; S_1, S_2)}{\partial \alpha})$ exists, by SLLN, we have $B_n \xrightarrow{a.s.} -FI(\alpha)$, as $n \to \infty$.
Following Serfling (1980) argument, we can show that $\hat{\alpha} \xrightarrow{a.s.} \alpha$. Since

$$0 = \frac{1}{n}\frac{\partial \log L(\alpha; S_1, S_2, Data)}{\partial \alpha} \mid_{\alpha=\hat{\alpha}} = A_n + B_n(\hat{\alpha} - \alpha) + \frac{1}{2}\xi^2 C_n(\hat{\alpha} - \alpha)^2,$$

we have

$$\sqrt{n}(\hat{\alpha} - \alpha) - \frac{-\sqrt{n}A_n}{B_n + \frac{1}{2}\xi^2 C_n(\hat{\alpha} - \alpha)} \xrightarrow{a.s} 0.$$

Also, since $\hat{\alpha} \xrightarrow{a.s} \alpha$, we have $B_n + \frac{1}{2}\xi^2 C_n(\hat{\alpha}_n - \alpha) \xrightarrow{a.s} -FI(\alpha)$; furthermore, we have $\sqrt{n}A_n \xrightarrow{d} N(0, FI(\alpha))$. By Slutsky's theorem,

$$\sqrt{n}(\hat{\alpha} - \alpha) \xrightarrow{d} N(0, FI^{-1}(\alpha)),$$

as $n \to \infty$.

Now we derive consistency, asymptotic normality and the variance estimator of pseudo-MLE as follows. 1. (*Consistency*) If S_1 and S_2 are unknown and are estimated by $\hat{S}_1$ and $\hat{S}_2$, respectively. Define

$$A_n^*(\alpha) = \frac{1}{n}\sum_{i=1}^{n} Q(\alpha; X_i, \hat{S}_1, \hat{S}_2) \triangleq \frac{1}{n}\sum_{i=1}^{n} a_i^*(\alpha),$$

,

$$B_n^*(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \frac{\partial Q(\alpha; X_i, \hat{S}_1, \hat{S}_2)}{\partial \alpha} \triangleq \frac{1}{n}\sum_{i=1}^{n} b_i^*(\alpha).$$

It is easy to verify through Taylor expansion that $\mathcal{C}^{ab}(\cdot, \cdot; \alpha)$, and $\mathcal{C}^{abc}(\cdot, \cdot; \alpha)$ still exist and are uniformtly bounded. So

$$\begin{aligned} A_n^* - A_n = &\frac{1}{n}\sum_{i=1}^{n} \phi_{1i}(S_1(X_i) - \hat{S}_1(X_i)) + \frac{1}{n}\sum_{i=1}^{n} \phi_{2i}(S_1(X_i) - \hat{S}_1(X_i)) \\ &+ \frac{1}{n}\sum_{i=1}^{n} o((S_1(X_i) - \hat{S}_1(X_i))^2 + (S_2(X_i) - \hat{S}_2(X_i))^2) \end{aligned} \tag{3.17}$$

where $\phi_{1i} = \partial Q(\theta; X_i, S_1, S_2)/\partial S_1(X)$, $\phi_{2i} = \partial Q(\theta; X_i, S_1, S_2)/\partial S_2(X)$. From regularity conditions, ϕ_{ki}'s are uniformly bounded ($k = 1, 2$), say by M. Since $\hat{S}_1 \to S_1$, and $\hat{S}_2 \to S_2$ uniformly, i.e., for $\forall\ \epsilon/2M > 0$, $\exists N$, s.t. for all $n > N$, $|S_1(X_i) - \hat{S}_1(X_i)| < \epsilon$, $|S_2(X_i) - \hat{S}_2(X_i)| < \epsilon$. Thus $|A_n^* - A_n| \xrightarrow{a.s.} 0$, so $Q(\alpha; \hat{S}_1, \hat{S}_2) \xrightarrow{a.s.} 0$. Similarly it can be shown that $|B_n^* - B_n| \xrightarrow{a.s.} 0$, and $B_n^* \xrightarrow{a.s} -B^*(\alpha)$, as $n \to \infty$. Following similar argument as above, we have

$$\hat{\alpha}^* \xrightarrow{a.s.} \alpha$$

2. (*Asymptotic Normality*) Define

$$\begin{aligned}
\omega_n &\triangleq \sqrt{n}(A_n^* - A_n) \\
&= \frac{1}{n}\sum_{i=1}^{n} \phi_{1i}\sqrt{n}(S_1(X_i) - \hat{S}_1(X_i)) \\
&+ \frac{1}{n}\sum_{i=1}^{n} \phi_{2i}\sqrt{n}(S_2(X_i) - \hat{S}_2(X_i)) \\
&+ \frac{1}{n}\sum_{i=1}^{n} \sqrt{n}o((S_1(X_i) - \hat{S}_1(X_i))^2 + (S_2(X_i) - \hat{S}_2(X_i))^2) \\
&\triangleq \omega_{nI} + \omega_{nII} + \Delta_{\omega_n}
\end{aligned} \tag{3.18}$$

It can be shown that $\Delta_{\omega_n} \xrightarrow{a.s} 0$ because of the uniform convergence of $\hat{S}_1$, $\hat{S}_2$. Given $\sqrt{n}(S_1 - \hat{S}_1)$ converges weakly to a zero-mean Gaussian process, denoted as $Z^{(I)}(\cdot)$, and $\sqrt{n}(S_2 - \hat{S}_2)$ converges weakly to a zero-mean Gaussian process, denoted as $Z^{(II)}(\cdot)$, by strong embedding theorem (Shorack & Wellner 2009), we could construct in another probablity space a sequence of stochastic processes $Z_n^{(I)}(\cdot)$, $Z_n^{(II)}(\cdot)$, such that

$$Z_n^{(I)}(t) \xrightarrow{a.s.} Z^{(I)}(t),$$

and

$$Z_n^{(II)}(t) \xrightarrow{a.s.} Z^{(II)}(t).$$

Therefore, now

$$\begin{aligned}
\omega_{nI} &= \frac{1}{n}\sum_{i=1}^{n} \phi_{1i}\{(Y_n^{(I)}(X_i) - Z_n^{(I)}(X_i)) + (Z_n^{(I)}(X_i) - Z^{(I)}(X_i)) + Z^{(I)}(X_i)\} \\
&\to \frac{1}{n}\sum_{i=1}^{n} \phi_{1i} Z^{(I)}(X_i) \\
&\sim N(0, V_{\omega I}^*)
\end{aligned} \tag{3.19}$$

since $\lim_{n\to\infty}(Y_n^{(I)}(X_i) - Z_n^{(I)}(X_i)) = 0$, and $\lim_{n\to\infty}(Z_n^{(I)}(X_i) - Z^{(I)}(X_i)) = 0$, where $Y_n^{(I)}(X_i) \triangleq \sqrt{n}(S_1(X_i) - \hat{S}_1(X_i))$, and $Y_n^{(II)}(X_i) \triangleq \sqrt{n}(S_2(X_i) - \hat{S}_2(X_i))$. Because $V_{\omega I}^* = \lim_{n\to\infty}\sum_{i=1}^n \phi_{1i}^2\sigma_I^2(X_i, X_i)/n$ and ϕ_{1i}'s are uniformly bounded and $\sigma_I^2(\cdot,\cdot)$ exists, thus $V_{\alpha I}^*$ exists.

Similarly, it can be shown that

$$\omega_{nII} \to \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^n \phi_{2i}Z^{(II)}(X_i) \sim N(0, V_{\omega II}^*)$$

.

Therefore,

$$\begin{aligned} AV_\omega^*(\alpha) &\triangleq \lim_{n\to\infty} Var(\omega_n) \\ &= V_{\omega I}^* + V_{\omega II}^* + 2\lim_{n\to\infty} Cov(\omega_{nI}, \omega_{nII}) \end{aligned} \tag{3.20}$$

By Cauchy Schwartz Inequality, it can be shown that the last terms exist too. Thus,

$$AV_{Q^*}(\alpha) = \lim_{n\to\infty} Var(\sqrt{n}A_n^*) = FI(\alpha) + AV_\omega^*(\alpha) + 2\lim_{n\to\infty}(\sqrt{n}A_n)(\sqrt{n}(A_n^* - A_n)).$$

So $\sqrt{n}A_n^* \xrightarrow{d} N(0, AV_{Q^*}(\alpha))$. Following a similar argument as above, we have

$$\sqrt{n}(\hat{\alpha}^* - \alpha) \xrightarrow{d} N(0, V^*),$$

where $V^* = [-B^*(\alpha)]^{-1}AV_Q^*(\alpha)[-B^*(\alpha)]^{-1}$.

3. (*Variance Estimation*) To estimate V^*, we use

$$\begin{aligned} \widehat{AV_{Q^*}(\alpha)} &= n\widehat{Var}(Q(\hat{\alpha}^*)) \\ &+ \frac{1}{n}\sum_{i=1}^n(\hat{\phi}_{1i}^2\hat{\sigma}_1^2(X_i)) + \frac{1}{n}\sum_{i=1}^n(\hat{\phi}_{2i}^2\hat{\sigma}_2^2(X_i)) + \frac{2}{n}\sum_{i=1}^n(\hat{\phi}_{1i}\hat{\sigma}_1(X_i))(\hat{\phi}_{2i}\hat{\sigma}_2(X_i)) \end{aligned} \tag{3.21}$$

and

$$\widehat{B^*(\alpha)} = \frac{1}{n}\sum_{i=1}^n b_i^*(\hat{\alpha}^*)$$

So the estimator for V^* is $[-\widehat{B^*(\alpha)}]^{-1}\widehat{AV_Q^*(\alpha)}[-\widehat{B^*(\alpha)}]^{-1}$. □

Theorem 3. *Under regularity conditions (R1) and (R2), as $n \to \infty$*

(a) $\hat{\theta}_j \xrightarrow{a.s.} \theta_j$, *for* $j = 1, 2$

(b) $\sqrt{n}(\hat{\theta}_j - \theta_j) \xrightarrow{d} N(0, V_j^*(\theta_j))$, *for* $j = 1, 2$

Proof. Theorem 3 is a direct application of lemma 1, by replacing S_1 with $S_{T_j\wedge D}$, and S_2 with S_D, and using Kaplan–Meier (KM) estimators to estimate $S_{T_j\wedge D}$ and S_D, respectively.

Since both $T_j \wedge D$ and D are subject to non-informative censoring, the KM estimator satisfies both uniform convergence and weak convergence requirements in lemma 1. □

Theorem 4. *Under regularity conditions (R1)-(R2), when $\mathcal{C}_j(\cdot,\cdot;\theta_j)$ belongs to the Archimedean family, then as $n \to \infty$,*

(a) $\hat{\theta} \xrightarrow{a.s.} \theta$

(b) $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, V^*(\theta))$

Proof. When $\mathcal{C}_j(\cdot,\cdot;\theta_j)$ belongs to the Archimedean family and regularity conditions (R3)-(R4) are satisfied, it can be easily shown that $\hat{S}_j(\cdot) = g(\hat{S}_{T_j^*}(\cdot), \hat{S}_D(\cdot); \hat{\theta}_j)$ converges to $S_j(\cdot)$ uniformly, and $\sqrt{n}(S_j(\cdot) - \hat{S}_j(\cdot))$ converges weakly to a mean zero Gaussian process. Applying lemma 1 again completes the proof. □

Because the estimated joint survival function $\hat{S}(\cdot)$ is obtained from $\hat{S}_{T_1}(\cdot), \hat{S}_{T_2}(\cdot), \hat{\theta}$, it can be easily shown that $\hat{S}_T(\cdot)$ is also consistent and converges to $S(\cdot)$.

Verification of density function $f(x)$

This section is to verify that

$$E(Q(\theta)) = 0.$$

The density function for $x \triangleq (u, c, \delta, \delta_D)$ can be written explicitly:

$$\begin{aligned} f_\theta(x) =& [\mathcal{C}^{11}(g(S_{T\wedge D}(u), S_D(u); \theta), S_D(c); \theta_j) f(u) f_D(c) S_A(c)]^{\delta\delta_D} \\ & [\mathcal{C}^{10}(g(S_{T\wedge D}(u), S_D(u); \theta), S_D(c); \theta_j) f(u) f_{C_A}(c)(-1)]^{\delta(1-\delta_D)} \\ & [\mathcal{C}^{01}(g(S_{T\wedge D}(u), S_D(u); \theta, S_D(c); \theta) f_D(c) f_{C_A}(c)(-1)]^{(1-\delta)\delta_D} \\ & [\mathcal{C}^{00}(g(S_{T\wedge D}(u), S_D(u); \theta, S_D(c); \theta) f_A(c)]^{(1-\delta)(1-\delta_D)} \end{aligned} \tag{3.22}$$

Decompose $X = (U, C, \Delta, \Delta_D)$ as $[Y, Z]$, with $Y = (U, C)$, a bivariate continuous random variable, and $Z = (\Delta, \Delta_D)$, a bivariate discrete random variable, with values taken as (0,0), (1,1), (0,1), (1,1) only. Then $[Y, Z] = [Y|Z][Z]$, then

$$\begin{aligned} E(Q(\theta)) &= \sum_{\forall z} \int_Y \frac{\partial f_{Y|Z}(y)}{\partial \theta} P(Z = z) dy \\ &= \frac{\partial}{\partial \theta} \sum_{\forall z} P(Z = z) \int_Y f_{Y|Z}(y) dy \\ &= \frac{\partial}{\partial \theta} (P(\delta = 1, \delta_D = 1) + P(\delta = 1, \delta_D = 0) + P(\delta = 0, \delta_D = 1) + P(\delta = 0, \delta_D = 0)) \\ &= 0 \end{aligned} \tag{3.23}$$

3.4.2 Asymptotics of a General Pseudo-MLE

This subsection presents derivations of the asymptotic properties listed in propositions 1, 2 and 3. We first introduce a general setting, and derive a lemma that establishes the consistency and asymptotic normality of a pseudo-MLE in the setting under the usual regularity conditions for the asymptotics of an MLE (e.g., Serfling 1980) combined with two additional conditions. We then adapt the regularity conditions, and outline proofs for the three propositions by applying the lemma.

Consider a K-dimensional random vector $(W_1, \ldots, W_K) \sim \mathcal{F}(G_1(w_1), \ldots, G_K(w_K); \alpha)$ with $G_k(\cdot)$ the marginal survivor function of W_k for $k = 1, \ldots, K$ and $\alpha \in \mathcal{A}$, where $\mathcal{A}$ is an open interval of the real line. Suppose there is a collection of n i.i.d. realizations on $\mathcal{X}(W_1, \ldots, W_K)$, a coarsened version of $(W_1, \ldots, W_K)$, denoted by

$$\text{General-Data} = \{X_i : X_i = \mathcal{X}(W_{1i}, \ldots, W_{Ki}), i = 1, \ldots, n\}. \tag{3.24}$$

Denote the loglikelihood function based on the data in (3.24) as

$$\log L(\alpha; G_1(\cdot), \ldots, G_K(\cdot)|\text{General-Data}) = \sum_{i=1}^{n} Q(X_i|G_1(\cdot), \ldots, G_K(\cdot); \alpha), \tag{3.25}$$

where $Q(\cdot)$ can be expressed as $Q_i(G_1(X_i^{(1)}), \ldots, G_K(X_i^{(K)}); \alpha)$. Here $(X_i^{(1)}, \ldots, X_i^{(K)})$ is a subcomponent vector of X_i, and $Q_i(\cdot)$ is determined by the distribution $\mathcal{F}(\cdot)$ and X_i.

Extend the notation introduced in section 3.3:
denote $\partial h^{(a_1+\ldots+a_K+b)}(r_1, \ldots, r_K; \alpha)/\partial r_1^{a_1} \ldots \partial r_K^{a_K} \partial \alpha^b$ by $h^{(a_1,\ldots,a_K;b)}(r_1, \ldots, r_K; \alpha)$ for a function $h(r_1, \ldots, r_K; \alpha)$ with well-defined partial derivatives. We adapt the conventional regularity conditions for the asymptotics of an MLE (e.g., Chp 4.2 of Serfling 1980) as follows.

Suppose $\theta \in \Theta$, an open interval of the real line, and $\mathcal{C}_{[J+1]}(r_1, \ldots, r_J, r_{J+1}; \theta)$ is a $(J+1)$-dimensional Archimedean copula function with its generator function $\psi(\cdot; \theta)$. Plus, let $S_j(\cdot)$ be a survivor function for $j = 1, \ldots, J+1$.

(RC1). $\mathcal{C}_{[J+1]}^{(a_1,\ldots,a_{J+1})}(r_1, \ldots, r_{J+1}; \theta)$ and $\mathcal{C}_{[J+1]}^{(a_1,\ldots,a_{J+1};b)}(r_1, \ldots, r_{J+1}; \theta)$ exist, and are continuous and uniformly bounded from above for $a_1, \ldots, a_{J+1}, b = 0, 1, 2, 3$ and $0 \leq r_j \leq 1$ for $j = 1, \ldots, J+1$.

(RC2). $0 < \mathrm{E}_\theta\{\mathcal{C}_{[J+1]}^{(a_1,\ldots,a_{J+1};b)}(R_1, \ldots, R_{J+1}; \theta)/\mathcal{C}_{[J+1]}^{(a_1,\ldots,a_{J+1})}(R_1, \ldots, R_{J+1}; \theta)\} < \infty$ for $a_j, b = 0, 1$ and $0 \leq R_j = S_j(T_j) \leq 1$ for $j = 1, \ldots, J+1$.

(RC3). With $g(r_j, r_{J+1}; \theta) = \psi^{-1}\{\psi(r_j; \theta) - \psi(r_{J+1}, \theta); \theta\}$, $g^{(a_1,a_2)}(r_j, r_{J+1}; \theta)$ and $g^{(a_1,a_2;b)}(r_j, r_{J+1}; \theta)$ exist and are continuous and uniformly bounded for $a_1, a_2, b = 0, 1, 2, 3$, $0 \leq r_j, r_{J+1} \leq 1$ for $j = 1, \ldots, J$.

(RC4). $0 < \mathrm{E}_\theta\{\mathcal{C}_{[2]}^{(a_1,a_2;b)}(R_j, R_{J+1};\theta) g^{(0,0;1)}(R_j^*, R_{J+1};\theta)/\mathcal{C}_{[2]}^{(a_1,a_2)}(R_j, R_{J+1};\theta)\} < \infty$ for $a_1, a_2, b = 0, 1$, $0 \le R_j = S_j(T_j) \le 1$ for $j = 1, \ldots, J+1$, and $0 < R_j^* = S_j^*(T_j^*)$ with $T_j^* = T_j \wedge T_{J+1}$ and $S_j^*(t) = P(T_1^* \ge t)$ for $j = 1, \ldots, J$.

It is easy to verify that the Clayton, Gumbel, and Frank copulas all satisfy the conditions (RC1)–(RC4).

We add the following conditions.

(AC1). Assume that there exists $\tilde{G}_n(\cdot)$, an estimator of the survivor function $G(\cdot)$ based on the data in (3.24), satisfying (i) $\tilde{G}_n(\cdot)$ converges uniformly to $G(\cdot)$ over $[0, w^*]$ and (ii) $\sqrt{n}(\tilde{G}_n(\cdot) - G(\cdot)) \xrightarrow{w} \mathcal{G}(\cdot)$ on $[0, w^*]$ with $\mathcal{G}(\cdot)$ a Gaussian process with mean zero and covariance function $\sigma^2(\cdot)$, where $\sigma^2(s_1 \wedge s_2) = cov(\mathcal{G}(s_1), \mathcal{G}(s_2))$.

(AC2). The derivative of $Q(X_i|G_1(\cdot), \ldots, G_K(\cdot);\alpha)$ in (3.25) with respect to α depends on $G_k(\cdot)$ through $G_k(X_i^{(k)})$ for $k = 1, \ldots, K$.

The following lemma establishes the consistency and asymptotic normality of the pseudo-MLE $\hat{\alpha}_n \triangleq \mathrm{argmax}_{\alpha\in\mathcal{A}} L(\alpha; \tilde{G}_1, \ldots, \tilde{G}_K|\text{General-Data})$ with the estimators $\tilde{G}_k(\cdot)$ for $G_k(\cdot)$ with $k = 1, \ldots, K$.

Lemma 2. *Assume conditions (AC1)–(AC2) in addition to the usual regularity conditions for the asymptotics of an MLE (e.g., Chp 4.2 Serfling 1980). The pseudo-MLE $\hat{\alpha}_n$ satisfies, as $n \to \infty$, (i) Consistency: $\hat{\alpha}_n \xrightarrow{a.s.} \alpha$, and (ii) Asymptotic Normality: $\sqrt{n}(\hat{\alpha}_n - \alpha) \xrightarrow{d} N(0, \Pi(\alpha))$ with $\Pi(\alpha)$ the asymptotic variance.*

Proof: Let $A_n(\alpha), B_n(\alpha)$ be respectively

$$\frac{1}{n}\sum_{i=1}^{n} \partial Q(X_i|G_1(\cdot), \ldots, G_K(\cdot);\alpha)/\partial\alpha, \quad \frac{1}{n}\sum_{i=1}^{n} \partial^2 Q(X_i|G_1(\cdot), \ldots, G_K(\cdot);\alpha)/\partial\alpha^2.$$

Step 1. When $G_k(\cdot)$ for $k = 1, \ldots, K$ are known: Following the standard arguments for MLE asymptotics such as those in Serfling (1980), we can establish the consistency and asymptotic normality of the MLE
$\hat{\alpha}_n = \mathrm{argmax}_{\alpha\in\mathcal{A}} \log L(\alpha; G_1(\cdot), \ldots, G_K(\cdot)|\text{General-Data})$ by noting that

$$0 = \frac{1}{n}\partial \log L(\hat{\alpha}; G_1, \ldots, G_K|\text{General-Data})/\partial\alpha = A_n(\alpha) + B_n(\alpha)(\hat{\alpha}_n - \alpha) + \frac{1}{2}C_n(\xi)(\hat{\alpha}_n - \alpha)^2,$$

where ξ is between α and $\hat{\alpha}_n$ and $C_n(\alpha) = \partial B_n(\alpha)/\partial\alpha$, and, as $n \to \infty$,

$$\sqrt{n}A_n(\alpha) \xrightarrow{d} N(0, FI(\alpha)); \quad B_n(\alpha) \to FI^{-1}(\alpha).$$

Here $\Pi(\alpha) = FI^{-1}(\alpha)$ is the limiting variance.

Step 2. When $G_j(\cdot)$ are estimated by $\tilde{G}_{jn}(\cdot)$: Let $A_n^*(\alpha), B_n^*(\alpha)$ be respectively

$$\frac{1}{n}\sum_{i=1}^{n}\partial Q(X_i|\tilde{G}_1(\cdot),\ldots,\tilde{G}_K(\cdot);\alpha)/\partial\alpha, \quad \frac{1}{n}\sum_{i=1}^{n}\partial^2 Q(X_i|\tilde{G}_1(\cdot),\ldots,\tilde{G}_K(\cdot);\alpha)/\partial\alpha^2.$$

By the Taylor expansion, $A_n^*(\alpha) - A_n(\alpha)$ is

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K}\phi_{ki}(\tilde{G}_k(X_i^{(k)}) - G_k(X_i^{(k)})) + \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} o((\tilde{G}_k(X_i^{(k)}) - G_k(X_i^{(k)}))^2 \tag{3.26}$$

with $\phi_{ki} = \partial^2 Q(X_i|G_1,\ldots,G_K;\alpha)/\partial(\alpha, G_k(X_i^{(k)}))$ for $k = 1,\ldots,K$. Thus, $|A_n^*(\alpha) - A_n(\alpha)| \xrightarrow{a.s.} 0$. Similarly, it can be shown that $|B_n^*(\alpha) - B_n(\alpha)| \xrightarrow{a.s.} 0$ and thus $B_n^*(\alpha) \xrightarrow{a.s} -B^*(\alpha)$ as $n \to \infty$. By step 1, we have the consistency of $\hat{\alpha}_n$. This proves that $\hat{\alpha}_n \xrightarrow{a.s.} \alpha$.

In the following, we show that $\sqrt{n}A_n^*(\alpha) \xrightarrow{d} N(0, AV_{A^*}(\alpha))$, which yields $\sqrt{n}(\hat{\alpha}_n - \alpha) \xrightarrow{d} N(0, V^*(\alpha))$ with $AV^*(\alpha) = [-B^*(\alpha)]^{-1}V_{A^*}(\alpha)[-B^*(\alpha)]^{-1}$ following step 1.

Denote $\sqrt{n}\{A_n^*(\alpha) - A_n(\alpha)\} = \sum_{k=1}^{K}\omega_{nk} + \Delta_{\omega_n}$ with the two terms corresponding to those in (3.26). By the strong embedding theorem (Shorack & Wellner 2009), we could construct another probability space such that $\sqrt{n}(\tilde{G}_{kn}(t) - G_k(t)) \xrightarrow{a.s.} Z_k(t)$, where $Z_k(\cdot)$ is a Gaussian process with mean zero. We can then show $\Delta_{\omega_n} \xrightarrow{a.s} 0$. In addition, we can show that, for $k = 1,\ldots,K$, ω_{nk} is asymptotically equivalent to $\frac{1}{n}\sum_{i=1}^{n}\phi_{ki}Z_k(X_i^{(k)})$, which converges in distribution to $N(0, V_{\omega k}^*)$. Furthermore, we can show $\Delta_{\omega_n} \xrightarrow{a.s} 0$. Thus, $\sqrt{n}A_n^*(\alpha) = \sqrt{n}\{A_n^*(\alpha) - A_n(\alpha)\} + \sqrt{n}A_n$ converges to $N(0, V_{A^*}(\alpha))$ as $n \to \infty$.

Proof of Proposition 1: For $j = 1,\ldots,J$, take $S_j^*(\cdot)$ for $j = 1,\ldots,J$ and $S_D(\cdot)$ as G_k for $k = 1,\ldots,K$ in lemma 2. We can use the Kaplan–Meier estimator with Observed-Data$_j$ in (3.2) to estimate $S_j^*(\cdot)$ and $S_D(\cdot)$. Denote the induced estimators by $\tilde{S}_j^*(\cdot)$ and $\tilde{S}_D(\cdot)$. Thus, Condition (AC1) is satisfied with $w^* = v_j$. Moreover, Condition (AC2) is satisfied with $X_i^{(j)} = T_{ji} \wedge D_i \wedge C_i$ for $j = 1,\ldots,J$ and $X_i^{(J+1)} = D_i \wedge C_i$. Thus, by Lemma 2, $\hat{\theta}_n$ is consistent and has asymptotic normality.

Proof of Proposition 2: Given the regularity conditions (RC1)–(RC4), together with the consistency of $\tilde{S}_j^*(\cdot)$, $\tilde{S}_D(\cdot)$, and $\hat{\theta}_n$, we can see that $\hat{S}_{jn}(t) = g(\tilde{S}_j^*(t), \tilde{S}_D(t); \hat{\theta}_n)$ as defined in (3.15) converges uniformly to $g_j(S_{T_j\wedge D}(t), S_D(t); \theta_j) = S_j(t)$ for $t \in [0, v_j^\star]$. Furthermore, by the weak convergence/asymptotic normality of $\tilde{S}_j^*(\cdot)$, $\tilde{S}_D(\cdot)$, and $\hat{\theta}_n$, $\sqrt{n}(\hat{S}_{jn}(t) - S_j(t))$ converges to a mean zero Gaussian process over $(0, v_j]$.

Proof of Proposition 3: Take $S_j(\cdot)$ as G_j for $j = 1,\ldots,J$ in lemma 2. The estimator $\hat{S}_{jn}(\cdot)$ in (3.15) satisfies the required condition (AC1) with $w^* = v_j$. By the regularity conditions (RC1)–(RC4), we can see that $\hat{S}_n(\underline{t}) = \mathcal{C}_{[J]}(\hat{S}_{1n}(t_1),\ldots,\hat{S}_{Jn}(t_J); \hat{\theta}_n)$ as defined in (3.16) converges uniformly to $\mathcal{C}_{[J]}(S_1(t_1),\ldots,S_J(t_J);\theta)$, and thus to $S(\underline{t})$ for $\underline{t} \in [0, v_1^\star] \times \ldots \times [0, v_J^\star]$.

Note that $\hat{S}_n(\underline{t}) = \psi^{-1}(\psi(\hat{S}_{1n}(t_1); \hat{\theta}_n) + \ldots + \psi(\hat{S}_{Kn}(t_K); \hat{\theta}_n); \hat{\theta}_n)$ with $\psi(\cdot)$ monotone. By the arguments similar to the one for the one-dimensional situation, the weak convergence

of $\hat{S}_n(\underline{t})$ yields from with the weak convergence of $\hat{S}_{jn}(\cdot)$ and the asymptotic normality of $\hat{\theta}_n$.

3.5 Simulation Study

Simulation studies were conducted to explore the finite-sample performance of the proposed approach in section 3.3. We started from $J = 1$ as a special case, and focused on the situations with $J = 2$ as in the breast cancer example. The observations from the simulations should apply to general situations with $J \geq 1$.

3.5.1 Data Generation

We simulated a study with n independent units where the primary outcome is the bivariate event times (T_1, T_2). The observations on (T_1, T_2) may be censored by either the terminating event time D or an administrative time C_A, whichever occurs first. That is, the study censoring time $C = D \wedge C_A$. As a preparation for the following simulation study, we started with generating bivariate event-times to mimic semi-competing risk data and verified consistency and robustness. The results are presented in 3.5.2. The thesis focuses more on simulation outcomes from trivariate event-times data as a more general scenario. To imitate potentially informative censoring due to a terminating event, the data were generated as follows:

Step (a). We independently generated the trivariate random variables (v_{1i}, v_{2i}, v_{Di}) for $i = 1, \ldots, n$ from an Archimedean copula model by the `R` package `copula` (Hofert et al. 2017, Yan 2007, Kojadinovic & Yan 2010, Hofert & Mächler 2011).

Step (b). We used the survivor functions of the Weibull distributions $S_j(\cdot)$ and $S_D(\cdot)$, where the scale and shape parameters mimic the corresponding event times and death times in the real example, to form the generated event times and terminating event times $t_{ji} = S_j^{-1}(v_{ji})$ with $j = 1, 2$ and $d_i = S_D^{-1}(v_{Di})$ for $i = 1, \ldots, n$.

Step (c). We generated the independent (administrative) censoring times c_{Ai} independently from (v_{1i}, v_{2i}, v_{Di}) from the exponential distribution with the parameter chosen to give a censoring rate of 25 percent. We then calculated $c_i = d_i \wedge c_{Ai}$ with the indicator $\delta_{Di} = I(d_i \leq c_{Ai})$ and $u_{ji} = t_{ji} \wedge c_i$ with the indicator $\delta_{ji} = I(t_{ji} \leq c_i)$.

Steps (a), (b), and (c) yield a generated observed-data: $\{[(u_{ji}, \delta_{ji}) : j = 1, 2] \bigcup [c_i, \delta_{Di}] : i = 1, \ldots, n\}$.

The data generation process was also applied to $J = 1$ case, where (v_i, v_{Di}) were generated from bivariate copula. We used the `R` functions `claytonCopula`, `gumbelCopula`, and `frankCopula` to generate trivariate variables from the Clayton, Gumbel, and Frank copulas, respectively, to exemplify Archimedean copulas. We considered $n = 500, 1000$, and 2000 to

generate medium to large studies. The value of the parameter θ was determined in each of the Archimedean copula examples according to the chosen Kendall's $\tau = 0.3, 0.6$ and 0.8 to generate weak, to moderate, to strong dependence between T_1, T_2 and D. We used the Kaplan–Meier estimator to obtain $\tilde{S}_j^*(\cdot)$ and $\tilde{S}_D(\cdot)$ in the estimation procedure.

3.5.2 Simulation Outcomes for $J = 1$

Consistency, Efficiency, and Variance Estimation

When $J = 1$ as in the semi-competing risk setting, we verified the consistency and efficiency of the pseudo-MLE θ, as well as the estimator for the survivor function of T, with five hundred simulated datasets for the bivariate Clayton, Gumbel and Frank copula models. We evaluated (i) the robust sandwich variance estimator described in section 3.3.2, (ii) the corresponding inverse Fisher information, and (iii) its bootstrap standard error estimator with resampling size 500 to examine the proposed variance estimation. In addition, we provided two references for comparison: the MLE of θ derived from the likelihood function (3.7) using the true survivor functions $S(\cdot)$ and $S_D(\cdot)$ and the naïve estimates obtained by maximizing (3.7) after substituting the marginal survivor functions by their Kaplan–Meier estimates. In the settings that this chapter focuses on, the MLE is in fact not applicable, and the naïve estimator can be biased because of the informative censoring.

Tables 3.1-3.3 present summaries of the simulation outcomes under the Clayton, Gumbel, and Frank copula model, respectively. The sample means of the pseudo-MLE and MLE estimates are close to the true parameter values, and the sample standard errors of the pseudo-MLE are comparable with their MLE counterparts. This verifies the consistency and efficiency in the semi-competing risk setting. However, the sample means of the naïve estimates are biased, which indicates the need to adjust for informative censoring. The sample means of the robust standard error estimates for the pseudo-MLE are similar to the bootstrap standard error estimates. However, the sample means of the conventional standard error estimates using the inverse Fisher information are rather different from the corresponding sample standard deviations associated with the pseudo-MLE estimates. This indicates the need to use the robust variance estimator with the pseudo-MLE.

Figure 3.1 shows the estimates of the marginal survivor function $S(\cdot)$ under Clayton copula with $\tau = 0.6$, and different sample sizes $n = 500, 1000, 2000$. Each plot contains the curve of the marginal survivor function, and two sets of estimates using proposed approach and the Kaplan-Meier estimator, or the naïve estimates, together with their confidence bands (CB). The true curve is fully covered under CBs of proposed estimates, but not the naïve one. The same patterns were observed for other scenarios with $\tau = 0.6$, and $\tau = 0.8$, and with Gumbel and Frank copula.

Robustness to Model Misspecification

To examine the pseudo-MLE's robustness to misspecification, we generated data under each of the three Archimedean copulas, the Clayton, Gumbel, and Frank copulas, and evaluated the pseudo-MLE and the MLE of the association parameter by the procedures with all the three copulas. In addition, we generated model from a non-Archimedean copula, bivariate Gaussian copula, and evaluated the performance of the proposed estimators.

Tables 3.4-3.6 present the estimates of τ with correctly specified and misspecified Archimedean copula models. Some biases were observed under model misspecification between Clayton and Gumbel copula, but Frank copula provides consistent estimates under model misspecification. This indicates that in real data analysis, one could choose to report results from Frank copula model.

Figure 3.2 shows the estimates of the marginal survivor function under Clayton copula with $\tau = 0.6$, with both correctly specified copula, and misspecified copulas, namely Gumbel and Frank in this case. There are some biases observed with misspecified copula but not significant. Robustness figures for other copulas and $\tau = 0.3, 0.8$ are also provided. As is observed in the τ estimates, it appears that Frank copula is more robust against model misspecification, compared to Clayton and Gumbel.

In addition, simulated data are generated from a non-Archimedean copula, Gaussian copula with $\tau = 0.3, 0.6,$ and 0.8, and estimates are shown in table 3.7 with Clayton, Gumbel, and Frank model specified. Figures 3.19-3.21 present the corresponding marginal survivor function estimates. Biases in the estimates of τ are observed when model is misspecified, but the marginal survivor function estimates are satisfactory, especially for Frank copula, the real curve is fully covered by the CB of marginal estimates.

3.5.3 Simulation Outcomes for $J = 2$

Consistency, Efficiency, and Variance Estimation

We evaluated the pseudo-MLE of the association parameter θ together with the estimator for the survivor function of T_j from section 3.3 with five hundred generated sets of data for the trivariate Clayton, Gumbel and Frank copula models. We evaluated (i) the robust sandwich variance estimator described in section 3.3.2, (ii) the corresponding inverse Fisher information, and (iii) its bootstrap standard error estimator with resampling size 500 to examine the proposed variance estimation. In addition, we provided two references for comparison: the MLE of θ derived from the likelihood function (3.7) using the true survivor functions $S_j(\cdot)$ and $S_D(\cdot)$ and the naïve estimates obtained by maximizing (3.7) after substituting the marginal survivor functions by their Kaplan–Meier estimates. In the settings that this chapter focuses on, the MLE is in fact not applicable, and the naïve estimator can be biased because of the informative censoring.

The sample means of the pseudo-MLE and MLE estimates are close to the true parameter values, especially when the sample size n is large. This verifies the consistency of the pseudo-MLE and the MLE. The sample means of the naïve estimates, on the other hand, are rather different from the true values. As expected, the difference becomes more obvious when the association parameter is larger. The sample standard errors of the pseudo-MLE are larger than but comparable with their MLE counterparts, which indicates that the pseudo-MLE has satisfactory efficiency. In addition, the sample means of the robust standard error estimates for the pseudo-MLE appear similar to the bootstrap standard error estimates, and both are close to the corresponding sample standard deviations associated with the pseudo-MLE estimates; however, the sample means of the conventional standard error estimates using the inverse Fisher information are rather different. This indicates the need to use the robust variance estimator with the pseudo-MLE. Tables 3.8-3.10 present a summary of the simulation outcomes based on the five hundred repetitions under the Clayton, Gumbel, and Frank copula model.

The six plots in figure 3.22 correspond to simulated studies under the trivariate Clayton copula model with the Kendall's $\tau = 0.6$ and different sample sizes: $n = 500, 1000$ and 2000, for $S_1(\cdot)$ and $S_2(\cdot)$. Each plot shows the true curve of the marginal survivor function $S_j(\cdot)$, $j = 1, 2$, and its two sets of estimates with the generated semicompeting-risks data, using the proposed pseudo-MLE or the naïve approach. The two sets of approximate 95% CBs for $S_j(\cdot)$ are also presented in the plots. The true $S_j(\cdot)$ curve is fully covered by the CB associated with the pseudo-MLE in every plot, and it is not within the CB associated with the naïve estimator, which requires the assumption of noninformative censoring and is in fact not valid in the simulation settings. This becomes clearer as the sample size increases. The same patterns are observed in the figures for $\tau = 0.8$ and $\tau = 0.3$ with the Clayton model. The simulation outcomes with the Gumbel and Frank copula models for all three different τ values are in agreement with the ones with the Clayton copula.

Robustness to Model Misspecification

To examine the pseudo-MLE's robustness to misspecification, we generated data under each of the three Archimedean copulas, the Clayton, Gumbel, and Frank copulas, and evaluated the pseudo-MLE and the MLE of the association parameter by the procedures with all the three copulas. Tables 3.11-3.13 summarize the sets of the pseudo-MLE and MLE estimates based on five hundred generated data sets and $\tau = 0.6$. Some biases with either the MLE or the pseudo-MLE occur across different simulated studies under both misspecified copulas. However, the biases of the resulting estimates compared to the true Kendall's τ values do not appear significant. It is especially so when the Frank copula is employed to evaluate the estimators. Similar observations were obtained for $\tau = 0.8$. This indicates that in practice one may choose Frank copula if no other information is available.

Table 3.1: Consistency Study. Estimation of Association Parameter with Simulated Data from Bivariate Clayton Copulas, Based on 500 Repetitions

	$\theta = 6/7(0.857)$			$\theta = 3$			$\theta = 8$		
sample size	($\tau = 0.3$)			($\tau = 0.6$)			($\tau = 0.8$)		
(n)	500	1000	2000	500	1000	2000	500	1000	2000
				Pseudo-MLE					
sm^*	0.82	0.84	0.85	2.90	2.95	2.96	7.65	7.77	7.91
$sse^\dagger$	.153	.098	.064	.264	.191	.137	.605	.385	.295
$sm^\ddagger_{s.se}$	.179	.112	.060	.249	.179	.129	.601	.397	.292
$sm^\ddagger_{c.se}$	.106	.077	.055	.190	.137	.097	.379	.274	.198
$sm^\ddagger_{b.se}$	.219	.124	.066	.277	.195	.137	.597	.418	.295
				MLE from (3.7) Using True Marginals					
sm	0.85	0.86	0.86	3.00	3.00	3.00	8.03	8.02	8.05
sse	.103	.083	.054	.193	.137	.100	.412	.273	.203
$sm_{c.se}$	.109	.077	.055	.198	.139	.098	.405	.286	.203
$sm_{b.se}$	.153	.086	.054	.197	.137	.097	.404	.283	.200
				Naive Estimates*					
sm	0.79	0.80	0.81	2.73	2.77	2.78	7.03	7.13	7.26
sse	.141	.095	.061	.256	.181	.130	.627	.390	.298
$sm_{c.se}$	.110	.080	.057	.187	.134	.095	.356	.257	.186
$sm_{b.se}$	.232	.132	.064	.262	.186	.131	.598	.428	.308

* sm: sample mean of estimates

† sse: sample standard error of estimates

‡ $sm_{r.se}$, $sm_{c.se}$, $sm_{b.se}$: sample mean of standard error estimates by robust (sandwich) variance estimator, by the Fisher information (the conventional, estimator), and by the bootstrap resampling, respectively

$\star$ obtained from maximizing (3.7) using KM estimates for the marginal

Table 3.2: Consistency Study. Estimation of Association Parameter with Simulated Data from Bivariate Gumbel Copulas, Based on 500 Repetitions

	$\theta = 10/7(1.429)$			$\theta = 2.5$			$\theta = 5$		
sample size	($\tau = 0.3$)			($\tau = 0.6$)			($\tau = 0.8$)		
(n)	500	1000	2000	500	1000	2000	500	1000	2000
				Pseudo-MLE					
sm^*	1.42	1.42	1.43	2.50	2.49	2.50	4.92	4.97	4.96
$sse^\dagger$	.073	.044	.033	.133	.102	.067	.283	.205	.144
$sm^\ddagger_{s.se}$	.069	.044	.033	.130	.099	.066	.275	.197	.140
$sm^\ddagger_{c.se}$	.084	.057	.040	.129	.091	.064	.229	.162	.115
$sm^\ddagger_{b.se}$	.069	.048	.034	.135	.094	.067	.276	.193	.138
				MLE from (3.7) Using True Marginals					
sm	1.43	1.43	1.43	2.51	2.50	2.51	5.00	5.00	5.00
sse	.060	.042	.030	.109	.083	.056	.198	.146	.102
$sm_{c.se}$	.083	.053	.033	.138	.097	.069	.243	.171	.121
$sm_{b.se}$	.061	.042	.030	.107	.076	.053	.206	.144	.102
				Naive Estimates*					
sm	1.40	1.40	1.40	2.47	2.46	2.47	4.91	4.96	4.95
sse	.070	.043	.032	.132	.102	.067	.279	.206	.144
$sm_{c.se}$	.089	.053	.033	.143	.100	.071	.242	.171	.121
$sm_{b.se}$	.066	.046	.033	.133	.093	.066	.276	.194	.138

* sm: sample mean of estimates

† sse: sample standard error of estimates

‡ $sm_{r.se}$, $sm_{c.se}$, $sm_{b.se}$: sample mean of standard error estimates by robust (sandwich) variance estimator, by the Fisher information (the conventional, estimator), and by the bootstrap resampling, respectively

$\star$ obtained from maximizing (3.7) using KM estimates for the marginal

Table 3.3: Consistency Study. Estimation of Association Parameter with Simulated Data from Bivariate Frank Copulas, Based on 500 Repetitions

	$\theta = 2.917$			$\theta = 7.930$			$\theta = 18.192$		
sample size	($\tau = 0.3$)			($\tau = 0.6$)			($\tau = 0.8$)		
(n)	500	1000	2000	500	1000	2000	500	1000	2000
				Pseudo-MLE					
sm^*	2.90	2.92	2.92	7.89	7.90	7.93	17.94	18.08	18.14
$sse^\dagger$	.390	.268	.194	.542	.369	.286	.891	.707	.525
$sm^\ddagger_{s.se}$	.375	.260	.189	.511	.353	.264	.933	.670	.483
$sm^\ddagger_{c.se}$	.081	.058	.038	.130	.091	.064	.227	.159	.112
$sm^\ddagger_{b.se}$	.388	.268	.188	.544	.374	.267	.902	.656	.466
				MLE from (3.7) Using True Marginals					
sm	2.94	2.94	2.93	7.96	7.94	7.96	18.20	18.22	18.21
sse	.373	.257	.194	.489	.332	.257	.864	.635	.478
$sm_{c.se}$	.085	.057	.041	.139	.097	.069	.235	.165	.116
$sm_{b.se}$	.369	.256	.181	.491	.341	.244	.852	.609	.430
				Naive Estimates*					
sm	2.77	2.78	2.77	7.56	7.56	7.58	17.03	17.18	17.23
sse	.379	.265	.192	.512	.352	.265	.928	.679	.506
$sm_{c.se}$	.085	.051	.036	.142	.099	.070	.234	.164	.115
$sm_{b.se}$	.377	.262	.184	.517	.357	.255	.940	.667	.473

* sm: sample mean of estimates

† sse: sample standard error of estimates

‡ $sm_{r.se}$, $sm_{c.se}$, $sm_{b.se}$: sample mean of standard error estimates by robust (sandwich) variance estimator, by the Fisher information (the conventional, estimator), and by the bootstrap resampling, respectively

⋆ obtained from maximizing (3.7) using KM estimates for the marginal

Table 3.4: Robustness Study. Estimation of Association Parameter with Simulated Data from Bivariate Archimedean Copulas with $\tau = 0.3$, Based on 500 Repetitions

Sample Size		MLE	Pseudo-MLE	MLE	Pseudo-MLE	MLE	Pseudo-MLE
(n)		by Clayton Copula		by Gumbel Copula		by Frank Copula	
		True Model: Trivariate Clayton Copula					
500	sm[†]	0.30	0.29	0.24	0.23	0.30	0.29
	sse[‡]	.025	.049	.038	.042	.034	.038
1000	sm	0.30	0.29	0.24	0.23	0.30	0.29
	sse	.020	.024	.026	.030	.025	.028
2000	sm	0.30	0.30	0.24	0.23	0.30	0.30
	sse	.013	.016	.018	.020	.015	.017
		True Model: Trivariate Gumbel Copula					
500	sm	0.21	0.19	0.30	0.30	0.30	0.29
	sse	.037	.040	.029	.036	.033	.035
1000	sm	0.21	0.18	0.30	0.30	0.30	0.30
	sse	.027	.030	.020	.022	.023	.023
2000	sm	0.22	0.18	0.30	0.30	0.30	0.30
	sse	.017	.022	.015	.016	.016	.016
		True Model: Trivariate Frank Copula					
500	sm	0.22	0.20	0.25	0.26	0.30	0.30
	sse	.036	.039	.032	.035	.033	.034
1000	sm	0.23	0.19	0.25	0.25	0.30	0.30
	sse	.023	.031	.024	.025	.023	.024
2000	sm	0.23	0.19	0.25	0.26	0.30	0.30
	sse	.018	.025	.018	.018	.017	.017

[†] sm: sample of parameter estimates

Table 3.5: Robustness Study. Estimation of Association Parameter with Simulated Data from Bivariate Archimedean Copulas with $\tau = 0.6$, Based on 500 Repetitions

Sample Size (n)		MLE	Pseudo-MLE	MLE	Pseudo-MLE	MLE	Pseudo-MLE
		by Clayton Copula		by Gumbel Copula		by Frank Copula	
		True Model: Trivariate Clayton Copula					
500	sm[†]	0.60	0.59	0.50	0.52	0.60	0.60
	sse[‡]	.015	.022	.029	.030	.019	.023
1000	sm	0.60	0.60	0.51	0.51	0.60	0.60
	sse	.011	.016	.020	.023	.015	.019
2000	sm	0.60	0.60	0.51	0.52	0.60	0.60
	sse	.008	.011	.014	.016	.010	.012
		True Model: Trivariate Gumbel Copula					
500	sm	0.45	0.43	0.60	0.60	0.60	0.59
	sse	.026	.029	.017	.021	.019	.023
1000	sm	0.45	0.43	0.60	0.60	0.60	0.59
	sse	.019	.023	.013	.016	.014	.017
2000	sm	0.45	0.43	0.60	0.60	0.60	0.59
	sse	.014	.015	.009	.011	.010	.012
		True Model: Trivariate Frank Copula					
500	sm	0.45	0.42	0.53	0.54	0.60	0.60
	sse	.028	.027	.024	.024	.018	.021
1000	sm	0.45	0.42	0.53	0.54	0.60	0.60
	sse	.020	.020	.017	.017	.012	.014
2000	sm	0.45	0.42	0.53	0.54	0.60	0.60
	sse	.014	.014	.012	.012	.010	.011

[†] sm: sample of parameter estimates

Table 3.6: Robustness Study. Estimation of Association Parameter with Simulated Data from Bivariate Archimedean Copulas with $\tau = 0.8$, Based on 500 Repetitions

Sample Size (n)		MLE by Clayton Copula	Pseudo-MLE by Clayton Copula	MLE by Gumbel Copula	Pseudo-MLE by Gumbel Copula	MLE by Frank Copula	Pseudo-MLE by Frank Copula
		True Model: Trivariate Clayton Copula					
500	sm[†]	0.80	0.79	0.70	0.71	0.79	0.79
	sse[‡]	.008	.013	.021	.021	.010	.013
1000	sm	0.80	0.80	0.70	0.71	0.80	0.80
	sse	.005	.008	.014	.015	.007	.009
2000	sm	0.80	0.80	0.70	0.71	0.80	0.80
	sse	.004	.006	.011	.010	.005	.006
		True Model: Trivariate Gumbel Copula					
500	sm	0.68	0.66	0.80	0.80	0.80	0.79
	sse	.019	.024	.008	.012	.010	.013
1000	sm	0.67	0.66	0.80	0.80	0.80	0.79
	sse	.014	.016	.006	.008	.007	.009
2000	sm	0.68	0.66	0.80	0.80	0.80	0.79
	sse	.010	.011	.004	.006	.005	.006
		True Model: Trivariate Frank Copula					
500	sm	0.65	0.64	0.73	0.74	0.80	0.80
	sse	.027	.024	.017	.014	.009	.009
1000	sm	0.65	0.64	0.73	0.74	0.80	0.80
	sse	.018	.017	.010	.011	.006	.007
2000	sm	0.65	0.64	0.73	0.74	0.80	0.80
	sse	.013	.013	.008	.007	.005	.005

[†] sm: sample of parameter estimates

Table 3.7: Robustness Study. Estimation of Association Parameter with Simulated Data from Bivariate Gaussian Copulas with $\tau = 0.3, 0.6,$ and 0.8, Based on 500 Repetitions

Sample Size (n)		MLE by Clayton Copula	Pseudo-MLE by Clayton Copula	MLE by Gumbel Copula	Pseudo-MLE by Gumbel Copula	MLE by Frank Copula	Pseudo-MLE by Frank Copula
				True Model: $\tau = 0.3$			
500	sm[†]	0.22	0.20	0.25	0.25	0.28	0.28
	sse[‡]	.032	.039	.027	.031	.028	.031
1000	sm	0.22	0.19	0.25	0.25	0.28	0.28
	sse	.024	.039	.024	.026	.024	.025
2000	sm	0.22	0.18	0.25	0.25	0.28	0.28
	sse	.017	.039	.016	.016	.016	.017
				True Model: $\tau = 0.6$			
500	sm	0.40	0.38	0.47	0.47	0.50	0.49
	sse	.025	.030	.024	.026	.024	.027
1000	sm	0.40	0.38	0.47	0.47	0.50	0.50
	sse	.018	.020	.017	.019	.017	.019
2000	sm	0.40	0.38	0.47	0.47	0.50	0.50
	sse	.012	.015	.013	.015	.012	.013
				True Model: $\tau = 0.8$			
500	sm	0.48	0.47	0.56	0.56	0.59	0.59
	sse	.023	.030	.021	.024	.020	.025
1000	sm	0.48	0.47	0.56	0.56	0.59	0.59
	sse	.016	.020	.013	.016	.013	.016
2000	sm	0.48	0.47	0.56	0.56	0.59	0.58
	sse	.012	.014	.010	.011	.009	.011

[†] sm: sample of parameter estimates

Table 3.8: Consistency Study. Estimation of Association Parameter with Simulated Data from Trivariate Clayton Copulas, Based on 500 Repetitions

	$\theta = 6/7(0.857)$			$\theta = 3$			$\theta = 8$		
sample size	($\tau = 0.3$)			($\tau = 0.6$)			($\tau = 0.8$)		
(n)	500	1000	2000	500	1000	2000	500	1000	2000
				Pseudo-MLE					
sm^*	0.83	0.85	0.85	2.93	2.95	2.97	7.68	7.83	7.93
$sse^\dagger$	.122	.070	.052	.239	.166	.120	.502	.362	.250
$sm^\ddagger_{s.se}$	.124	.074	.053	.232	.169	.121	.504	.368	.258
$sm^\ddagger_{c.se}$	.074	.052	.036	.132	.094	.067	.269	.193	.139
$sm^\ddagger_{b.se}$	.117	.073	.052	.234	.165	.117	.501	.365	.259
				MLE from (3.7) Using True Marginals					
sm	0.86	0.86	0.86	2.99	3.00	3.00	7.99	7.99	8.01
sse	.078	.054	.038	.145	.100	.067	.279	.202	.139
$sm_{c.se}$	.077	.055	.039	.137	.097	.069	.280	.198	.140
$sm_{b.se}$	.083	.055	.039	.137	.097	.069	.280	.198	.140
				Naive Estimates*					
sm	0.81	0.83	0.83	2.64	2.66	2.69	6.49	6.60	6.68
sse	.114	.067	.050	.213	.152	.107	.492	.362	.259
$sm_{c.se}$	.083	.058	.041	.130	.092	.066	.237	.170	.121
$sm_{b.se}$	.120	.071	.050	.214	.151	.108	.480	.355	.257

* sm: sample mean of estimates

† sse: sample standard error of estimates

‡ $sm_{r.se}$, $sm_{c.se}$, $sm_{b.se}$: sample mean of standard error estimates by robust (sandwich) variance estimator, by the Fisher information (the conventional, estimator), and by the bootstrap resampling, respectively

$\star$ obtained from maximizing (3.7) using KM estimates for the marginal survivor functions

Table 3.9: Consistency Study. Estimation of Association Parameter with Simulated Data from Trivariate Gumbel Copulas, Based on 500 Repetitions

	$\theta = 10/7(1.429)$			$\theta = 2.5$			$\theta = 5$		
sample size	($\tau = 0.3$)			($\tau = 0.6$)			($\tau = 0.8$)		
(n)	500	1000	2000	500	1000	2000	500	1000	2000
				Pseudo-MLE					
sm^*	1.40	1.42	1.43	2.39	2.42	2.48	4.91	4.93	4.95
$sse^\dagger$	.058	.036	.022	.131	.081	.057	.233	.169	.118
$sm^\ddagger_{s.se}$	.057	.035	.022	.130	.081	.058	.235	.167	.118
$sm^\ddagger_{c.se}$	.040	.023	.014	.086	.056	.039	.154	.102	.073
$sm^\ddagger_{b.se}$	.057	.033	.020	.130	.082	.058	.230	.163	.116
				MLE from (3.7) Using True Marginals					
sm	1.43	1.43	1.43	2.49	2.51	2.50	4.99	5.00	5.00
sse	.029	.021	.015	.090	.056	.039	.156	.110	.072
$sm_{c.se}$	.030	.021	.014	.089	.056	.039	.154	.104	.073
$sm_{b.se}$	.032	.022	.014	.090	.055	.039	.153	.104	.073
				Naive Estimates*					
sm	1.37	1.39	1.40	2.37	2.39	2.40	4.72	4.76	4.79
sse	.042	.026	.018	.107	.073	.052	.204	.160	.115
$sm_{c.se}$	.031	.022	.014	.079	.053	.038	.132	.099	.070
$sm_{b.se}$	.046	.028	.018	.104	.074	.053	.198	.160	.114

* sm: sample mean of estimates

† sse: sample standard error of estimates

‡ $sm_{r.se}$, $sm_{c.se}$, $sm_{b.se}$: sample mean of standard error estimates by robust (sandwich) variance estimator, by the Fisher information (the conventional, estimator), and by the bootstrap resampling, respectively

$\star$ obtained from maximizing (3.7) using KM estimates for the marginal survivor functions

Table 3.10: Consistency Study. Estimation of Association Parameter with Simulated Data from Trivariate Frank Copulas, Based on 500 Repetitions

	$\theta = 2.917$			$\theta = 7.930$			$\theta = 18.192$		
sample size	($\tau = 0.3$)			($\tau = 0.6$)			($\tau = 0.8$)		
(n)	500	1000	2000	500	1000	2000	500	1000	2000
				Pseudo-MLE					
sm^*	2.88	2.90	2.91	7.61	7.65	7.67	17.85	18.01	18.04
$sse^\dagger$	.324	.186	.147	.431	.309	.209	.963	.531	.343
$sm^\ddagger_{s.se}$	.322	.184	.147	.431	.312	.210	.964	.527	.346
$sm^\ddagger_{c.se}$	.180	.142	.109	.312	.249	.176	.694	.424	.300
$sm^\ddagger_{b.se}$	.318	.181	.146	.426	.339	.215	.965	.510	.357
				MLE from (3.7) Using True Marginals					
sm	2.91	2.91	2.92	7.95	7.94	7.94	18.20	18.20	18.17
sse	.204	.158	.112	.385	.260	.180	.613	.453	.288
$sm_{c.se}$	.202	.157	.111	.363	.248	.176	.611	.422	.299
$sm_{b.se}$	.206	.158	.111	.367	.247	.175	.614	.432	.305
				Naive Estimates*					
sm	2.81	2.83	2.84	7.30	7.35	7.36	16.18	16.25	16.28
sse	.315	.182	.144	.344	.274	.190	.703	.544	.377
$sm_{c.se}$	.209	.160	.118	.303	.239	.169	.479	.391	.277
$sm_{b.se}$	.320	.183	.144	.345	.269	.190	.699	.531	.375

* sm: sample mean of estimates

† sse: sample standard error of estimates

‡ $sm_{r.se}$, $sm_{c.se}$, $sm_{b.se}$: sample mean of standard error estimates by robust (sandwich) variance estimator, by the Fisher information (the conventional, estimator), and by the bootstrap resampling, respectively

$\star$ obtained from maximizing (3.7) using KM estimates for the marginal survivor functions

Table 3.11: Robustness Study. Estimation of Association Parameter with Simulated Data from Trivariate Archimedean Copulas with $\tau = 0.3$, Based on 500 Repetitions

Sample Size		MLE	Pseudo-MLE	MLE	Pseudo-MLE	MLE	Pseudo-MLE
(n)		by Clayton Copula		by Gumbel Copula		by Frank Copula	
		True Model: Trivariate Clayton Copula					
500	sm[†]	0.30	0.29	0.26	0.25	0.31	0.30
	sse[‡]	.019	.046	.027	.029	.025	.028
1000	sm	0.30	0.30	0.26	0.25	0.31	0.30
	sse	.013	.017	.018	.021	.017	.020
2000	sm	0.30	0.30	0.26	0.25	0.31	0.30
	sse	.009	.013	.013	.015	.012	.014
		True Model: Trivariate Gumbel Copula					
500	sm	0.24	0.22	0.30	0.30	0.29	0.28
	sse	.031	.033	.014	.016	.021	.026
1000	sm	0.24	0.23	0.30	0.30	0.30	0.29
	sse	.023	.027	.011	.014	.017	.021
2000	sm	0.24	0.23	0.30	0.30	0.30	0.30
	sse	.018	.022	.008	.011	.013	.017
		True Model: Trivariate Frank Copula					
500	sm	0.23	0.20	0.27	0.26	0.30	0.29
	sse	.032	.034	.021	.024	.021	.027
1000	sm	0.23	0.21	0.27	0.27	0.30	0.29
	sse	.027	.031	.017	.020	.017	.024
2000	sm	0.23	0.21	0.27	0.27	0.30	0.30
	sse	.020	.022	.012	.014	.012	.015

[†] sm: sample of parameter estimates

Table 3.12: Robustness Study. Estimation of Association Parameter with Simulated Data from Trivariate Archimedean Copulas with $\tau = 0.6$, Based on 500 Repetitions

Sample Size (n)		MLE	Pseudo-MLE	MLE	Pseudo-MLE	MLE	Pseudo-MLE
		by Clayton Copula		by Gumbel Copula		by Frank Copula	
		True Model: Trivariate Clayton Copula					
500	sm[†]	0.60	0.59	0.51	0.53	0.60	0.56
	sse[‡]	.012	.020	.021	.024	.015	.211
1000	sm	0.60	0.59	0.52	0.54	0.60	0.60
	sse	.008	.014	.016	.017	.010	.014
2000	sm	0.60	0.60	0.52	0.54	0.60	0.60
	sse	.005	.010	.011	.012	.007	.010
		True Model: Trivariate Gumbel Copula					
500	sm	0.45	0.42	0.60	0.58	0.60	0.57
	sse	.021	.025	.012	.019	.013	.077
1000	sm	0.45	0.41	0.60	0.59	0.60	0.58
	sse	.015	.019	.009	.014	.010	.015
2000	sm	0.45	0.41	0.60	0.59	0.60	0.58
	sse	.010	.012	.006	.010	.007	.010
		True Model: Trivariate Frank Copula					
500	sm	0.44	0.40	0.53	0.53	0.60	0.59
	sse	.023	.026	.016	.020	.013	.017
1000	sm	0.44	0.40	0.53	0.53	0.60	0.59
	sse	.016	.018	.013	.015	.010	.012
2000	sm	0.44	0.40	0.53	0.53	0.60	0.59
	sse	.011	.011	.009	.009	.007	.008

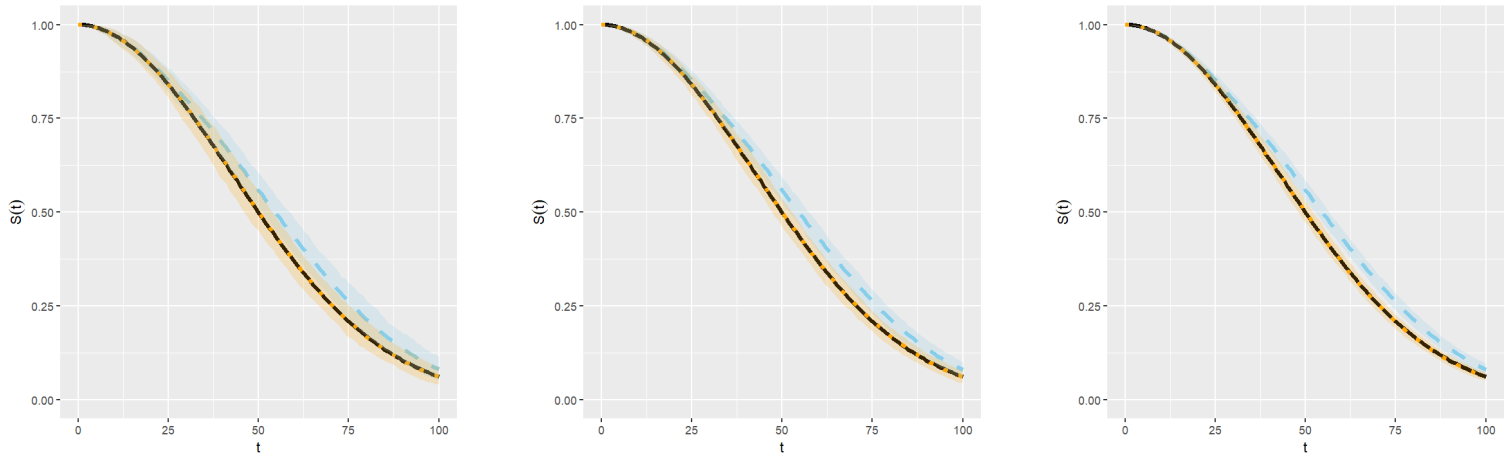
[†] sm: sample of parameter estimates

[‡] sse: sample standard errors of parameter estimates

Table 3.13: Robustness Study. Estimation of Association Parameter with Simulated Data from Trivariate Archimedean Copulas with $\tau = 0.8$, Based on 500 Repetitions

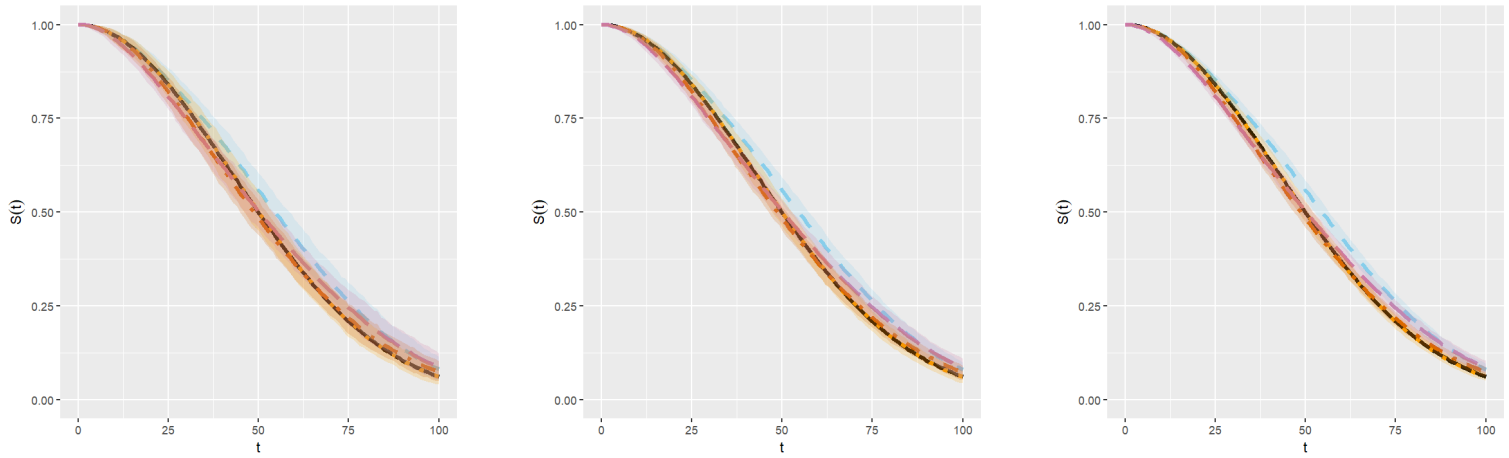
Sample Size (n)		MLE by Clayton Copula	Pseudo-MLE by Clayton Copula	MLE by Gumbel Copula	Pseudo-MLE by Gumbel Copula	MLE by Frank Copula	Pseudo-MLE by Frank Copula
		True Model: Trivariate Clayton Copula					
500	sm[†]	0.80	0.79	0.70	0.71	0.80	0.77
	sse[‡]	.008	.014	.016	.021	.010	.018
1000	sm	0.80	0.79	0.71	0.73	0.80	0.80
	sse	.004	.008	.012	.012	.006	.008
2000	sm	0.80	0.80	0.71	0.73	0.80	0.80
	sse	.003	.005	.008	.008	.004	.006
		True Model: Trivariate Gumbel Copula					
500	sm	0.67	0.67	0.80	0.80	0.80	0.78
	sse	.017	.019	.007	.010	.008	.011
1000	sm	0.67	0.66	0.80	0.80	0.80	0.79
	sse	.011	.013	.004	.007	.005	.008
2000	sm	0.67	0.66	0.80	0.80	0.80	0.79
	sse	.008	.009	.003	.005	.004	.006
		True Model: Trivariate Frank Copula					
500	sm	0.67	0.66	0.80	0.80	0.80	0.80
	sse	.013	.018	.005	.008	.007	.010
1000	sm	0.67	0.66	0.80	0.80	0.80	0.79
	sse	.008	.009	.003	.005	.004	.006
2000	sm	0.64	0.63	0.80	0.80	0.80	0.80
	sse	.013	.010	0.002	0.004	.003	.003

[‡] sse: sample standard errors of parameter estimates



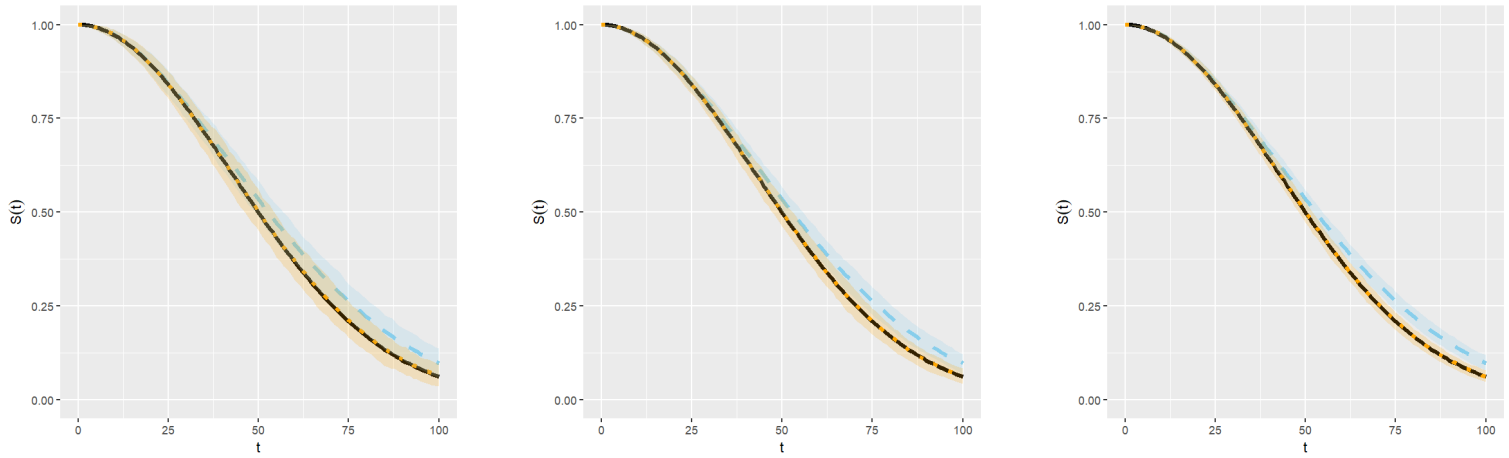
(a) n=500 (b) n=1000 (c) n=2000

Figure 3.1: Consistency Study. Estimates of Marginal Survivor Function $S(\cdot)$ by Proposed Approach and Kaplan-Meier Estimator Using Simulated Data with Different Sizes under Bivariate Clayton Copula: $\tau = 0.6$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted.



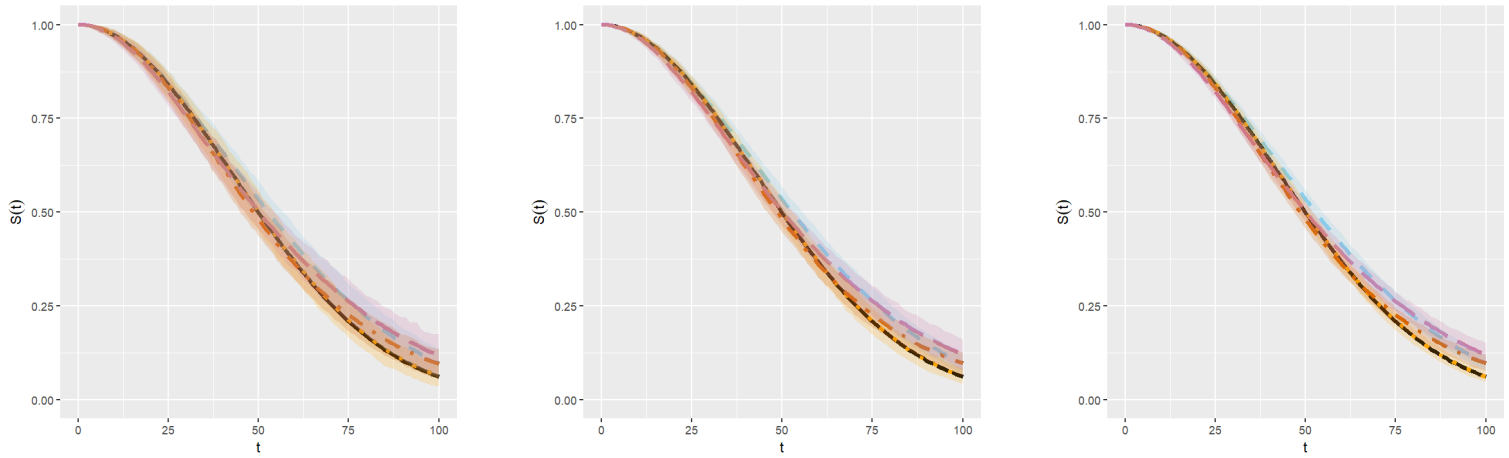
(a) n=500 (b) n=1000 (c) n=2000

Figure 3.2: Robustness Study. Estimates of Marginal Survivor Function $S(\cdot)$ by Proposed Approach Using Simulated Data with Different Sizes under Different Bivariate Copula. True: Clayton with $\tau = 0.6$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted. Gumbel: reddish purple longdash. Frank: vermillion dotdash.



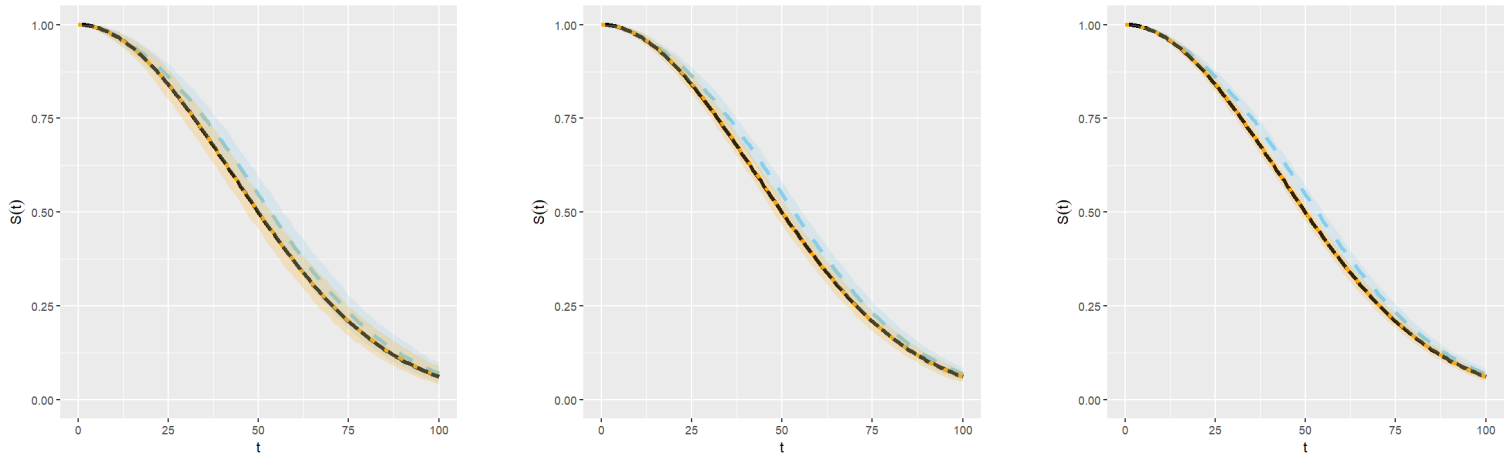
(a) n=500 (b) n=1000 (c) n=2000

Figure 3.3: Consistency Study. Estimates of Marginal Survivor Function $S(\cdot)$ by Proposed Approach and Kaplan-Meier Estimator Using Simulated Data with Different Sizes under Bivariate Clayton Copula: $\tau = 0.3$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted.



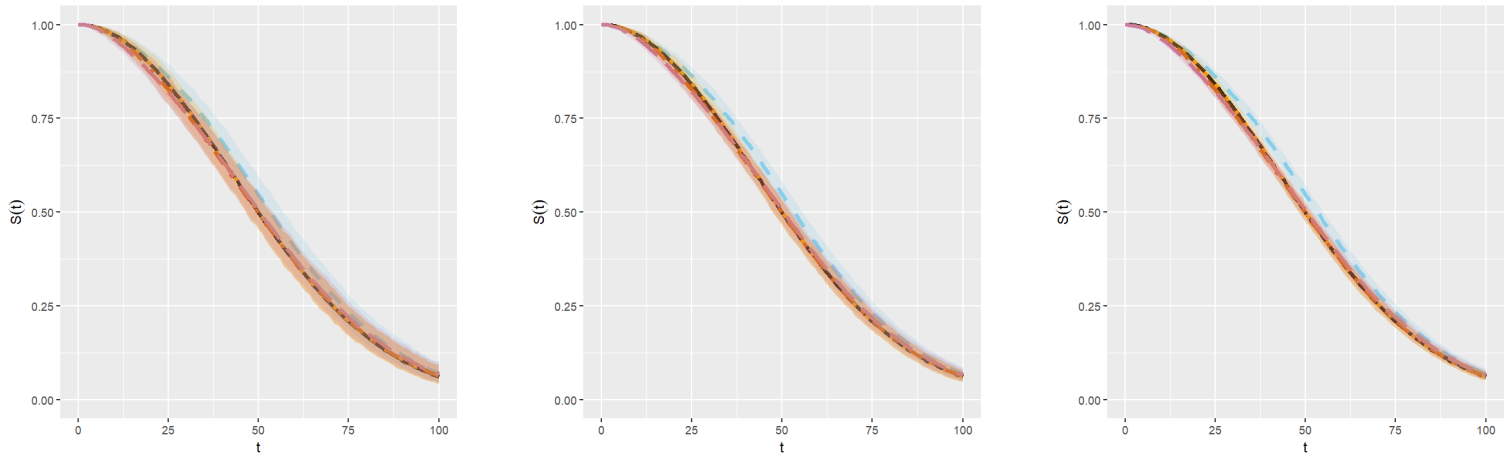
(a) n=500 (b) n=1000 (c) n=2000

Figure 3.4: Robustness Study. Estimates of Marginal Survivor Function $S(\cdot)$ by Proposed Approach Using Simulated Data with Different Sizes under Different Bivariate Copula. True: Clayton with $\tau = 0.3$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted. Gumbel: reddish purple longdash. Frank: vermillion dotdash.



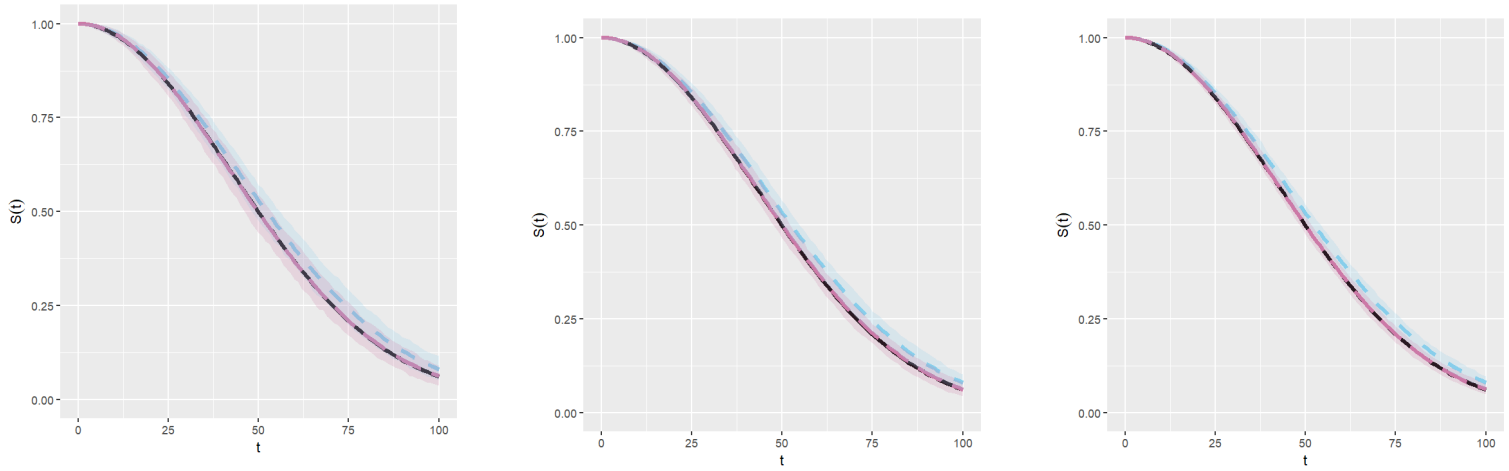
(a) n=500 (b) n=1000 (c) n=2000

Figure 3.5: Consistency Study. Estimates of Marginal Survivor Function $S(\cdot)$ by Proposed Approach and Kaplan-Meier Estimator Using Simulated Data with Different Sizes under Bivariate Clayton Copula: $\tau = 0.8$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted.



(a) n=500 (b) n=1000 (c) n=2000

Figure 3.6: Robustness Study. Estimates of Marginal Survivor Function $S(\cdot)$ by Proposed Approach Using Simulated Data with Different Sizes under Different Bivariate Copula. True: Clayton with $\tau = 0.8$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted. Gumbel: reddish purple longdash. Frank: vermillion dotdash.



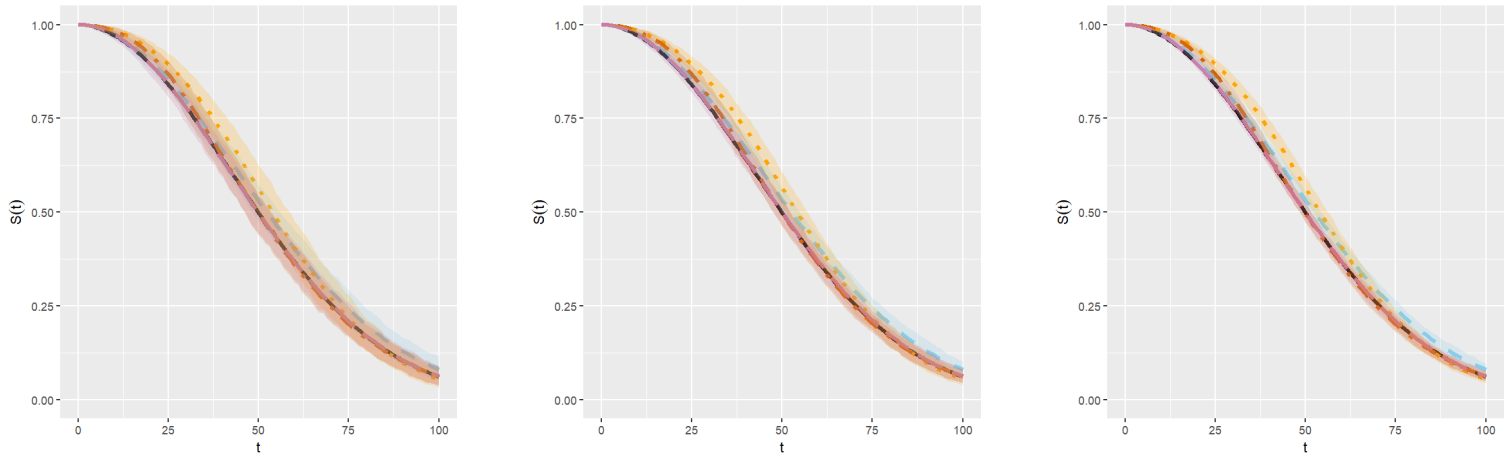
(a) n=500

(b) n=1000

(c) n=2000

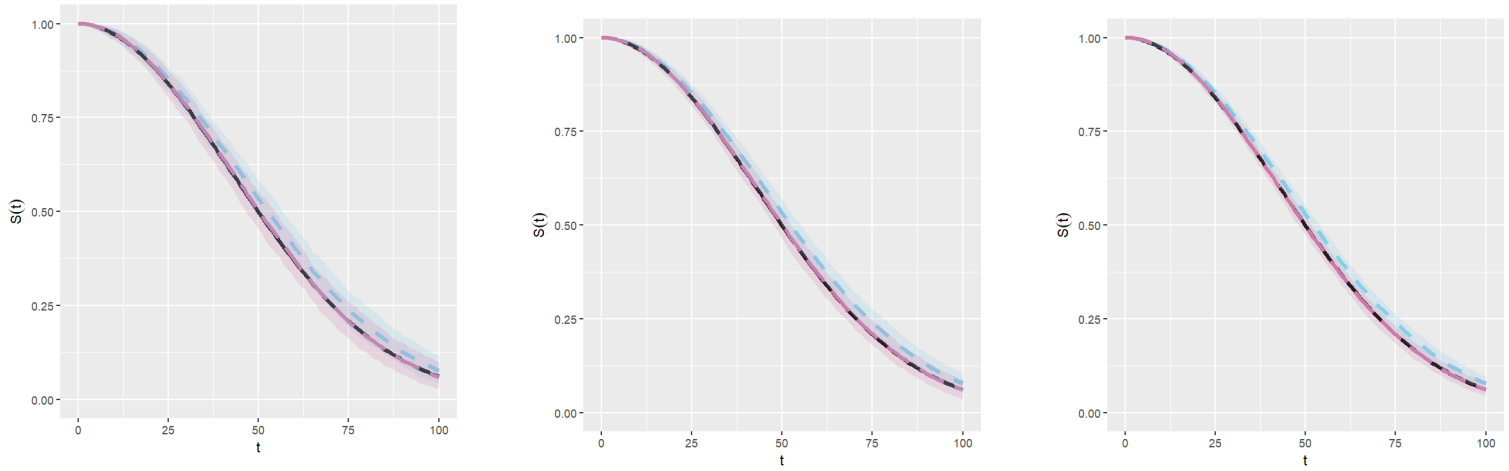
l

Figure 3.7: Consistency Study. Estimates of Marginal Survivor Function $S(\cdot)$ by Proposed Approach and Kaplan-Meier Estimator Using Simulated Data with Different Sizes under Bivariate Gumbel Copula: $\tau = 0.6$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Gumbel: reddish purple longdash.



(a) n=500 (b) n=1000 (c) n=2000

Figure 3.8: Robustness Study. Estimates of Marginal Survivor Function $S(\cdot)$ by Proposed Approach Using Simulated Data with Different Sizes under Different Bivariate Copula. True: Gumbel with $\tau = 0.6$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted. Gumbel: reddish purple longdash. Frank: vermillion dotdash.



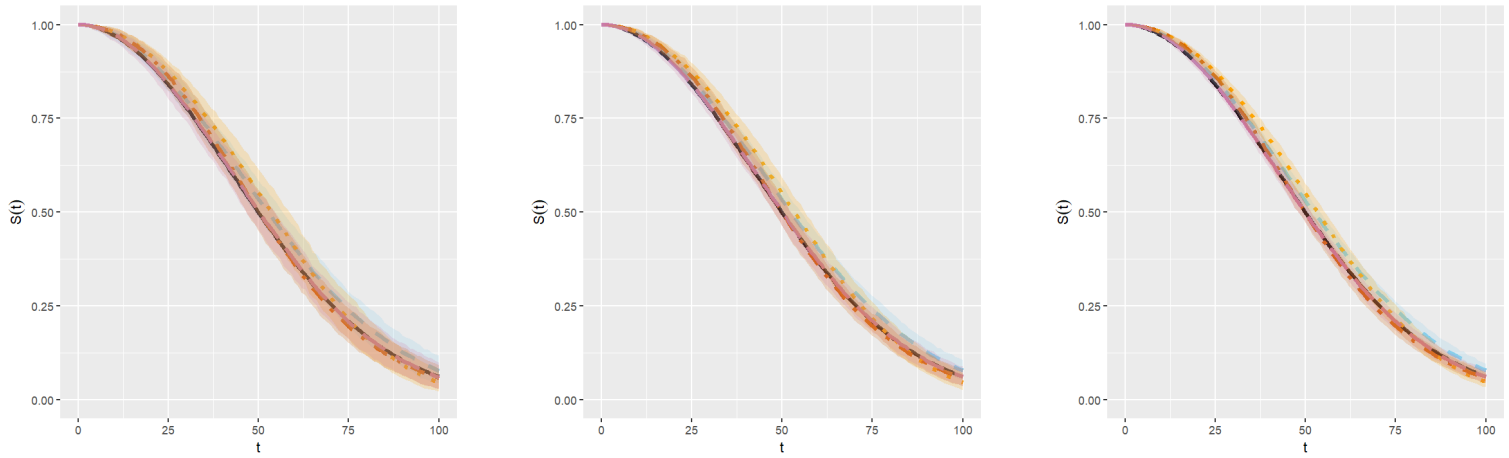
(a) n=500

(b) n=1000

(c) n=2000

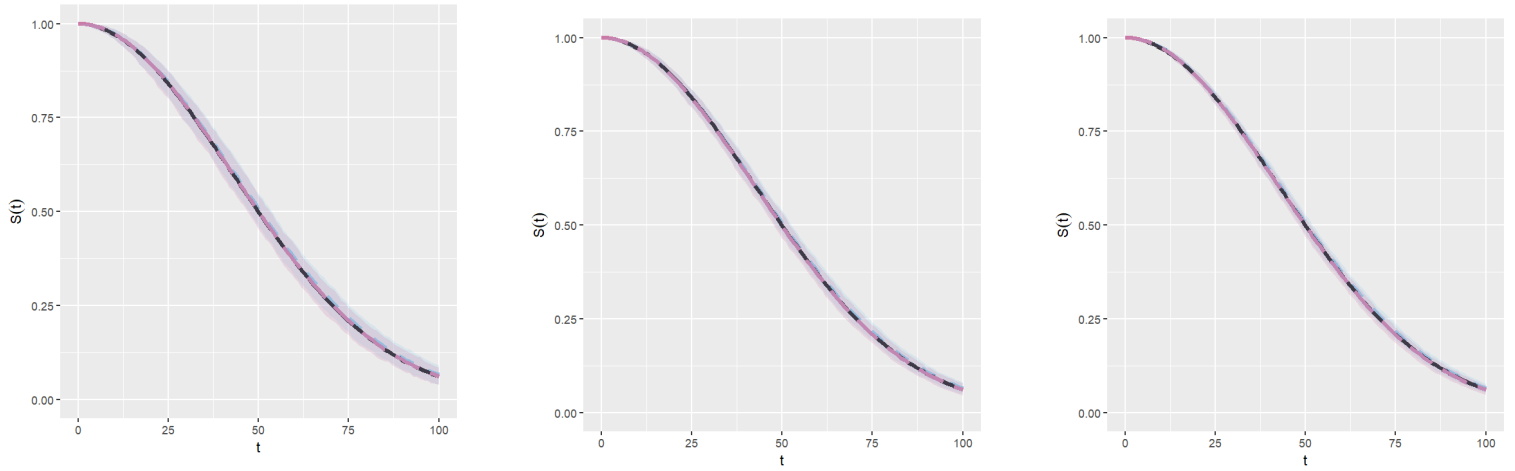
l

Figure 3.9: Consistency Study. Estimates of Marginal Survivor Function $S(\cdot)$ by Proposed Approach and Kaplan-Meier Estimator Using Simulated Data with Different Sizes under Bivariate Gumbel Copula: $\tau = 0.3$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Gumbel: reddish purple longdash.



(a) n=500 (b) n=1000 (c) n=2000

Figure 3.10: Robustness Study. Estimates of Marginal Survivor Function $S(\cdot)$ by Proposed Approach Using Simulated Data with Different Sizes under Different Bivariate Copula. True: Gumbel with $\tau = 0.3$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted. Gumbel: reddish purple longdash. Frank: vermillion dotdash.



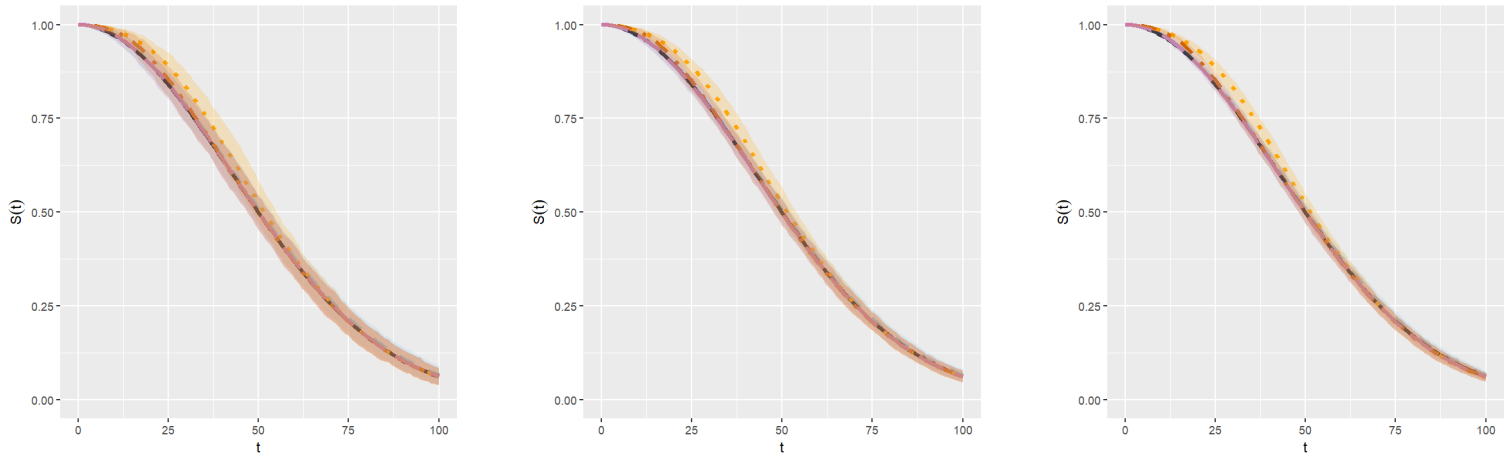
(a) n=500

(b) n=1000

(c) n=2000

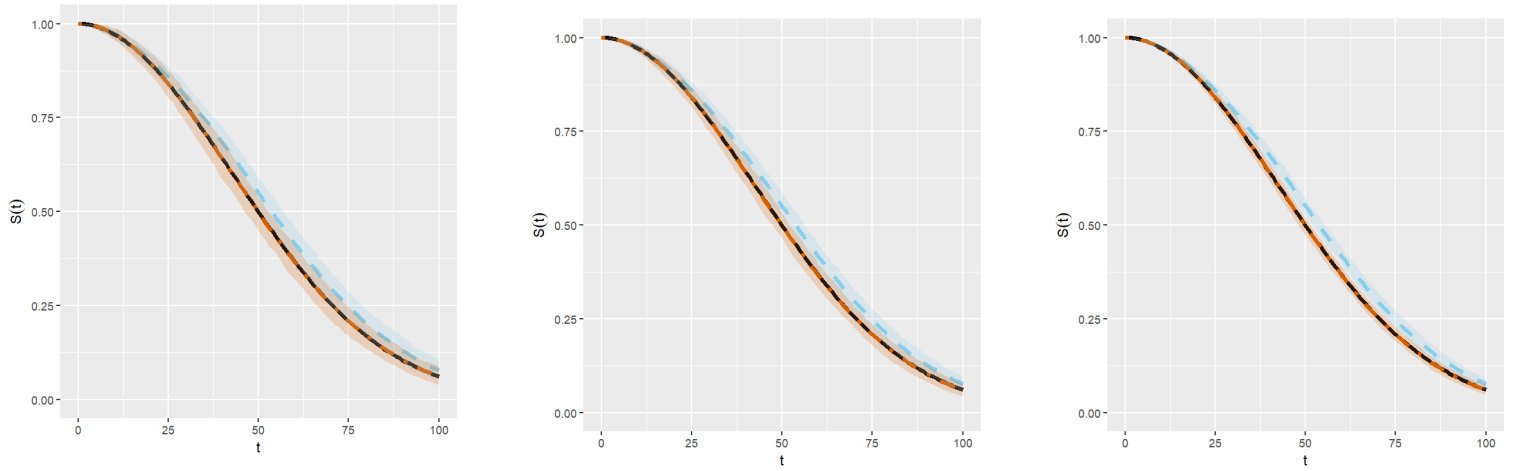
l

Figure 3.11: Consistency Study. Estimates of Marginal Survivor Function $S(\cdot)$ by Proposed Approach and Kaplan-Meier Estimator Using Simulated Data with Different Sizes under Bivariate Gumbel Copula: $\tau = 0.8$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Gumbel: reddish purple longdash.



(a) n=500 (b) n=1000 (c) n=2000

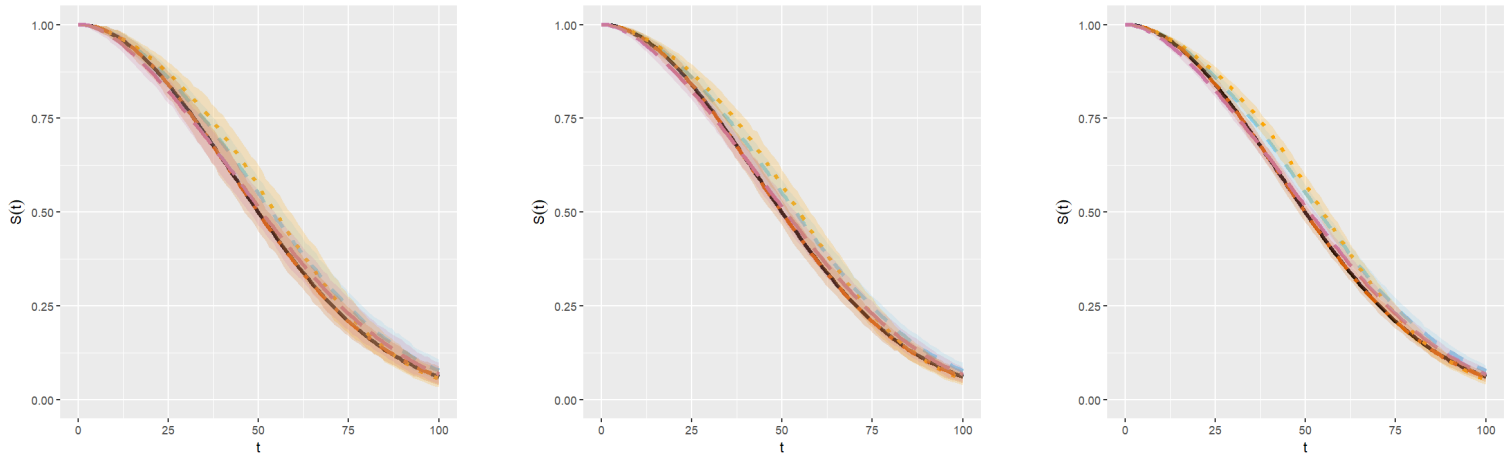
Figure 3.12: Robustness Study. Estimates of Marginal Survivor Function $S(\cdot)$ by Proposed Approach Using Simulated Data with Different Sizes under Different Bivariate Copula. True: Gumbel with $\tau = 0.8$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted. Gumbel: reddish purple longdash. Frank: vermillion dotdash.



(a) n=500 (b) n=1000 (c) n=2000

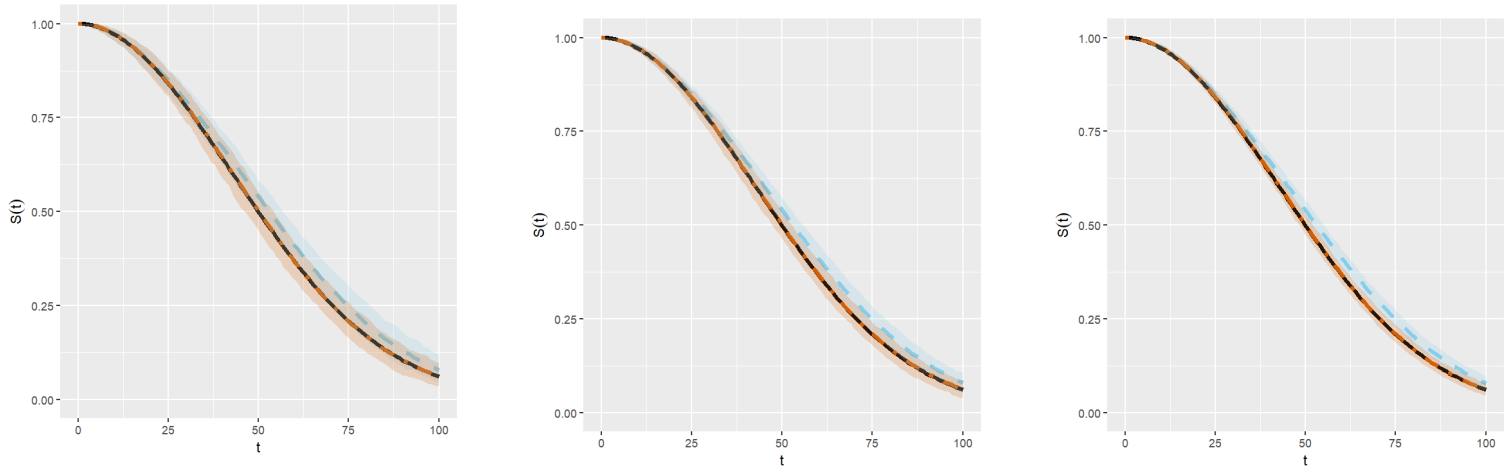
l

Figure 3.13: Consistency Study. Estimates of Marginal Survivor Function $S(\cdot)$ by Proposed Approach and Kaplan-Meier Estimator Using Simulated Data with Different Sizes under Bivariate Frank Copula: $\tau = 0.6$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Frank: vermillion dotdash.



(a) n=500 (b) n=1000 (c) n=2000

Figure 3.14: Robustness Study. Estimates of Marginal Survivor Function $S(\cdot)$ by Proposed Approach Using Simulated Data with Different Sizes under Different Bivariate Copula. True: Frank with $\tau = 0.6$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted. Gumbel: reddish purple longdash. Frank: vermillion dotdash.



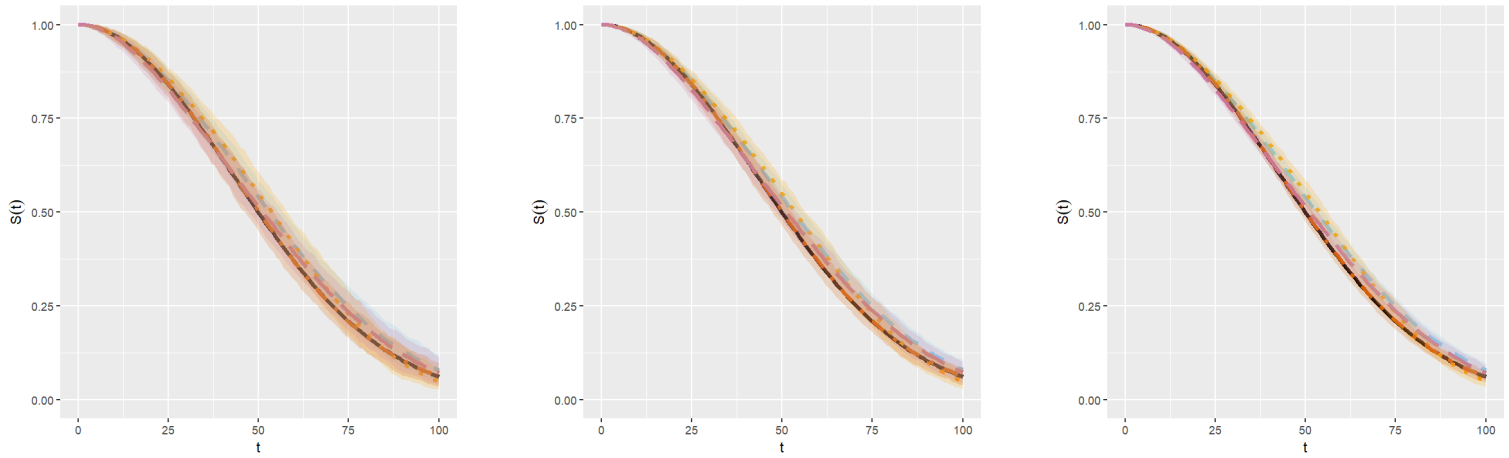
(a) n=500

(b) n=1000

(c) n=2000

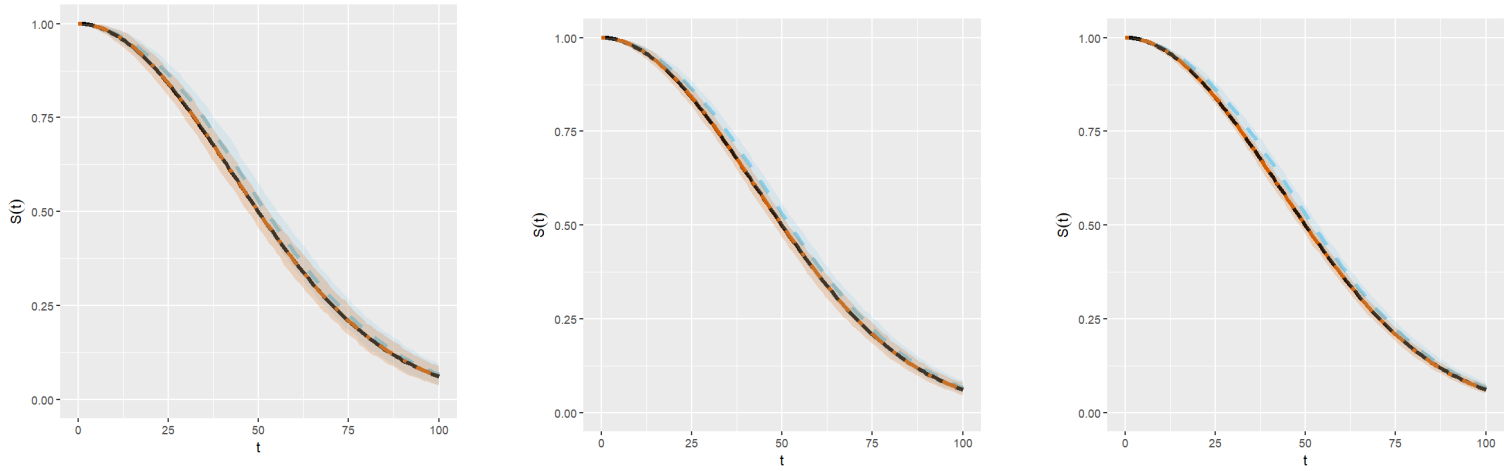
l

Figure 3.15: Consistency Study. Estimates of Marginal Survivor Function $S(\cdot)$ by Proposed Approach and Kaplan-Meier Estimator Using Simulated Data with Different Sizes under Bivariate Frank Copula: $\tau = 0.3$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Frank: vermillion dotdash.



(a) n=500 (b) n=1000 (c) n=2000

Figure 3.16: Robustness Study. Estimates of Marginal Survivor Function $S(\cdot)$ by Proposed Approach Using Simulated Data with Different Sizes under Different Bivariate Copula. True: Frank with $\tau = 0.3$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted. Gumbel: reddish purple longdash. Frank: vermillion dotdash.



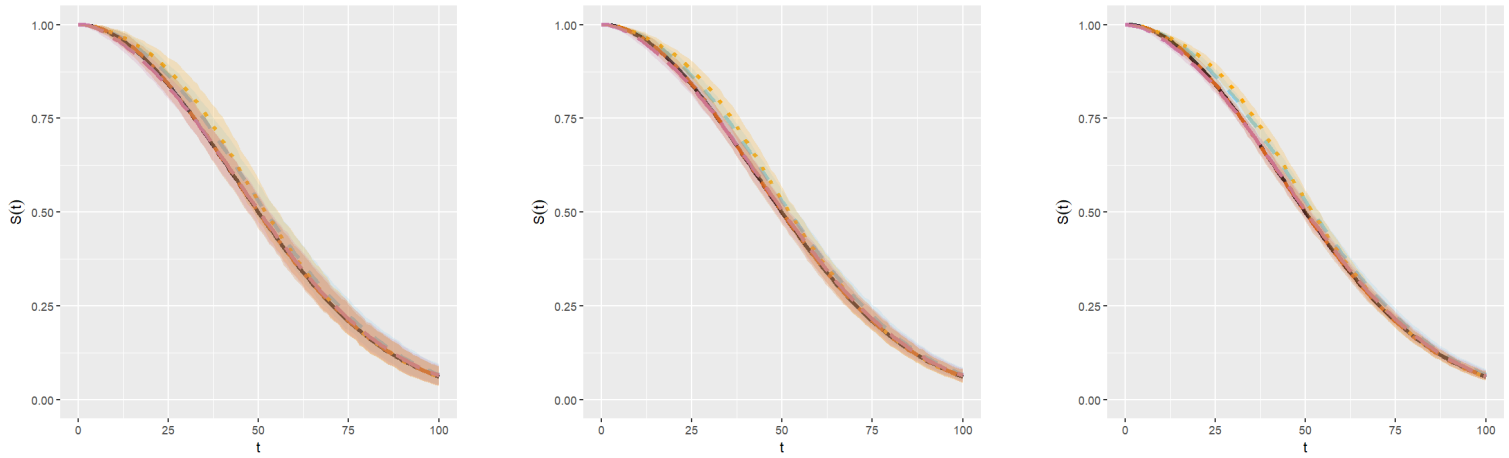
(a) n=500

(b) n=1000

(c) n=2000

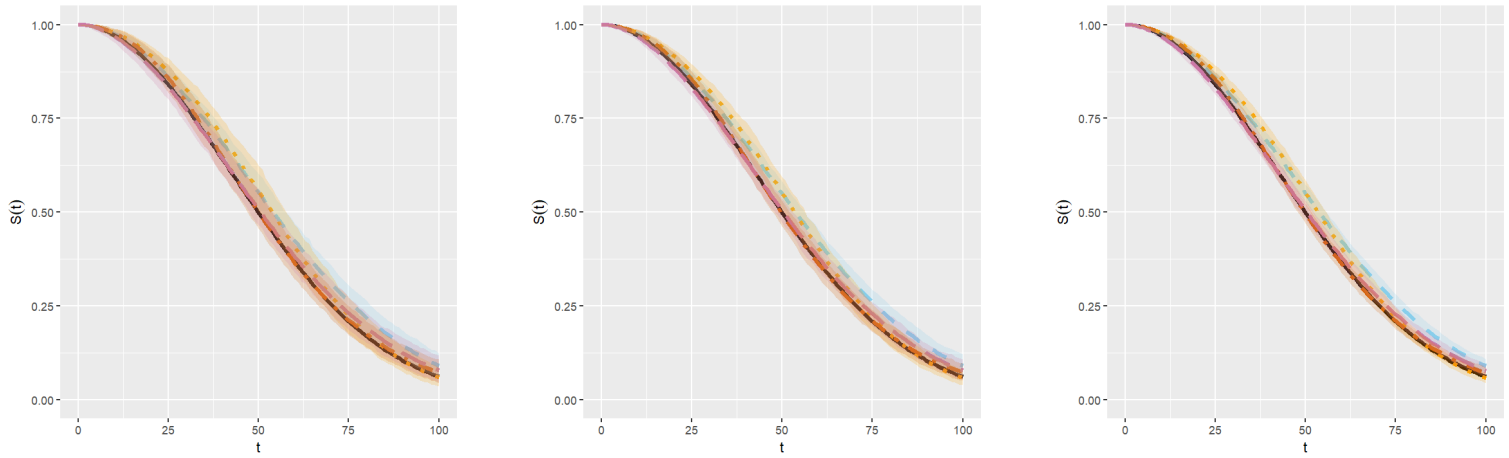
l

Figure 3.17: Consistency Study. Estimates of Marginal Survivor Function $S(\cdot)$ by Proposed Approach and Kaplan-Meier Estimator Using Simulated Data with Different Sizes under Bivariate Frank Copula: $\tau = 0.8$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Frank: vermillion dotdash.



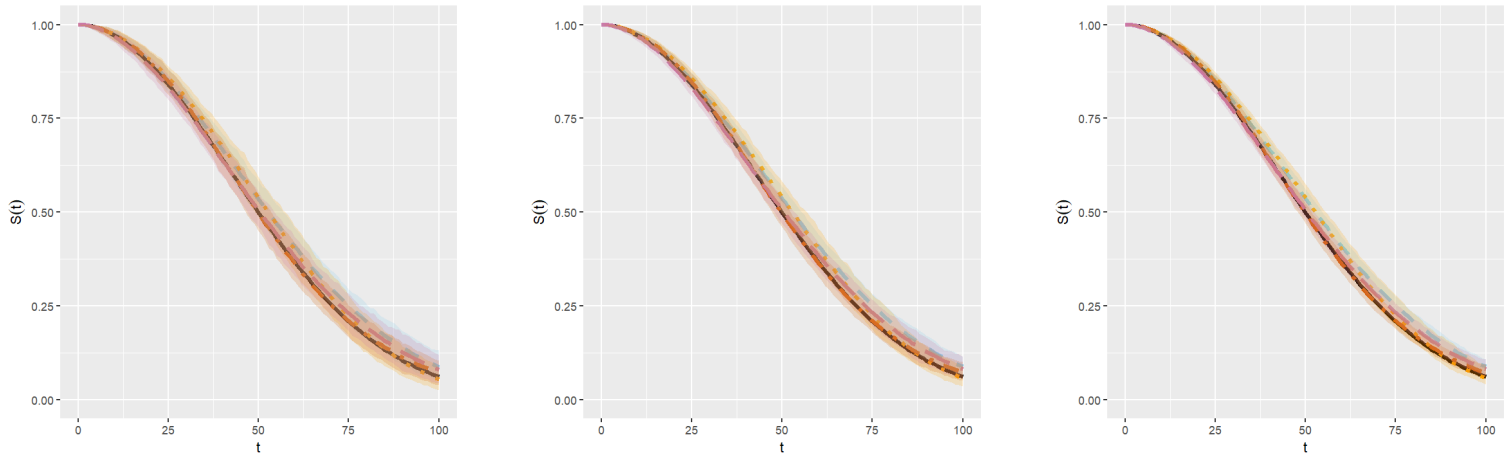
(a) n=500 (b) n=1000 (c) n=2000

Figure 3.18: Robustness Study. Estimates of Marginal Survivor Function $S(\cdot)$ by Proposed Approach Using Simulated Data with Different Sizes under Different Bivariate Copula. True: Frank with $\tau = 0.8$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted. Gumbel: reddish purple longdash. Frank: vermillion dotdash.



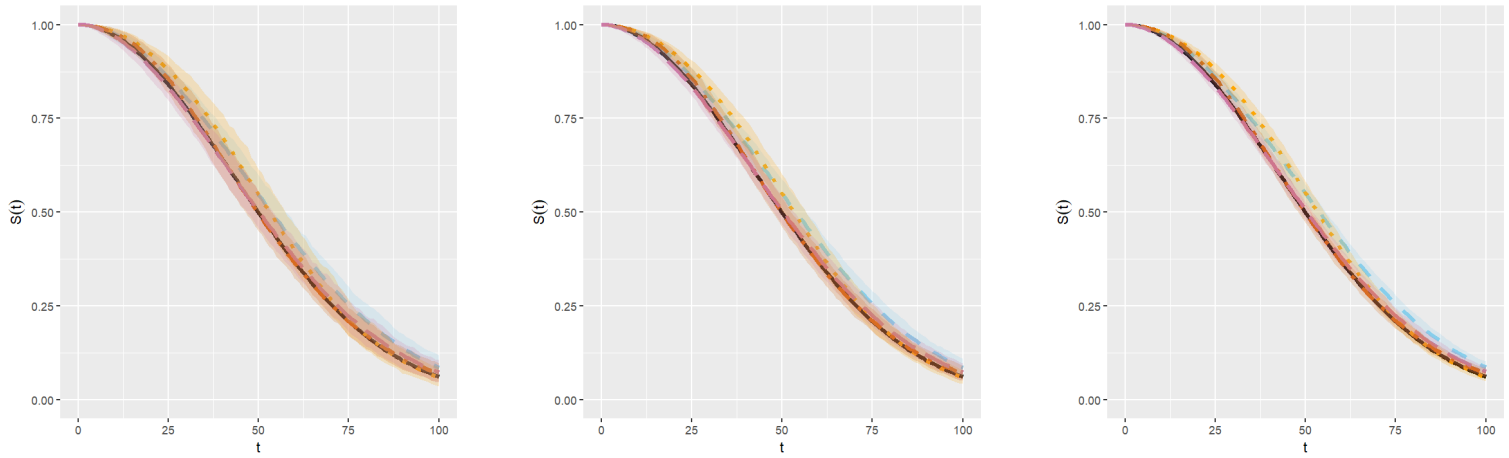
(a) n=500 (b) n=1000 (c) n=2000

Figure 3.19: Robustness Study. Estimates of Marginal Survivor Function $S(\cdot)$ by Proposed Approach Using Simulated Data with Different Sizes under Different Bivariate Copula. True: Gaussian with $\tau = 0.6$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted. Gumbel: reddish purple longdash. Frank: vermillion dotdash.



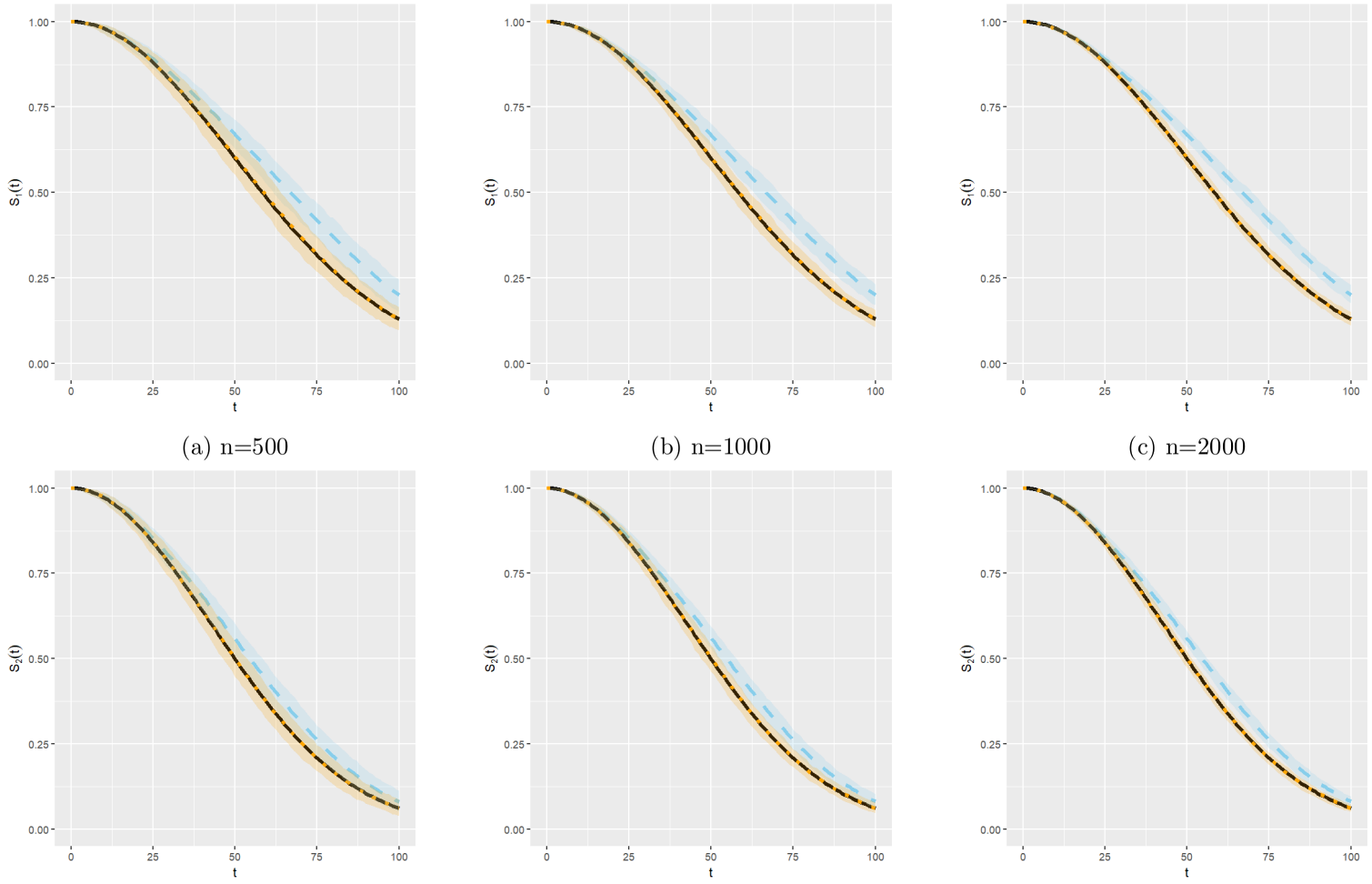
(a) n=500 (b) n=1000 (c) n=2000

Figure 3.20: Robustness Study. Estimates of Marginal Survivor Function $S(\cdot)$ by Proposed Approach Using Simulated Data with Different Sizes under Different Bivariate Copula. True: Gaussian with $\tau = 0.3$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted. Gumbel: reddish purple longdash. Frank: vermillion dotdash.



(a) n=500 (b) n=1000 (c) n=2000

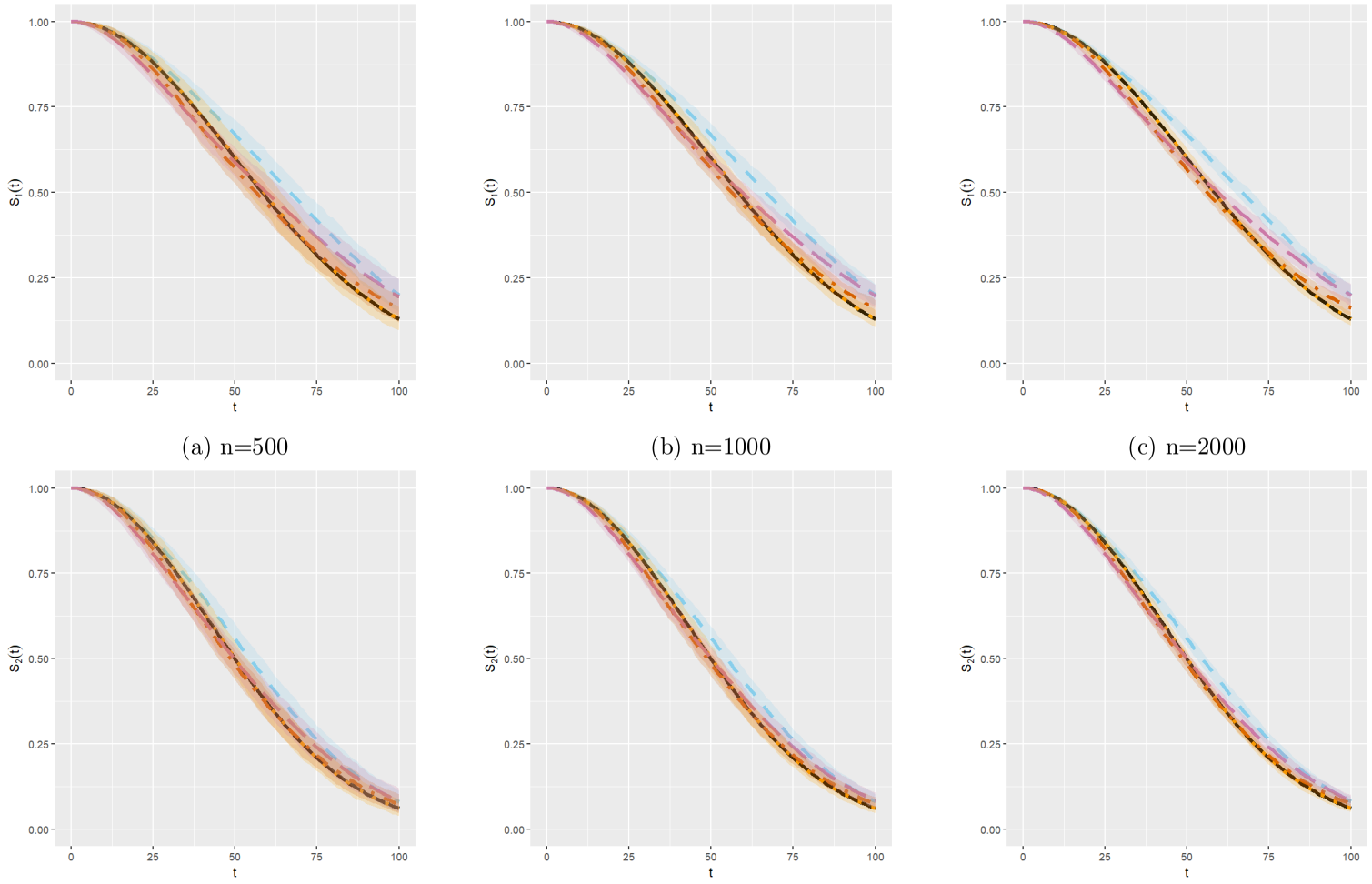
Figure 3.21: Robustness Study. Estimates of Marginal Survivor Function $S(\cdot)$ by Proposed Approach Using Simulated Data with Different Sizes under Different Bivariate Copula. True: Gaussian with $\tau = 0.8$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted. Gumbel: reddish purple longdash. Frank: vermillion dotdash.



(a) n=500 (b) n=1000 (c) n=2000

(d) n=500 (e) n=1000 (f) n=2000

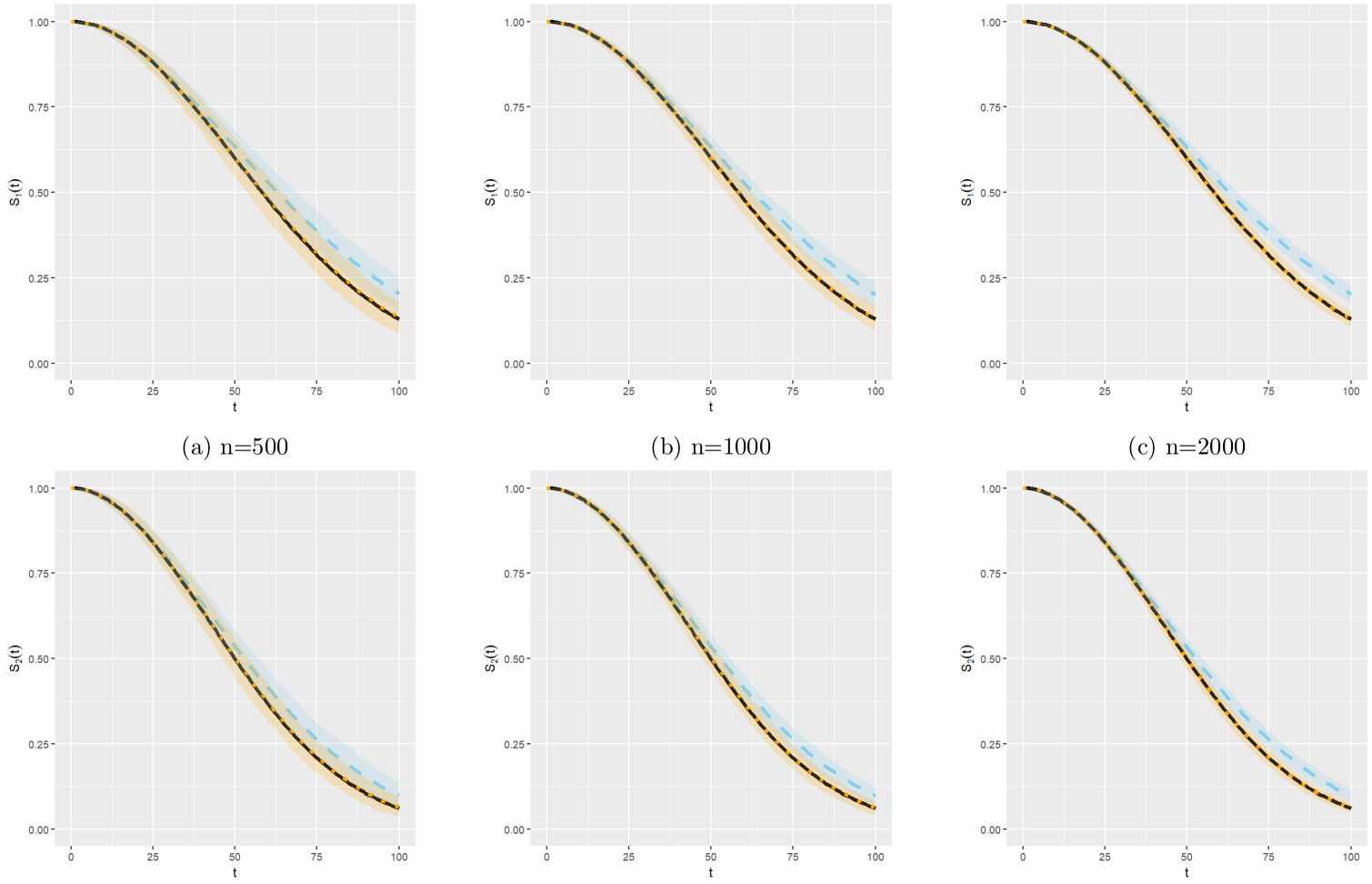
Figure 3.22: Consistency Study. Estimates of Marginal Survivor Function $S_1(\cdot)$ and $S_2(\cdot)$ by Proposed Approach and Kaplan-Meier Estimator Using Simulated Data with Different Sizes under Trivariate Clayton Copula: $\tau = 0.6$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted.



(a) n=500 (b) n=1000 (c) n=2000

(d) n=500 (e) n=1000 (f) n=2000

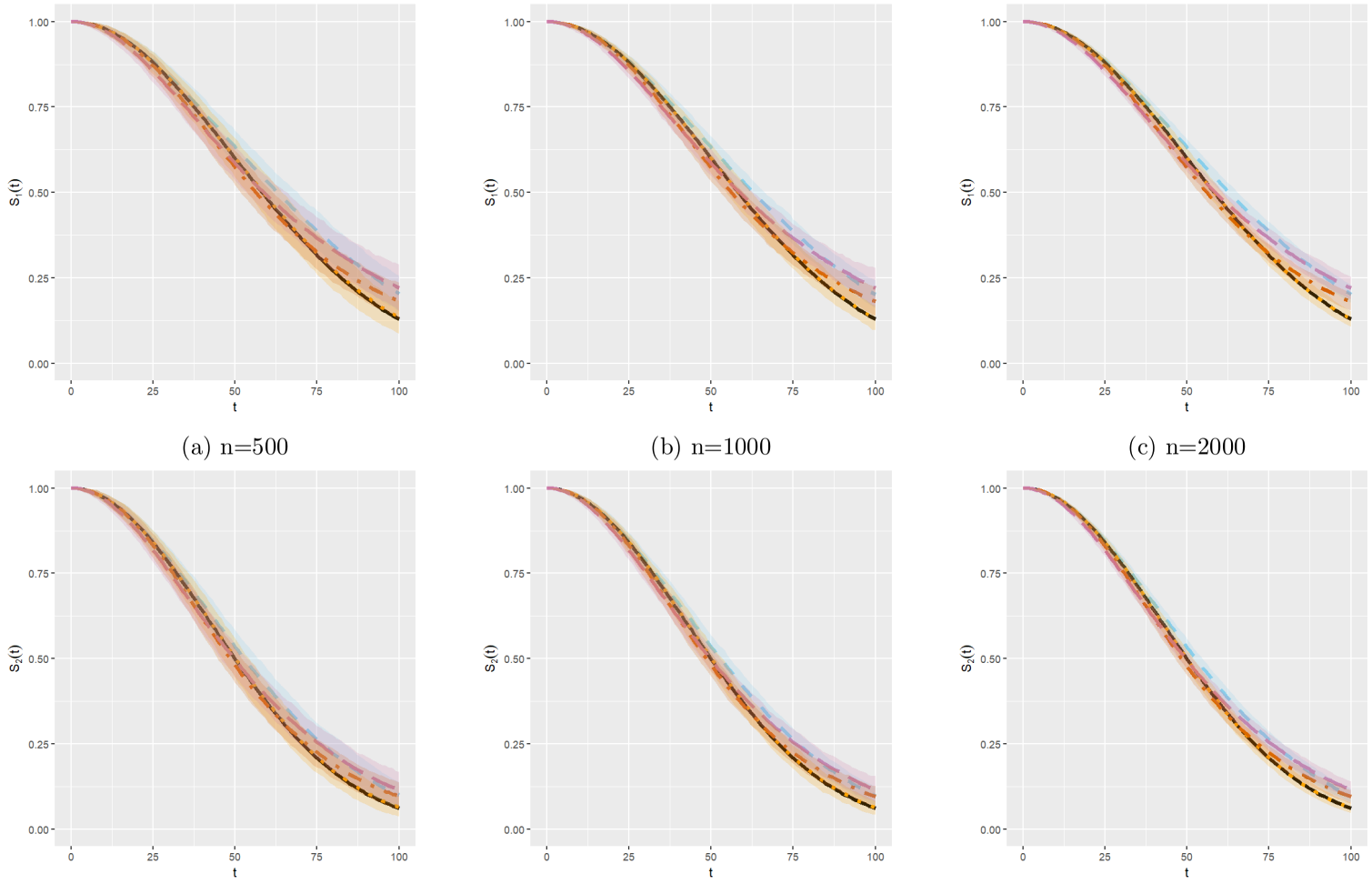
Figure 3.23: Robustness Study. Estimates of Marginal Survivor Function $S_1(\cdot)$ and $S_2(\cdot)$ by Proposed Approach Using Simulated Data with Different Sizes under Different Trivariate Copula. True: Clayton with $\tau = 0.6$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted. Gumbel: reddish purple longdash. Frank: vermillion dotdash.



(a) n=500 (b) n=1000 (c) n=2000

(d) n=500 (e) n=1000 (f) n=2000

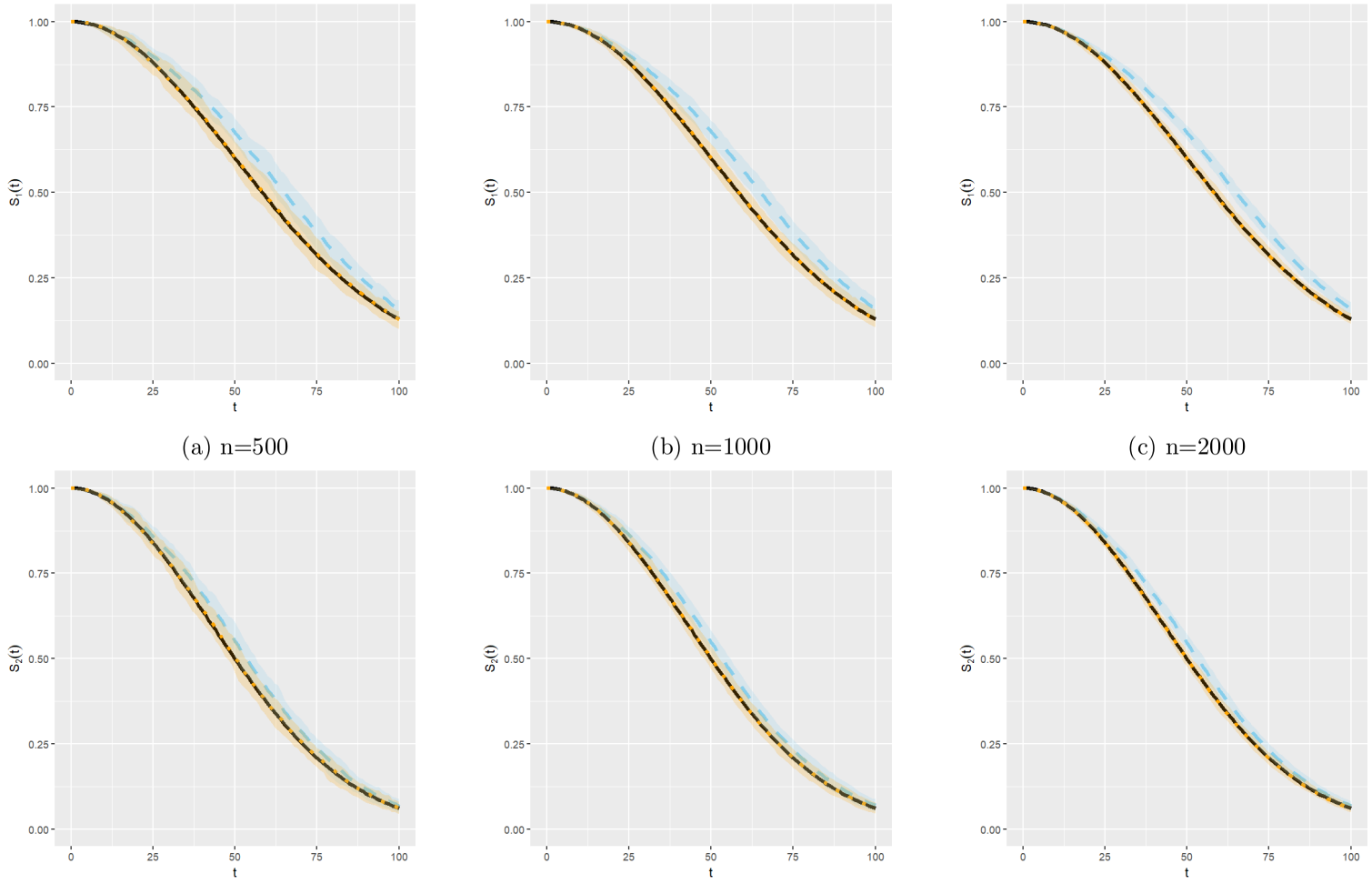
Figure 3.24: Consistency Study. Estimates of Marginal Survivor Function $S_1(\cdot)$ and $S_2(\cdot)$ by Proposed Approach and Kaplan-Meier Estimator Using Simulated Data with Different Sizes under Trivariate Clayton Copula: $\tau = 0.3$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted.



(a) n=500 (b) n=1000 (c) n=2000

(d) n=500 (e) n=1000 (f) n=2000

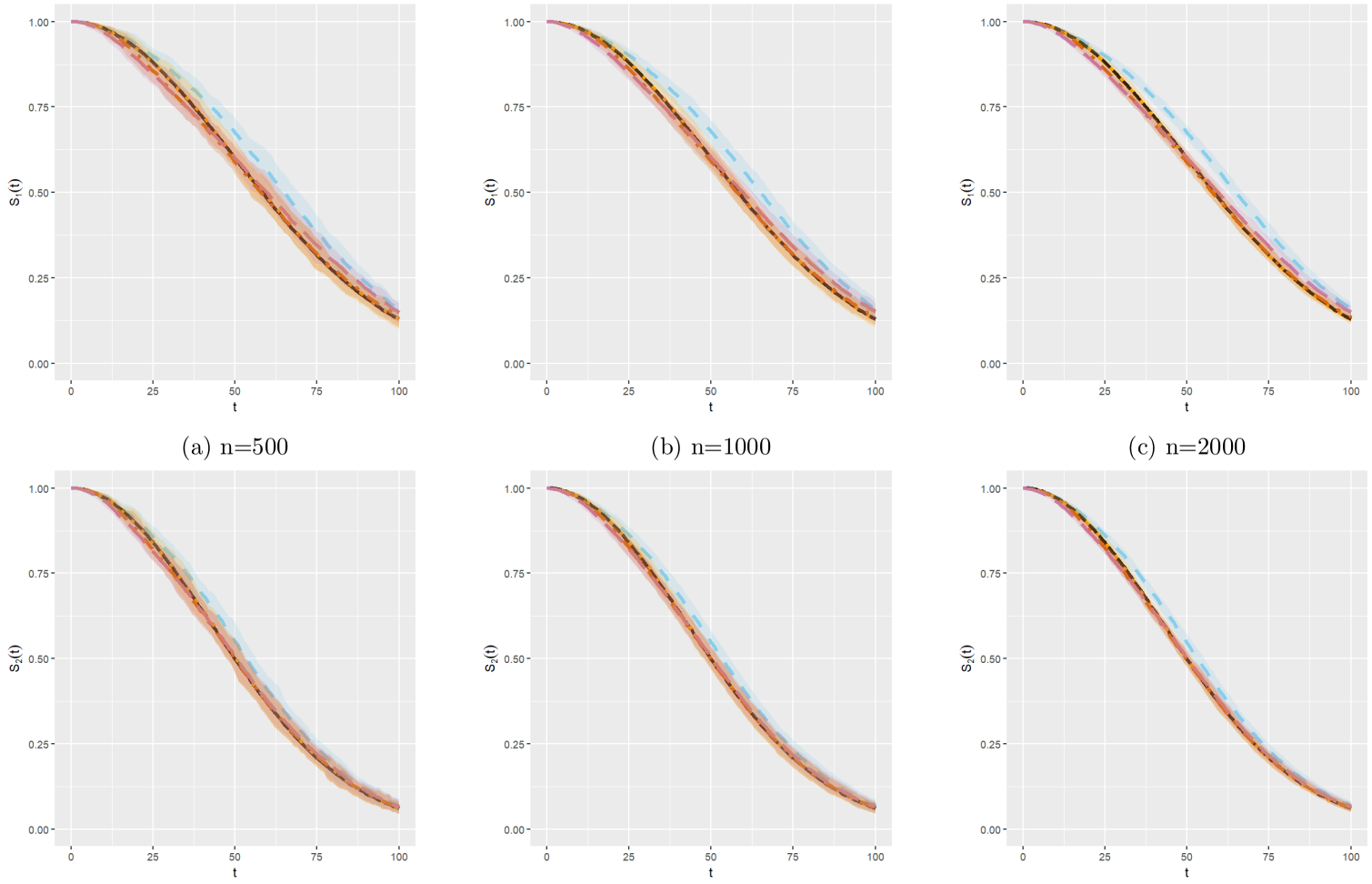
Figure 3.25: Robustness Study. Estimates of Marginal Survivor Function $S_1(\cdot)$ and $S_2(\cdot)$ by Proposed Approach Using Simulated Data with Different Sizes under Different Trivariate Copula. True: Clayton with $\tau = 0.3$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted. Gumbel: reddish purple longdash. Frank: vermillion dotdash.



(a) n=500 (b) n=1000 (c) n=2000

(d) n=500 (e) n=1000 (f) n=2000

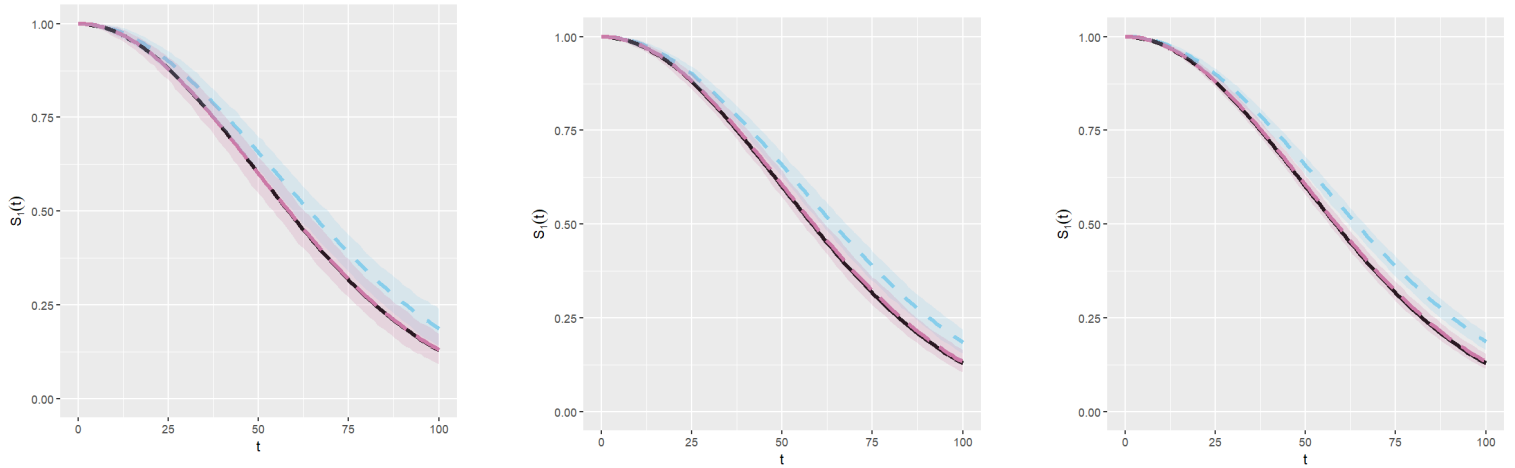
Figure 3.26: Consistency Study. Estimates of Marginal Survivor Function $S_1(\cdot)$ and $S_2(\cdot)$ by Proposed Approach and Kaplan-Meier Estimator Using Simulated Data with Different Sizes under Trivariate Clayton Copula: $\tau = 0.8$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted.



(a) n=500 (b) n=1000 (c) n=2000

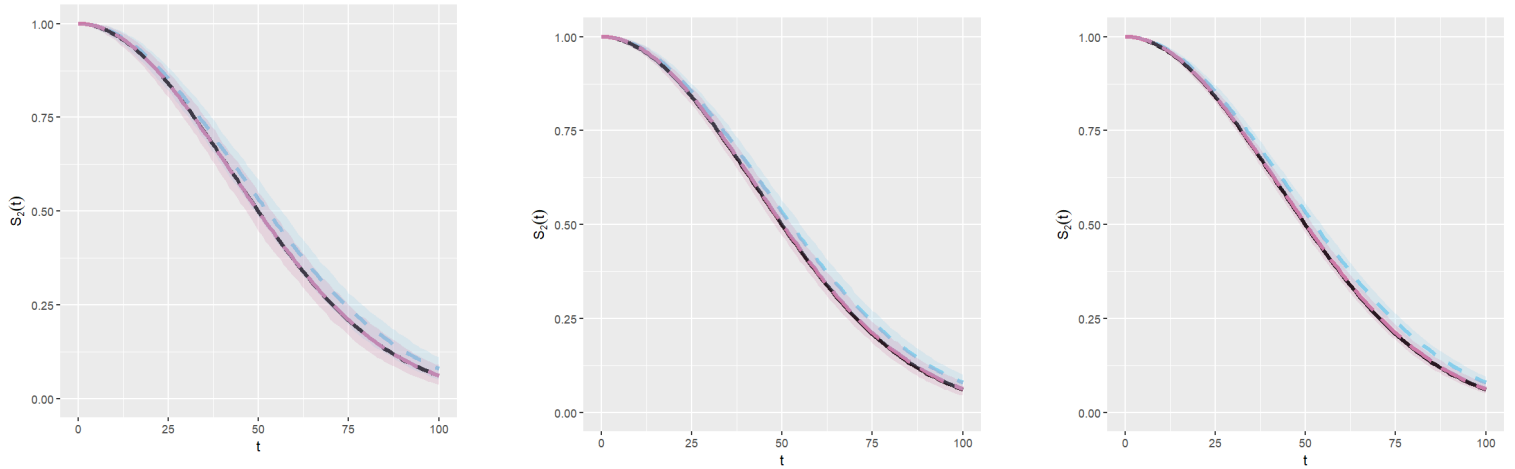
(d) n=500 (e) n=1000 (f) n=2000

Figure 3.27: Robustness Study. Estimates of Marginal Survivor Function $S_1(\cdot)$ and $S_2(\cdot)$ by Proposed Approach Using Simulated Data with Different Sizes under Different Trivariate Copula. True: Clayton with $\tau = 0.8$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted. Gumbel: reddish purple longdash. Frank: vermillion dotdash.



(a) n=500 (b) n=1000 (c) n=2000

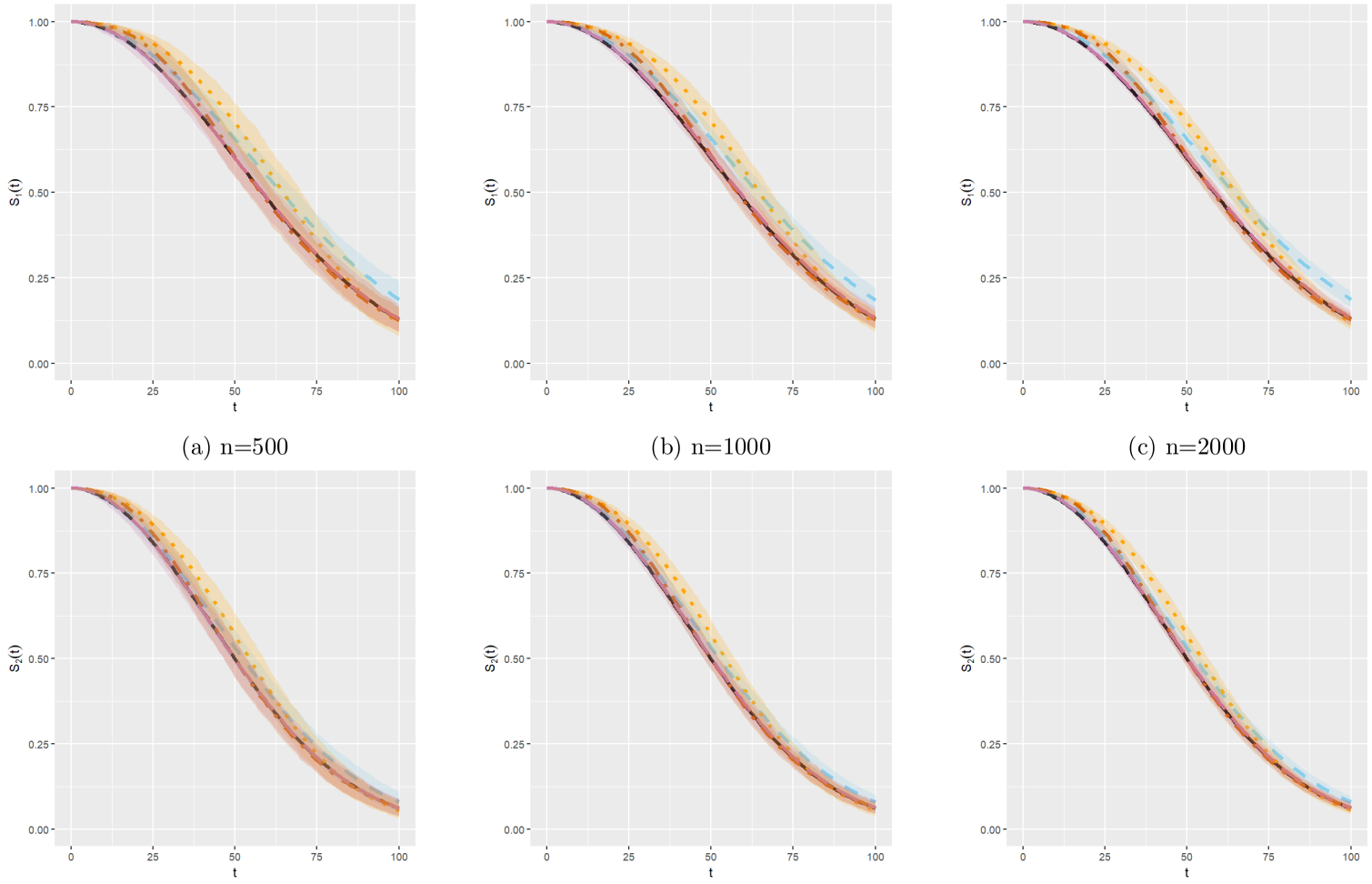
l



(d) n=500 (e) n=1000 (f) n=2000

l

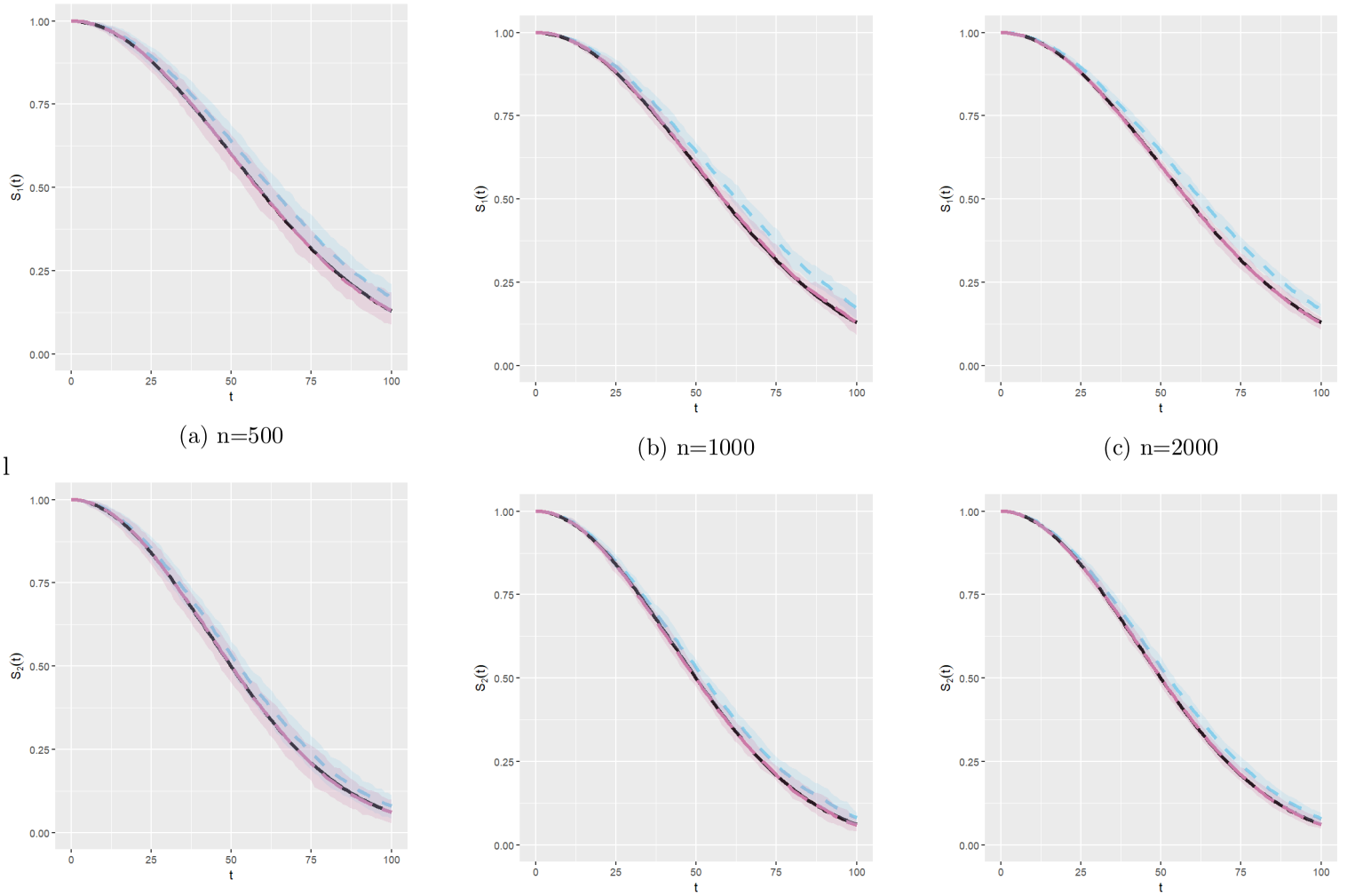
Figure 3.28: Consistency Study. Estimates of Marginal Survivor Function $S_1(\cdot)$ and $S_2(\cdot)$ by Proposed Approach and Kaplan-Meier Estimator Using Simulated Data with Different Sizes under Trivariate Gumbel Copula: $\tau = 0.6$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Gumbel: reddish purple longdash.



(a) n=500 (b) n=1000 (c) n=2000

(d) n=500 (e) n=1000 (f) n=2000

Figure 3.29: Robustness Study. Estimates of Marginal Survivor Function $S_1(\cdot)$ and $S_2(\cdot)$ by Proposed Approach Using Simulated Data with Different Sizes under Different Trivariate Copula. True: Gumbel with $\tau = 0.6$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted. Gumbel: reddish purple longdash. Frank: vermillion dotdash.



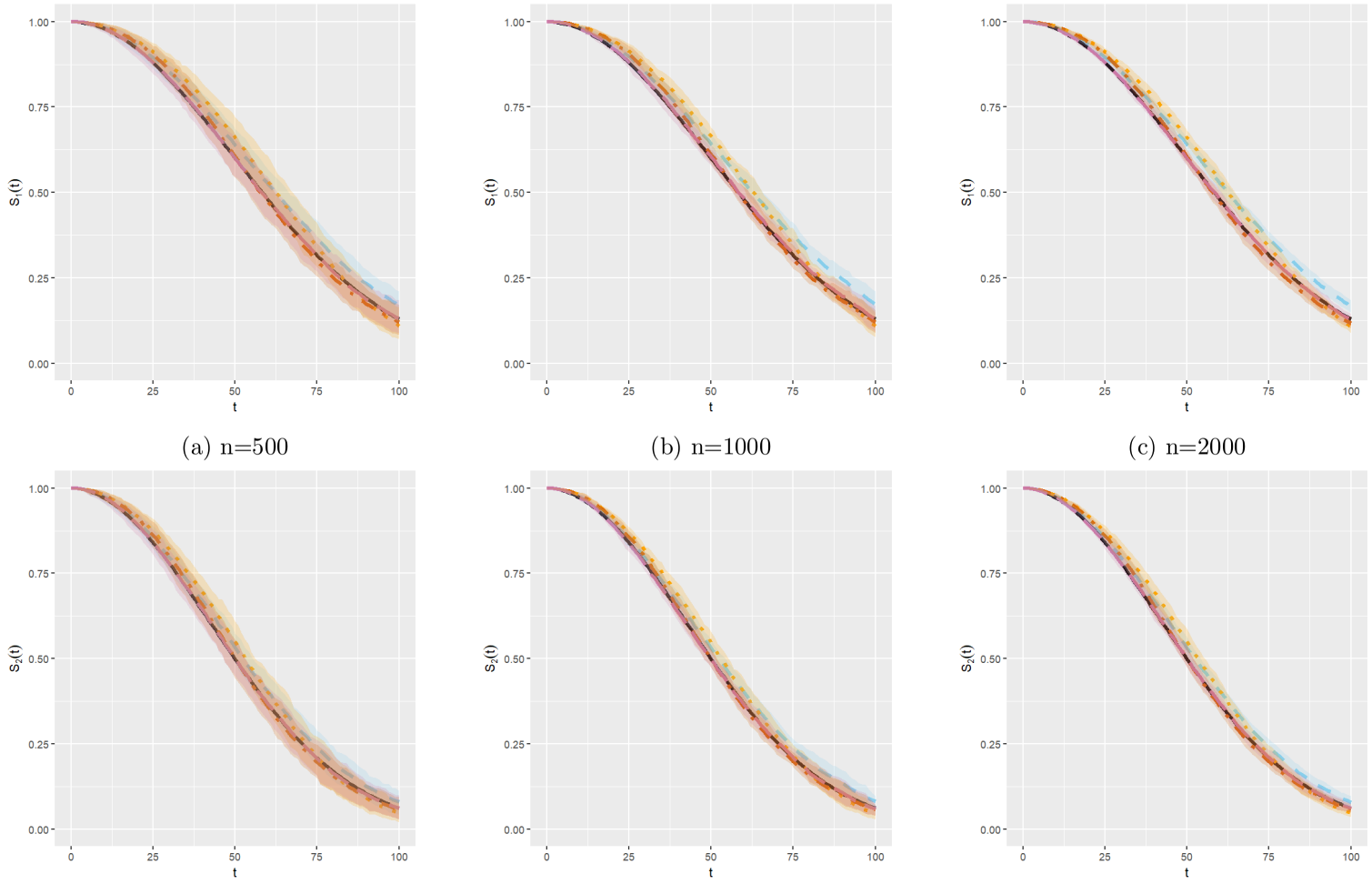
(a) n=500 (b) n=1000 (c) n=2000

l

(d) n=500 (e) n=1000 (f) n=2000

l

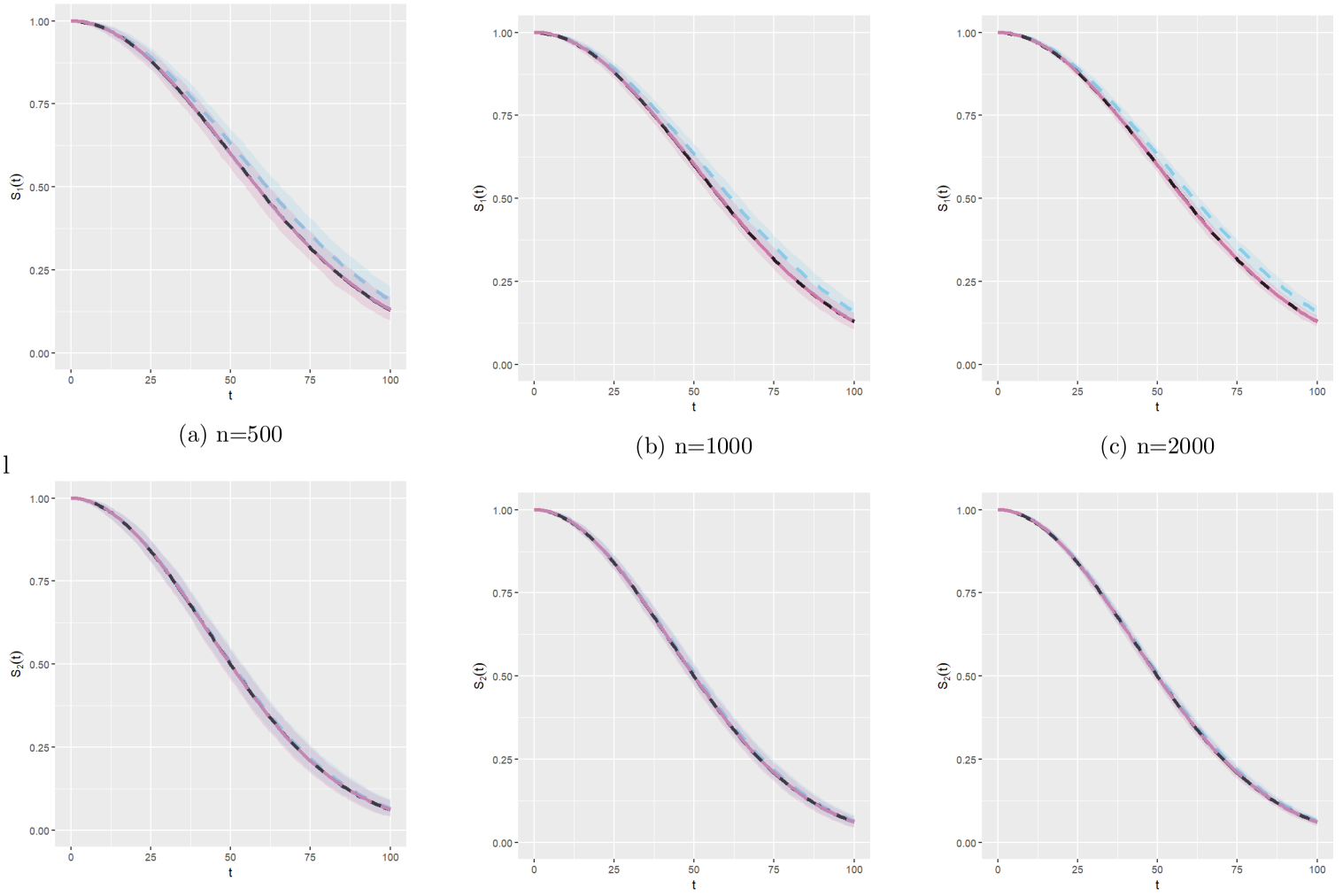
Figure 3.30: Consistency Study. Estimates of Marginal Survivor Function $S_1(\cdot)$ and $S_2(\cdot)$ by Proposed Approach and Kaplan-Meier Estimator Using Simulated Data with Different Sizes under Trivariate Gumbel Copula: $\tau = 0.3$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Gumbel: reddish purple longdash.



(a) n=500 (b) n=1000 (c) n=2000

(d) n=500 (e) n=1000 (f) n=2000

Figure 3.31: Robustness Study. Estimates of Marginal Survivor Function $S_1(\cdot)$ and $S_2(\cdot)$ by Proposed Approach Using Simulated Data with Different Sizes under Different Trivariate Copula. True: Gumbel with $\tau = 0.3$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted. Gumbel: reddish purple longdash. Frank: vermillion dotdash.



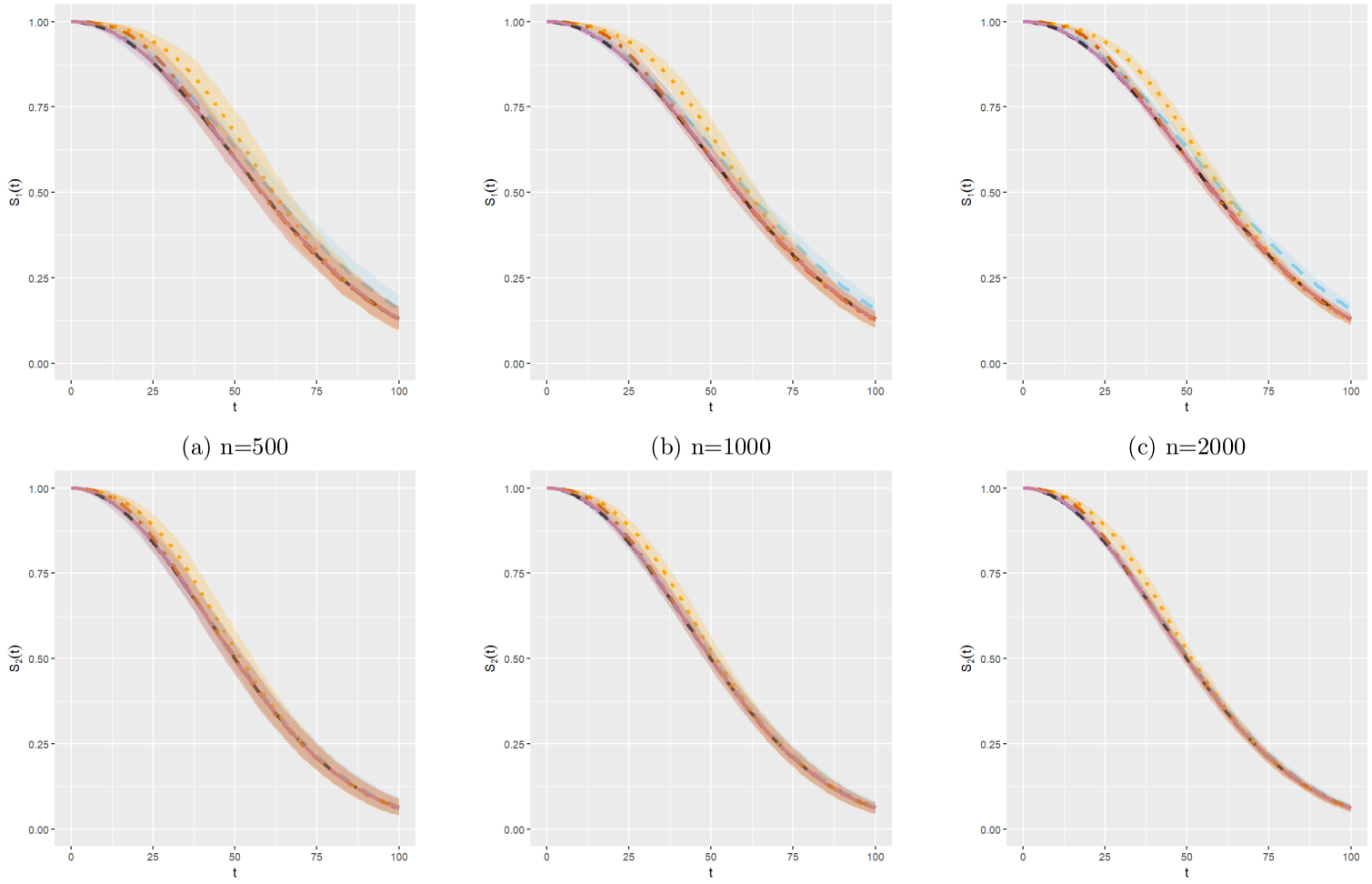
(a) n=500 (b) n=1000 (c) n=2000

l

(d) n=500 (e) n=1000 (f) n=2000

l

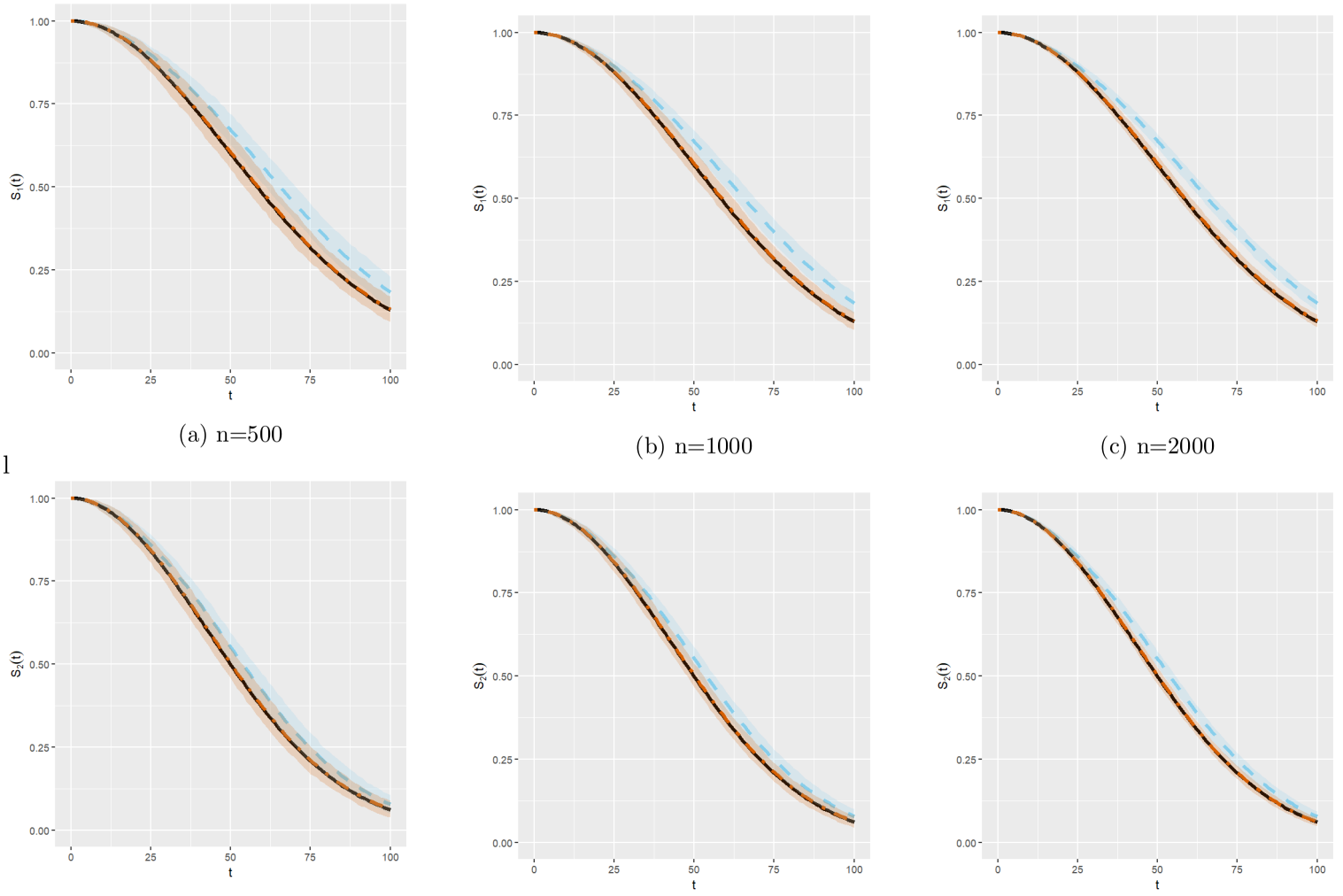
Figure 3.32: Consistency Study. Estimates of Marginal Survivor Function $S_1(\cdot)$ and $S_2(\cdot)$ by Proposed Approach and Kaplan-Meier Estimator Using Simulated Data with Different Sizes under Trivariate Gumbel Copula: $\tau = 0.8$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Gumbel: reddish purple longdash.



(a) n=500 (b) n=1000 (c) n=2000

(d) n=500 (e) n=1000 (f) n=2000

Figure 3.33: Robustness Study. Estimates of Marginal Survivor Function $S_1(\cdot)$ and $S_2(\cdot)$ by Proposed Approach Using Simulated Data with Different Sizes under Different Trivariate Copula. True: Gumbel with $\tau = 0.8$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted. Gumbel: reddish purple longdash. Frank: vermillion dotdash.



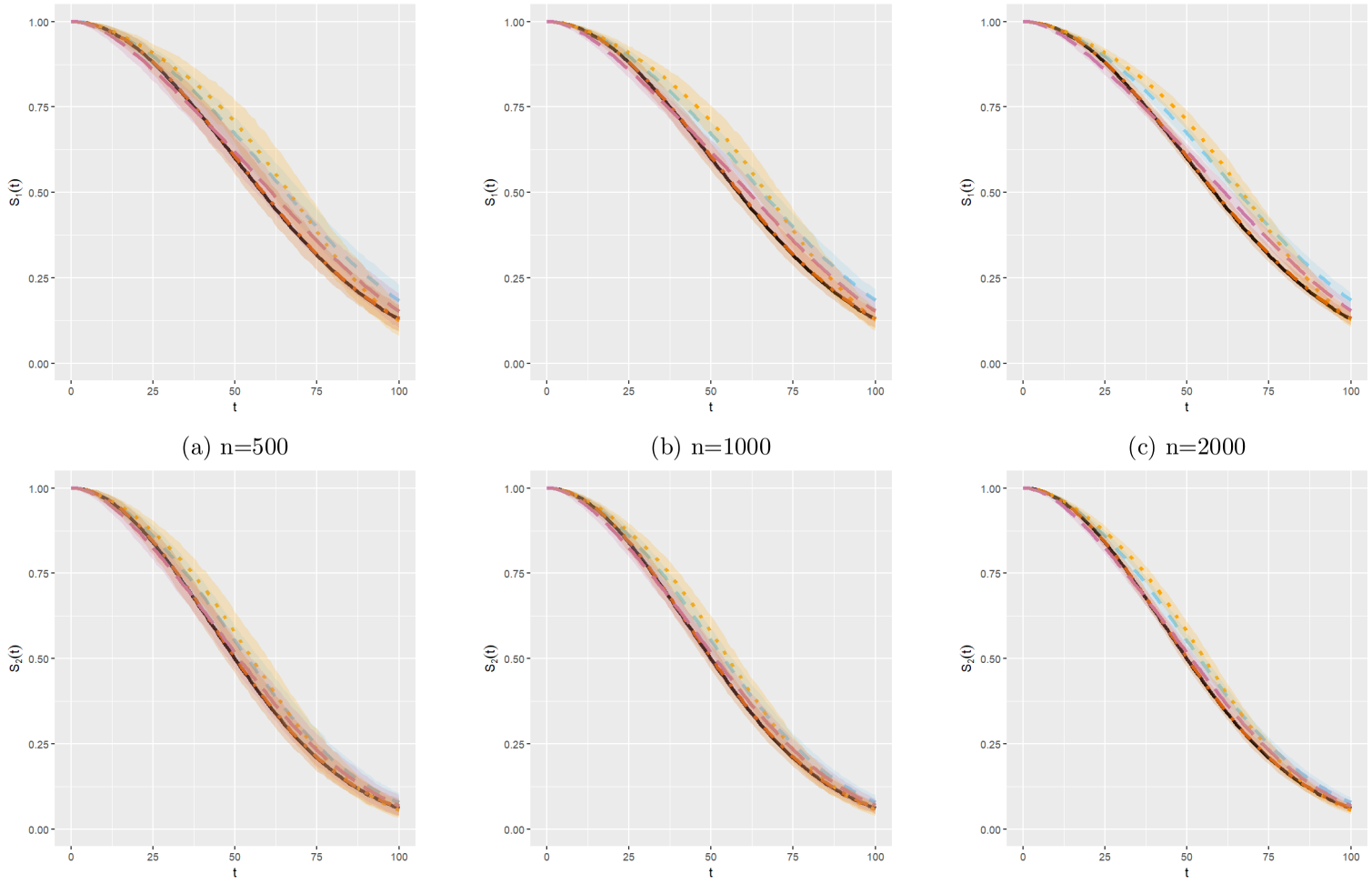
(a) n=500 (b) n=1000 (c) n=2000

l

(d) n=500 (e) n=1000 (f) n=2000

l

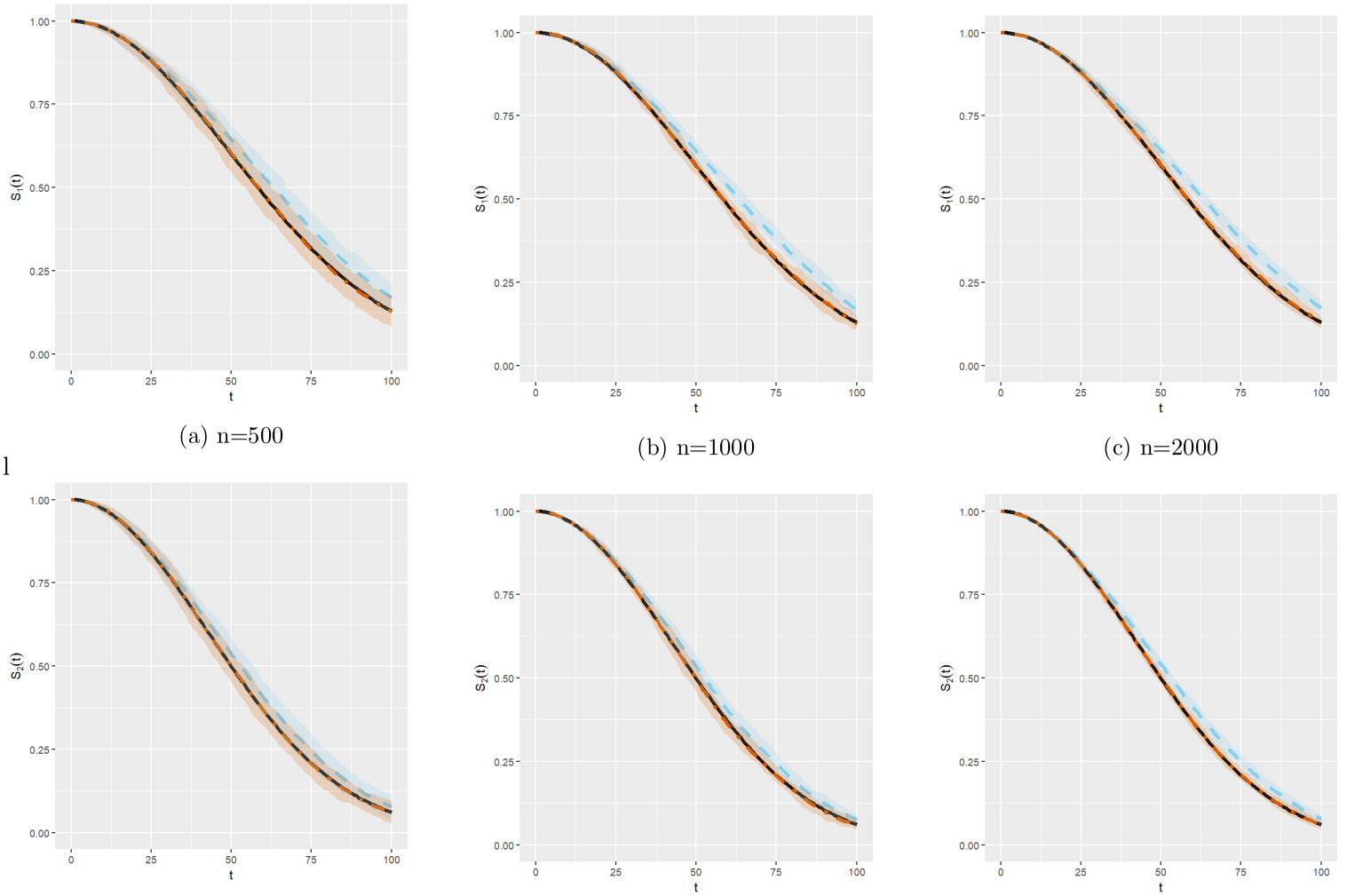
Figure 3.34: Consistency Study. Estimates of Marginal Survivor Function $S_1(\cdot)$ and $S_2(\cdot)$ by Proposed Approach and Kaplan-Meier Estimator Using Simulated Data with Different Sizes under Trivariate Frank Copula: $\tau = 0.6$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Frank: vermillion dotdash.



(a) n=500 (b) n=1000 (c) n=2000

(d) n=500 (e) n=1000 (f) n=2000

Figure 3.35: Robustness Study. Estimates of Marginal Survivor Function $S_1(\cdot)$ and $S_2(\cdot)$ by Proposed Approach Using Simulated Data with Different Sizes under Different Trivariate Copula. True: Frank with $\tau = 0.6$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted. Gumbel: reddish purple longdash. Frank: vermillion dotdash.



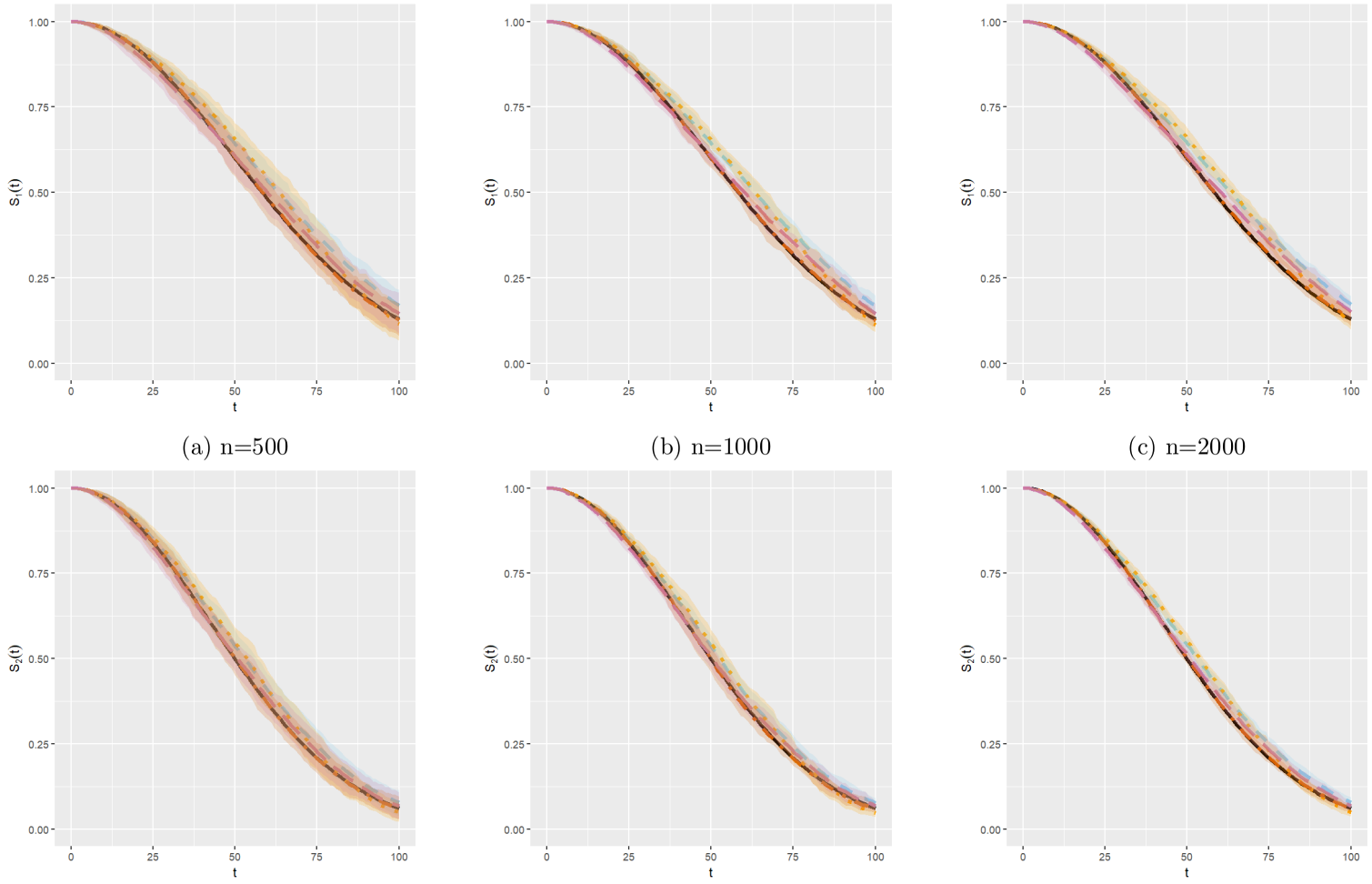
(a) n=500 (b) n=1000 (c) n=2000

l

(d) n=500 (e) n=1000 (f) n=2000

l

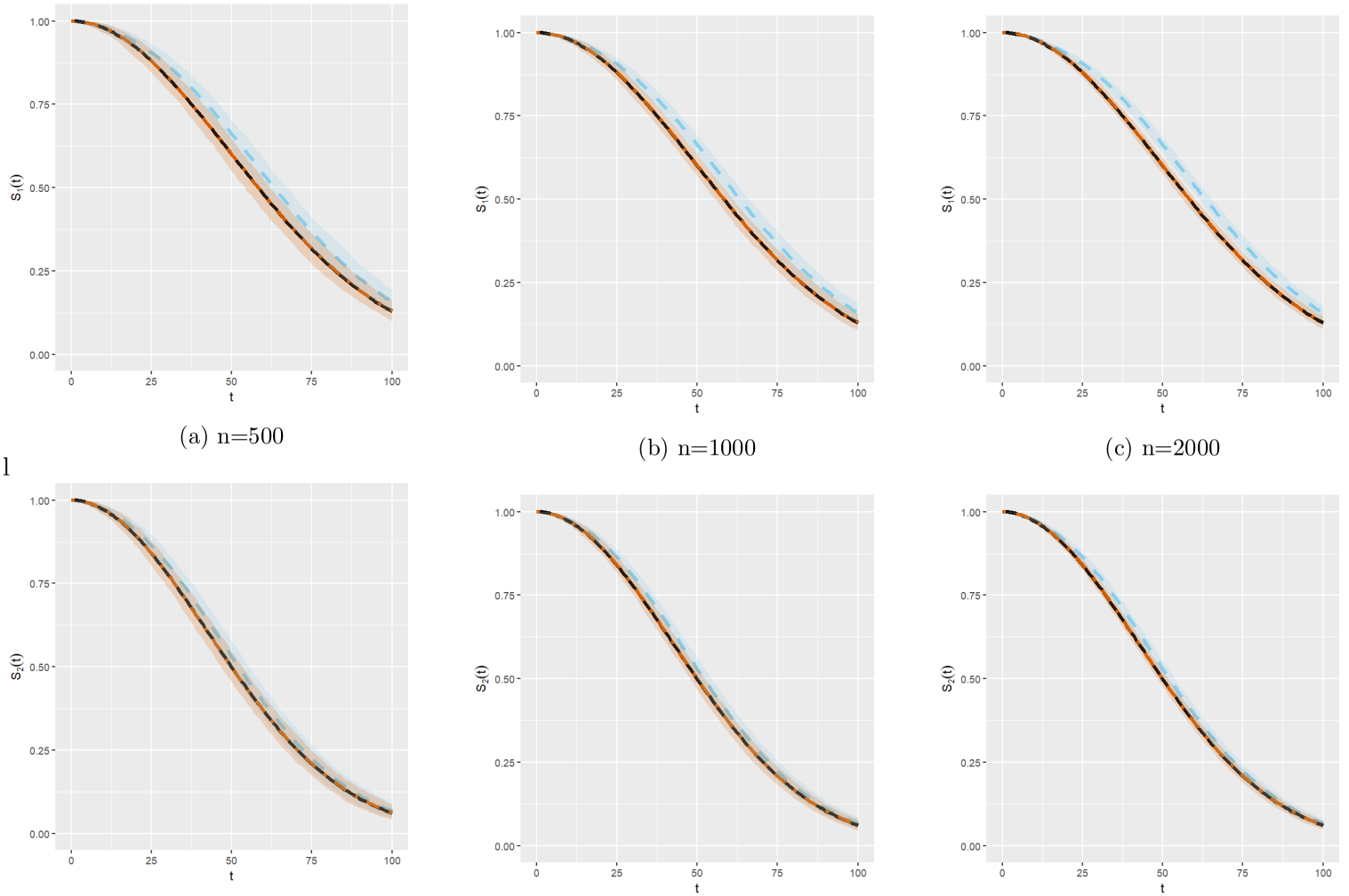
Figure 3.36: Consistency Study. Estimates of Marginal Survivor Function $S_1(\cdot)$ and $S_2(\cdot)$ by Proposed Approach and Kaplan-Meier Estimator Using Simulated Data with Different Sizes under Trivariate Frank Copula: $\tau = 0.3$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Frank: vermillion dotdash.



(a) n=500 (b) n=1000 (c) n=2000

(d) n=500 (e) n=1000 (f) n=2000

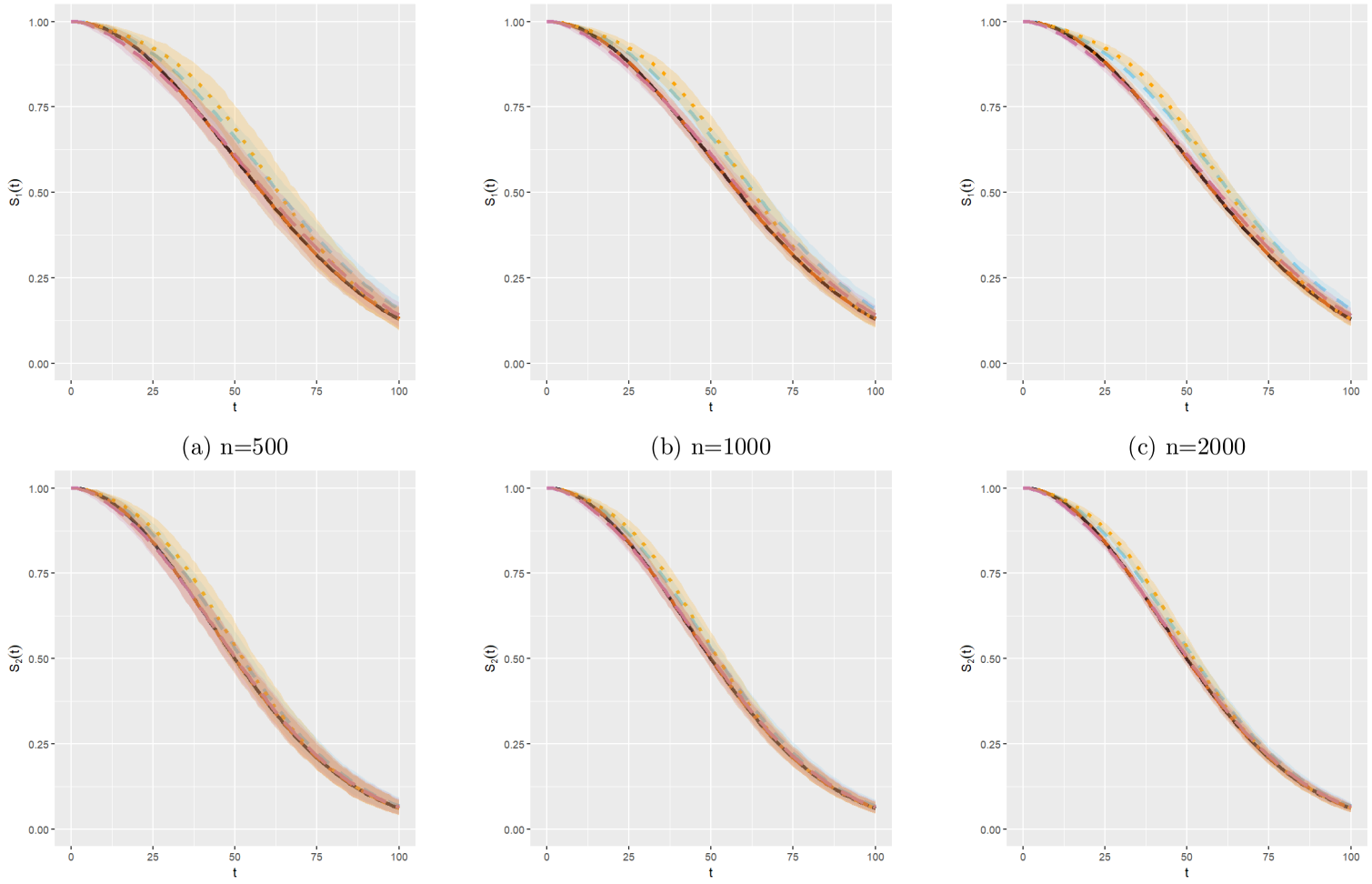
Figure 3.37: Robustness Study. Estimates of Marginal Survivor Function $S_1(\cdot)$ and $S_2(\cdot)$ by Proposed Approach Using Simulated Data with Different Sizes under Different Trivariate Copula. True: Frank with $\tau = 0.3$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted. Gumbel: reddish purple longdash. Frank: vermillion dotdash.



(a) n=500 (b) n=1000 (c) n=2000

(d) n=500 (e) n=1000 (f) n=2000

Figure 3.38: Consistency Study. Estimates of Marginal Survivor Function $S_1(\cdot)$ and $S_2(\cdot)$ by Proposed Approach and Kaplan-Meier Estimator Using Simulated Data with Different Sizes under Trivariate Frank Copula: $\tau = 0.8$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Frank: vermillion dotdash.



(a) n=500 (b) n=1000 (c) n=2000

(d) n=500 (e) n=1000 (f) n=2000

Figure 3.39: Robustness Study. Estimates of Marginal Survivor Function $S_1(\cdot)$ and $S_2(\cdot)$ by Proposed Approach Using Simulated Data with Different Sizes under Different Trivariate Copula. True: Frank with $\tau = 0.8$. True: black solid. Kaplan-Meier (Naïve): skyblue dashed. Clayton: orange dotted. Gumbel: reddish purple longdash. Frank: vermillion dotdash.

3.6 Analysis of BC-BRCAS Data (II)

To illustrate our approach, we present an analysis using the BC-BRCAS data (McBride et al. 2016).

3.6.1 Study Description

Each subject's date of cancer diagnosis is taken as her time origin. The first event time T_1 is the time to RSC and the second time T_2 is the time to the first CVD-related hospitalization after the diagnosis. The information on T_1 and T_2 is subject to censoring by death and end of administrative data extraction. Each subject's censoring time is formulated as $C = D \wedge C_A$, where D is the time to death and C_A is the time at the end of the administrative data extraction window. Table 3.14 presents a summary of the available data on T_1, T_2, and D. As mentioned in Chapter 1, under the time since diagnosis scale, one can reasonably assume that C_A is independent of T_1 and T_2 conditional on stage. Thus, we conducted the analysis stratified by stage at diagnosis. We also ran the analyses using the whole cohort (overall), and by age at diagnosis (< 40, ≥ 40), as well as treatment (Chemo and Surgery and Treatment, Other, Unknown).

Since T_1 cannot be defined for stage IV subjects, and there were about 20 percent of patients with unknown stage in the original dataset who they turned out to be 'non-referred' patients, we thus chose to include only study new patients that are referred to BC Cancer and removed stage IV breast cancer patients in the subsequent analysis.

3.6.2 Correlations amongst Event Times and Death

The analysis of the BC-BRCAS data aims to evaluate the correlation between a breast cancer patient's time to RSC and her time to a CVD event after the cancer diagnosis. Using the proposed approach, Kendall's τ was estimated under a copula model for the joint survivor function of (T_1, T_2) with the available data.

Specifically, to account for the potential informative censoring of the observations on (T_1, T_2) due to death, the distribution of (T_1, T_2, D) is assumed to follow an Archimedean copula with the association parameter θ. The pseudo-MLE procedure was implemented as described in section 3.3 to estimate θ, the associated standard error, and the marginal survivor functions $S_j(\cdot)$ using the data from subgroups based on age at diagnosis, stage at diagnosis, and treatment, as well as from the full cohort.

For the comparison of estimates from different models, we converted the estimated $\hat{\theta}$ into estimates of the corresponding Kendall's τ. Table 3.15 presents the estimates of τ under the Clayton, Gumbel, and Frank copulas. Based on the estimated τ, the associations between the death time D and the event times T_j (the time to RSC or CVD) all appear strongly positive across different subgroups (age at diagnosis, stage, treatment) and the whole cohort, regardless of the copula model used in the estimation. This is further evidence that

informative censoring occurred in the observations of the two event times. For comparison, we have also presented the estimates obtained by the naïve approach that ignore the informative censoring and used the Kaplan–Meier estimates of $S_j(\cdot)$ in the pseudo-MLE procedure of section 3.3. These estimates seem to have underestimated the association parameter compared to the pseudo-MLE estimates.

The estimated marginal survivor functions $\hat{S}_j(\cdot)$ and approximate 95 percent CIs are shown in figure 3.40 for the early and late stage subgroups. They appear rather different from the corresponding naïve estimates. From simulation studies in section 3.5, one could choose Frank copula in practice because of its robustness, although estimated curves from Clayton and Frank copula are similar using BC-BRCAS study.

Table 3.14: Summary Statistics of BC-BRCAS Data $\mathcal{P}_{\text{referred}}$

		Total	$N(T_1^{\text{obs}})^\dagger$	$\overline{T_1^{\text{obs}}}^\ddagger$	$N(T_2^{\text{obs}})$	$\overline{T_2^{\text{obs}}}$	$N(D^{\text{obs}})$	$\overline{D^{\text{obs}}}$
Overall		40147	11763	5.12	5747	6.32	12216	7.3
Diagnosis Age Group	<40	2801	1070	4.49	110	7.1	759	5.76
	40+	37346	10693	5.19	5637	6.31	11457	7.41
Stage	Early (I and II)	35339	9717	5.5	5184	6.52	9983	7.85
	Late (III)	4486	1942	3.19	500	4.29	2062	4.62
	Unknown	322	104	6.11	63	6.16	171	7.55
Treatment	Chemo and Rad	11210	3540	4.94	904	6.41	2773	6.22
	Chemo	2883	952	4.28	251	6.19	739	5.8
	Rad	14560	4059	6.14	2587	6.96	4341	8.71
	No Chemo or Rad	8347	2564	4.87	1781	5.85	3616	7.6
	Unknown	3147	648	1.96	224	2.42	747	3.19

$^\dagger N(T_1^{\text{obs}}), N(T_2^{\text{obs}}), N(D^{\text{obs}})$: numbers of individuals with times to RSC, CVD, death (T_1, T_2, D) observed, respectively

$^\ddagger \overline{T_1^{\text{obs}}}, \overline{T_2^{\text{obs}}}, \overline{D^{\text{obs}}}$: sample means of the observed times to RSC, CVD, death (T_1, T_2, D), respectively

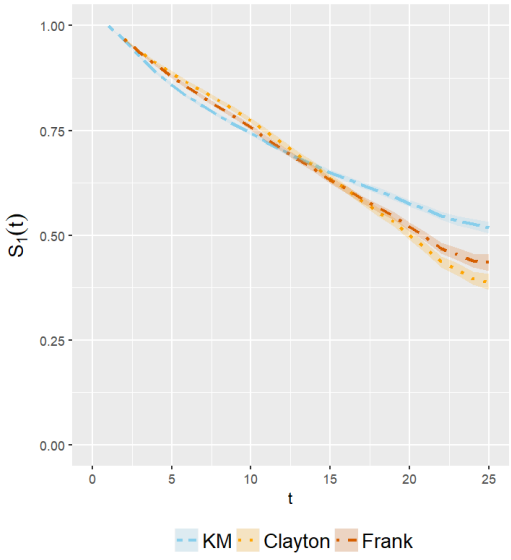
Table 3.15: Kendall's τ Estimates with BC-BRCAS Data $\mathcal{P}_{\text{referred}}$ Using Trivariate Archimedean Copulas on (T_1, T_2, D)

Approach	**Assumed Copula**		**Overall**	Age at diagnosis		Stage at diagnosis			Treatment		
				<40	**40+**	**Early)**	**Late**	**Unknown**	**CRS***	**Other**	**Unknown**
Pseudo-MLE	3-D Clayton	$\hat{\tau}$	0.62	0.82	0.60	0.60	0.70	0.54	0.77	0.54	0.60
		se^*	.004	.009	.004	.004	.010	.011	.007	.005	.010
	3-D Frank	$\hat{\tau}$	0.54	0.75	0.52	0.51	0.65	0.52	0.65	0.46	0.51
		se	.004	.011	.004	.004	.009	.012	.008	.005	.010
Naive†	3-D Clayton	$\hat{\tau}$	0.51	0.70	0.53	0.50	0.61	0.31	0.71	0.50	0.37
		se	.003	.010	.004	.004	.009	.007	.006	.005	.007
	3-D Frank	$\hat{\tau}$	0.42	0.62	0.44	0.43	0.50	0.28	0.62	0.41	0.30
		se	.003	.011	.003	.004	.008	.006	.007	.004	.006

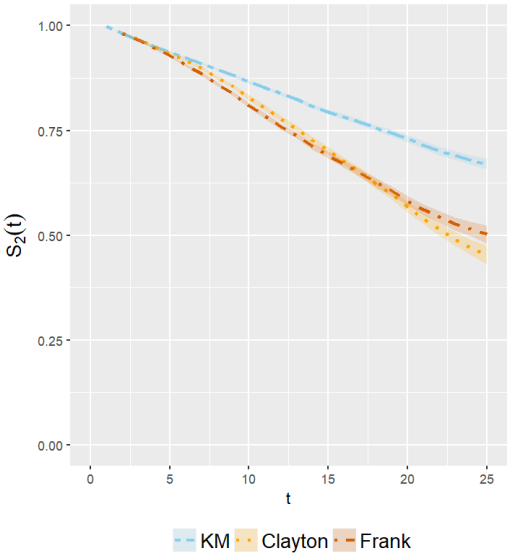
*: estimated standard error

†: naive estimates using Kaplan-Meier estimates for all the three marginal survivor functions

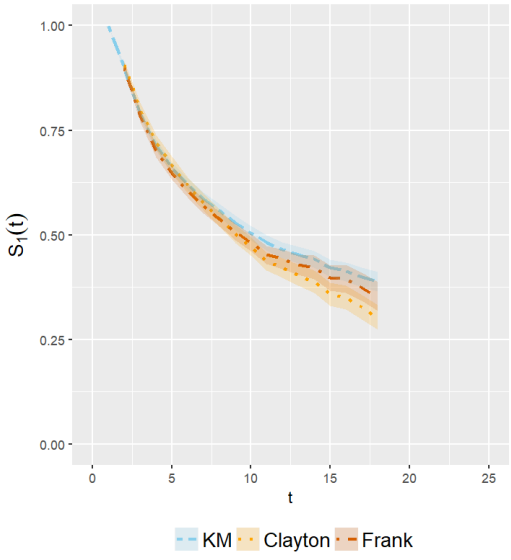
* : Chemo+Radiation+Surgery



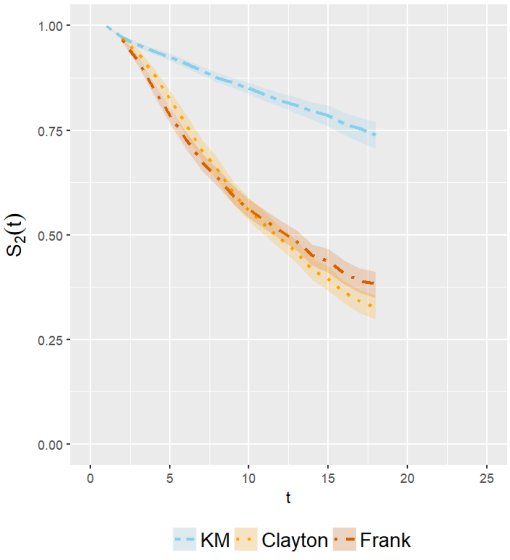
(a1) time to RSC in younger group



(a2) time to CVD in younger group



(b1) time to RSC in older group



(b2) time to CVD in older group

Figure 3.40: Estimated Marginal Survivor Functions $S_1(\cdot)$ and $S_2(\cdot)$ of Times to RSC and CVD by Proposed Approach with Different Copulas and Kaplan-Meier Estimator Using the BC-BRCAS Data $\mathcal{P}_{\text{referred}}$: Early vs. Late Age at Diagnosis

3.7 Discussion

This chapter has proposed a modeling approach by Archimedean copula and the associated pseudolikelihood-based procedure for the analysis of multiple event times in the presence of informative censoring due to a terminating event. The approach allows us to account for the informative censoring and to estimate validly the joint distribution of the multiple event times. It also has the inference convenience associated with a copula model.

We have studied the proposed estimator asymptotically and numerically, and it is easy to implement. In addition, the procedure for estimating the survivor function of an event time can be used to provide a valid estimator for semicompeting-risks data. The proposed modeling requires the same association between the event times, and between them jointly and the time to the terminating event. Its association parameter may be taken as an average of the associations with varying magnitude between different pairs of event times.

Since the time scale used in the analysis is the time since diagnosis, the administrative censoring time C_A is likely dependent of the event times. This dependence may be captured by the stage at diagnosis. We thus assumed that C_A is independent of T_1, T_2 and D conditional on the diagnosis stage when interpreting the analysis. In particular, we would report to the research team mainly based on the subgroup analysis according to the stage at diagnosis. On the other hand, this consideration has partly motivated our next research topic, to analyze the event times with adjustment for potential covariates, including the diagnosis stage.

The real data analysis (II) estimates showed strong positive associations between the two event times, and each of them with death time amongst early and late stage subgroups. This indicates that extension to a regression setting is useful. We consider the regression extension in Chapter 5 and Chapter 6. On the other hand, the three individual association parameters do not necessarily appear the same. This indicates that an alternative modeling would be desirable to allow different event time pairs to have different association parameters, or different association structures.

Chapter 4

Multiple Event Times in the Presence of Informative Censoring Using Copula - Part Two: a Flexible Approach

This chapter considers modeling two event times jointly when the observations of them are subject to informative censoring caused by a terminating event. We formulate the correlation of the bivariate event time with the censoring time by embedding the bivariate distribution in a bivariate copula model. This allows the convenience of inference under the conventional copula model. At the same time, the proposed model is more flexible, and thus potentially more appropriate in many practical situations than modeling the event times and the associated censoring time jointly by a single multivariate copula. Adapting the commonly used two-stage estimation procedure under a copula model, we develop an easy-to-implement estimator for the joint survivor function of the two event times. A by-product of the approach is an estimator for the marginal distribution of a single event time with semicompeting-risks data. We conduct asymptotic and simulation studies to examine the consistency, efficiency, and robustness of the proposed approach. The breast cancer project that motivated this research is employed to illustrate the method.

Compared to that in Chapter 3, the proposed model is more flexible and thus potentially more feasible in many practical situations than modeling the event times and the associated censoring time jointly by a single copula.

4.1 Modeling

4.1.1 Model Specification

We assume that the administrative censoring time C_A is independent of the bivariate event time (T_1, T_2) and the time to the terminating event D. Moreover, to specify the correlation of (T_1, T_2) with D, we embed the bivariate survivor function of (T_1, T_2) in a bivariate Archimedean copula model (e.g., Joe 1997) and assume the joint survivor function with D equal to

$$Pr(T_1 \geq t_1, T_2 \geq t_2, D \geq d) = \mathcal{A}_{[2]}(S_{12}(t_1, t_2), S_D(d); \theta). \tag{4.1}$$

The association parameter θ characterizes the correlation between (T_1, T_2) and D. Note that $S_{12}(t_1, 0) = P(T_1 \geq t_1)$ and $S_{12}(0, t_2) = P(T_2 \geq t_2)$ are the marginal survivor functions of T_1 and T_2, respectively. Let $S_j(t) = P(T_j \geq t)$ for $j = 1, 2$. The model in (4.1) induces the joint model of T_j and D:

$$Pr(T_j \geq t, D \geq d) = \mathcal{A}_{[2]}(S_j(t), S_D(d); \theta). \tag{4.2}$$

Denote $\delta_{1i} + \delta_{2i}$ by $\delta_{\cdot i}$, and let $\dot{h}(r)$ be $dh(r)/dr$ for a function $h(r)$ and $h^{(a_1,a_2)}(r_1, r_2; \phi)$ be $\partial h^{(a_1+a_2)}(r_1, r_2; \phi)/\partial r_1^{a_1} \partial r_2^{a_2}$ for a function $h(r_1, r_2; \phi)$ with well-defined (partial) derivatives. The likelihood function with the available data under the copula model (4.1) is

$$\begin{aligned} & L(S_{12}(\cdot), S_D(\cdot), \theta | \text{Observed-Data}) \\ = & \prod_{i=1}^{n} \left\{ (-1)^{\delta_{\cdot i} + \delta_{Di}} \frac{\partial^{\delta_{\cdot i}} \mathcal{A}_{[2]}^{(0,\delta_{Di})}(S_{12}(u_{1i}, u_{2i}), S_D(c_i); \theta)}{\partial u_1^{\delta_{1i}} \partial u_2^{\delta_{2i}}} \dot{S}_D(c_i)^{\delta_{Di}} \right\}. \end{aligned} \tag{4.3}$$

If the joint survivor function $S_{12}(\cdot)$ is specified upon a finite-dimensional parameter θ_{12}, i.e., $S_{12}(\cdot) = S_{12}(\cdot; \theta_{12})$, maximizing (4.3) with respect to θ, θ_{12}, and $S_D(\cdot)$ yields the maximum likelihood estimator (MLE) of θ, θ_{12} and thus the MLE of the joint distribution of (T_1, T_2).

Assume that the current observations on D are subject to noninformative (administrative) right-censoring with the censoring time C_A. There is a readily available consistent estimator for $S_D(\cdot)$, e.g., the Kaplan–Meier estimator, denoted by $\tilde{S}_D(\cdot)$. Following the idea of the pseudolikelihood estimation procedure under a copula model (e.g., Lawless & Yilmaz 2011), we may consider a pseudo-MLE of θ, θ_{12} by maximizing $L(S_{12}(\cdot; \theta_{12}), \tilde{S}_D(\cdot), \theta | \text{Observed-Data})$, which is proportional to

$$\prod_{i=1}^{n} \left\{ (-1)^{\delta_{\cdot i} + \delta_{Di}} \frac{\partial^{\delta_{\cdot i}} \mathcal{A}_{[2]}^{(0,\delta_{Di})}(S_{12}(u_{1i}, u_{2i}; \theta_{12}), \tilde{S}_D(c_i); \theta)}{\partial u_1^{\delta_{1i}} \partial u_2^{\delta_{2i}}} \right\}, \tag{4.4}$$

with respect to θ_{12} jointly with θ only. The partial derivative in (4.4) is

$$\frac{\partial^{\delta_{\cdot i}} \mathcal{A}_{[2]}^{(0,\delta_{Di})}(S_{12}(u_{1i},u_{2i};\theta_{12}),\tilde{S}_D(c_i);\theta)}{\partial u_1^{\delta_{1i}} \partial u_2^{\delta_{2i}}}$$

$$= \begin{cases} \mathcal{A}_{[2]}^{(0,\delta_{Di})}(S(u_{1i},u_{2i};\theta_{12}),\tilde{S}_D(c_i);\theta), & \delta_{1i}=\delta_{2i}=0 \\ \mathcal{A}_{[2]}^{(1,\delta_{Di})}(S(u_{1i},u_{2i};\theta_{12}),\tilde{S}_D(c_i);\theta) S_{12}^{(\delta_{1i},\delta_{2i})}(u_{1i},u_{2i};\theta_{12}), & \delta_{1i}\neq\delta_{2i} \\ \mathcal{A}_{[2]}^{(2,\delta_{Di})}(S(u_{1i},u_{2i};\theta_{12}),\tilde{S}_D(c_i);\theta) S_{12}^{(1,0)}(u_{1i},u_{2i};\theta_{12}) S_{12}^{(0,1)}(u_{1i},u_{2i};\theta_{12}) & \\ \quad +\mathcal{A}_{[2]}^{(1,\delta_{Di})}(S(u_{1i},u_{2i};\theta_{12}),\tilde{S}_D(c_i);\theta) S_{12}^{(1,1)}(u_{1i},u_{2i};\theta_{12}), & \delta_{1i}=\delta_{2i}=1. \end{cases}$$

The resulting estimator, with the trade-off of some efficiency loss, can be much easier to implement than its MLE counterpart.

However, often the bivariate survivor function $S_{12}(\cdot)$ cannot be confidently specified via a parametric model. We therefore consider a semiparametric model for the joint distribution of T_1, T_2:

$$S_{12}(t_1,t_2;\theta_{12}) = \mathcal{C}_{[2]}(S_1(t_1),S_2(t_2);\theta_{12}), \tag{4.5}$$

where the univariate marginal survivor functions $S_j(\cdot)$ are unspecified, and $\mathcal{C}_{[2]}(\cdot;\theta_{12})$ is specified up to θ_{12}, defined on $[0,1]^2$, and valued over $[0,1]$.

In principle, one may maximize (4.3) under model (4.1) coupled with model (4.5) with respect to $\theta, S_j(\cdot), \theta_{12}$, and $S_D(\cdot)$ to obtain their MLE, which leads to the semiparametric MLE of the joint survivor function $S_{12}(\cdot;\theta_{12})$. This, however, requires rather intensive computing. Furthermore, the counterpart of the pseudo-MLE approach for parametric $S_{12}(\cdot)$ is not directly applicable since there is no readily available consistent estimator for $S_j(\cdot)$ with the current semicompeting-risks data on T_j. These considerations motivate the two procedures in Section 4.2 for estimating the joint survivor function $S_{12}(\cdot;\theta_{12})$ under model (4.5).

4.1.2 More on Modeling

It is of interest in many situations to estimate the marginal survivor function of the event times T_j $(j=1,2)$ with the semicompeting-risks data, Observed-Data$_j$ in (3.2). When the copula function $\mathcal{A}_{[2]}(\cdot;\theta)$ in (4.1) is an Archimedean copula with the generator $\psi(\cdot;\theta)$, the induced model (4.2) for the joint survivor function of T_j and D yields

$$S_j(t) = g(S_j^*(t),S_D(t);\theta) = \psi^{-1}\{\psi(S_j^*(t);\theta)-\psi(S_D(t);\theta);\theta\}, \tag{4.6}$$

for $j=1,2$, where $S_j^*(t)=P(T_j^*\geq t)$ is the survivor function of $T_j^*=T_j\wedge D$.

Useful examples for $\mathcal{C}_{[2]}(\cdot;\theta_{12})$ in (4.5) include commonly used bivariate parametric copula functions (e.g., Diao & Cook 2014). When the copula function $\mathcal{C}_{[2]}(\cdot)$ in model (4.5) is assumed to equal the bivariate Archimedean copula $\mathcal{A}_{[2]}(\cdot)$ in model (4.1), the joint survivor function of the trivariate event times T_1, T_2, D in (4.1) becomes $\mathcal{A}_{[3]}(S_1(t_1),S_2(t_2),S_D(d);\theta)$.

In fact, in such situations, the joint survivor function of each pair of T_1, T_2, D is the same bivariate Archimedean copula $\mathcal{A}_{[2]}(\cdot)$ with the two indices equal to the appropriate univariate marginal survivor functions (e.g., Li et al. 2018).

Another example for $\mathcal{C}_{[2]}(\cdot)$ is

$$\mathcal{C}_{[2]}(w_1, w_2; \theta_{12}) = \int w_1^{\exp(\theta_1 \xi)} w_2^{\exp(\theta_2 \xi)} \eta(\xi; \theta_\eta) d\xi \tag{4.7}$$

with $w_1, w_2 \in [0, 1]$, $\theta_{12} = (\theta_1, \theta_2, \theta_\eta)$, and $\eta(\cdot; \theta_\eta)$ a probability density function with parameter θ_η. This second type of function results from assuming T_1 and T_2 to be independent conditional on a random variable $\xi \sim \eta(\xi; \theta_\eta)$ and, for $j = 1, 2$, $Pr(T_j \geq t|\xi) = S_j(t)^{\exp(\theta_j \xi)}$, which is a Cox proportional hazards model conditional on ξ, with ξ being a frailty variable.

We remark that this chapter allows the bivariate function $\mathcal{C}_{[2]}(\cdot)$ in model (4.5) to be different from the bivariate Archimedean copula $\mathcal{A}_{[2]}(\cdot)$ in model (4.1). We may choose $\mathcal{C}_{[2]}(\cdot)$ in (4.5) to be a commonly used non-Archimedean copula or a bivariate function such as (4.7). This leads to additional modeling flexibility. More discussion of this is given with the numerical studies reported in sections 4.3 and 4.4.

4.2 Pseudolikelihood-Based Estimation Procedures

Using the idea underlying two-stage estimation procedures with a copula model (e.g., Oakes 1994), we estimate $S_{12}(\cdot)$, the joint survivor function of (T_1, T_2), under model (4.1) with model (4.5) embedded. The estimation procedure yields a consistent estimator for the marginal survivor function of each of the two event times as a by-product. We also present the asymptotic properties of the estimators.

4.2.1 Estimating Association Parameters with the Observed-Data

Under model (4.2), as given in (4.6), the marginal survivor function $S_j(t) = g(S_j^*(t), S_D(t); \theta)$, a known function of the marginal survivor function of $T_j^* = T_j \wedge D$ and the marginal survivor function of D up to θ for $j = 1, 2$. With known $S_j^*(t)$ and $S_D(t)$, $S_j(t)$ is known only up to the parameter θ.

In addition, note that the likelihood function in (4.3) becomes

$$\prod_{i=1}^{n} \left\{ (-1)^{\delta_{\cdot i} + \delta_{Di}} \frac{\partial^{\delta_{\cdot i}} \mathcal{A}_{[2]}^{(0, \delta_{Di})}(S_{12}(u_{1i}, u_{2i}), S_D(c_i); \theta)}{\partial S_1(u_1)^{\delta_{1i}} \partial S_2(u_2)^{\delta_{2i}}} \left[\dot{S}_1(u_{1i})^{\delta_{1i}} \dot{S}_2(u_{2i})^{\delta_{2i}} \right] \dot{S}_D(c_i)^{\delta_{Di}} \right\},$$

which is proportional to

$$\begin{aligned} & L(\theta, \theta_{12}; S_1(\cdot), S_2(\cdot), S_D(\cdot) | \text{Observed-Data}) \\ = & \prod_{i=1}^{n} \left\{ (-1)^{\delta_{\cdot i} + \delta_{Di}} \frac{\partial^{\delta_{\cdot i}} \mathcal{A}_{[2]}^{(0, \delta_{Di})}(S_{12}(u_{1i}, u_{2i}), S_D(c_i); \theta)}{\partial S_1(u_1)^{\delta_{1i}} \partial S_2(u_2)^{\delta_{2i}}} \right\} \end{aligned} \tag{4.8}$$

when $S_j(\cdot)$ and $S_D(\cdot)$ are known. This leads to the following estimation procedure.

Given consistent estimators for $S_j(\cdot)$ and $S_D(\cdot)$, we maximize the resulting pseudolikelihood function of $\boldsymbol{\theta} = (\theta, \theta_{12})$ or, equivalently, its log-transformation with respect to the parameters $\boldsymbol{\theta} = (\theta, \theta_{12})$, to derive a pseudo-MLE:

$$\hat{\boldsymbol{\theta}}_n = \text{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta} | \tilde{S}_1(\cdot; \theta), \tilde{S}_2(\cdot; \theta), \tilde{S}_D(\cdot); \text{Observed-Data}). \tag{4.9}$$

This pseudo-MLE procedure is computationally easy to implement. We present below an iterative algorithm to calculate $\hat{\boldsymbol{\theta}}_n$.

ALGORITHM Using the Kaplan–Meier estimates $\tilde{S}_j^*(\cdot)$ and $\tilde{S}_D(\cdot)$ together with the current estimate $\boldsymbol{\theta}^{(k-1)}$ and $S_j^{(k-1)}(\cdot)$ for $j = 1, 2$ and with $k \geq 1$:

Step 1. Obtain the updated estimate for $\boldsymbol{\theta}$ via

$$\boldsymbol{\theta}^{(k)} = \text{argmax}_{\boldsymbol{\theta}} L(\boldsymbol{\theta} | S_1^{(k-1)}(\cdot), S_2^{(k-1)}(\cdot), \tilde{S}_D(\cdot); \text{Observed-Data});$$

Step 2. Obtain the updated estimates for $S_j(\cdot)$ via $S_j^{(k)}(t) = \tilde{S}_j(t; \boldsymbol{\theta}^{(k)}) = g(\tilde{S}_j^*(t), \tilde{S}_D(t); \boldsymbol{\theta}^{(k)})$ for $j = 1, \ldots, J$.

Repeat steps 1 and 2 until the sequence $\{\boldsymbol{\theta}^{(k)} : k = 0, 1, \ldots\}$ converges. The limit is $\hat{\boldsymbol{\theta}}_n$ defined in (4.9). The initial estimate $\boldsymbol{\theta}^{(0)}$ is in fact not needed. The Kaplan–Meier estimates of $S_j(\cdot)$ may be used as the initial estimates $S_j^{(0)}(\cdot)$ for $j = 1, 2$.

The following proposition establishes the consistency and asymptotic normality of the resulting estimator.

Proposition 4. *Under the regularity conditions (RC1)–(RC4) presented in Chapter 3 and provided $\tilde{S}_j^*(t)$ and $\tilde{S}_D(t)$ satisfy condition (AC1), as $n \to \infty$, $\hat{\boldsymbol{\theta}}_n \xrightarrow{a.s.} \boldsymbol{\theta}$ and $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{d} N(0, AV(\boldsymbol{\theta}))$, where the asymptotic variance is*

$$AV(\boldsymbol{\theta}) = V_B(\boldsymbol{\theta})^{-1} V_A(\boldsymbol{\theta}) V_B(\boldsymbol{\theta})^{-1} \tag{4.10}$$

with $V_B(\boldsymbol{\theta})$ and $V_A(\boldsymbol{\theta})$ the limits of

$$-\frac{1}{n} \sum_{i=1}^{n} \partial^2 \log L(\boldsymbol{\theta} | \tilde{S}_1(\cdot; \theta), \tilde{S}_2(\cdot; \theta), \tilde{S}_D(\cdot); \textit{Observed-Data}) / \partial \boldsymbol{\theta}^2 \tag{4.11}$$

and

$$\frac{1}{n} Var\Big\{ \sum_{i=1}^{n} \partial \log L(\boldsymbol{\theta}|\tilde{S}_1(\cdot;\theta), \tilde{S}_2(\cdot;\theta), \tilde{S}_D(\cdot); \textit{Observed-Data})/\partial\boldsymbol{\theta} \Big\}, \tag{4.12}$$

respectively, and $\tilde{S}_j(t;\theta) = g(\tilde{S}_j^*(t), \tilde{S}_D(t);\theta)$.

As mentioned in Chapter 3, one may estimate the variance of $\hat{\boldsymbol{\theta}}_n$ by a bootstrap approach (e.g., Lawless & Yilmaz 2011). A natural and practical variance estimator evaluates (4.10) at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_n$ and uses (4.11) and (4.12) to replace their limits. The resulting variance estimator is often referred to as Huber's robust sandwich estimator (Huber 1967). Note that when $S_j(\cdot)$ for $j = 1, 2$ are known and used to estimate $\boldsymbol{\theta} = (\theta, \theta_{12})$, $\hat{\boldsymbol{\theta}}_n$ is an MLE, and $V_A(\boldsymbol{\theta})$ and $V_B(\boldsymbol{\theta})$ in (4.11) and (4.12) are the same as the corresponding inverse Fisher information matrix.

4.2.2 Resulting Estimators for Marginal and Joint Survivor Function

Substituting $\hat{\theta}_n$, $\tilde{S}_j^*(t)$, and $\tilde{S}_D(t)$ from the above section into (4.6) gives a natural estimator for the marginal survivor function $S_j(\cdot)$:

$$\hat{S}_{jn}(t) = g(\tilde{S}_j^*(t), \tilde{S}_D(t); \hat{\theta}_n). \tag{4.13}$$

for $j = 1, 2$.

Proposition 5. *Under the regularity conditions (RC1)–(RC4) presented in the Chapter 3 and provided* $\tilde{S}_j^*(\cdot)$ *and* $\tilde{S}_D(\cdot)$ *satisfy condition (AC1) in the Appendix, as* $n \to \infty$, $\hat{S}_{jn}(t) \xrightarrow{a.s.} S_j(t)$ *uniformly and* $\sqrt{n}(\hat{S}_{jn}(t) - S_j(t)) \xrightarrow{w} \mathcal{G}_j(t)$ *with* $t \in [0, v_j^\star]$, *where* $\mathcal{G}_j(t)$ *is a Gaussian process with mean zero and variance function* $\sigma_j^2(t)$ *as defined in, for example, Andersen et al. (1993).*

When the sample size is large and the censoring rate is not too high, we may choose to ignore the variation of the Kaplan–Meier estimates $\tilde{S}_j^*(\cdot)$ and $\tilde{S}_D(\cdot)$. This then yields an approximate confidence band (CB) for $S_j(\cdot)$ based on (4.6) with $\hat{\theta}_n$ substituted in, and using the proposed variance estimator of $\hat{\theta}_n$ in the above section.

In addition, the joint survivor function $S_{12}(t_1, t_2)$ of (T_1, T_2) based on (4.5):

$$\hat{S}_{12n}(t_1, t_2) = \mathcal{C}_{[2]}(\hat{S}_{1n}(t_1), \hat{S}_{2n}(t_2); \hat{\theta}_{12n}). \tag{4.14}$$

The following proposition establishes the consistency and asymptotic normality/weak convergence of the resulting estimator.

Proposition 6. *Under the regularity conditions (RC1)–(RC4) presented in Chapter 3 and provided* $\tilde{S}_j^*(t)$ *and* $\tilde{S}_D(t)$ *satisfy condition (AC1) in the Appendix,* *as* $n \to \infty$, $\hat{S}_n(t_1, t_2) \xrightarrow{a.s.} S(t_1, t_2)$ *uniformly and* $\sqrt{n}(\hat{S}_n(t_1, t_2) - S(t_1, t_2)) \xrightarrow{w} \mathcal{G}(t_1, t_2)$ *with*

$t_1, t_2 \in [0, v_1^\star] \times \times [0, v_2^\star]$, *where* $\mathcal{G}(t_1, t_2)$ *is a Gaussian field with mean zero and variance function* $\sigma^2(t_1, t_2)$.

4.3 Simulation Study

We conducted simulation studies to explore the finite-sample performance of the approach in section 4.2.

4.3.1 Simulation Setting and Data Generation

We simulated a study with n independent units where the primary outcome is the bivariate event time (T_1, T_2). The observations on (T_1, T_2) may be censored by either the terminating event time D or an administrative time C_A, whichever occurs first. That is, the study censoring time $C = D \wedge C_A$. We allow the association between T_1 and T_2 and that between (T_1, T_2) jointly with D to be different.

We simulated two main general settings to verify the performance of the proposed estimators and one additional setting to verify the performance in (4.7) as an example of non-copula formulation for the correlation between T_1 and T_2.

> *Setting 1:* We generated data from a nested Archimedean copula (see, e.g., Joe 1997) that allows the association parameter in the bivariate Archimedean copula ("outer" copula) that links (T_1, T_2) and D to be different from the association parameter in the bivariate copula ("inner" copula) that links T_1 and T_2.
>
> *Setting 2:* Since real-world data may not always fit an Archimedean copula family, we generated data from a non-Archimedean trivariate copula as part of the robustness check.
>
> *Setting 3:* Since the formulation of $\mathcal{C}_{[2]}(\cdot)$ does not necessarily have to be through copula functions, we generated data from a gamma frailty model to show the flexibility on model specification for (T_1, T_2) using our proposed approach.

To imitate potentially informative censoring due to a terminating event, we generated the data as follows:

For *setting 1* and *setting 2*,

> Step (a). We independently generated the trivariate random variables $((v_{1i}, v_{2i}), v_{3i})$ for $i = 1, \ldots, n$ from a nested Archimedean copula model using the `R` package `copula` (Hofert & Mächler 2011) for *setting 1* and from a trivariate Gaussian copula model for *setting 2*.

Step (b). We used the survivor functions of the Weibull distributions $S_j(\cdot)$ and $S_D(\cdot)$, where the scale and shape parameters mimic the corresponding event times and death times in the real example, to form the generated event times and terminating event times $t_{ji} = S_j^{-1}(v_{ji})$ with $j = 1, 2$ and $d_i = S_D^{-1}(v_{3i})$ for $i = 1, \ldots, n$.

Step (c). We generated the independent (administrative) censoring times c_{Ai} independently from (v_{1i}, v_{2i}, v_{3i}) from the exponential distribution with the parameter chosen to give a censoring rate of 25 percent. We then calculated $c_i = d_i \wedge c_{Ai}$ with the indicator $\delta_{Di} = I(d_i \leq c_{Ai})$ and $u_{ji} = t_{ji} \wedge c_i$ with the indicator $\delta_{ji} = I(t_{ji} \leq c_i)$.

Steps (a), (b), and (c) yield generated observed-data: $\{[(u_{ji}, \delta_{ji}) : j = 1, 2] \bigcup [c_i, \delta_{Di}] : i = 1, \ldots, n\}$

For *Setting 3*,

Step (a). We independently generated trivariate random variables $((v_{1i}, v_{2i}), v_{3i})$ for $i = 1, \ldots, n$ from a nested Archimedean copula model with $\theta = 1$ and $\theta_{12} = 0.8$.

Step (b). We fixed $(\theta_1, \theta_2, \theta_\eta)$ to specify the formulation in (4.7), such that the frailty ξ follows gamma distribution with shape and scale parameter equal to 1. Letting $v_{ji} = S_j(t_{ji}) = \int S_{0j}(t_{ji})^{\exp(\theta_1 \xi)} \eta(\xi) d\xi$, we can solve for $S_{0j}(t_{ji})$, denoted as w_{ji}, for $j = 1, 2$, where $S_{0j}(t)$ are the baseline survival function for T_1 and T_2. In addition, we let $v_{3i} = S_D(d_i)$. We used the survivor functions of the Weibull distributions for $S_{0j}(\cdot)$ and $S_D(\cdot)$, to form the generated event times and terminating event times $t_{ji} = S_{0j}^{-1}(w_{ji})$, and $d_i = S_D^{-1}(v_{3i})$.

Step (c) generated data following the same step (c) as above for *setting 1* and *setting 2*.

We considered the sample sizes $n = 500, 1000$, and 2000 to generate medium to large studies. In *setting 1*: The values of the outer and inner copula parameters (θ, θ_{12}) were set so that the corresponding Kendall's (τ, τ_{12}) are $(0.4, 0.5)$ and $(0.3, 0.8)$, representing weak and moderate-to-strong dependence structures. In *setting 2*: The values of the trivariate Gaussian parameters (θ, θ_{12}) were set so that Kendall's (τ, τ_{12}) are $(0.6, 0.8)$, $(0.8, 0.8)$, and $(0.8, 0.6)$ respectively, representing the moderate-to-strong dependence observed in the real-data example. We used the Kaplan–Meier estimator to obtain $\tilde{S}_j^*(\cdot)$ and $\tilde{S}_D(\cdot)$ in the estimation procedure. In *setting 3*: The frailty ξ follows gamma distribution with shape and scale parameter as 1. θ_1 and θ_2 are fixed at 1 but can be changed to any value as needed. θ is fixed as 0.8.

4.3.2 Consistency and Efficiency

We evaluated the pseudo-MLE of the association parameters (θ, θ_{12}) and the estimator for the marginal survivor function of T_j from section 4.2 with five hundred generated sets

of data from the nested Archimedean copula models. For comparison, we also found the MLE of (θ, θ_{12}) derived from likelihood function (4.3) using the true survivor functions $S_j(\cdot)$ and $S_D(\cdot)$, and the naïve estimates obtained by maximizing (4.3) after substituting the marginal survivor functions by their Kaplan–Meier estimates. In the settings that this chapter focuses on, the MLE is not applicable, and the naïve estimator can be biased because of the informative censoring.

Table 4.1 presents a summary of the simulation outcomes based on the five hundred repetitions under the nested Clayton and the nested Frank models, with two different combinations of values for θ and θ_{12}. The sample means of the pseudo-MLE and MLE estimates are close to the true parameter values, especially when n is large. This verifies the consistency of the pseudo-MLE and the MLE. The sample means of the naïve estimates, on the other hand, are quite different from the true values. The sample standard errors of the pseudo-MLE are larger than but comparable with their MLE counterparts, which indicates that the pseudo-MLE has satisfactory efficiency in the simulation settings.

The six plots in figure 4.1 correspond to simulated studies under the nested Clayton copula model with Kendall's $\tau = 0.4$ for the outer copula and $\tau_{12} = 0.5$ for the inner copula; the sample sizes are $n = 500, 1000$, and 2000. The three plots in the upper and lower rows correspond to the curves for $S_1(\cdot)$ and $S_2(\cdot)$, respectively. Each plot shows the true curve of the marginal survivor function $S_j(\cdot)$ and the two sets of estimates with the generated semicompeting-risks data, using the proposed pseudo-MLE or the naïve approach. The two sets of approximate 95% CBs for $S_j(\cdot)$ are also presented. The true $S_j(\cdot)$ curve is fully covered by the CB associated with the pseudo-MLE in every plot. It is not within the CB associated with the naïve estimator, which requires the assumption of noninformative censoring which is in fact not valid in the simulation settings. This pattern becomes clearer as the sample size increases. The same patterns are observed when $\tau = 0.3$, $\tau_{12} = 0.8$ in the nested Clayton model. The simulation outcomes with the nested Frank models for all the combinations of (τ, τ_{12}) agree with those for the nested Clayton copula.

4.3.3 Robustness to Model Misspecification

To examine the pseudo-MLE's robustness to model misspecification, we generated data under nested Clayton copulas and nested Frank copulas and evaluated the pseudo-MLE. To compare the estimates, we present Kendall's τ and τ_{12} since it is a universal metric of dependence for different models. Table 4.2-4.3 summarizes the sets of pseudo-MLE estimates based on five hundred generated data sets. Some biases occur across different simulated studies under both misspecified copula models. However, the biases of the resulting estimates compared to the true Kendall's τ and τ_{12} values appear to be insignificant, especially when the nested Frank copula is used to evaluate the estimators.

The six plots in figure 4.5 correspond to sets of estimates of the marginal survivor functions $S_1(\cdot)$, using correct and misspecified inner and outer copulas. The naïve estimates

of the marginal survivor functions are biased, but the estimates from the other misspecified models are close to the true marginals, especially when the outer copula model is correctly specified. Plots from other types of copula for both $S_1(\cdot)$ and $S_2(\cdot)$ are shown from figure 4.6 to figure 4.12.

In addition, we used the data sets generated from *setting 2*, i.e., from trivariate Gaussian copula models, to evaluate the robustness of the proposed approach when the true model is not Archimedean. Table 4.4- Table 4.6 summarize the sets of pseudo-MLE estimates based on five hundred generated data sets. Some biases were observed under misspecified copula models. However, the biases appear to be insignificant for most of the nested Archimedean models, especially for the nested Frank copula model, and when the inner copula is Gaussian and the outer copula is Frank. Similar to the conclusion in Chapter 3, Frank copula seems to be a flexible Archimedean copula to use in practice.

4.3.4 Flexibility on Modeling Correlation Between Event Times

Table 4.7 presents the sample mean and sample standard error of the estimates for the parameter set $(\theta_1, \theta_2, \theta_\eta, \theta)$ in model (4.7), for different sample sizes, based on five hundred generated datasets. Figure 4.13 shows six plots corresponding to the marginals for $S_1(\cdot)$ and $S_2(\cdot)$. Each plot shows the true curves and the estimated marginals with confidence bands (CB). The sample means of the parameter estimates are close to the corresponding true values for the parameters. The true curves for marginals are covered the CBs in every plot.

Table 4.1: Consistency Study. Estimation of Association Parameters τ and τ_{12} with Simulated Data from Nested Archimedean Copulas, Based on 500 Repetitions.

			Nested Clayton			**Nested Frank**		
Parameter	True Value	Estimates	$n = 500$	$n = 1000$	$n = 2000$	$n = 500$	$n = 1000$	$n = 2000$
					Pseudo-MLE			
τ_{12}	0.50	mean	0.49	0.49	0.50	0.50	0.50	0.50
		se	.027	.019	.013	.024	.016	.012
τ	0.40	mean	0.39	0.39	0.40	0.40	0.40	0.40
		se	.029	.019	.013	.025	.017	.013
τ_{12}	0.80	mean	0.79	0.79	0.80	0.80	0.80	0.80
		se	.013	.010	.006	.010	.007	.005
τ	0.30	mean	0.29	0.29	0.30	0.29	0.30	0.30
		se	.032	.024	.016	.033	.023	.016
					MLE using True Marginals			
τ_{12}	0.50	mean	0.50	0.50	0.50	0.50	0.50	0.50
		se	.023	.015	.010	.023	.014	.011
τ	0.40	mean	0.40	0.40	0.40	0.39	0.39	0.39
		se	.021	.013	.010	.020	.014	.010
τ_{12}	0.80	mean	0.80	0.80	0.80	0.80	0.80	0.80
		se	.009	.006	.005	.010	.007	.004
τ	0.30	mean	0.30	0.30	0.30	0.30	0.30	0.30
		se	.026	.019	.013	.031	.023	.016
					Naive			
τ_{12}	0.50	mean	0.51	0.53	0.54	0.51	0.51	0.51
		se	.045	.024	.015	.024	.017	.012
τ	0.40	mean	0.35	0.37	0.37	0.37	0.37	0.38
		se	.031	.018	.012	.023	.016	.012
τ_{12}	0.80	mean	0.79	0.80	0.81	0.80	0.80	0.80
		se	.027	.015	.009	.010	.007	.005
τ	0.30	mean	0.25	0.27	0.27	0.28	0.28	0.28
		se	.033	.021	.015	.032	.023	.016

Table 4.2: Robustness Study. Estimation of Kendall's τ and τ_{12} with Simulated Data from Nested Archimedean Copulas Using Different Copulas for Robustness Check with $\mathcal{A}_{[2]}$ = Clayton, Based on 500 Repetitions.

				$\mathcal{A}_{[2]}$: Clayton								
				$\mathcal{C}_{[2]}$: Clayton			$\mathcal{C}_{[2]}$: Frank			$\mathcal{C}_{[2]}$: Gaussian		
$\mathcal{A}_{[2]}$	$\mathcal{C}_{[2]}$			n=500	n=1000	n=2000	n=500	n=1000	n=2000	n=500	n=1000	n=2000
Clayton	Clayton	$\tau_{12} = 0.5$	τ_{12}	0.49	0.49	0.50	0.49	0.48	0.48	0.46	0.46	0.46
			se	.026	.018	.011	.029	.022	.014	.031	.022	.015
		$\tau = 0.4$	τ	0.39	0.39	0.40	0.35	0.34	0.35	0.39	0.38	0.38
			se	.029	.019	.012	.043	.032	.021	.045	.035	.031
		$\tau_{12} = 0.8$	τ_{12}	0.79	0.79	0.80	0.78	0.78	0.78	0.72	0.71	0.71
			se	.013	.009	.007	.015	.011	.009	.016	.014	.010
		$\tau = 0.3$	τ	0.30	0.29	0.30	0.26	0.26	0.26	0.27	0.26	0.26
			se	.034	.023	.017	.033	.024	.018	.033	.023	.018
Frank	Frank	$\tau_{12} = 0.5$	τ_{12}	0.33	0.32	0.32	0.47	0.47	0.47	0.42	0.42	0.42
			se	.037	.030	.026	.032	.019	.017	.031	.023	.020
		$\tau = 0.4$	τ	0.28	0.29	0.28	0.27	0.27	0.27	0.35	0.30	0.29
			se	.051	.048	.045	.048	.036	.030	.066	.065	.056
		$\tau_{12} = 0.8$	τ_{12}	0.68	0.68	0.67	0.80	0.80	0.80	0.74	0.74	0.74
			se	.029	.048	.014	.024	.017	.005	.015	.013	.006
		$\tau = 0.3$	τ	0.20	0.18	0.19	0.15	0.15	0.19	0.19	0.18	0.19
			se	.026	.042	.013	.025	.024	.013	.026	.021	.013

Table 4.3: Robustness Study. Estimation of Kendall's τ and τ_{12} with Simulated Data from Nested Archimedean Copulas Using Different Copulas for Robustness Check with $\mathcal{A}_{[2]}$ = Frank, Based on 500 Repetitions.

				$\mathcal{A}_{[2]}$: Frank								
				$\mathcal{C}_{[2]:Clayton}$			$\mathcal{C}_{[2]:Frank}$			$\mathcal{C}_{[2]:Gaussian}$		
$\mathcal{A}_{[2]}$	$\mathcal{C}_{[2]}$			n=500	n=1000	n=2000	n=500	n=1000	n=2000	n=500	n=1000	n=2000
Clayton	Clayton	$\tau_{12} = 0.5$	τ_{12}	0.52	0.52	0.52	0.50	0.50	0.50	0.46	0.46	0.47
			se	.038	.033	.031	.027	.019	.012	.043	.029	.011
		$\tau = 0.4$	τ	0.42	0.43	0.44	0.40	0.41	0.41	0.40	0.39	0.39
			se	.064	.063	.068	.032	.019	.014	.059	.039	.014
		$\tau_{12} = 0.8$	τ_{12}	0.79	0.79	0.79	0.77	0.77	0.78	0.71	0.70	0.71
			se	.016	.009	.007	.015	.011	.008	.016	.014	.010
		$\tau = 0.3$	τ	0.28	0.28	0.28	0.30	0.29	0.30	0.28	0.27	0.27
			se	.039	.026	.020	.040	.027	.022	.040	.027	.022
Frank	Frank	$\tau_{12} = 0.5$	τ_{12}	0.39	0.35	0.34	0.50	0.50	0.50	0.46	0.45	0.47
			se	.132	.029	.025	.027	.015	.012	.044	.043	.018
		$\tau = 0.4$	τ	0.45	0.49	0.51	0.39	0.40	0.40	0.40	0.42	0.38
			se	.066	.048	.032	.031	.018	.012	.066	.069	.028
		$\tau_{12} = 0.8$	τ_{12}	0.73	0.73	0.73	0.80	0.80	0.80	0.75	0.75	0.75
			se	.021	.017	.011	.011	.007	.005	.013	.012	.006
		$\tau = 0.3$	τ	0.26	0.25	0.26	0.30	0.29	0.30	0.28	0.27	0.28
			se	.028	.023	.016	.028	.025	.016	.027	.025	.016

Table 4.4: Robustness Study. Estimation of Kendall's τ and τ_{12} with Simulated Data from Trivariate Gaussian Copulas with $\tau = 0.8$ and$\tau_{12} = 0.8$. Based on 500 Repetitions.

			True Model: $\mathcal{A}_{[2]}$:Gaussian, $\mathcal{C}_{[2]}$:Gaussian					
		n	500		1000		2000	
			τ	τ_{12}	τ	τ_{12}	τ	τ_{12}
$\mathcal{A}_{[2]}$: Clayton	$\mathcal{C}_{[2]}$: Clayton	*sm*	0.70	0.68	0.69	0.68	0.69	0.68
		sse	.027	.026	.020	.020	.016	.017
	$\mathcal{C}_{[2]}$: Gaussian	*sm*	0.71	0.76	0.7	0.76	0.7	0.76
		sse	.029	.016	.021	.012	.016	.009
	$\mathcal{C}_{[2]}$: Frank	*sm*	0.67	0.78	0.67	0.78	0.67	0.78
		sse	.022	.016	.016	.011	.012	.008
$\mathcal{A}_{[2]}$: Frank	$\mathcal{C}_{[2]}$: Clayton	*sm*	0.79	0.68	0.79	0.68	0.79	0.68
		sse	.014	.021	.011	.015	.009	.011
	$\mathcal{C}_{[2]}$: Gaussian	*sm*	0.80	0.77	0.80	0.78	0.80	0.78
		sse	.013	.016	.010	.011	.009	.009
	$\mathcal{C}_{[2]}$: Frank	*sm*	0.78	0.78	0.78	0.78	0.78	0.79
		sse	.016	.020	.014	.016	.010	.012

Table 4.5: Robustness Study. Estimation of Kendall's τ and τ_{12} with Simulated Data from Trivariate Gaussian Copulas with $\tau = 0.6$ and$\tau_{12} = 0.8$. Based on 500 Repetitions.

			True Model: $\mathcal{A}_{[2]}$:Gaussian, $\mathcal{C}_{[2]}$:Gaussian					
		n	500		1000		2000	
			τ	τ_{12}	τ	τ_{12}	τ	τ_{12}
$\mathcal{A}_{[2]}$: Clayton	$\mathcal{C}_{[2]}$: Clayton	*sm*	0.45	0.69	0.45	0.70	0.45	0.70
		sse	.026	.020	.018	.014	.014	.010
	$\mathcal{C}_{[2]}$: Gaussian	*sm*	0.44	0.78	0.44	0.79	0.44	0.79
		sse	.030	.011	.019	.008	.014	.006
	$\mathcal{C}_{[2]}$: Frank	*sm*	0.42	0.78	0.42	0.78	0.42	0.79
		sse	.072	.026	.025	.010	.019	.008
$\mathcal{A}_{[2]}$: Frank	$\mathcal{C}_{[2]}$: Clayton	*sm*	0.54	0.74	0.54	0.74	0.54	0.74
		sse	.035	.020	.017	.011	.011	.008
	$\mathcal{C}_{[2]}$: Gaussian	*sm*	0.58	0.79	0.58	0.80	0.58	0.80
		sse	.048	.020	.021	.010	.013	.006
	$\mathcal{C}_{[2]}$: Frank	*sm*	0.65	0.73	0.64	0.74	0.63	0.75
		sse	.075	.059	.071	.058	.065	.058

Table 4.6: Robustness Study. Estimation of Kendall's τ and τ_{12} with Simulated Data from Trivariate Gaussian Copulas with $\tau = 0.8$ and$\tau_{12} = 0.6$. Based on 500 Repetitions.

			True Model: $\mathcal{A}_{[2]}$:Gaussian, $\mathcal{C}_{[2]}$:Gaussian					
		n	500		1000		2000	
			τ	τ_{12}	τ	τ_{12}	τ	τ_{12}
$\mathcal{A}_{[2]}$: Clayton	$\mathcal{C}_{[2]}$: Clayton	sm	0.75	0.48	0.76	0.49	0.76	0.49
		sse	.014	.025	.011	.019	.008	.014
	$\mathcal{C}_{[2]}$: Gaussian	sm	0.75	0.59	0.75	0.59	0.75	0.6
		sse	.013	.019	.010	.014	.007	.010
	$\mathcal{C}_{[2]}$: Frank	sm	0.72	0.6	0.72	0.61	0.72	0.61
		sse	.015	.021	.011	.014	.008	.011
$\mathcal{A}_{[2]}$: Frank	$\mathcal{C}_{[2]}$: Clayton	sm	0.80	0.52	0.80	0.52	0.80	0.52
		sse	.009	.023	.006	.016	.005	.012
	$\mathcal{C}_{[2]}$: Gaussian	sm	0.82	0.59	0.82	0.59	0.82	0.59
		sse	.009	.020	.007	.014	.006	.011
	$\mathcal{C}_{[2]}$: Frank	sm	0.83	0.59	0.83	0.60	0.83	0.60
		sse	.010	.021	.006	.014	.005	.010

Table 4.7: Estimation of $(\theta_1, \theta_2, \theta_\eta, \theta)$ with Simulated Data From (4.7).

		θ_1	θ_2	θ_η	θ_0
Real		1.00	1.00	1.00	0.80
$n = 500$	$sm^\dagger$	0.98	0.95	0.997	0.79
	$sse^\ddagger$	0.147	0.151	0.159	0.14
$n = 1000$	sm	0.97	0.96	1.002	0.79
	sse	0.124	0.118	0.149	0.172
$n = 2000$	sm	0.99	0.98	0.988	0.77
	sse	0.115	0.103	0.121	0.145

$sm^\dagger$: sample mean

$sse^\ddagger$: sample standard error

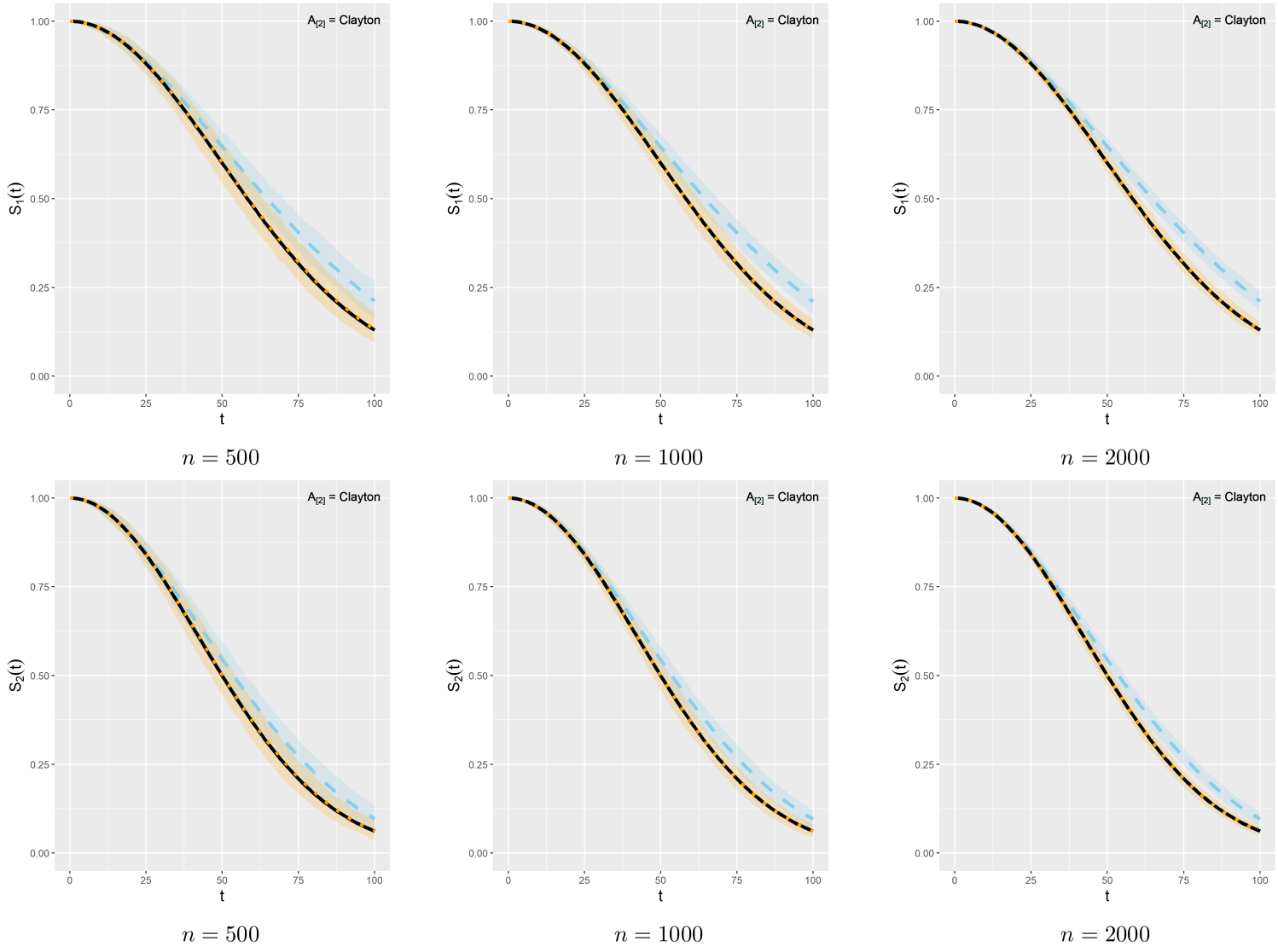


Figure 4.1: Consistency Study of Marginal Estimates with Data Simulated from Nested Clayton Copula with $\tau = 0.4, \tau_{12} = 0.5$. Upper row: S_1; bottom row: S_2. Skyblue dashed: naïve. Black solid: true. Orange dotted: Clayton.

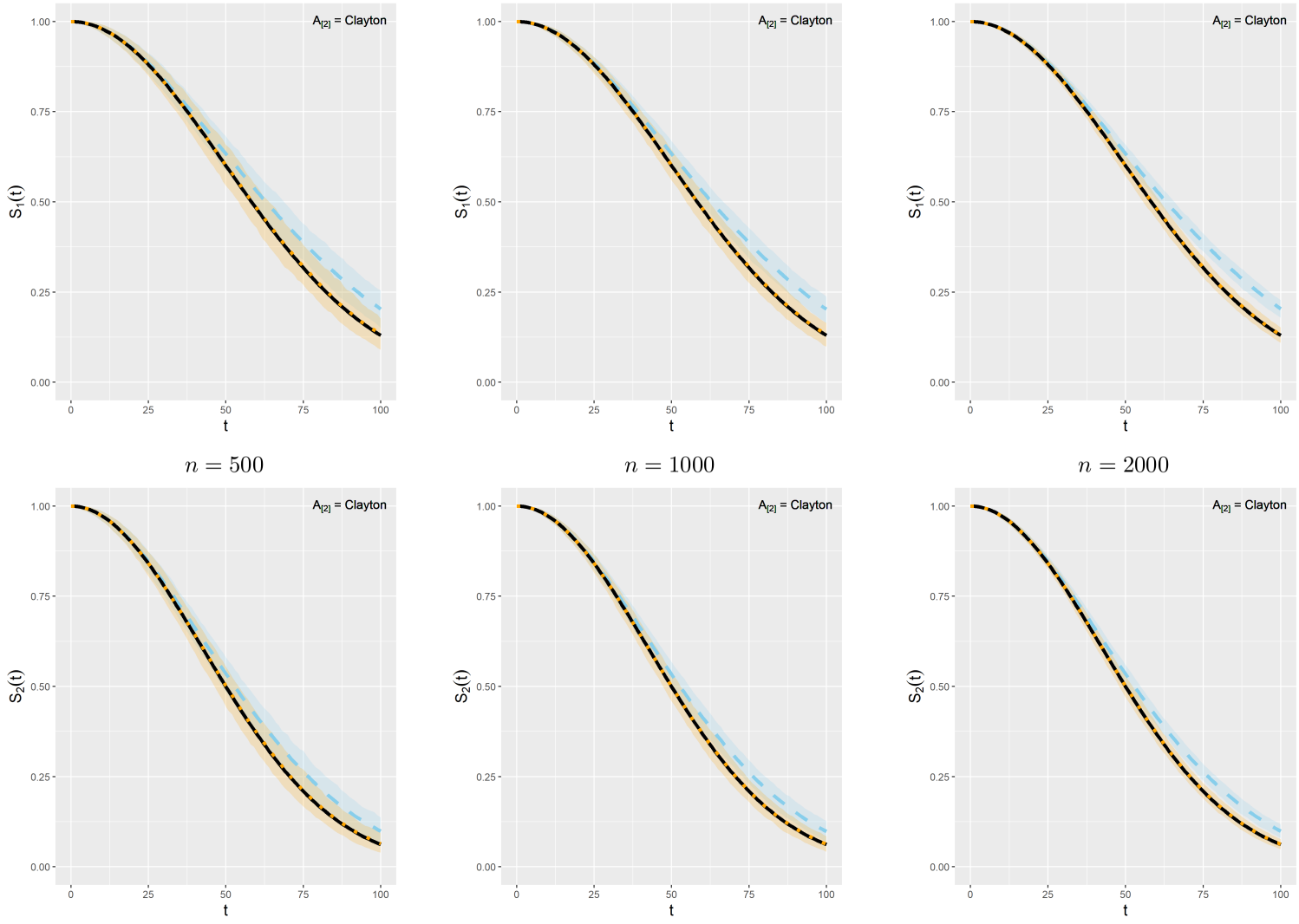


Figure 4.2: Consistency Study of Marginal Estimates with Data Simulated from Nested Clayton Copula with $\tau = 0.3, \tau_{12} = 0.8$. Upper Row: S_1; Bottom Row: S_2. Skyblue dashed: naïve. Black solid: true. Orange dotted: Clayton.

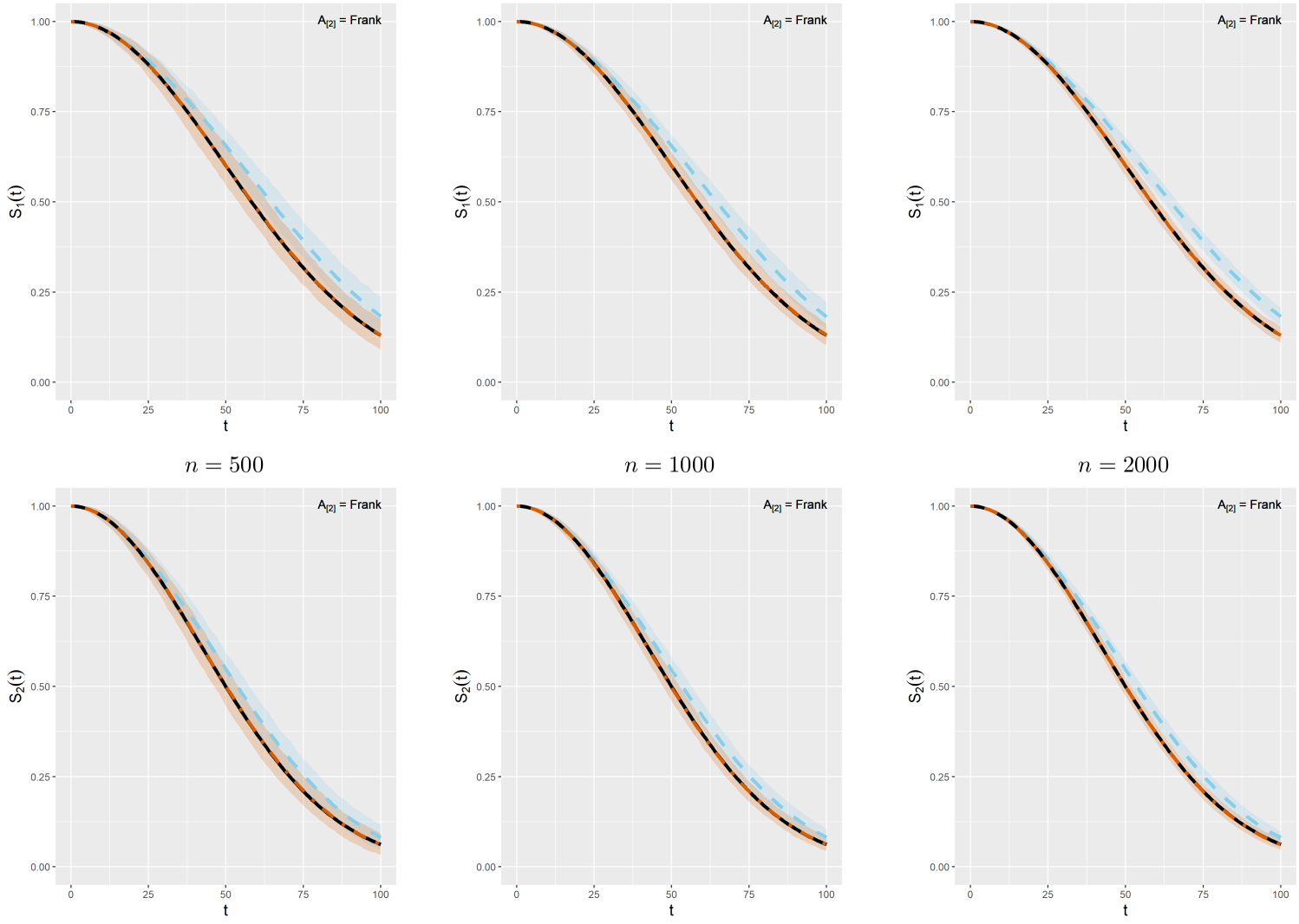


Figure 4.3: Consistency Study of Marginal Estimates with Data Simulated from Nested Frank Copula with $\tau = 0.4, \tau_{12} = 0.5$. Upper Row: S_1; Bottom Row: S_2. Skyblue dashed: naïve. Black solid: true. Vermillion dotdash: Frank.

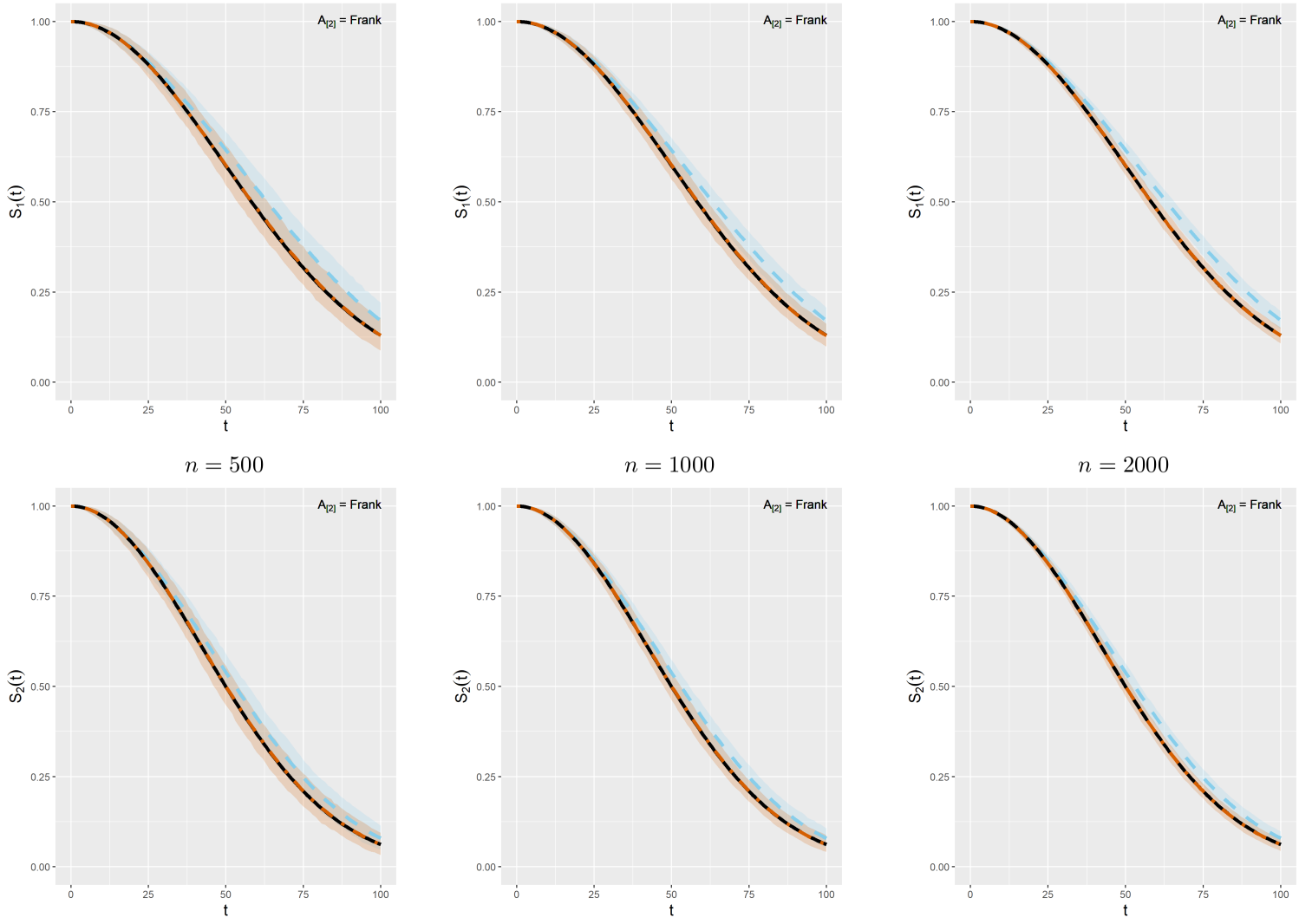
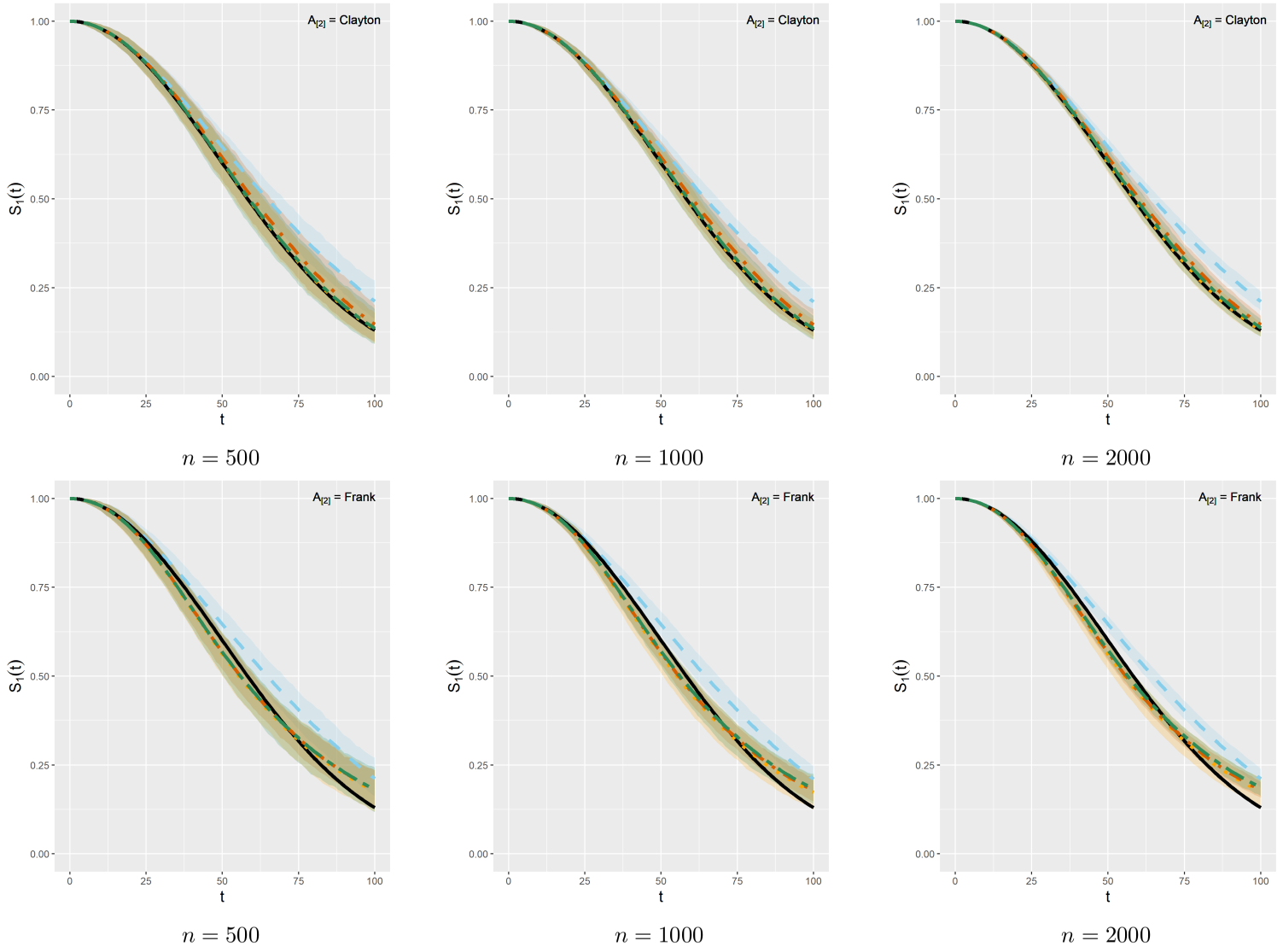


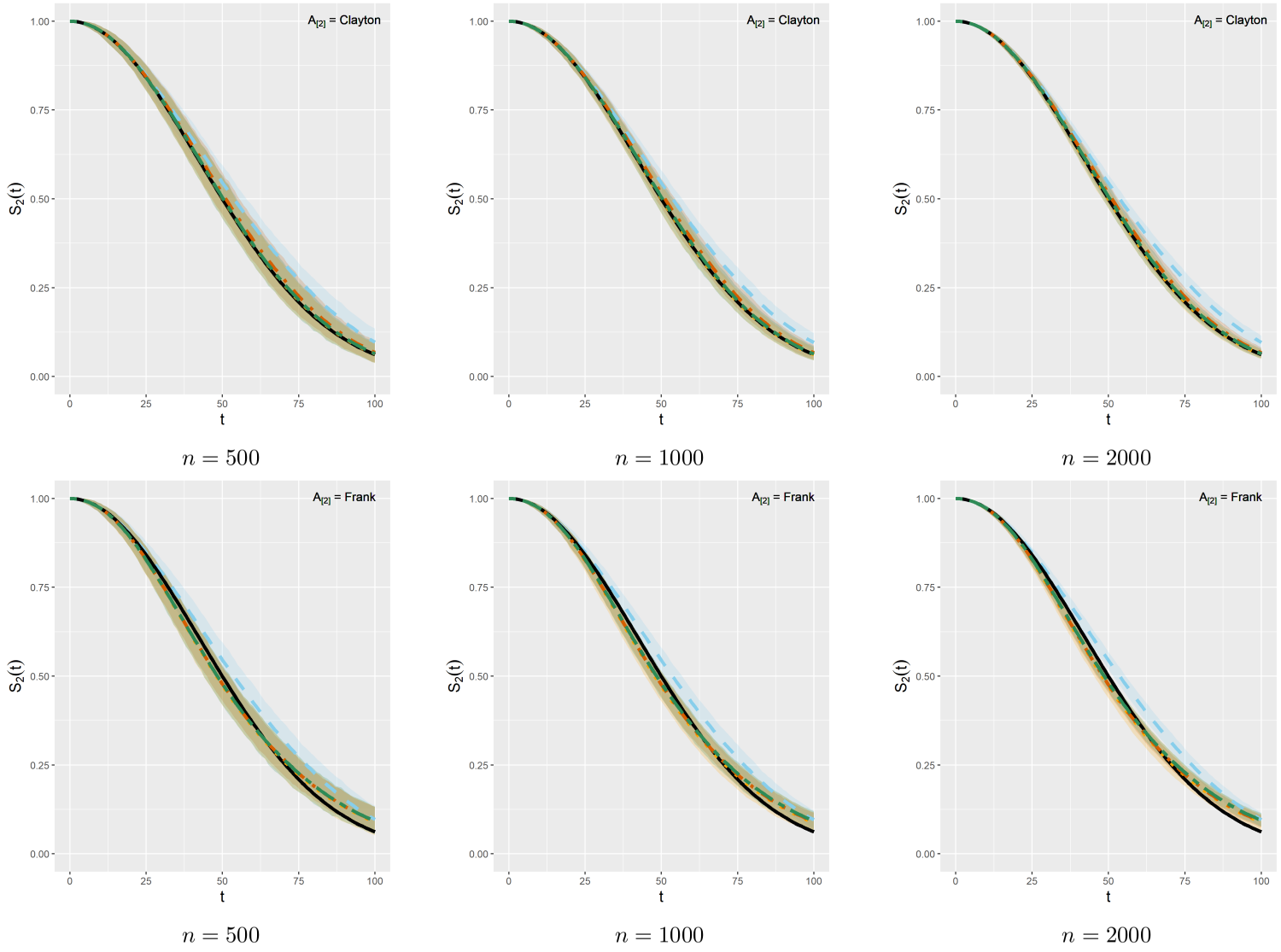
Figure 4.4: Consistency Study of Marginal Estimates with Data Simulated from Nested Frank Copula with $\tau = 0.3, \tau_{12} = 0.8$. Upper Row: S_1; Bottom Row: S_2. Skyblue dashed: naïve. Black solid: true. Vermillion dotdash: Frank.



$n = 500$ $n = 1000$ $n = 2000$

$n = 500$ $n = 1000$ $n = 2000$

Figure 4.5: Robustness to Model Misspecification for Marginal Function Estimates of $S_1(\cdot)$. Data Simulated from Nested Clayton with $\tau = 0.4, \tau_{12} = 0.5$. Skyblue dashed: naïve. Black solid: true. Orange dotted: Clayton. Vermillion dotdash: Frank. Seagreen twodash: Gaussian.



$n = 500$ $n = 1000$ $n = 2000$

$n = 500$ $n = 1000$ $n = 2000$

Figure 4.6: Robustness to Model Misspecification for Marginal Function Estimates of $S_2(\cdot)$. Data Simulated from Nested Clayton Coupla with $\tau = 0.4, \tau_{12} = 0.5$. Skyblue dashed: naïve. Black solid: true. Orange dotted: Clayton. Vermillion dotdash: Frank. Seagreen twodash: Gaussian.

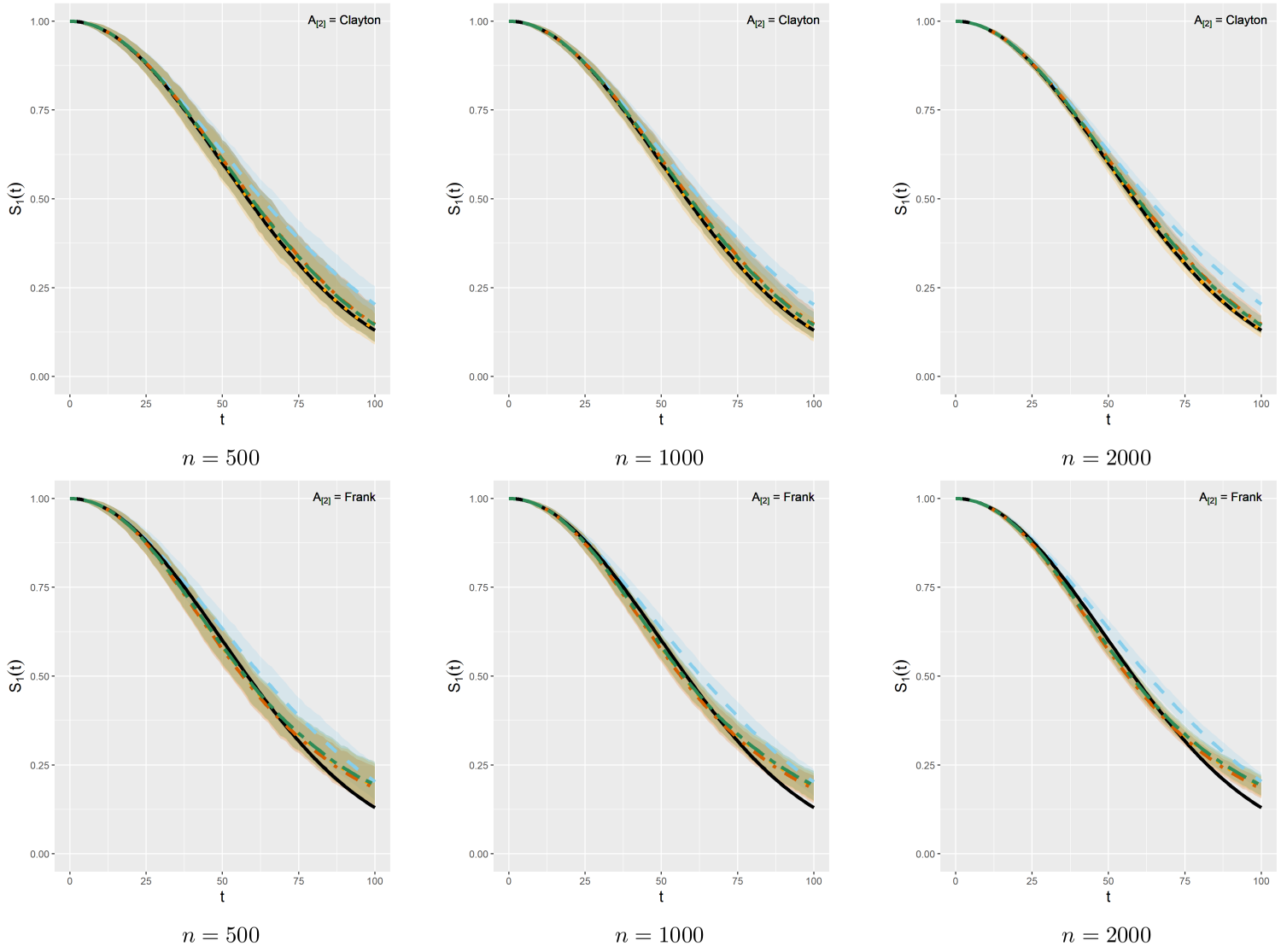


Figure 4.7: Robustness to Model Misspecification for Marginal Function Estimates of $S_1(\cdot)$. Data Simulated from Nested Clayton with $\tau = 0.3, \tau_{12} = 0.8$. Skyblue dashed: naïve. Black solid: true. Orange dotted: Clayton. Vermillion dotdash: Frank. Seagreen twodash: Gaussian.

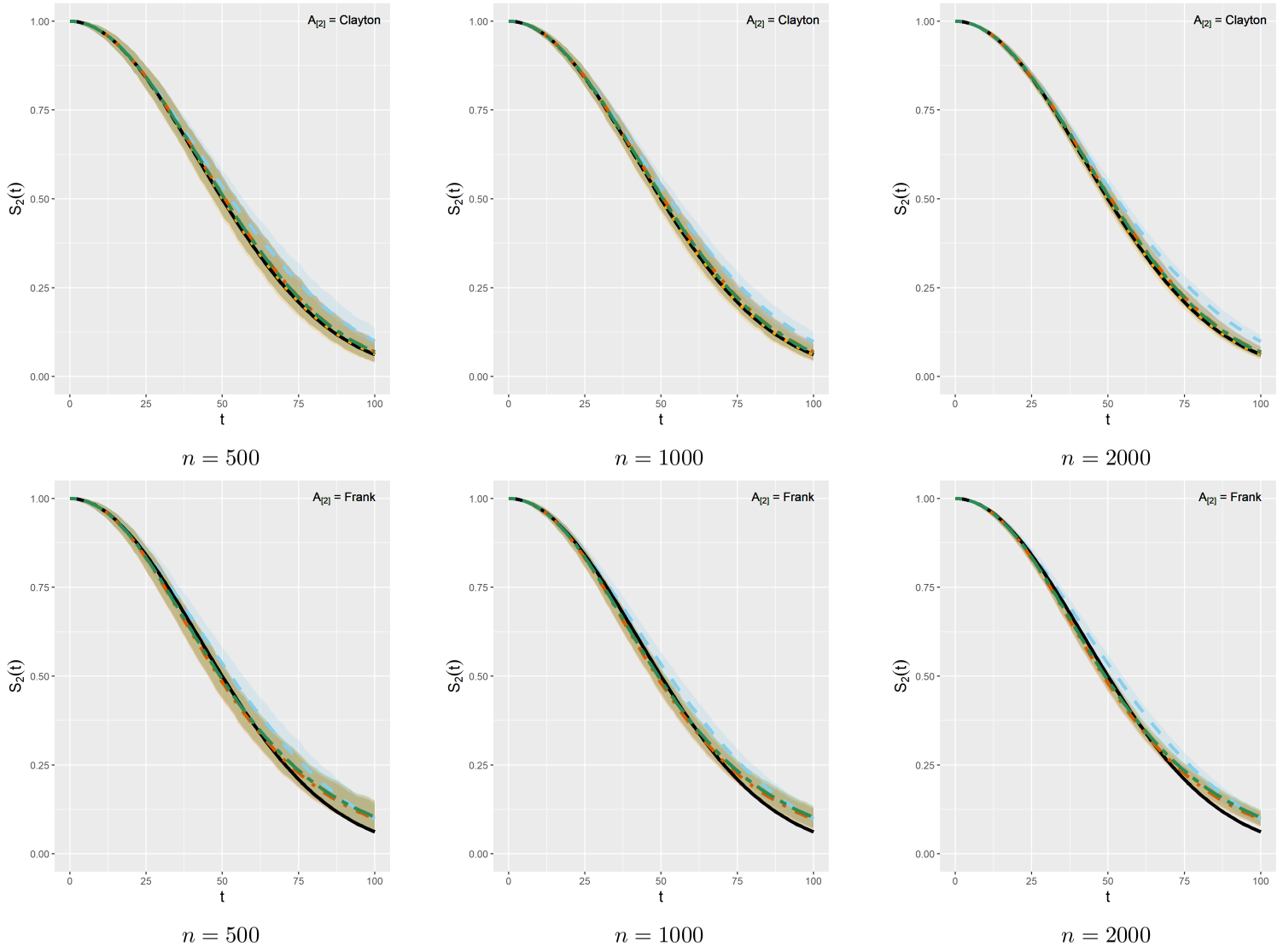


Figure 4.8: Robustness to Model Misspecification for Marginal Function Estimates of $S_2(\cdot)$. Data Simulated from Nested Clayton with $\tau = 0.3, \tau_{12} = 0.8$. Skyblue dashed: naïve. Black solid: true. Orange dotted: Clayton. Vermillion dotdash: Frank. Seagreen twodash: Gaussian.

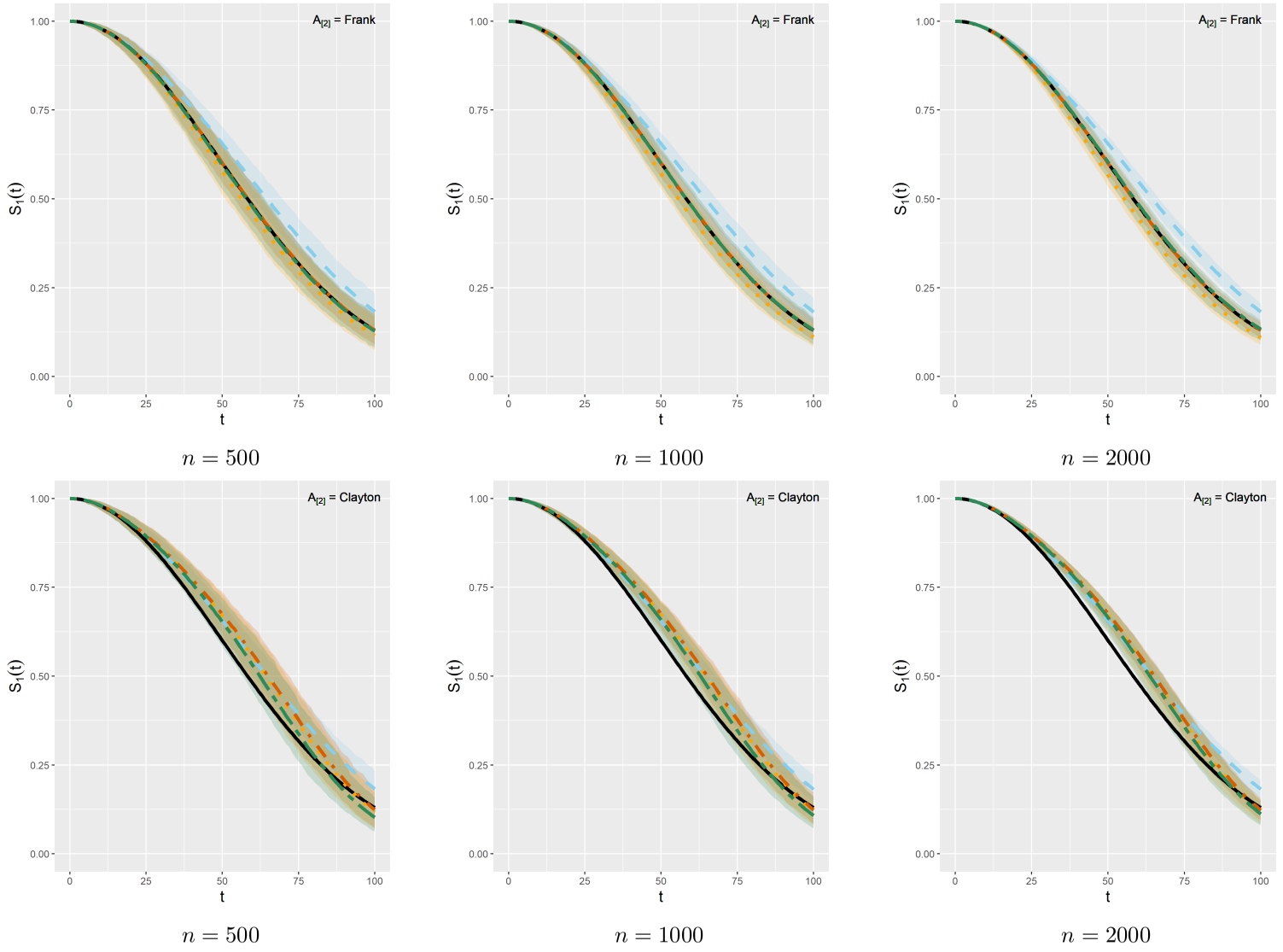


Figure 4.9: Robustness to Model Misspecification for Marginal Function Estimates of $S_1(\cdot)$. Data Simulated from Nested Frank with $\tau = 0.4, \tau_{12} = 0.5$. Skyblue dashed: naïve. Black solid: true. Orange dotted: Clayton. Vermillion dotdash: Frank. Seagreen twodash: Gaussian.

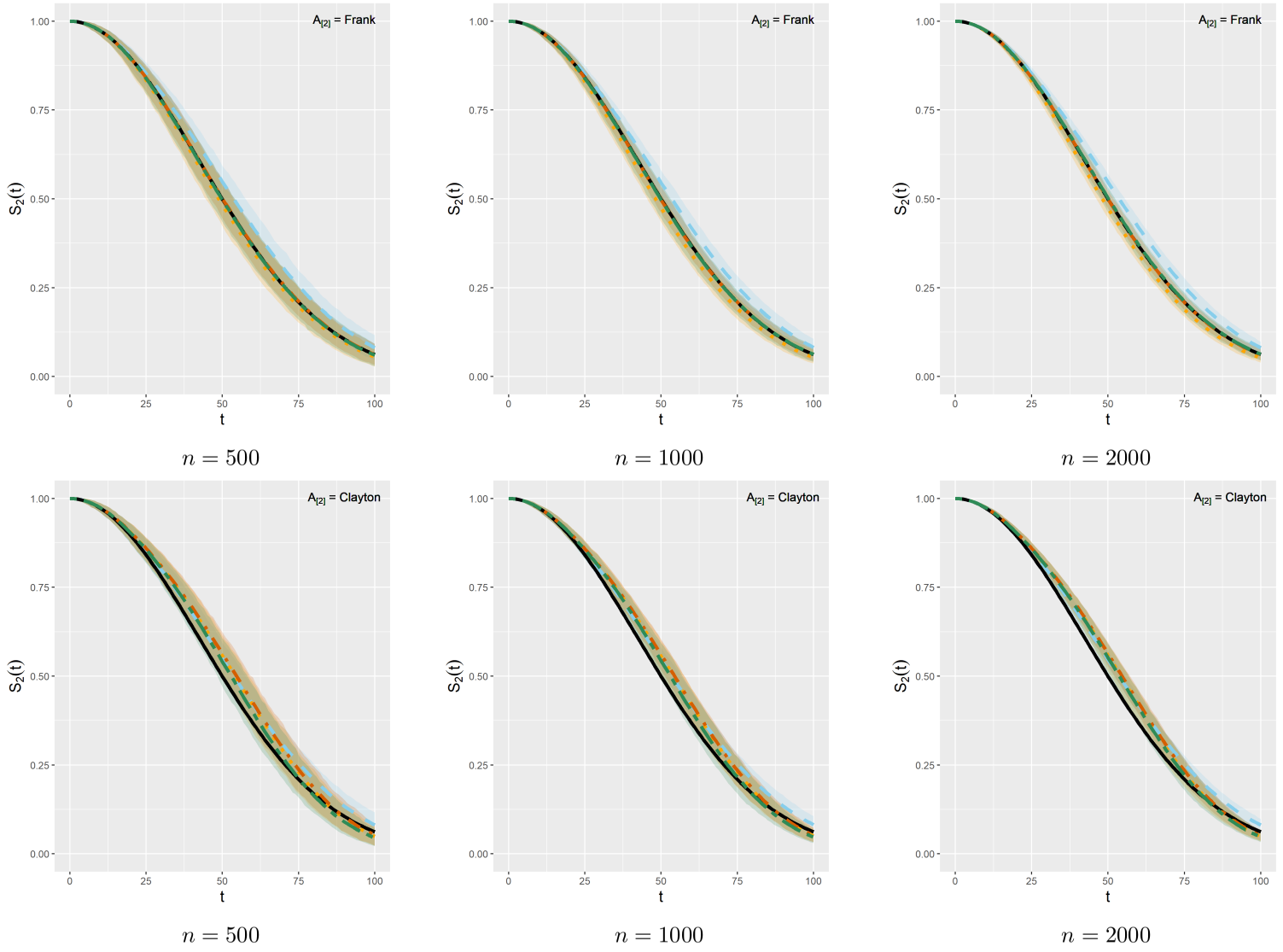
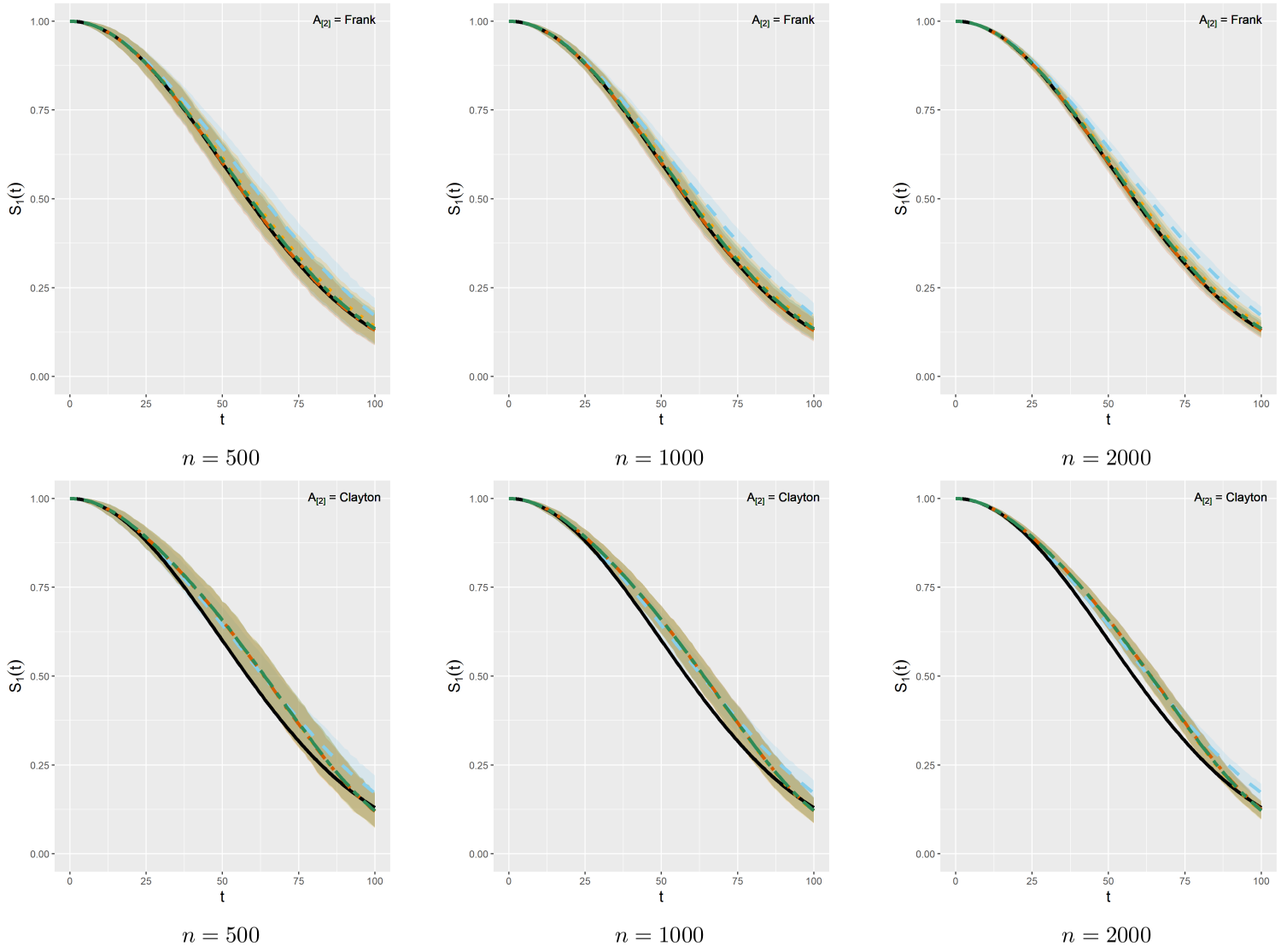


Figure 4.10: Robustness to Model Misspecification for Marginal Function Estimates of $S_2(\cdot)$. Data Simulated from Nested Frank with $\tau = 0.4, \tau_{12} = 0.5$. Skyblue dashed: naïve. Black solid: true. Orange dotted: Clayton. Vermillion dotdash: Frank. Seagreen twodash: Gaussian.



$n = 500$ $n = 1000$ $n = 2000$

$n = 500$ $n = 1000$ $n = 2000$

Figure 4.11: Robustness to Model Misspecification for Marginal Function Estimates of $S_1(\cdot)$. Data Simulated from Nested Frank with $\tau = 0.3, \tau_{12} = 0.8$. Skyblue dashed: naïve. Black solid: true. Orange dotted: Clayton. Vermillion dotdash: Frank. Seagreen twodash: Gaussian.

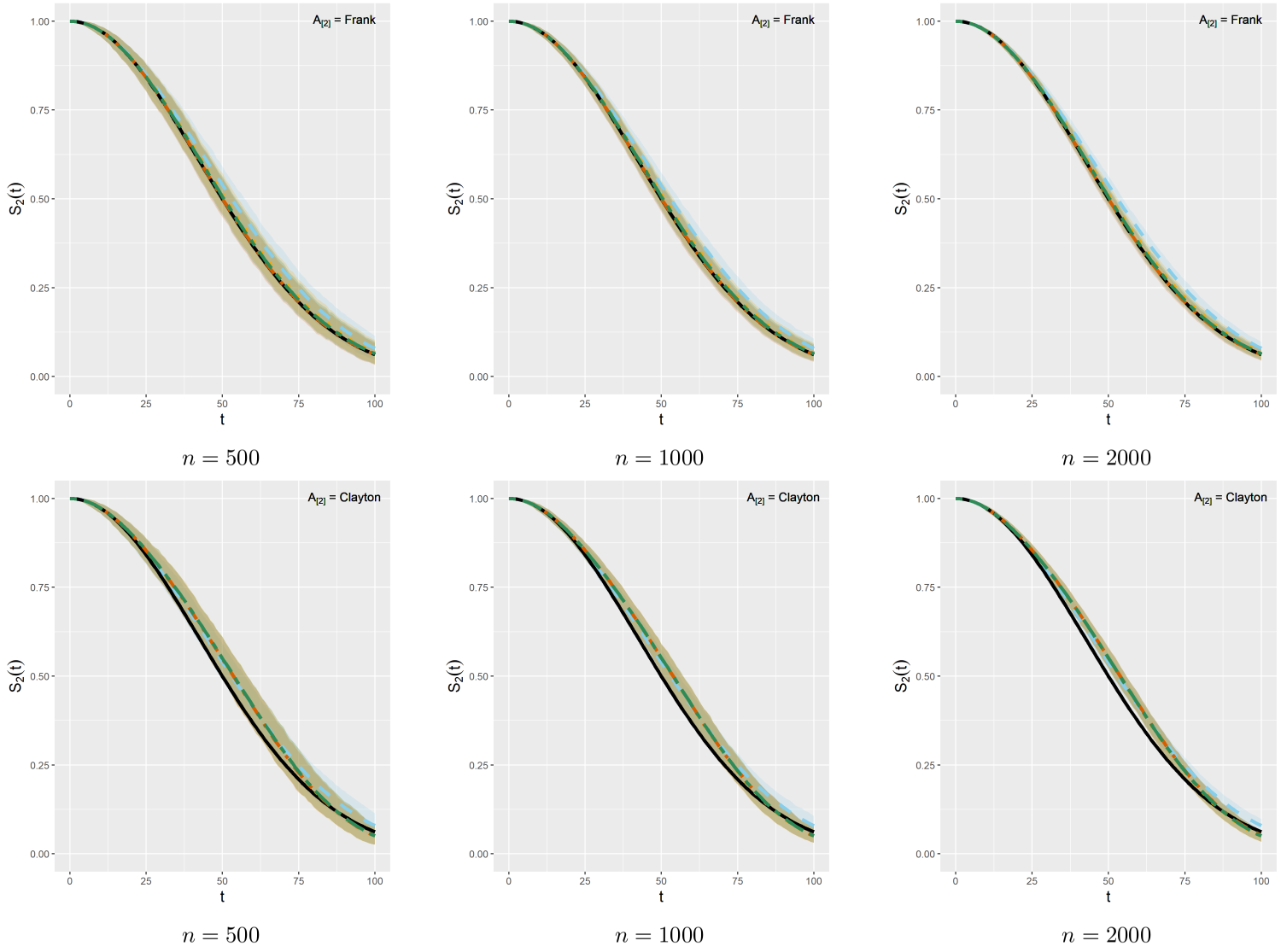


Figure 4.12: Robustness to Model Misspecification for Marginal Function Estimates of $S_2(\cdot)$. Data Simulated from Nested Frank with $\tau = 0.3, \tau_{12} = 0.8$. Skyblue dashed: naïve. Black solid: true. Orange dotted: Clayton. Vermillion dotdash: Frank. Seagreen twodash: Gaussian.

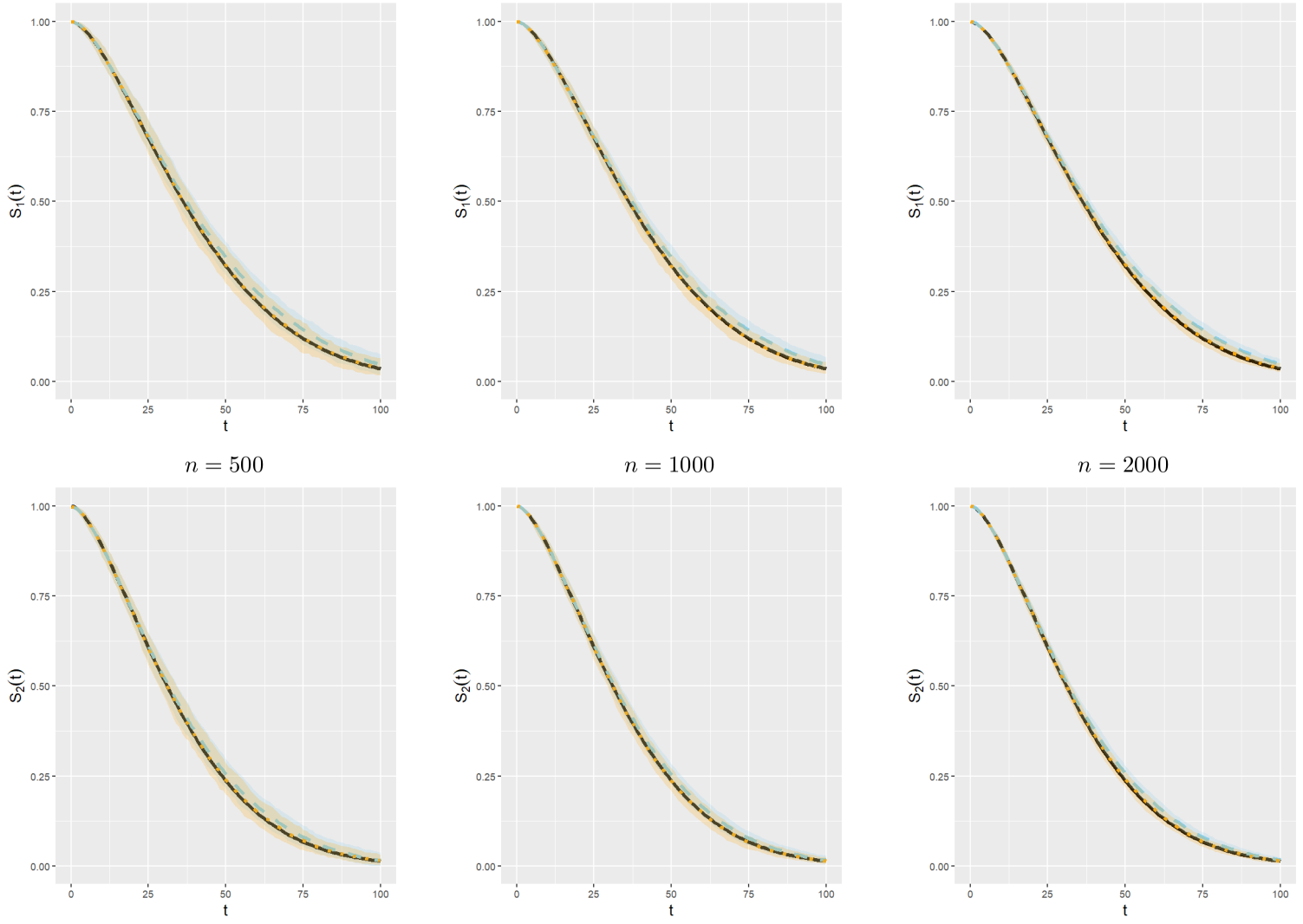


Figure 4.13: Marginal Function Estimates with Simulated Data Generated from Model (4.7). Upper Row: S_1; Bottom Row: S_2. Skyblue dashed: naïve. Black solid: true. Orange dotted: Proposed

4.4 Analysis of BC-BRCAS Data (III)

To illustrate the approach proposed in this chapter, we present an analysis of data from the BC-BRCAS study (McBride et al. 2016).

4.4.1 Study Description

The study subjects are the same as those in Chapter 3. Table 3.14 presents a summary of the available data on T_1, T_2, and D.

4.4.2 Estimates of Correlations between Event Times

Recall that our analysis of the BC-BRCAS study aims to evaluate the correlation between a breast cancer patient's time to RSC (T_1) and her time to a CVD event (T_2) after the cancer diagnosis. In this section, we estimated Kendall's τ_{12}, the measure of association between T_1 and T_2, under a copula model embedded within an Archimedean copula model.

Specifically, we specified the joint survivor function of (T_1, T_2) with the available study data through a copula model ("inner" copula) with parameter θ_{12}, and we specified the joint survivor function of (T_1, T_2) with D through another copula model ("outer" copula) belonging to the Archimedean copula family with parameter θ. The inner copula does not need to be an Archimedean copula, and the association parameter can be different from that for the outer copula. We implemented the pseudo-MLE procedure described in section 4.2 to estimate θ_{12} and θ, the associated standard error, and the marginal survivor functions $S_j(\cdot)$ using the data from subgroups based on age at diagnosis, stage at diagnosis, and treatment, as well as from the full cohort. We note here that under the current time scale, we assume that C_A is noninformative given stage at diagnosis, and results for other subgroups (age at diagnosis, treatment) are for exploratory purpose.

To compare the estimates from different models, we converted $\hat{\theta}_{12}$ and $\hat{\theta}$ into estimates of the corresponding Kendall's τ_{12} and τ. Table 4.8 presents the estimates of τ_{12} and τ under different combinations of copula models. $\mathcal{A}_{[2]}$ corresponds to the outer Archimedean copula, and $\mathcal{C}_{[2]}$ corresponds to the inner copula, which can be Archimedean (e.g., Clayton or Frank) or not (e.g., Gaussian).

Based on the estimated τ, the associations between the death time D and the event times (T_1, T_2) (the time to RSC or CVD) all appear strongly positive for both early and late stage at diagnosis, regardless of the copula model used in the estimation. This is further evidence that informative censoring occurred in the observations for the two event times.

The estimates of τ_{12} appear to be low to moderate, depending on the subgroup. Those diagnosed at late stage (stage III) have a higher association between T_1 and T_2. For comparison, we also obtained the estimates by the naïve approach (table 4.9) which ignores the informative censoring and uses the Kaplan–Meier estimates of $S_j(\cdot)$ in the pseudo-MLE

procedure of section 4.2. These values underestimate the association parameter between the bivariate event times compared to the pseudo-MLE estimates.

The estimated marginal survivor functions $\hat{S}_j(\cdot)$ and approximate 95% CIs are shown in figure 4.14 and figure 4.15 respectively, for the early and late stages at diagnosis. The estimates from different models appear quite similar, and all the estimates are significantly different from the corresponding naïve estimates. The marginal survivor functions are different among the two subgroups.

Table 4.8: Kendall's τ and τ_{12} Estimates with BC-BRCAS Data $\mathcal{P}_{\text{referred}}$ Using Flexible Modeling Approach

Model			**Overall**	**Age**		**Stage**			**Treatment**				
$\mathcal{A}_{[2]}$	$\mathcal{C}_{[2]}$			Young	Old	Early	Late	Unknown	Both	Chemo	Rad	Neither	Unknown
Clayton	Clayton	$sm^{\dagger}(\tau_{12})$	0.47	0.73	0.47	0.42	0.61	0.45	0.65	0.61	0.38	0.42	0.46
		$se^{\ddagger}(\tau_{12})$	.013	.018	.016	.018	.014	.058	.014	.031	.029	.015	.037
		$sm(\tau)$	0.76	0.88	0.76	0.76	0.81	0.69	0.85	0.83	0.73	0.71	0.82
		$se(\tau)$	.007	.007	.006	.006	.008	.037	.004	.008	.009	.008	.015
	Frank	$sm(\tau_{12})$	0.36	0.62	0.39	0.36	0.54	0.41	0.54	0.49	0.28	0.35	0.42
		$se(\tau_{12})$	.014	.021	.014	.014	.015	.048	.014	.029	.020	.014	.037
		$sm(\tau)$	0.77	0.88	0.76	0.76	0.80	0.68	0.86	0.83	0.72	0.70	0.81
		$se(\tau)$	.005	.007	.007	.006	.011	.027	.006	.008	.008	.008	.014
	Gaussian	$sm(\tau_{12})$	0.34	0.48	0.27	0.24	0.46	0.38	0.42	0.36	0.23	0.26	0.32
		$se(\tau_{12})$	.020	.024	.022	.022	.017	.097	.018	.028	.059	.021	.045
		$sm(\tau)$	0.76	0.88	0.75	0.74	0.79	0.80	0.85	0.83	0.71	0.70	0.78
		$se(\tau)$	.012	.006	.011	.010	.016	.135	.006	.008	.026	.015	.031
Frank	Clayton	$sm(\tau_{12})$	0.46	0.73	0.42	0.39	0.62	0.44	0.64	0.60	0.32	0.38	0.39
		$se(\tau_{12})$	.014	.018	.014	.016	.015	.053	.017	.031	.021	.017	.041
		$sm(\tau)$	0.70	0.82	0.70	0.67	0.76	0.60	0.80	0.76	0.65	0.65	0.74
		$se(\tau)$	.006	.009	.006	.007	.009	.040	.006	.011	.008	.007	.014
	Frank	$sm(\tau_{12})$	0.35	0.62	0.33	0.29	0.54	0.42	0.55	0.49	0.23	0.32	0.33
		$se(\tau_{12})$	.042	.020	.039	.047	.016	.043	.013	.027	.039	.029	.068
		$sm(\tau)$	0.67	0.82	0.66	0.66	0.75	0.63	0.79	0.76	0.62	0.62	0.72
		$se(\tau)$	.020	.010	.019	.017	.019	.059	.006	.014	.019	.027	.027
	Gaussian	$sm(\tau_{12})$	0.33	0.52	0.31	0.05	0.08	0.07	0.08	0.07	0.04	0.05	0.34
		$se(\tau_{12})$	.016	.017	.042	.019	.019	.086	.024	.028	.067	.029	.079
		$sm(\tau)$	0.70	0.81	0.68	0.81	0.65	0.72	0.92	0.58	0.61	0.69	0.73
		$se(\tau)$	.009	.009	.009	.009	.013	. 078	.007	.013	..026	.029	.037

$^{\dagger}sm$: sample mean of parameter estimate

$^{\ddagger}se$: sample standard error of parameter estimate

Table 4.9: Naïve Estimates of Kendall's τ with BC-BRCAS Data $\mathcal{P}_{\text{referred}}$ Using Flexible Modeling Approach

Model			**Overall**	**Age**		**Stage**			**Treatment**				
$\mathcal{A}_{[2]}$	$\mathcal{C}_{[2]}$			Young	Old	Early	Late	Unknown	Both	Chemo	Rad	Neither	Unknown
Clayton	Clayton	$sm^{\dagger}(\tau_{12})$	0.03	-0.25	0.03	-0.01	-0.07	-0.32	0.18	0.03	0.03	-0.16	-0.68
		$se^{\ddagger}(\tau_{12})$	.026	.144	.023	.023	.046	.163	.075	.087	.030	.033	.251
		$sm(\tau)$	0.82	0.85	0.82	0.81	0.82	0.79	0.87	0.84	0.79	0.76	0.81
		$se(\tau)$	.004	.007	.004	.004	.007	.033	.006	.008	.008	.007	.012
	Frank	$sm(\tau_{12})$	0.05	-0.04	0.03	0.02	-0.04	-0.24	0.05	0.02	0.03	-0.12	-0.47
		$se(\tau_{12})$	.016	.090	.016	.017	.032	.116	.048	.048	.020	.026	.090
		$sm(\tau)$	0.83	0.86	0.82	0.82	0.82	0.79	0.87	0.84	0.79	0.76	0.79
		$se(\tau)$	.004	.008	.004	.005	.007	.032	.006	.008	.007	.008	.017
	Gaussian	$sm(\tau_{12})$	0.04	-0.18	0.07	0.00	-0.04	-0.25	0.10	0.02	0.03	-0.11	-0.47
		$se(\tau_{12})$	.025	.105	.025	.029	.037	.135	.065	.076	.035	.030	.156
		$sm(\tau)$	0.82	0.86	0.84	0.81	0.82	0.79	0.87	0.84	0.79	0.76	0.79
		$se(\tau)$	.004	.008	.004	.005	.007	.033	.006	.008	.008	.008	.015
Frank	Clayton	$sm(\tau_{12})$	-0.05	-0.29	-0.05	-0.08	-0.15	-0.38	0.07	-0.04	-0.09	-0.22	-0.92
		$se(\tau_{12})$	.034	.257	.036	.053	.050	.246	.062	.087	.041	.046	.075
		$sm(\tau)$	0.75	0.79	0.75	0.73	0.77	0.59	0.81	0.77	0.70	0.70	0.74
		$se(\tau)$	.006	.010	.007	.008	.008	.061	.006	.010	.009	.012	.017
	Frank	$sm(\tau_{12})$	-0.02	-0.24	-0.02	-0.04	-0.09	-0.27	0.04	-0.12	-0.02	-0.12	-0.53
		$se(\tau_{12})$	.016	.113	.019	.023	.036	.134	.044	.106	.025	.028	.055
		$sm(\tau)$	0.76	0.79	0.74	0.74	0.77	0.76	0.81	0.75	0.71	0.70	0.74
		$se(\tau)$	.005	.009	.006	.006	.008	.030	.006	.011	.008	.007	.017
	Gaussian	$sm(\tau_{12})$	0.02	-0.14	0.01	0.00	-0.08	-0.24	-0.01	-0.02	-0.02	-0.13	-0.54
		$se(\tau_{12})$	.027	.196	.028	.064	.060	.211	.053	.096	.037	.041	.076
		$sm(\tau)$	0.76	0.78	0.76	0.74	0.77	0.77	0.80	0.76	0.70	0.71	0.73
		$se(\tau)$	.007	.010	.007	.007	.009	.035	.006	.011	.009	.008	.017

$^{\dagger}sm$: sample mean of parameter estimates

$^{\ddagger}se$: sample standard error of parameter estimates

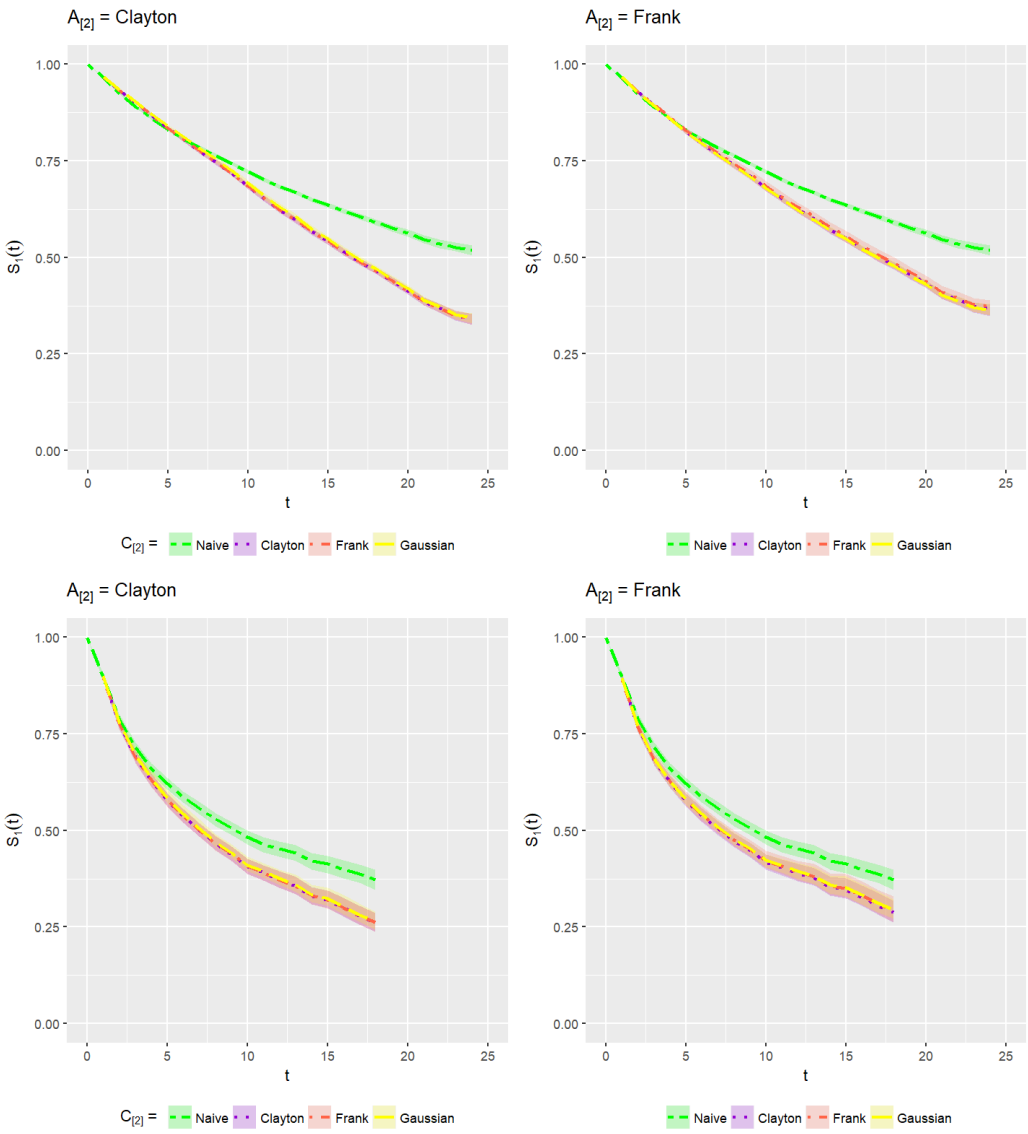


Figure 4.14: Estimates of Marginal Survivor Functions $S_1(\cdot)$ on Time to RSC Using Proposed Approach with Different Copulas and Using Kaplan–Meier Estimator with BCBRCAS Data $\mathcal{P}_{\text{referred}}$. Early Stage at Diagnosis (Upper Row) vs. Late Stage at Diagnosis (Lower Row).

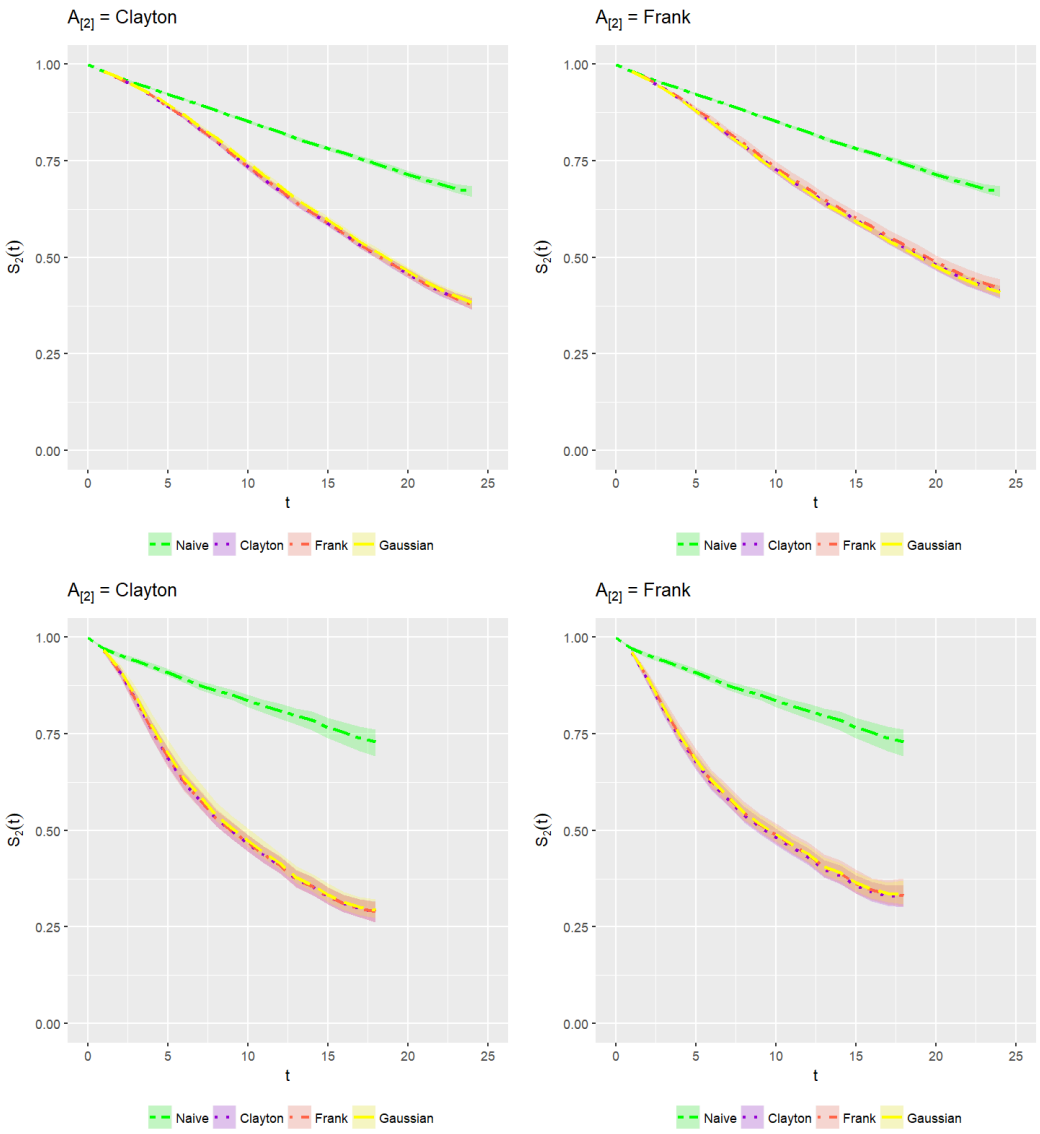


Figure 4.15: Estimates of Marginal Survivor Functions of $S_2(\cdot)$ on Time to CVD Using Proposed Approach with Different Copulas and Using Kaplan–Meier Estimator with BC-BRCAS Data $\mathcal{P}_{\text{referred}}$. Early Stage at Diagnosis (Upper Row) vs. Late Stage at Diagnosis (Lower Row).

4.5 Discussion

We have proposed a flexible modeling approach and the associated pseudolikelihood-based inference procedure to deal with informative censoring in the analysis of bivariate event times. The approach models bivariate event times jointly with the informative censoring time through a bivariate Archimedean copula ("outer" copula) function. The joint survivor function of the bivariate event times can be modeled through a copula that is different from the outer copula. This allows the association between the bivariate event times and their dependence on the informative censoring time to be different. This provides flexibility and is more appropriate for practical studies. In addition, the joint survivor function can also be modeled through other bivariate functions such as (4.7), which leads to additional flexibility. Our approach can be extended to multiple (≥ 3) event times.

Our procedure is pseudolikelihood-based and is computationally more feasible than the likelihood-based counterpart. The trade-off is its potential loss of inference efficiency, which is shown by comparisons such as that in Lawless & Yilmaz (2011). We may consider maximizing the likelihood based on the full data set (3.1) with respect to all the association parameters θ_{12} and θ jointly with the survivor functions $S_j(\cdot)$ for $j = 1, 2$ and $S_D(\cdot)$. Alternatively, to ease the computational intensity of the maximum semiparametric likelihood procedure, we may adopt a data-smoothing technique to handle the unknown survivor functions. In this chapter we provided an iterative algorithm to find the pseudo-MLE without having to estimate the density of the marginal survivor functions.

The real-data analysis in section 4.4 has shown that the association between the bivariate event times is weaker than their dependence on the informative censoring time. This confirms the usefulness of our flexible modeling approach. Our analysis of the real data shows that the survival patterns and the association patterns are different among the subgroups. This suggests an important and useful extension of our approach: if we include the potential covariates in the model, we can assess the covariate effects on the survivor functions and the association between the event times. The extended approach could be applied to compare the breast cancer patients with the general population in terms of CVD-related health issues, which is the primary goal of the analysis of the breast cancer study.

In summary, compared to Chapter 3, this approach allows the dependence between the event times, and between them jointly and the time to the terminating event to be different. We examine the performance through numerical studies and theoretical justification. The pseudo MLE inference procedure is easy to implement. We applied the approach to real data example and verified that the association parameter in Chapter 3 may be taken as an average of the associations with varying magnitude between different pairs of event times.

Chapter 5

Regression Analysis of Bivariate Event Time with Observations on Response Subject to Informative Censoring

Studies on association between two event times and how covariates affect the association are often of interest to researchers. Conventional marginal or conditional approaches on multivariate event times assume noninformative censoring which could lead to biased inference if the assumption is violated. In this chapter, we extend the proposed approach in Chapter 4 in a regression setting, where we formulated the joint distribution of the bivariate event times together with the informative censoring time through embedded copula functions, which allows for flexible dependence amongst event times. In the meantime, we incorporated covariate effects by specifying the association parameters as functions of the covariates. A by-product of our approach is the estimator of conditional survivor functions for each of the event time in presence of informative censoring. We developed an easy-to-implement pseudolikelihood-based inference procedure. Simulation studies are conducted to examine the performance of the proposed modeling and inference procedure. Asymptotic properties of the proposed estimator are established. We applied the proposed approach to analyze the motivating real-data example which attempts to evaluate how clinical factors (e.g. treatment) affect the time to the first cardiovascular disease amongst breast cancer patients who have experienced relapse or second cancer.

5.1 Introduction

Association or dependence structures between event times are often of interest in practical studies. Conventional statistical approaches focus mainly on regression analysis with mul-

tivariate event times data using marginal or conditional models, and specify the association between event times by a frailty or a copula parameter (see, for example, Clayton 1978, Oakes 1994, Genest et al. 1995, Shih & Louis 1995, Lawless & Yilmaz 2011, Diao & Cook 2014, Zhong & Cook 2016). There is also some literature on the bivariate association. For example, Ning et al. (2015) propose the use of rate ratio to assess the local dependence between two types of recurrent event processes by modeling the rate ratio as a parametric function of time. Fine & Jiang (2000) consider estimation of the cross ratio in Clayton's (1978) copula in which covariates are incorporated into the marginal distributions via semiparametric accelerated life regression models. Other works on bivariate association include Bandeen-Roche & Ning (2008), Fan et al. (2000), Cheng et al. (2007), amongst others. These approaches assume noninformative censoring on the bivariate event times, and the covariates effects are incorporated through the marginal functions.

The breast cancer study, a recent cancer survivorship project (Davis et al. 2014) at the BC Cancer Agency (`www.bccancer.ca`) investigated the association between the time to RSC and the time to CVD, and how the clinical (e.g. age, treatment, stage) and sociodemographic (e.g. socioeconomic status), and health system factors (e.g. health authority at diagnosis) might affect this association. This is an important health issue for cancer survivors. The study's observations on the times to CVD and RSC are heavily censored because of either the study follow-up time limit or death. The time to death is likely correlated to the two event times of interest. Thus, conventional event time analysis methods, such as Cox PH model for regression analysis are not directly applicable. Statistical inference with such data requires us to formulate the potential dependence amongst the multiple event times and, at the same time, their dependence on the informative censoring. However, often we cannot confidently specify either the correlation structures, or the distributions of the event times and censoring times. Conjectures about the dependence structures, in particular, can be many and varied. In addition, it is desirable to have explicit visualization or interpretations of the covariates effect on the association between CVD and RSC.

This chapter focuses on a semiparametric analysis of bivariate event times with observations on two event times informatively censored due to a terminating event. Leaving the marginal distribution of the terminating event unspecified, we model the correlation of the two events with the terminating event via a bivariate copula model, and we model the two event times of interest via another bivariate copula model. Covariate effects are incorporated through specifying the copula association parameters as functions of the covariates. The two-step estimation procedure with a copula model can then naturally be adapted to the proposed model. On the other hand, the model may adopt the deemed structure of the multivariate event time distribution in any form, not necessarily that of the bivariate copula model.

We motivate the model and illustrate the associated estimation procedure using the breast cancer study. The methodology, however, has broader applicability. The rest of this

chapter is organized as follows. Section 5.2 presents the models after introducing the notation and framework. We propose in section 5.3 a pseudolikelihood-based semiparametric procedure to estimate the model parameters. We then derive the asymptotic properties of the resulting estimators, and in particular the maximum pseudolikelihood estimator (pseudo-MLE) for the model parameter that measures the association between the event times. Section 5.4 reports a simulation study that evaluated the finite-sample performance of the estimation procedure in terms of consistency, efficiency, and robustness. Section 5.5 presents an analysis of the real data from the breast cancer study, and section 5.6 provides concluding remarks.

5.2 Notation and Modeling

5.2.1 Notation

We aim to estimate the joint survivor function $S_{12|Z}(\cdot)$ with the study's right-censored bivariate event times when T_1 and T_2 are potentially correlated with D, given covariates Z. Adopting the conventional notation, let Δ_D be the indicator $I\{D \leq C_A\}$, and $U_j = T_j \wedge C$ with $\Delta_j = I\{T_j \leq C\}$ for $j = 1, 2$. Suppose that the study data are n independent realizations of $\{(U_1, \Delta_1), (U_2, \Delta_2), (C, \Delta_D); Z\}$, denoted by

$$\text{Observed-Data} = \bigcup_{i=1}^{n} \Big\{ \Big[\{(u_{ji}, \delta_{ji}) : j = 1, 2\} \bigcup \{(c_i, \delta_{Di})\} \Big]; z_i : i = 1, \ldots, n \Big\}. \tag{5.1}$$

This is the union of the two semicompeting-risks data sets on T_1 and T_2 associated with $C = D \wedge C_A$, together with the observed covariates Z_i: for $j = 1, 2$,

$$\text{Observed-Data}_j = \{(u_{ji}, \delta_{ji}, c_i, \delta_{Di}; Z_i) : i = 1, \ldots, n\}. \tag{5.2}$$

We perform inference on the distributions of the event times T_j over the intervals $[0, v_j]$ with v_j chosen to be slightly smaller than $\max_i\{u_{ji}\}$ for $j = 1, 2$.

5.2.2 Model Specification

We extended the modeling proposed in Chapter 4, by assuming that the administrative censoring time C_A is independent of the event times T_1, T_2 and the time to the terminating event D, conditional on Z. Furthermore, to specify the conditional correlation of (T_1, T_2) with D, we embedded the bivariate survivor function of $(T_1, T_2)|Z$ in a bivariate Archimedean copula model (e.g., Joe 1997). That is, we assumed the joint survivor function of (T_1, T_2) with D conditional on Z to be

$$Pr(T_1 \geq t_1, T_2 \geq t_2, D \geq d|Z) = \mathcal{A}_{[2]}(S_{12}(t_1, t_2|Z), S_D(d|Z); \theta(Z)). \tag{5.3}$$

where association parameter $\theta(Z)$ is an unknown function of Z which characterizes the correlation between $S_{12}(t_1, t_2|Z)$ and $S_D(d|Z)$. Note that $S_{12}(t_1, 0|Z) = P(T_1 \geq t_1|Z)$ and $S_{12}(0, t_2|Z) = P(T_2 \geq t_2|Z)$ are the conditional marginal survivor functions of T_1 and T_2, respectively. Let $S_j(t_j|Z) = P(T_j \geq t_j|Z)$ for $j = 1, 2$. The model in (5.3) induces the joint model of T_j and D conditional on Z:

$$Pr(T_j \geq t_j, D \geq d|Z) = \mathcal{A}_{[2]}\big(S_j(t_j|Z), S_D(d|Z); \theta(Z)\big). \tag{5.4}$$

Often the bivariate survivor function $S_{12}(\cdot|Z)$ in (5.3) cannot be confidently specified using a parametric model. We consider a semiparametric model for the conditional joint distribution of T_1, T_2:

$$S_{12}(t_1, t_2|Z) = \mathcal{C}_{[2]}\big(S_1(t_1|Z), S_2(t_2|Z); \theta_{12}(Z)\big), \tag{5.5}$$

where the univariate marginal survivor functions $S_j(\cdot|Z)$ are unspecified, and $\mathcal{C}_{[2]}(\cdot; \theta_{12}(\cdot))$ is a known bivariate copula function upon $\theta_{12}(\cdot)$. We remark that here it allows the bivariate function $\mathcal{C}_{[2]}(\cdot)$ in model (5.5) to be different from the bivariate Archimedean copula $\mathcal{A}_{[2]}(\cdot)$ in model (5.3). We may choose $\mathcal{C}_{[2]}(\cdot)$ in (5.5) to be a commonly-used non-Archimedean copula or a bivariate function. This leads to additional modeling flexibility. Since the motivating question is to address the association between event times, and copula models provide a feasible measure of association, we thus focus on copula modeling in this chapter. More discussion on it is provided with numerical studies reported in sections 5.4 and 5.5.

Denote $\delta_{1i} + \delta_{2i}$ by $\delta_{.i}$. Let $\dot{h}(r)$ be $dh(r)/dr$ for a function $h(r)$ and $h^{(a_1,a_2)}(r_1, r_2; \phi)$ be $\partial h^{(a_1+a_2)}(r_1, r_2; \phi)/\partial r_1^{a_1} \partial r_2^{a_2}$ for a function $h(r_1, r_2; \phi)$ with well-defined partial derivatives. The likelihood function with the available data under the copula model (5.3) with model (5.5) embedded in is

$$\begin{aligned}
& L(S_1(\cdot|Z), S_2(\cdot|Z), S_D(\cdot|Z), \theta(\cdot), \theta_{12}(\cdot)|\text{Observed-Data}) \\
= & \prod_{i=1}^{n} \left\{ (-1)^{\delta_{.i}+\delta_{Di}} \frac{\partial^{\delta_{.i}} \mathcal{A}_{[2]}^{(0,\delta_{Di})}\big(S_{12}(u_{1i}, u_{2i}|Z), S_D(c_i|Z); \theta(Z)\big)}{\partial u_1^{\delta_{1i}} \partial u_2^{\delta_{2i}}} \dot{S}_D(c_i|Z)^{\delta_{Di}} \right\}. \qquad (5.6) \\
= & \prod_{i=1}^{n} \left\{ (-1)^{\delta_{.i}+\delta_{Di}} \frac{\partial^{\delta_{.i}} \mathcal{A}_{[2]}^{(0,\delta_{Di})}\big(\mathcal{C}_{[2]}(S_1(u_{1i}|Z), S_2(u_{2i}|Z); \theta_{12}(Z)), S_D(c_i|Z); \theta(Z)\big)}{\partial u_1^{\delta_{1i}} \partial u_2^{\delta_{2i}}} \dot{S}_D(c_i|Z)^{\delta_{Di}} \right\}
\end{aligned}$$

It is not easy to directly maximize (5.6) with respect to $\theta(Z), \theta_{12}(Z)$, $S_D(\cdot|Z)$, and $S_j(\cdot|Z)$, $j = 1, 2$. When all covariates in Z are discrete, denoted Z^* with finite categories, it is straightforward to estimate $\theta(Z^*)$ and $\theta_{12}(Z^*)$ as a set of finite dimensional parameters. When continuous covariate is present, denoted X, we model $\theta(X), \theta_{12}(X)$ as linear combinations of cubic B-spline basis functions (see for example Rosenberg 1995) in the real data analysis. Specifically, we used $\theta(X) = \boldsymbol{\alpha}'\mathcal{B}$, and $\theta_{12}(X) = \boldsymbol{\alpha_{12}}'\mathcal{B}$ where $\mathcal{B}$ are B-spline basis functions of degree = 3, number of knots = 1, location of knots at the median of X .

In other words, the estimation of the two unknown functions $\theta(\cdot)$, $\theta_{12}(\cdot)$ can be specified into estimation of finite dimensional parameters, denoted by $\boldsymbol{\alpha}$, and $\boldsymbol{\alpha}_{12}$, respectively. Now maximizing (5.6) is simplified into maximizing

$$\begin{aligned} & L(S_1(\cdot|Z), S_2(\cdot|Z), S_D(\cdot|Z), \boldsymbol{\alpha}, \boldsymbol{\alpha}_{12}|\text{Observed-Data}) \qquad (5.7) \\ = & \prod_{i=1}^{n} \left\{ (-1)^{\delta_{\cdot i}+\delta_{Di}} \frac{\partial^{\delta_{\cdot i}} \mathcal{A}_{[2]}^{(0,\delta_{Di})}(\mathcal{C}_{[2]}(S_1(u_{1i}|Z), S_2(u_{2i}|Z); \boldsymbol{\alpha}_{12}), S_D(c_i|Z); \boldsymbol{\alpha})}{\partial u_1^{\delta_{1i}} \partial u_2^{\delta_{2i}}} \dot{S}_D(c_i|Z)^{\delta_{Di}} \right\} \end{aligned}$$

with three unknown functions $S_D(\cdot|Z)$ and $S_j(\cdot|Z)$, $j = 1, 2$ and two sets of parameters $\boldsymbol{\alpha}_{12}$, $\boldsymbol{\alpha}$.

5.2.3 More on Modeling

The current observations on D are right-censored with the noninformative censoring time C_A. There is a readily available consistent estimator for $S_D(\cdot)$, e.g., the Kaplan–Meier estimator, denoted as $\tilde{S}_D(\cdot)$. In regression setting, there are also readily available models and estimating procedures to obtain conditional $S_D(\cdot|Z)$, denoted as $\tilde{S}_D(\cdot|Z)$, including parametric models (e.g. Weibull, Exponential, Lognormal) or the Cox PH model. In the real data analysis in section (5.5), we applied the Cox model:

$$S_D(t|Z) = \exp\{H_{0D}(t)e^{\beta_D Z}\} \qquad (5.8)$$

where β_D can be estimated using standard partial likelihood inference procedure, and $H_{0D}(t)$ can be estimated by the Breslow estimator. We denote the estimated conditional survivor function as $\tilde{S}_D(t|Z)$.

In addition, estimating the conditional survivor function of the event times T_j ($j = 1, 2$) with the semicompeting-risks data, Observed-Data$_j$ in (5.2), is of interest in many situations. When the copula function $\mathcal{A}_{[2]}(\cdot; \theta(Z))$ in (5.3) is an Archimedean copula with its generator $\psi(\cdot; \theta(Z))$, the induced model (5.4) for the joint survivor function of T_j and D yields

$$S_j(t|Z) = g(S_j^*(t|Z), S_D(t|Z); \theta(Z)) = \psi^{-1}\{\psi(S_j^*(t|Z); \theta(Z)) - \psi(S_D(t|Z); \theta(Z)); \theta(Z)\}, \qquad (5.9)$$

where $S_j^*(t|Z) = P(T_j^* \geq t|Z)$ is the survivor function of $T_j^* = T_j \wedge D$ conditional on Z. Similar to D, T_j^* is only censored by C_A, the noninformative censoring time. In this chapter, we apply the semiparametric Cox model:

$$S_j^*(t|Z) = \exp\{H_{0j}^*(t)e^{\beta_j^* Z}\} \qquad (5.10)$$

where β_j^* and H_{0j}^* can be estimated through standard approach for the Cox model. We denote the estimated conditional survivor function as $\tilde{S}_j^*(t|Z)$. Plugging in the estimates

$\tilde{S}_D^*(t|Z)$ and $\tilde{S}_j^*(t|Z)$ into (5.9), and with $\theta(\cdot)$ specified with parameter $\boldsymbol{\alpha}$, we have

$$\tilde{S}_j(t|Z;\boldsymbol{\alpha}) = g(\tilde{S}_j^*(t|Z), \tilde{S}_D(t|Z);\boldsymbol{\alpha}) \tag{5.11}$$

Following the idea of the pseudolikelihood estimation procedure under a copula model (e.g., Lawless & Yilmaz 2011), we may consider a pseudo-MLE of $\boldsymbol{\alpha}, \boldsymbol{\alpha}_{12}$ by plugging (5.11) into (5.7), and now the likelihood function is proportional to:

$$\begin{aligned} & L(\boldsymbol{\alpha}, \boldsymbol{\alpha}_{12}|\text{Observed-Data}) \qquad (5.12)\\ = & \prod_{i=1}^{n} \left\{ (-1)^{\delta_{\cdot i}+\delta_{Di}} \frac{\partial^{\delta_{\cdot i}} \mathcal{A}_{[2]}^{(0,\delta_{Di})}(\mathcal{C}_{[2]}(\tilde{S}_1(u_{1i}|Z;\boldsymbol{\alpha}), \tilde{S}_2(u_{2i}|Z;\boldsymbol{\alpha});\boldsymbol{\alpha}_{12}), \tilde{S}_D(c_i|Z);\boldsymbol{\alpha})}{\partial u_1^{\delta_{1i}} \partial u_2^{\delta_{2i}}} \right\} \end{aligned}$$

with respect to $\boldsymbol{\alpha}_{12}$ and $\boldsymbol{\alpha}$ only. Here the partial derivative in (5.12) is

$$\begin{aligned} & \frac{\partial^{\delta_{\cdot i}} \mathcal{A}_{[2]}^{(0,\delta_{Di})}(\mathcal{C}_{[2]}(\tilde{S}_{1|Z}(u_{1i};\boldsymbol{\alpha}), \tilde{S}_{2|Z}(u_{2i};\boldsymbol{\alpha});\boldsymbol{\alpha}_{12}), \tilde{S}_{D|Z}(c_i);\boldsymbol{\alpha})}{\partial u_1^{\delta_{1i}} \partial u_2^{\delta_{2i}}} \\ = & \begin{cases} \mathcal{A}_{[2]}^{(0,\delta_{Di})}(\mathcal{C}_{[2]}(\tilde{S}_{1|Z}(u_{1i};\boldsymbol{\alpha}), \tilde{S}_{2|Z}(u_{2i};\boldsymbol{\alpha});\boldsymbol{\alpha}_{12}), \tilde{S}_{D|Z}(c_i);\boldsymbol{\alpha}), & \delta_{1i}=\delta_{2i}=0 \\ \mathcal{A}_{[2]}^{(1,\delta_{Di})}(\mathcal{C}_{[2]}(\tilde{S}_{1|Z}(u_{1i};\boldsymbol{\alpha}), \tilde{S}_{2|Z}(u_{2i};\boldsymbol{\alpha});\boldsymbol{\alpha}_{12}), \tilde{S}_{D|Z}(c_i);\boldsymbol{\alpha})\times & \\ \mathcal{C}_{[2]}^{(\delta_{1i},\delta_{2i})}(\tilde{S}_{1|Z}(u_{1i};\boldsymbol{\alpha}), \tilde{S}_{2|Z}(u_{2i};\boldsymbol{\alpha});\boldsymbol{\alpha}_{12}), & \delta_{1i}\neq\delta_{2i} \\ \mathcal{A}_{[2]}^{(2,\delta_{Di})}(\mathcal{C}_{[2]}(\tilde{S}_{1|Z}(u_{1i};\boldsymbol{\alpha}), \tilde{S}_{2|Z}(u_{2i};\boldsymbol{\alpha});\boldsymbol{\alpha}_{12}), \tilde{S}_{D|Z}(c_i);\boldsymbol{\alpha})\times & \\ \mathcal{C}_{[2]}^{(1,0)}(\tilde{S}_{1|Z}(u_{1i};\boldsymbol{\alpha}), \tilde{S}_{2|Z}(u_{2i};\boldsymbol{\alpha});\boldsymbol{\alpha}_{12})\times & \\ \mathcal{C}_{[2]}^{(0,1)}(\tilde{S}_{1|Z}(u_{1i};\boldsymbol{\alpha}), \tilde{S}_{2|Z}(u_{2i};\boldsymbol{\alpha});\boldsymbol{\alpha}_{12}) & \\ +\mathcal{A}_{[2]}^{(1,\delta_{Di})}(\mathcal{C}_{[2]}(\tilde{S}_{1|Z}(u_{1i};\boldsymbol{\alpha}), \tilde{S}_{2|Z}(u_{2i};\boldsymbol{\alpha});\boldsymbol{\alpha}_{12}), \tilde{S}_{D|Z}(c_i);\boldsymbol{\alpha})\times & \\ \mathcal{C}_{[2]}^{(1,1)}(\tilde{S}_{1|Z}(u_{1i};\boldsymbol{\alpha}), \tilde{S}_{2|Z}(u_{2i};\boldsymbol{\alpha});\boldsymbol{\alpha}_{12}) & \delta_{1i}=\delta_{2i}=1. \end{cases} \end{aligned}$$

The resulting estimator with the trade-off of some efficiency loss, can be much easier to implement than its MLE counterpart.

In principle, one may maximize (5.6) under model (5.3) coupled with model (5.5) with respect to $\boldsymbol{\alpha}, S_j(\cdot), \boldsymbol{\alpha}_{12}$, and $S_D(\cdot)$ to obtain their MLE, which leads to the semiparametric MLE of the joint survivor function $S_{12|Z}(\cdot;\theta_{12})$. This, however, requires quite intensive computing. Furthermore, the counterpart of the pseudo-MLE approach for parametric $S_{12|Z}(\cdot)$ is not directly applicable since there is no readily available consistent estimator for $S_{j|Z}(\cdot)$ with the current semicompeting-risks data on T_j. These considerations motivate the two procedures in section 5.3 for estimating the joint survivor function $S_{12}(\cdot|Z;\boldsymbol{\alpha}_{12})$ under model (5.5).

5.3 Pseudolikelihood-Based Estimation Procedures

Using the idea underlying two-stage estimation procedures with a copula model (e.g., Oakes 1994, Genest et al. 1995), we estimate $S_{12}(\cdot|Z)$, the joint survivor function of (T_1, T_2), under the model (5.3) with model (5.5) embedded in. The estimation procedure yields a consistent estimator for the marginal survivor function of each of the two event times as a by-product. We also present the asymptotic properties of the estimators.

5.3.1 Estimating the Parameters with the Observed-Data

Under model (5.4), as it is given in (5.9), the marginal survivor function $S_j(t|Z) = g(S_j^*(t|Z), S_D(t|Z); \theta(Z))$, a known function of the marginal survivor function of $T_j^* = T_j \wedge D$ and the marginal survivor function of D upon $\theta(Z)$ for $j = 1, 2$. With known $S_j^*(t|Z)$ and $S_D(t|Z)$, $S_j(t|Z)$ is known only upon the parameter $\theta(Z)$. In this chapter, the functions $\theta(\cdot)$ and $\theta_{12}(\cdot)$ can be estimated by parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}_{12}$, respectively.

Furthermore, provided with consistent estimators for $S_j^*(\cdot|Z)$ and $S_D(\cdot|Z)$, we maximize the pseudolikelihood in (5.12) with respect to $\boldsymbol{\eta} = (\boldsymbol{\alpha}', \boldsymbol{\alpha}_{12}')'$ or, equivalently, its log-transformation with respect to the parameters $\boldsymbol{\eta} = (\boldsymbol{\alpha}', \boldsymbol{\alpha}_{12}')'$, to derive a pseudo-MLE:

$$\hat{\boldsymbol{\eta}}_n = \text{argmax}_{\boldsymbol{\eta}} L(\boldsymbol{\eta}|\tilde{S}_{1|Z}(\cdot), \tilde{S}_{2|Z}(\cdot), \tilde{S}_{D|Z}(\cdot); \text{Observed-Data}). \tag{5.13}$$

This pseudo-MLE procedure is computationally easy to implement. We present below an iterative algorithm to calculate $\hat{\boldsymbol{\eta}}_n$.

ALGORITHM. Using the estimated $\tilde{S}_j^*(\cdot|Z)$ and $\tilde{S}_D(\cdot|Z)$ together with the current estimate $\boldsymbol{\eta}^{(k-1)}$ and $S_j^{(k-1)}(\cdot|Z)$ for $j = 1, 2$ and with $k \geq 1$,

Step 1. obtain the updated estimate for $\boldsymbol{\eta}$ as

$$\boldsymbol{\eta}^{(k)} = \text{argmax}_{\boldsymbol{\eta}} L(\boldsymbol{\eta}|S_1^{(k-1)}(\cdot|Z), S_2^{(k-1)}(\cdot|Z), \tilde{S}_D(\cdot|Z); \text{Observed-Data});$$

Step 2. obtain the updated estimates for $S_j(\cdot|Z)$ as $S_j^{(k)}(t|Z) = \tilde{S}_j(t|Z; \boldsymbol{\alpha}^{(k)}) = g(\tilde{S}_j^*(t|Z), \tilde{S}_D(t|Z); \boldsymbol{\alpha}^{(k)})$ for $j = 1, 2$.

Repeat steps 1 and 2 until the sequence $\{\boldsymbol{\eta}^{(k)} : k = 0, 1, \ldots\}$ converges. The limit is $\hat{\boldsymbol{\eta}}_n$ defined in (5.13).

5.3.2 Resulting Estimators for Marginal and Joint Survivor Function

Plugging $\hat{\boldsymbol{\alpha}}_n(\cdot)$, $\tilde{S}_j^*(t|Z)$, and $\tilde{S}_D(t|Z)$ from the above section in (5.9) gives a natural estimator for the marginal survivor function $S_j(\cdot|Z)$:

$$\hat{S}_{jn}(t|Z) = g(\tilde{S}_j^*(t|Z), \tilde{S}_D(t|Z); \hat{\boldsymbol{\alpha}}_n). \tag{5.14}$$

Moreover, the joint survivor function $S_{12}(t_1, t_2|Z)$ of (T_1, T_2) based on (5.5):

$$\hat{S}_{12n}(t_1, t_2|Z) = \mathcal{C}_{[2]}(\hat{S}_{1n}(t_1|Z), \hat{S}_{2n}(t_2|Z); \hat{\boldsymbol{\alpha}}_{12n}). \tag{5.15}$$

5.3.3 Asymptotic Properties

The following proposition establishes the consistency and asymptotic normality of the resulting estimator. Define the following regularity conditions:

(RC5.1) Suppose $\theta(X)$ is a smooth function defined on $\mathbb{R}$, and $\mathcal{C}^{ab}(r_1, r_2; \theta)$, $\mathcal{C}^{abc}(r_1, r_2; \theta)$ exist and are continuous and uniformly bounded by some constant M for $a, b, c \in$ (0,1,2,3), $0 \le r_1 \le 1$, $0 \le r_2 \le 1$, $0 \le v_1 \le 1$, $0 \le v_2 \le 1$.

(RC5.2) For each x, $0 < E_\theta \frac{\mathcal{C}^{abc}(r_1, r_2; \theta(x))}{\mathcal{C}^{ab}(r_1, r_2; \theta(x))} \le \infty$ for $a, b, c \in$ (0,1), $0 \le r_1 \le 1$, $0 \le r_2 \le 1$, $0 \le v_1 \le 1$, $0 \le v_2 \le 1$.

(RC5.3) Under Archimedean copula, $g^{ab}(v_1, v_2; \theta)$ and $g^{abc}(v_1, v_2; \theta)$ exist and are continuous and uniformly bounded by some constant M for $a, b, c \in$ (0,1,2,3), $0 \le r_1 \le 1$, $0 \le r_2 \le 1$, $0 \le v_1 \le 1$, $0 \le v_2 \le 1$.

(RC5.4) Under Archimedean copula, for each x, $0 < E_\theta \frac{\mathcal{C}^{abc}(r_1, r_2; \theta(x)) g^{001}(v_1, v_2; \theta)}{\mathcal{C}^{ab}(r_1, r_2; \theta(x))} \le \infty$ for $a, b, c \in$ (0,1), $0 \le r_1 \le 1$, $0 \le r_2 \le 1$, $0 \le v_1 \le 1$, $0 \le v_2 \le 1$.

In addition, if the estimator functions $\hat{S}_1$ and $\hat{S}_2$ of S_1, S_2 satisfy the following two conditions:

AC5.1 $\hat{S}_1$ and $\hat{S}_2$ converge uniformly to S_1 and S_2, respectively.

AC5.2 $\sqrt{n}(\hat{S}_1 - S_1) \xrightarrow{w} \mathcal{G}_1$, and $\sqrt{n}(\hat{S}_2 - S_2) \xrightarrow{w} \mathcal{G}_2$, where $\mathcal{G}_1$ and $\mathcal{G}_2$ are two mean zero Gaussian processes with limiting covariance $cov(G_j(s_1), G_j(s_2)) = \sigma_j^2(s_1 \wedge s_2)$ for $j = 1, 2$, with σ_j^2 defined as in Andersen et al. (1993). For simplification, we denote $\sigma_1^2(\cdot)$ and $\sigma_2^2(\cdot)$ as the limiting variance function for $\sqrt{n}(\hat{S}_1 - S_1)$ and $\sqrt{n}(\hat{S}_2 - S_2)$, respectively.

Proposition 7. *Under the regularity conditions (RC5.1)–(RC5.4) presented above and provided $\tilde{S}_1^*(t|Z)$, $\tilde{S}_2^*(t|Z)$, and $\tilde{S}_{D|Z}(t)$ satisfy condition (AC5.1), then as $n \to \infty$, $\hat{\boldsymbol{\eta}}_n \xrightarrow{a.s.} \boldsymbol{\eta}$ and $\sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}) \xrightarrow{d} N(0, AV(\boldsymbol{\eta}))$, where the asymptotic variance is*

$$AV(\boldsymbol{\eta}) = V_B(\boldsymbol{\eta})^{-1} V_A(\boldsymbol{\eta}) V_B(\boldsymbol{\eta})^{-1} \tag{5.16}$$

with $V_B(\boldsymbol{\eta})$ and $V_A(\boldsymbol{\eta})$ the limits of

$$-\frac{1}{n}\sum_{i=1}^{n} \partial^2 \log L(\boldsymbol{\eta}|\tilde{S}_{1|Z}(\cdot; \boldsymbol{\alpha}), \tilde{S}_2(\cdot|Z; \boldsymbol{\alpha}), \tilde{S}_D(\cdot|Z); \textit{Observed-Data})/\partial \boldsymbol{\eta}^2 \tag{5.17}$$

and

$$\frac{1}{n} Var\Big\{ \sum_{i=1}^{n} \partial \log L(\boldsymbol{\eta}|\tilde{S}_1(\cdot|Z;\theta), \tilde{S}_2(\cdot|Z;\theta), \tilde{S}_D(\cdot|Z); \textit{Observed-Data})/\partial \boldsymbol{\eta} \Big\}, \tag{5.18}$$

respectively, and $\tilde{S}_j(t|Z;\boldsymbol{\alpha}) = g(\tilde{S}^*_j(t|Z), \tilde{S}_D(t|Z);\boldsymbol{\alpha})$.

Similar to previous two chapters, one may estimate the variance of $\hat{\boldsymbol{\eta}}_n$ by a bootstrap approach (e.g., Lawless & Yilmaz 2011). A natural and practical variance estimator evaluates (5.16) at $\boldsymbol{\eta} = \hat{\boldsymbol{\eta}}_n$ and uses (5.17) and (5.18) to replace their limits, that is, the Huber's robust sandwich estimator (Huber 1967). Note that when $S_{j|Z}(\cdot)$ for $j = 1, 2$ are known and used to estimate $\boldsymbol{\eta} = (\boldsymbol{\alpha}', \boldsymbol{\alpha}'_{12})'$, $\hat{\boldsymbol{\eta}}_n$ is an MLE, and $V_A(\boldsymbol{\eta})$ and $V_B(\boldsymbol{\eta})$ in (5.17) and (5.18) are the same as the corresponding inverse Fisher information matrix.

Proposition 8. *Under the regularity conditions (RC5.1)–(RC5.4) and provided* $\tilde{S}^*_j(\cdot|Z)$ *and* $\tilde{S}_D(\cdot|Z)$ *satisfy condition (AC5.1) , as* $n \to \infty$, $\hat{S}_{jn}(t|Z) \xrightarrow{a.s.} S_j(t|Z)$ *uniformly and* $\sqrt{n}(\hat{S}_{jn}(t|Z) - S_j(t|Z)) \xrightarrow{w} \mathcal{G}_j(t|Z)$ *with* $t \in [0, v^\star_j]$, *where* $\mathcal{G}_j(t|Z)$ *is a Gaussian process with mean zero and variance function* $\sigma^2_j(t)$ *as defined in, for example, Andersen et al. (1993).*

When the sample size is large and the censoring rate is not too high, we may choose to ignore the variation of $\tilde{S}^*_j(\cdot|Z)$ and $\tilde{S}_D(\cdot|Z)$. It then yields an approximate confidence band (CB) for $S_j(\cdot|Z)$ based on (5.9) with $\hat{\boldsymbol{\alpha}}_n$ plugged in, and using the proposed variance estimator of $\hat{\boldsymbol{\alpha}}_n$ in the above section.

The following proposition establishes the consistency and asymptotic normality/weak convergence of the resulting estimator.

Proposition 9. *Under the regularity conditions (RC1)–(RC4) and provided* $\tilde{S}^*_{j|Z}(t)$ *and* $\tilde{S}_D(t|Z)$ *satisfy condition (AC1), as* $n \to \infty$, $\hat{S}_{12n}(t_1, t_2|Z) \xrightarrow{a.s.} S(t_1, t_2|Z)$ *uniformly and* $\sqrt{n}(\hat{S}_{12n}(t_1, t_2|Z) - S(t_1, t_2|Z)) \xrightarrow{w} \mathcal{G}(t_1, t_2|Z)$ *with* $t_1, t_2 \in [0, v^\star_1] \times \times [0, v^\star_2]$, *where* $\mathcal{G}(t_1, t_2)$ *is a Gaussian field with mean zero and variance function* $\sigma^2(t_1, t_2)$.

The proofs follow the steps give in Section 3.4. The additional step is that the estimated $\theta(\cdot) = \boldsymbol{\alpha}'\mathcal{B}$ using spline is consistent, which has been shown in (de Boor 1978).

5.4 Simulation

Simulation studies were conducted to explore the finite-sample performance of the proposed approach in section 5.3.

5.4.1 Setting and Data Generation

We simulated a study with n independent units where the primary outcome is the bivariate event times (T_1, T_2) conditional on Z. The observations on (T_1, T_2) may be censored by

either the terminating event time D or an administrative time C_A, whichever occurs first. That is, the study censoring time $C = D \wedge C_A$. We allow the dependence parameter $\theta_{12}(Z)$ between (T_1, T_2) and the dependence parameter $\theta(Z)$ between (T_1, T_2) jointly with D to be different functions of Z.

We generated data from nested Archimedean copula (Joe 1997) which allows the outer and inner copula association parameters to be different, representing different strengths of dependence. To imitate potentially informative censoring due to a terminating event, the data were generated as follows:

Step (a). We independently generated z_i for $i = 1, \ldots, n$ from Uniform $[0, 1]$, and calculated the dependence parameters for inner and outer copula respectively as $\theta_{12}(z_i) = \exp(\sin(2\pi z_i)) + 1$, $\theta(z_i) = \exp(\sin(\frac{3}{2}\pi z_i)) + 4$.

Step (b). We generated the trivariate random variables (v_{1i}, v_{2i}, v_{3i}) conditional on z_i for $i = 1, \ldots, n$ from an nested Archimedean copula model with parameter $\theta_{12}(z_i)$ for the inner copula between (v_{1i}, v_{2i}), and $\theta(z_i)$ for the outer copula, by the `R` package `copula` (Hofert & Mächler 2011).

Step (c). Letting $S_j(t_{ji}|z_i) = v_{ji} = g(S^*_{0j}(t_{ji})^{e^{\beta^*_j z_i}}, S_{0D}(t_{ji})^{e^{\beta_D z_i}}, \theta(z_j))$, for $j = 1, 2$, we used the survivor functions of the Weibull distributions $S^*_{0j}(\cdot)$ and $S_{0D}(\cdot)$, and we solved for the t_{ji}. The scale and shape parameters, together with the regression coefficients are pre-determined. Letting $S_D(d_i|z_i) = v_{3i} = S_{0D}(d_i)^{e^{\beta_D z_i}}$, we obtained the terminating event time d_i for $i = 1, \ldots, n$. Thus we have formed the generated event times and terminating event times.

Step (d). We generated the independent (administrative) censoring times c_{Ai} independently from (v_{1i}, v_{2i}, v_{3i}) from the exponential distribution with the parameter chosen to give a censoring rate of 25 percent. We then calculated $c_i = d_i \wedge c_{Ai}$ with the indicator $\delta_{Di} = I(d_i \leq c_{Ai})$ and $u_{ji} = t_{ji} \wedge c_i$ with the indicator $\delta_{ji} = I(t_{ji} \leq c_i)$.

Steps (a)-(d) yield a generated observed-data: $\{[(u_{ji}, \delta_{ji}) : j = 1, 2] \bigcup [c_i, \delta_{Di}] \bigcup [z_i] : i = 1, \ldots, n\}$

We considered $n = 500, 1000$ and 2000 to generate medium to large studies. The functions of outer and inner copula parameters ($\theta(\cdot)$ and $\theta_{12}(\cdot)$) were determined such that the corresponding Kendall's $(\tau(\cdot), \tau_{12}(\cdot))$ ranges between $(0.2, 0.5)$, and $(0.6, 0.8)$ respectively. to represent weak and moderate-to-strong dependence structures. We used the Cox PH model to obtain $\tilde{S}^*_j(\cdot|Z)$ and $\tilde{S}_D(\cdot|Z)$ in the estimation procedure.

5.4.2 Simulation Outcomes

We conducted estimation under four scenarios. *Scenario (I)*: We assumed both $\theta_{12}(\cdot)$ and $\theta(\cdot)$ to be scalars and obtained the pseudo-MLE. This set of estimates can be viewed as the

average estimate of dependence . *Scenario (II)*: We assumed $\theta(\cdot)$ to be a scalar and modeled the function $\theta_{12}(\cdot)$ through B-spline basis functions. *Scenario (III)*: We assumed $\theta_{12}(\cdot)$ to be a scalar and modeled the function $\theta(\cdot)$ through B-spline basis functions. *Scenario (IV)*: We kept both $\theta_{12}(\cdot)$ and $\theta(\cdot)$ as unknown functions and estimated both through B-spline basis functions. The first three scenarios can be viewed as reference estimates for scenario (IV). We evaluated the pseudo-MLE of the association parameters (θ, θ_{12}), the coefficient estimates of the B-spline functions, and the estimated conditional marginal survivor functions, with one thousand generated sets of data.

Table 5.1 presents a summary of the estimates for the Cox regression coefficients, for $S_j^*(\cdot|Z)$, $j = 1, 2$ and $S_D(\cdot|Z)$, based on one thousand repetitions under the nested Clayton model. The sample means of the estimates are close to the true parameter values.

Table 5.2 and figures 5.1 to 5.3 show the estimates of coefficients for the spline approximation function and the plots for the sets of estimated $\theta_{12}(\cdot)$, and/or $\theta(\cdot)$ for the four scenarios described earlier. Table 5.3 provides the τ and τ_{12} estimates for scenarios I, II and III. The estimated $\tau_{12}(\cdot)$ and $\tau(\cdot)$ are close to the true functions, which verifies the consistency of the proposed estimators. The true function curves are fully covered by the CB associated with the pseudo-MLE in every plot.

Figures 5.4-5.15 display the estimated marginal survivor functions $S_1(\cdot|Z)$ and $S_2(\cdot|Z)$ for $Z = 0.3, 0.5, 0.7$ for scenarios (I) to (IV) respectively. Each plot provides three curves representing the real marginal survivor function, the estimated survivor function, and the estimates from naïve Cox model ignoring informative censoring. It is clear that the naïve estimates are biased and the proposed approach provided consistent estimates, when sample size increases the estimate gets closer to the true marginal and CB gets narrower. This confirms that the proposed approach provides consistent estimators for the marginal survivor functions. It is shown that proposed approach provides better estimate than the naïve estimator for all scenarios. Estimates for scenario (I) and scenario (II) are just slightly biased for certain values of Z, but this is expected as θ is treated as a scalar for all values of Z. The average estimates of the marginals across all Z values are unbiased. Scenario (III) provides the closest estimates to scenario (IV) because it assumes $\theta(\cdot)$ to be a function of Z. This provides a guideline that in practice if the association parameter does not vary systematically across the covariates, and the main goal is to estimate the marginal survivor functions, then assuming $\theta(\cdot)$ and/or $\theta_{12}(\cdot)$ to be a scalar will ease the computation, while still providing good estimates for marginal survivor function on average.

Table 5.1: Estimates of Coefficients in Cox PH Model for $S_j^*(\cdot|Z)$, $j = 1, 2$, and $S_D(\cdot|Z)$ with Simulated Data

n		β_1^*	β_2^*	β_D
	True	2	2	2
500	sm^*	1.935	1.937	1.932
	$\hat{se}$	0.402	0.404	0.402
1000	sm	1.936	1.936	1.941
	$\hat{se}$	0.380	0.380	0.381
2000	sm	1.933	1.934	1.934
	se	0.368	0.369	0.369

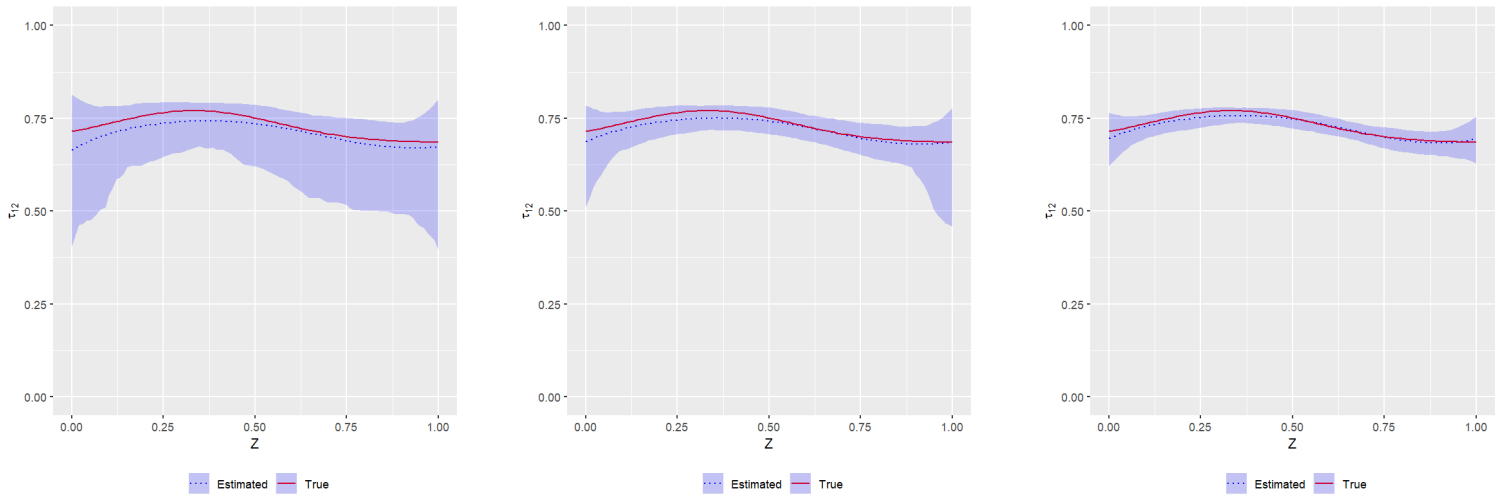
sm^* sample mean of estimates

Table 5.3: Estimates of (τ, τ_{12}) with Simulated Data for Scenarios I, II and III

n		Scenario	I	II	III	IV
500	τ	sm	0.693	-	0.698	-
		se	0.129	-	0.130	-
	τ_0	sm	0.480	0.483	-	-
		se	0.092	0.093	-	-
1000	τ	sm	0.699	-	0.703	-
		se	0.130	-	0.130	-
	τ_0	sm	0.483	0.483	-	-
		se	0.091	0.091	-	-
2000	τ	sm	0.702	-	0.706	-
		se	0.130	-	0.131	-
	τ_0	sm	0.486	0.486	-	-
		se	0.091	0.091	-	-

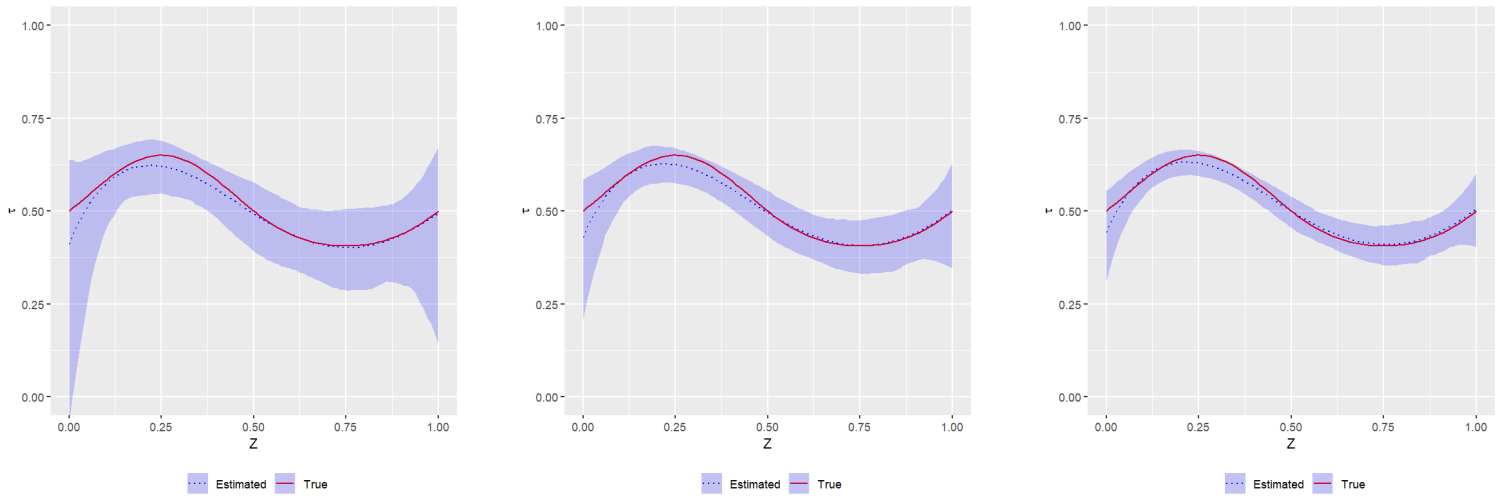
Table 5.2: Estimates of Coefficients $(\boldsymbol{\alpha}, \boldsymbol{\alpha}_{12})$ for B-spline Basis Functions with Simulated Data

			$\theta(\cdot)$					$\theta_{12}(\cdot)$				
n	Scenario	coefficient	$\alpha^{(1)}$	$\alpha^{(2)}$	$\alpha^{(3)}$	$\alpha^{(4)}$	$\alpha^{(5)}$	$\alpha_{12}^{(1)}$	$\alpha_{12}^{(2)}$	$\alpha_{12}^{(3)}$	$\alpha_{12}^{(4)}$	$\alpha_{12}^{(5)}$
500	I	*sm*										
		sse										
	II	*sm*						1.650	1.900	2.053	1.553	1.662
		sse						0.395	0.364	0.472	0.344	0.320
	III	*sm*	0.902	1.859	0.861	0.746	1.109					
		sse	0.403	0.377	0.511	0.391	0.374					
	IV	*sm*	1.667	2.000	2.031	1.548	1.694	0.956	1.852	0.946	0.702	1.127
		sse	0.318	0.331	0.452	0.353	0.285	0.353	0.374	0.493	0.385	0.309
1000	I	*sm*										
		sse										
	II	*sm*						1.705	1.937	2.074	1.571	1.698
		sse						0.268	0.254	0.334	0.248	0.247
	III	*sm*	0.931	1.871	0.864	0.761	1.114					
		sse	0.260	0.247	0.335	0.255	0.245					
	IV	*sm*	1.699	2.035	2.041	1.588	1.706	0.938	1.895	0.922	0.703	1.122
		sse	0.226	0.235	0.318	0.243	0.190	0.238	0.245	0.325	0.248	0.209
2000	I	*sm*										
		sse										
	II	*sm*						1.730	1.964	2.108	1.565	1.719
		sse						0.188	0.183	0.236	0.169	0.159
	III	*sm*	0.958	1.880	0.878	0.763	1.123					
		sse	0.176	0.173	0.240	0.195	0.230					
	IV	*sm*	1.722	2.043	2.077	1.575	1.713	0.949	1.917	0.926	0.708	1.125
		sse	0.147	0.164	0.221	0.161	0.129	0.158	0.169	0.226	0.174	0.143



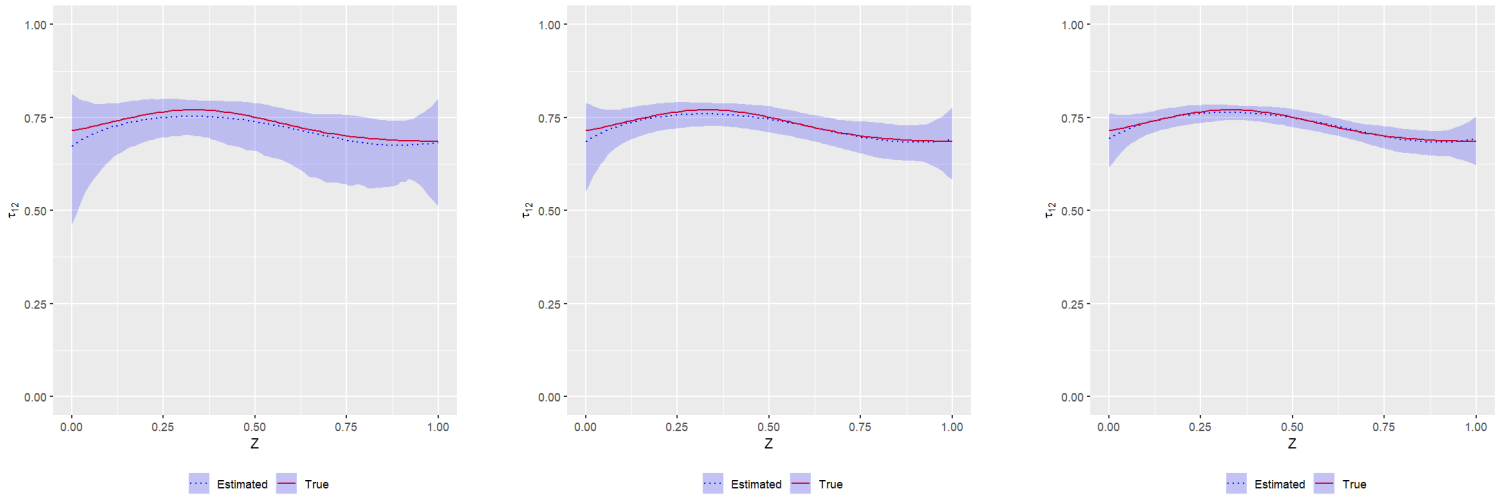
$n = 500$ $n = 1000$ $n = 2000$

Figure 5.1: Estimates of $\tau_{12}(Z)$ for Scenario (II) with Simulated Data

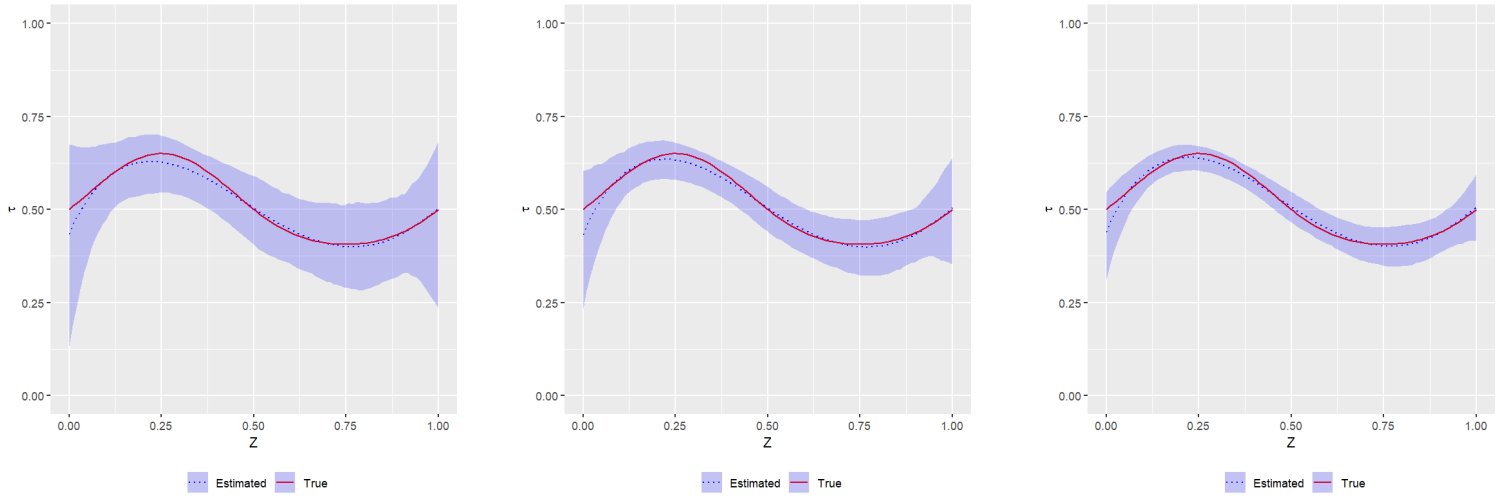


$n = 500$ $n = 1000$ $n = 2000$

Figure 5.2: Estimates of $\tau(Z)$ for Scenario (III) with Simulated Data



$n = 500$ $n = 1000$ $n = 2000$



$n = 500$ $n = 1000$ $n = 2000$

Figure 5.3: Estimates of $\tau_{12}(Z)$ and $\tau(Z)$ for Scenario (IV) with Simulated Data

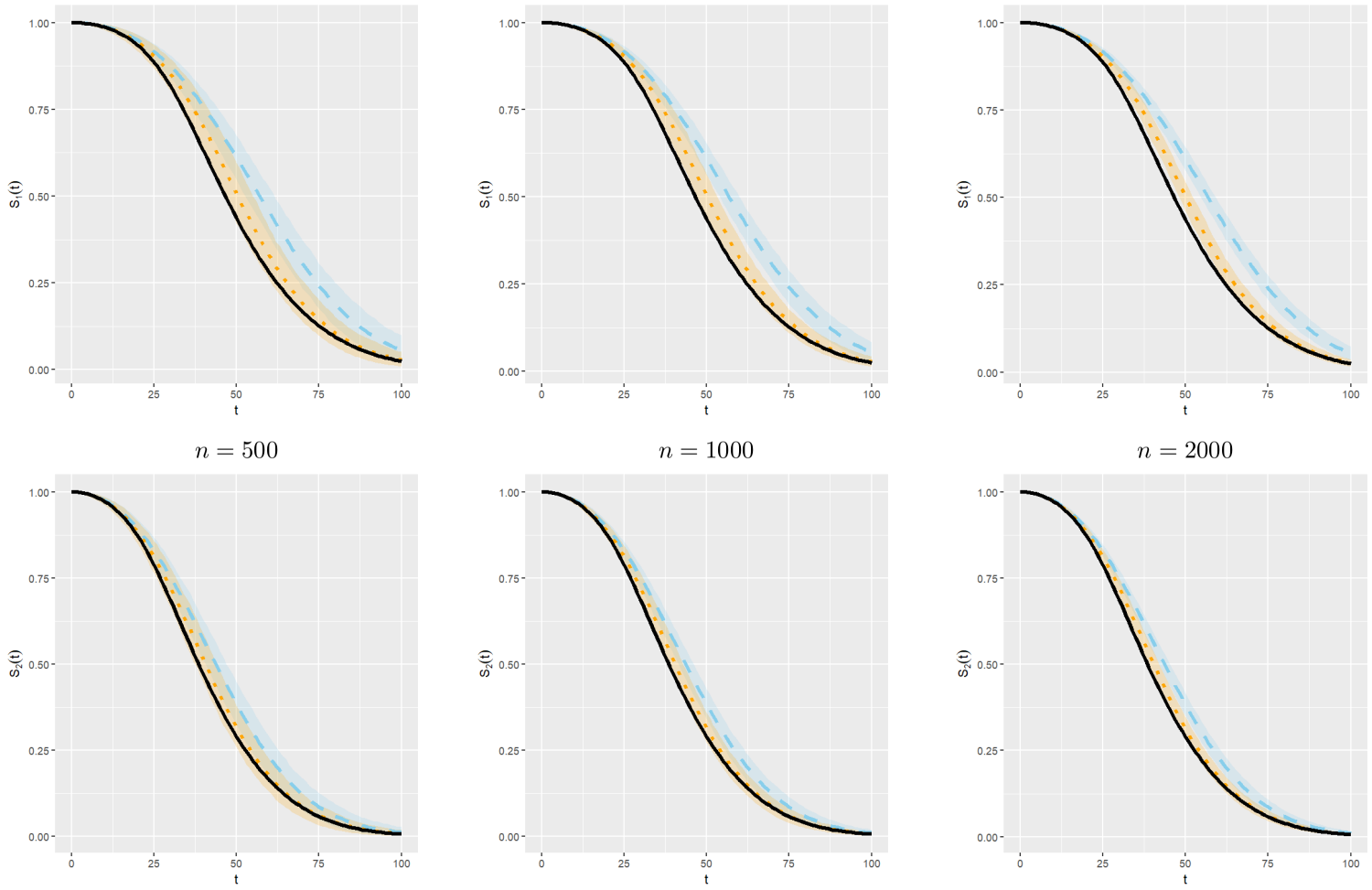


Figure 5.4: Estimates of Marginal Survivor Functions $S_1(t|Z)$ and $S_2(t|Z)$ when $Z = 0.3$ with Simulated Data. Scenario (I). Orange dotted: naïve. Skyblue dashed: proposed. Black solid: true.

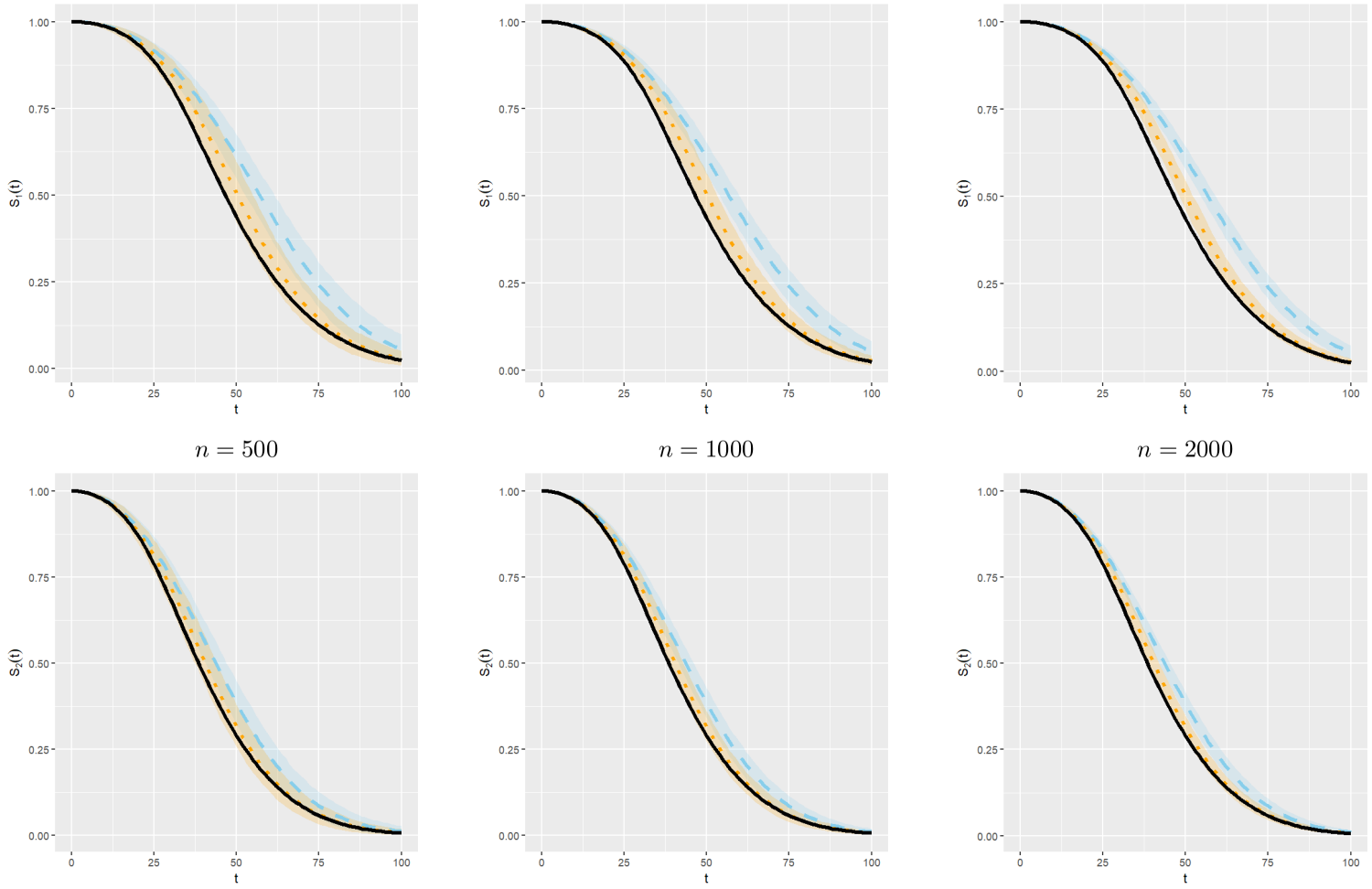


Figure 5.5: Estimates of Marginal Survivor Functions $S_1(t|Z)$ and $S_2(t|Z)$ when $Z = 0.3$ with Simulated Data. Scenario (II). Orange dotted: naïve. Skyblue dashed: proposed. Black solid: true.

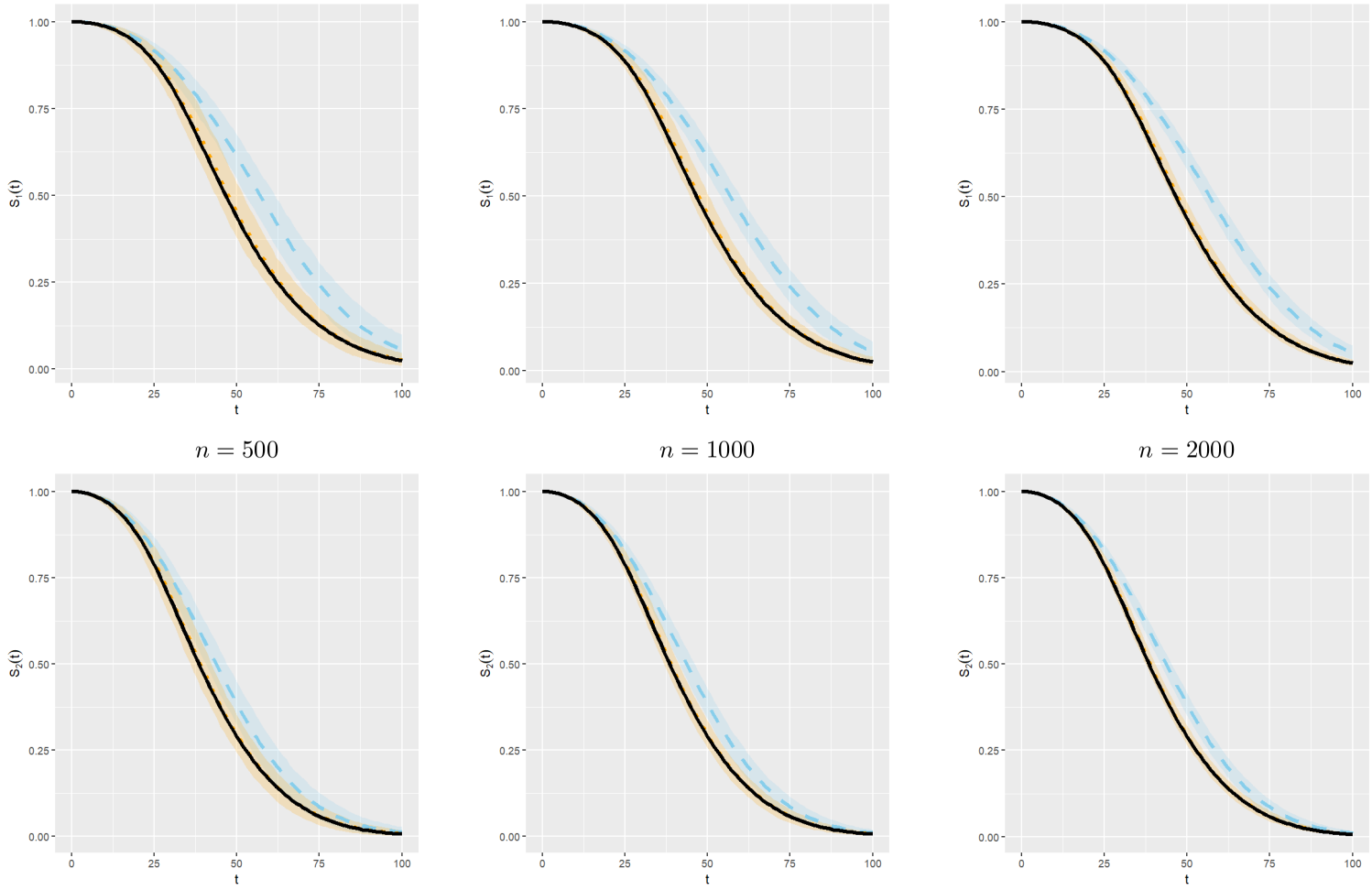


Figure 5.6: Estimates of Marginal Survivor Functions $S_1(t|Z)$ and $S_2(t|Z)$ when $Z = 0.3$ with Simulated Data. Scenario (III). Orange dotted: naïve. Skyblue dashed: proposed. Black solid: true.

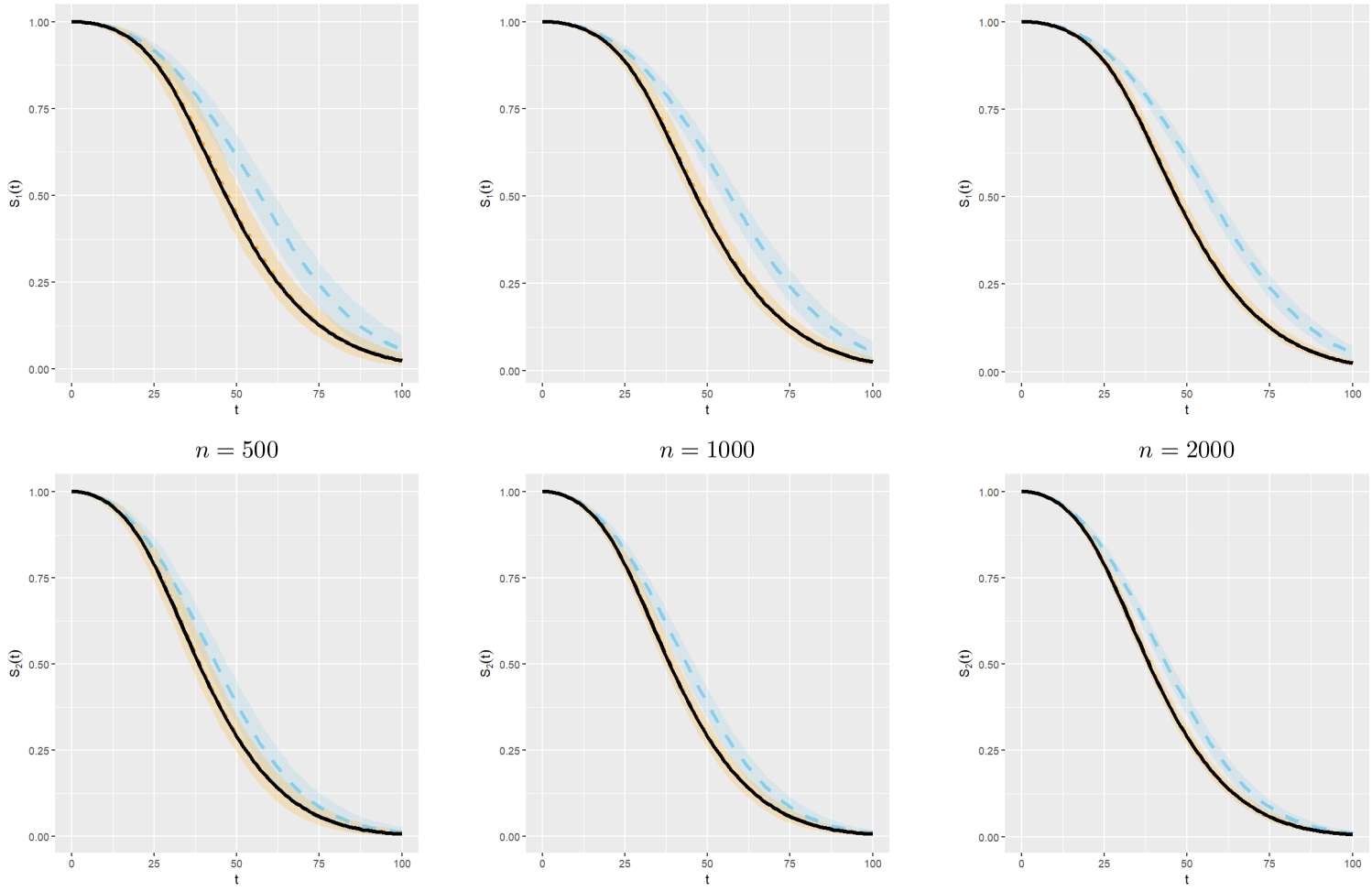


Figure 5.7: Estimates of Marginal Survivor Functions $S_1(t|Z)$ and $S_2(t|Z)$ when $Z = 0.3$ with Simulated Data. Scenario (IV). Orange dotted: naïve. Skyblue dashed: proposed. Black solid: true.

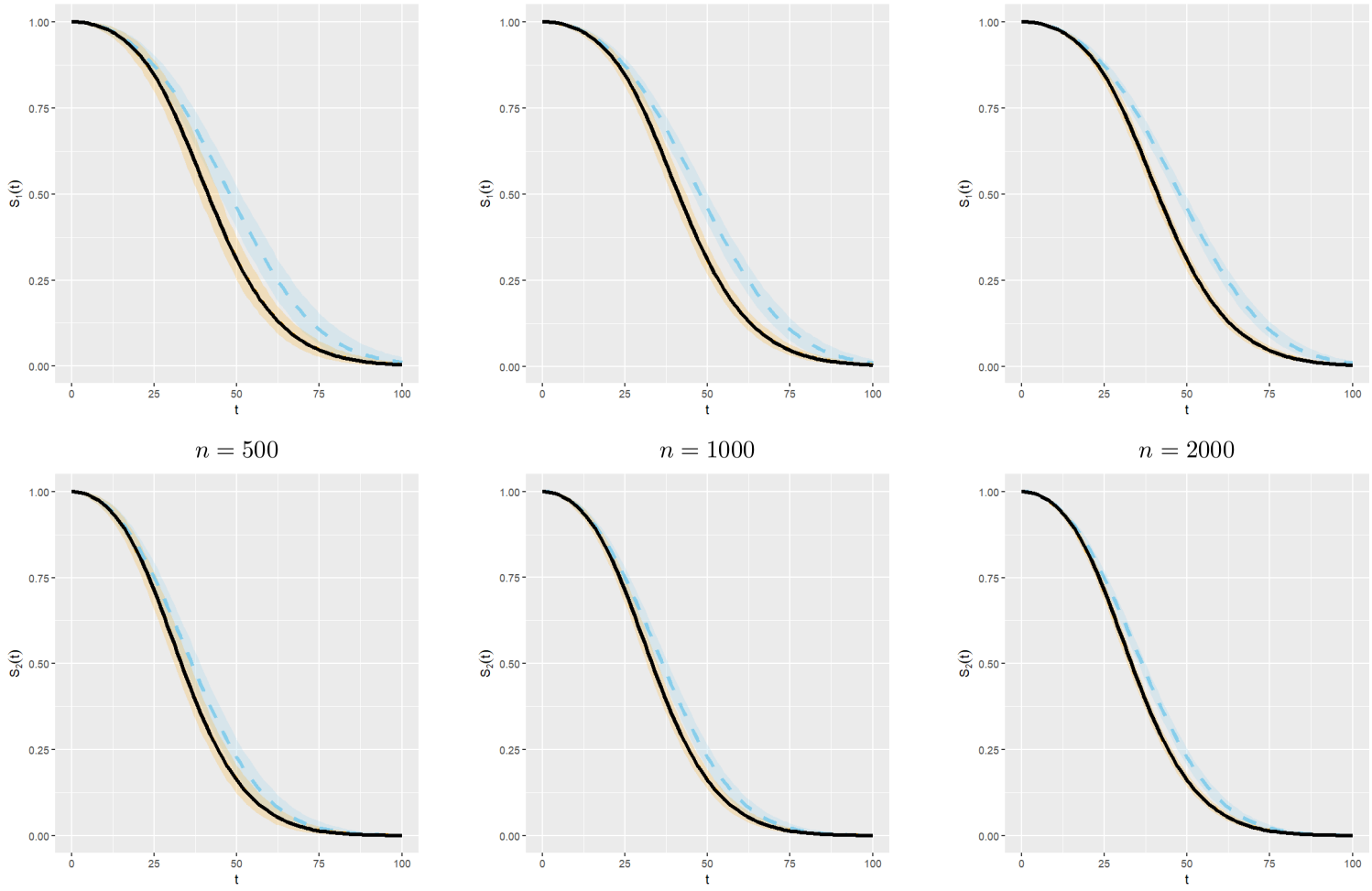


Figure 5.8: Estimates of Marginal Survivor Functions $S_1(t|Z)$ and $S_2(t|Z)$ when $Z = 0.5$ with Simulated Data. Scenario (I). Orange dotted: naïve. Skyblue dashed: proposed. Black solid: true.

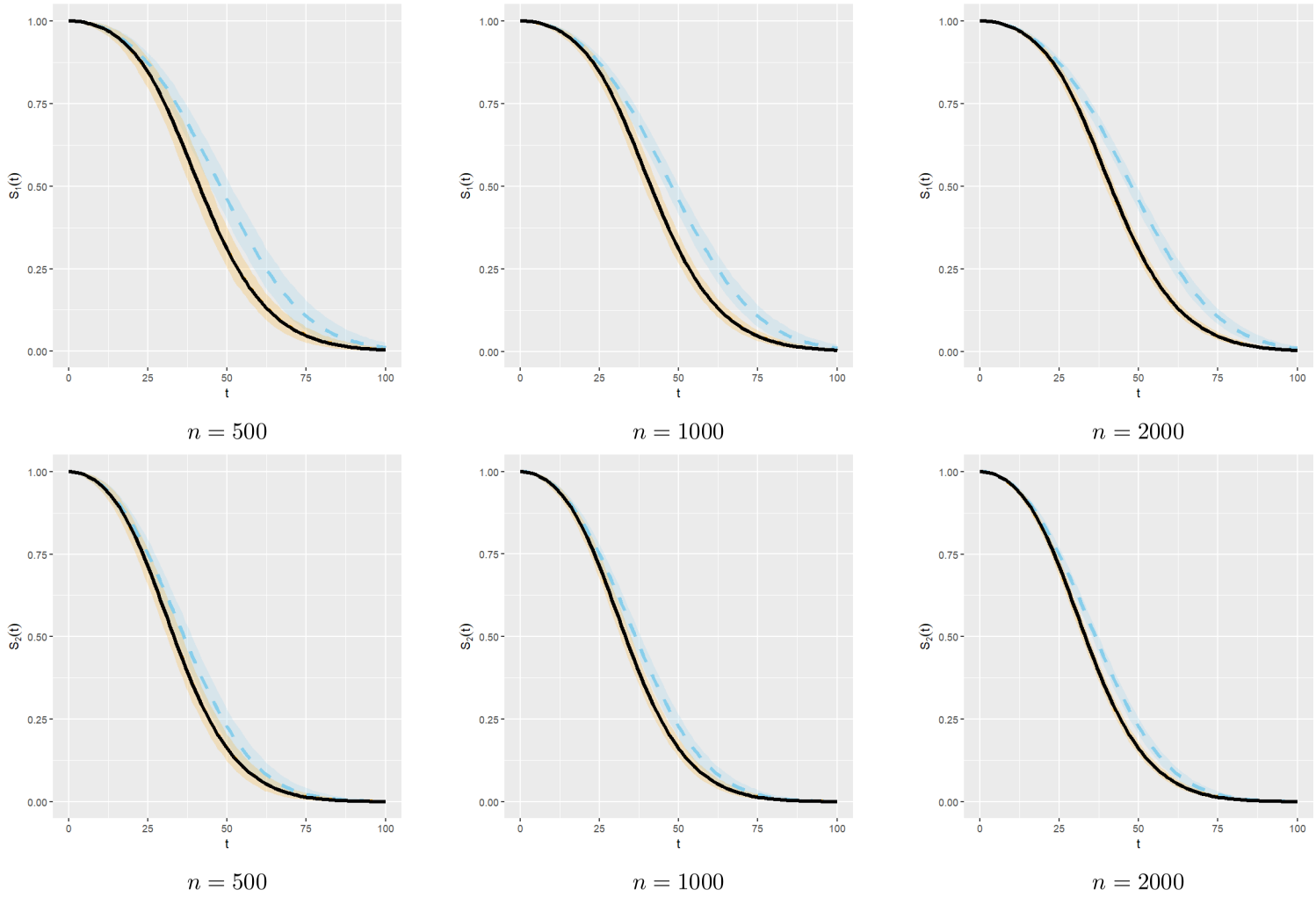


Figure 5.9: Estimates of Marginal Survivor Functions $S_1(t|Z)$ and $S_2(t|Z)$ when $Z = 0.5$ with Simulated Data. Scenario (II). Orange dotted: naïve. Skyblue dashed: proposed. Black solid: true.

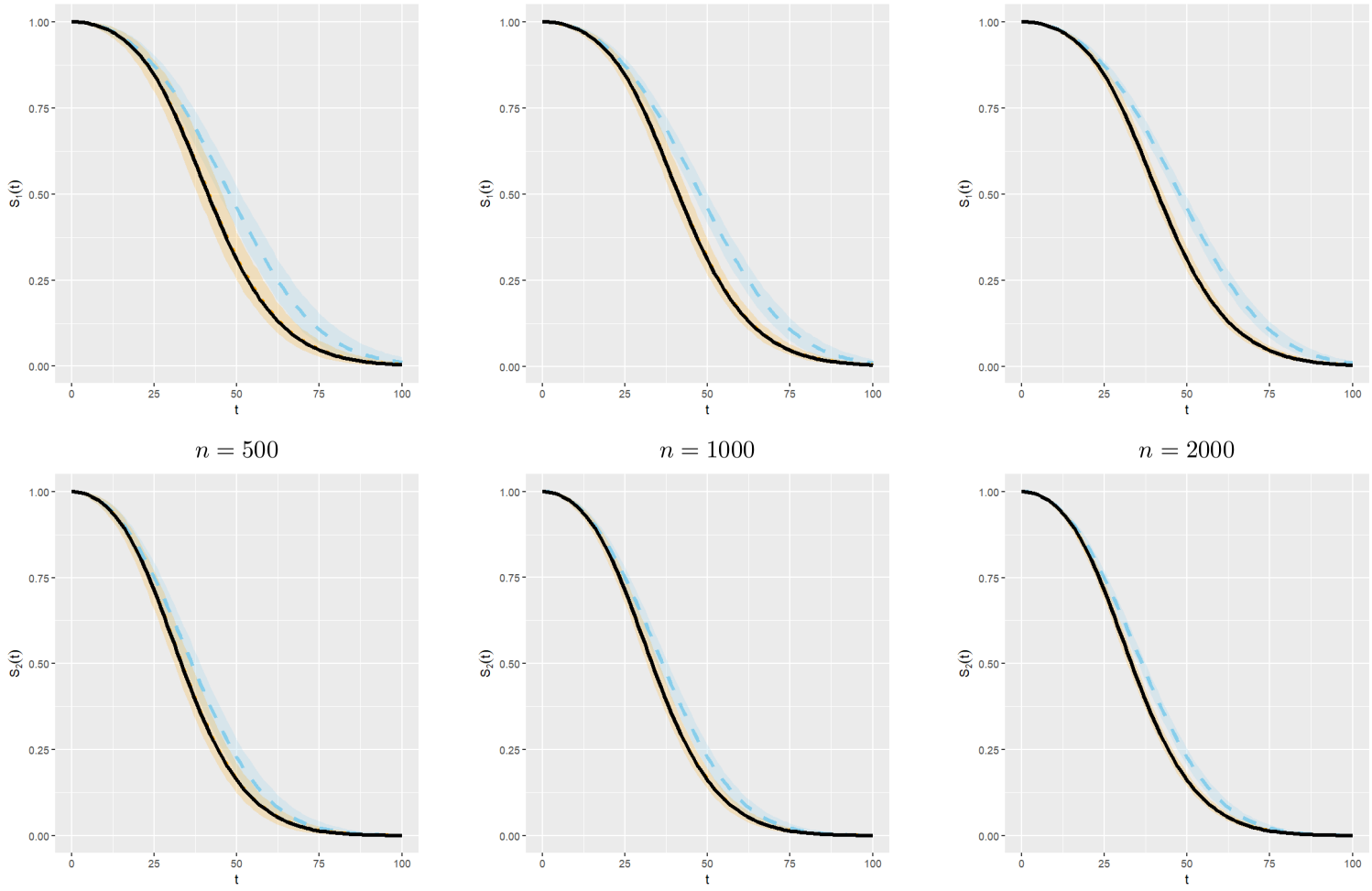


Figure 5.10: Estimates of Marginal Survivor Functions $S_1(t|Z)$ and $S_2(t|Z)$ when $Z = 0.5$ with Simulated Data. Scenario (III). Orange dotted: naïve. Skyblue dashed: proposed. Black solid: true.

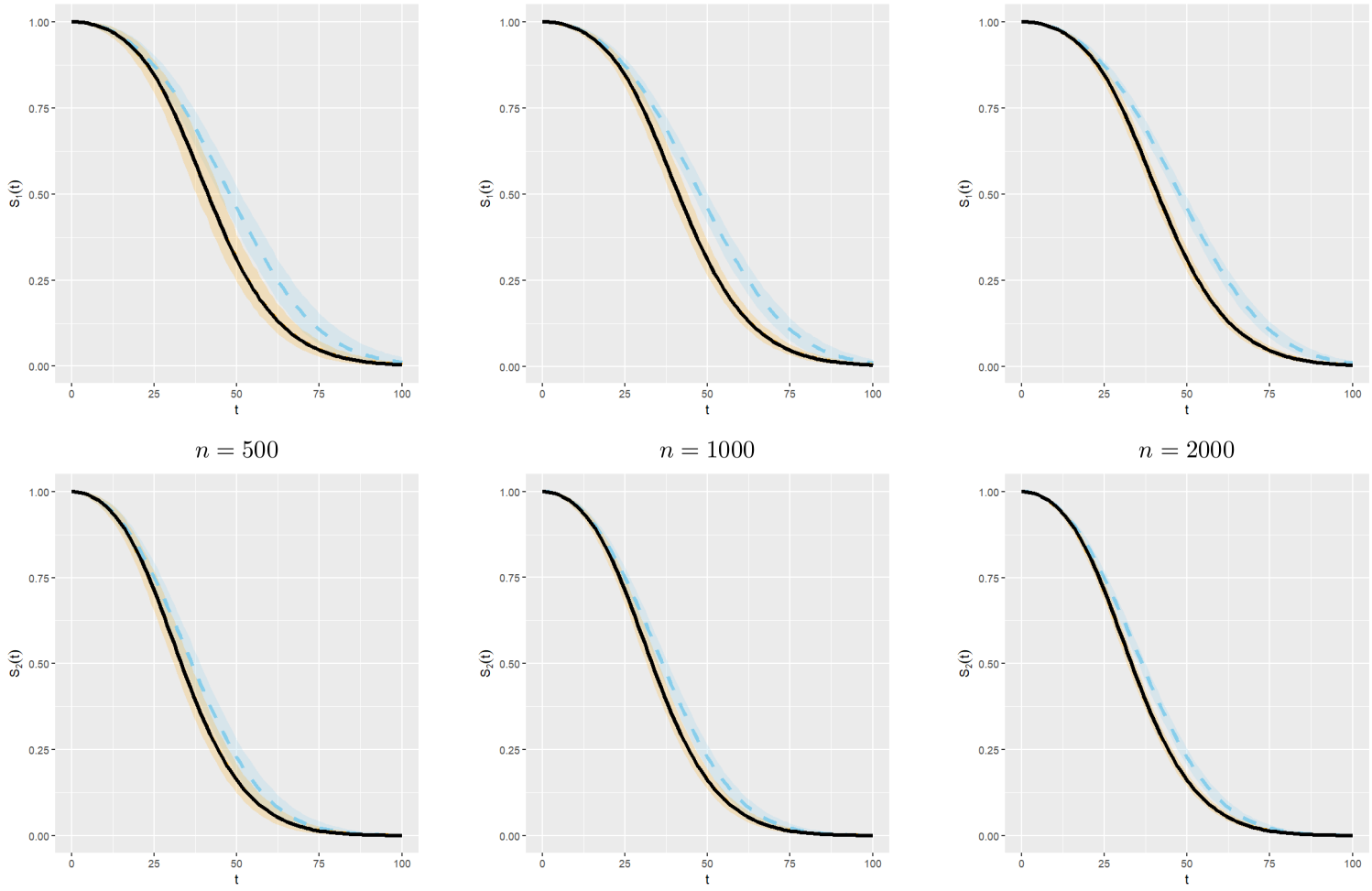


Figure 5.11: Estimates of Marginal Survivor Functions $S_1(t|Z)$ and $S_2(t|Z)$ when $Z = 0.5$ with Simulated Data. Scenario (IV). Orange dotted: naïve. Skyblue dashed: proposed. Black solid: true.

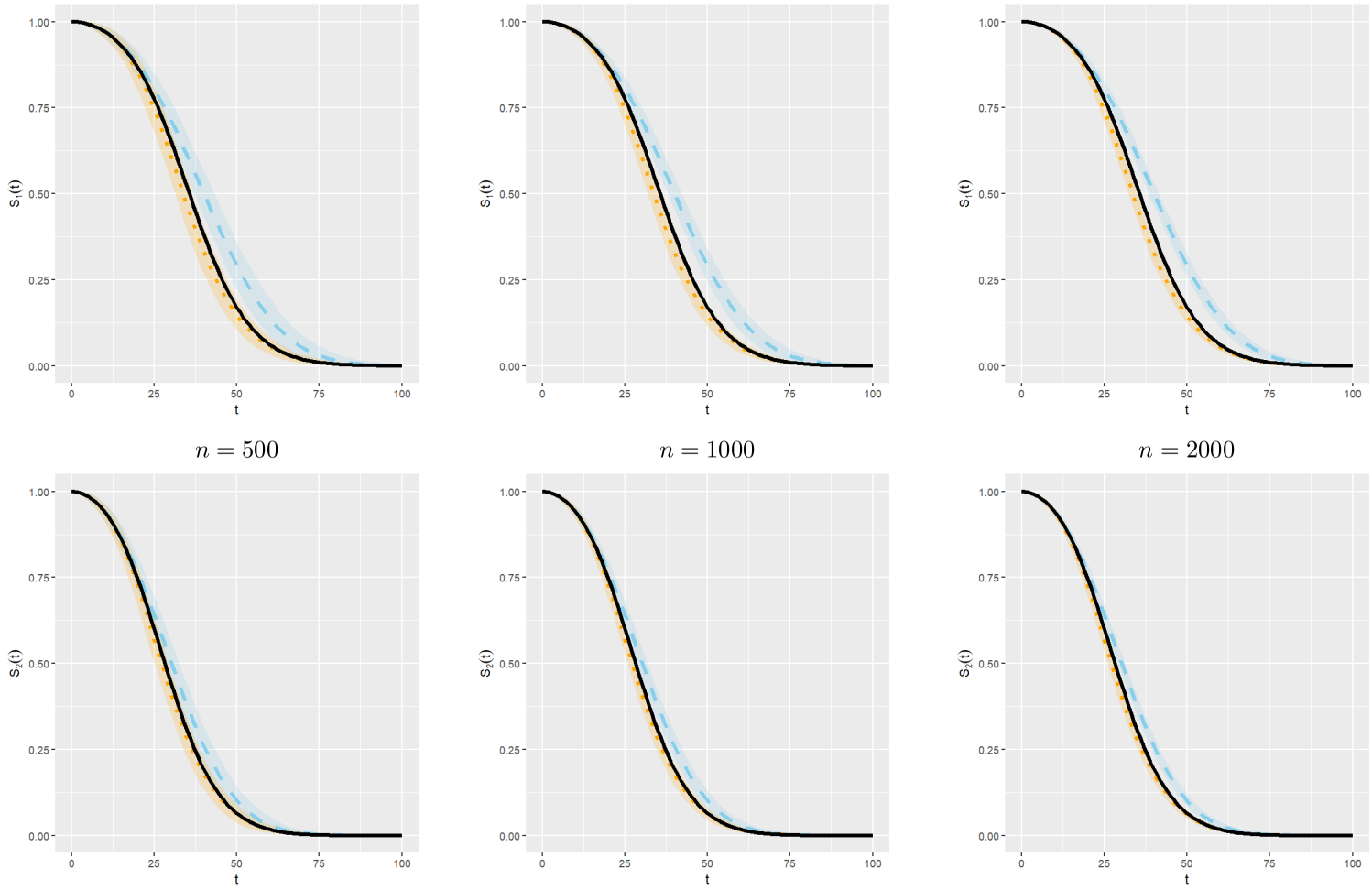


Figure 5.12: Estimates of Marginal Survivor Functions $S_1(t|Z)$ and $S_2(t|Z)$ when $Z = 0.7$ with Simulated Data. Scenario (I). Orange dotted: naïve. Skyblue dashed: proposed. Black solid: true.

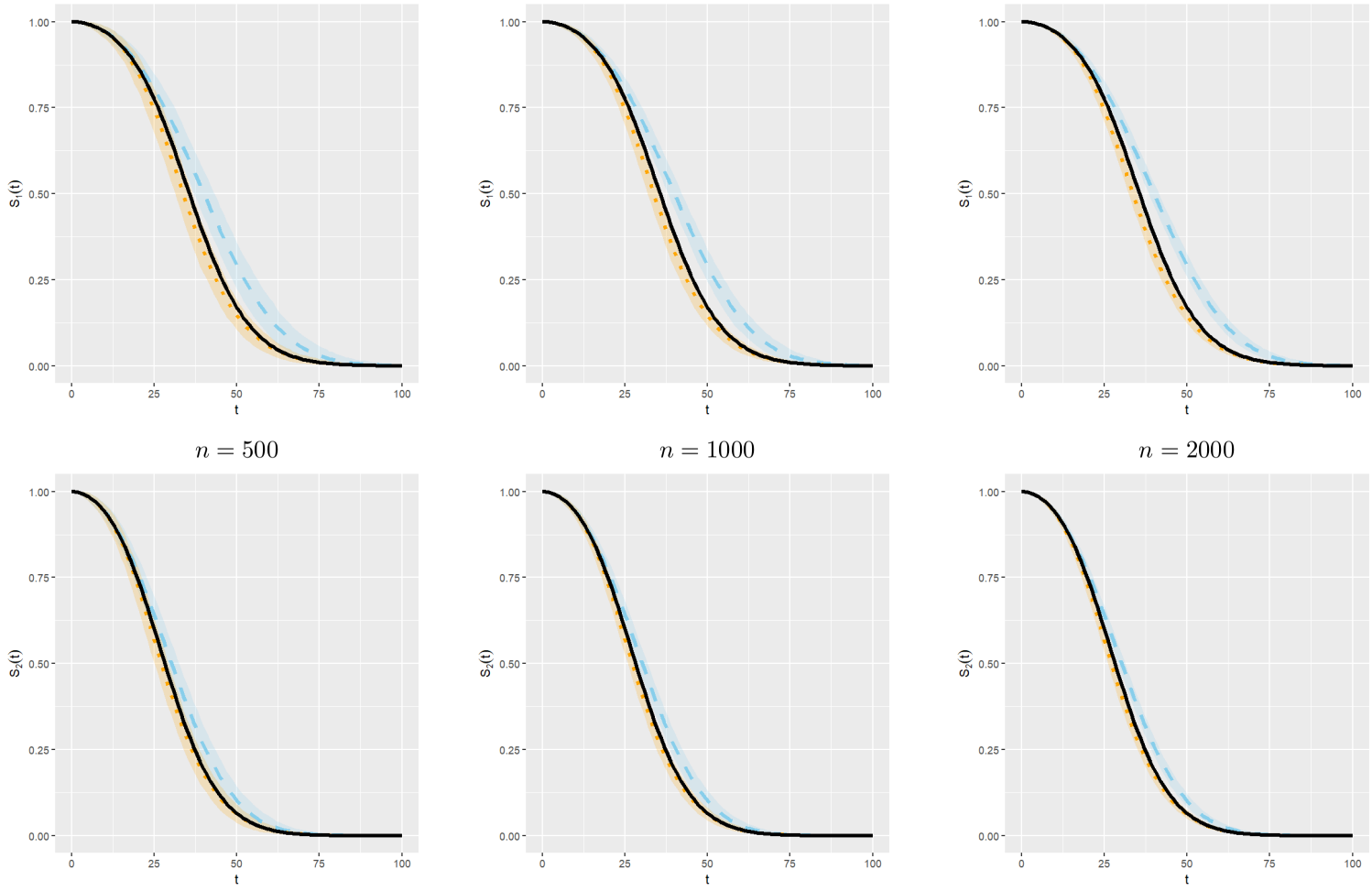


Figure 5.13: Estimates of Marginal Survivor Functions $S_1(t|Z)$ and $S_2(t|Z)$ when $Z = 0.7$ with Simulated Data. Scenario (II). Orange dotted: naïve. Skyblue dashed: proposed. Black solid: true.

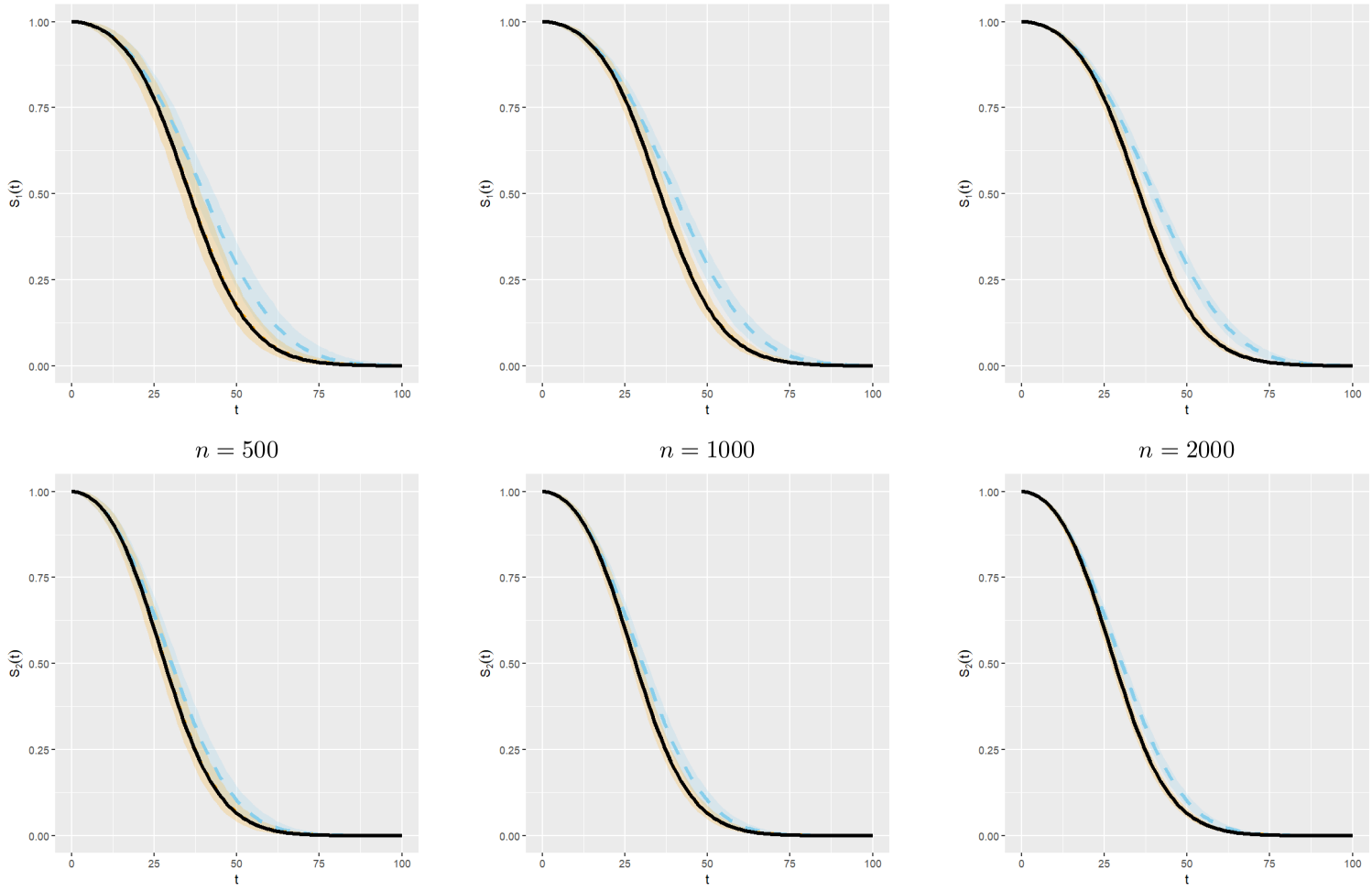


Figure 5.14: Estimates of Marginal Survivor Functions $S_1(t|Z)$ and $S_2(t|Z)$ when $Z = 0.7$ with Simulated Data. Scenario (III). Orange dotted: naïve. Skyblue dashed: proposed. Black solid: true.

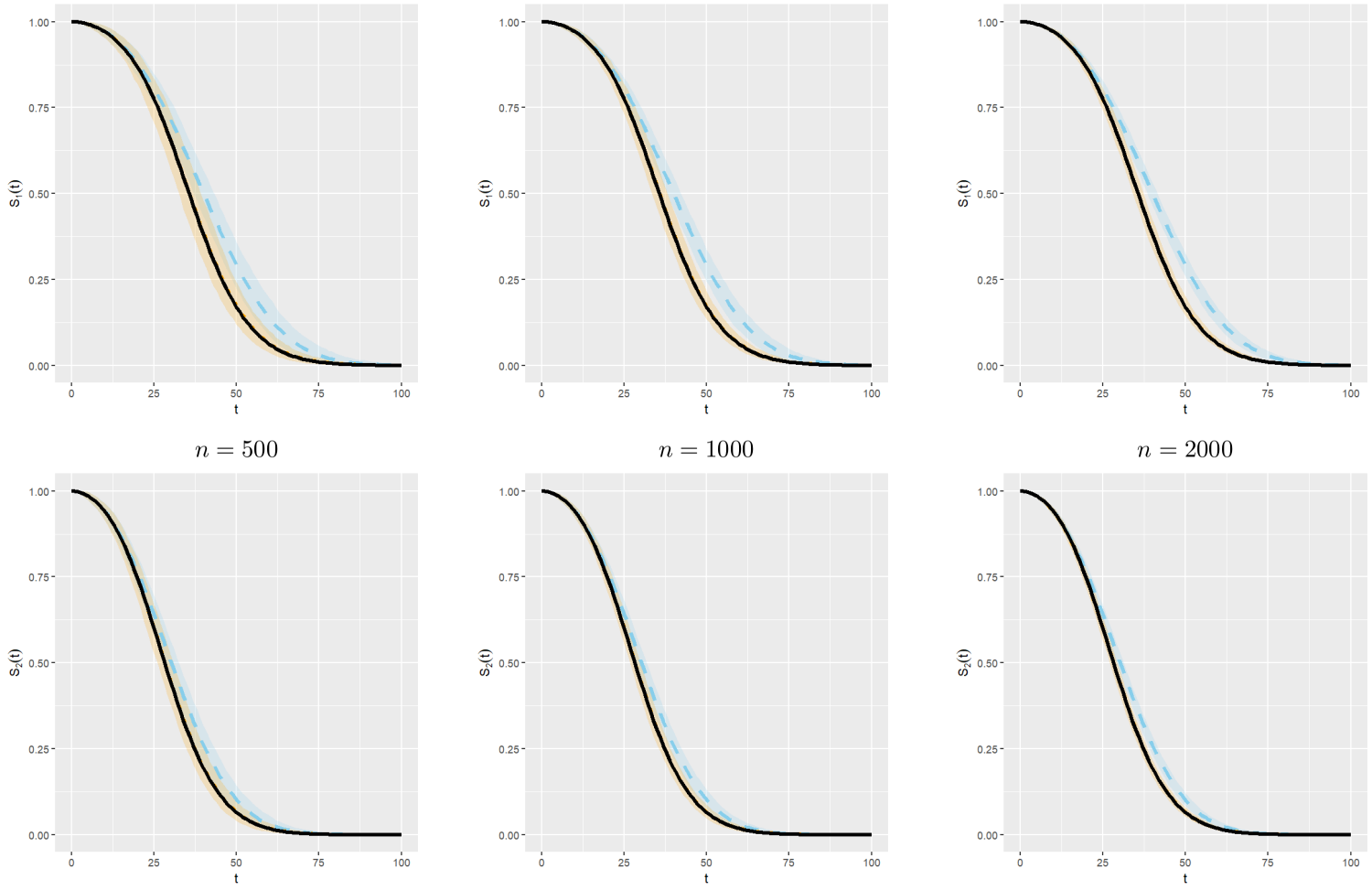


Figure 5.15: Estimates of Marginal Survivor Functions $S_1(t|Z)$ and $S_2(t|Z)$ when $Z = 0.7$ with Simulated Data. Scenario (IV). Orange dotted: naïve. Skyblue dashed: proposed. Black solid: true.

5.5 Analysis of BC-BRCAS Data (IV)

In this section we present an analysis using data from the BC-BRCAS program (McBride et al. 2016).

5.5.1 Study Description

The study subjects are those from the cohort as defined in Chapter 2, the cohort $\mathcal{P}_{\text{final}}$. We took each subject's date of cancer diagnosis as her time origin. We considered the first event time T_1 as the time to RSC and the second time T_2 as the time to the first CVD-related hospitalization after the diagnosis. The availability of information on T_1 and T_2 is subject to censoring by death or the end of administrative data extraction, whichever occurs sooner. Thus, we formulated each subject's censoring time as $C = D \wedge C_A$, where D is the time to death and C_A is the time to the end of the administrative data extraction window. In the analysis, we only included the new patients who were referred to the BC Cancer Agency for treatment and with known stage and treatment information. The covariates Z include one continuous variable *age*, denoted X, and three discrete variables, namely *stage*,*treatment*, and *birth era*, denoted Z^*. Preliminary exploratory analysis shows that other sociodemographic factors did not have any significant effect, so we did not include them in the final analysis. Table 5.4 shows descriptive information on the study subjects. We discretize *age* in this table for descriptive and exploratory data analyses.

5.5.2 Estimates of Correlations between Event Times

We modeled S_j^* and S_D through marginal (Kaplan Meier estimator) and conditional approaches (Cox PH model). Table 5.5 shows a summary of the estimates of regression coefficients for the Cox PH model of $S_j^*(\cdot|Z)$, $j = 1, 2$, and $S_D(\cdot|Z)$. Table 5.6 shows the estimates of τ_{12} and τ for scenarios (I)-(III), as described in section 5.4. Figure 5.20 and figure 5.21 show the plots of the estimated functions for $\tau_{12}(\cdot)$ and $\tau(\cdot)$ with confidence bands, for scenario (IV).

Late stage at diagnosis seems to be a risk factor for the increased association between T_1 and T_2. It appears that chemo treatment is associated with higher dependence between T_1 and T_2. Treatment of radiation, on the other hand, did not appear to increase the association. Further investigation would be desirable to examine the effect of specific types of chemo or dose and location of radiation. Although it seems that those born in later era have an increased risk of association between T_1 and T_2, one needs to be careful when interpreting these results because potential informative left truncation exists. Those who were born in an earlier era and experienced T_2 may not have been included in the study. In addition, because of left censoring due to the administrative starting time, the T_2 observed for those born in earlier era may not be the first time to cardiovascular disease.

As shown in figure 5.20 , the Kendall's τ_{12} between T_1 and T_2 is stronger among those diagnosed at a late stage, while the association τ between T_j and D are roughly the same for both early and late stages. The plot on $\tau(\cdot)$ in figure 5.20 also confirms that informative censoring exists due to the terminating event. One needs to take informative censoring into consideration when dealing with dependence among multivariate event times.

Figure 5.16 and figure 5.19 present marginal survivor function estimates for $S_1(t|Z)$ and $S_2(t|Z)$, respectively, among those diagnosed at age 49 and born in era II, for 4 different treatment groups and 2 different stages at diagnosis. As expected, compared to early stage at diagnosis, advanced stage at diagnosis is a risk factor for earlier time to RSC and time to CVD. Treatment of only radiation seems to be beneficial compared to other treatment options. For comparison, figure 5.17 and figure 5.18 show the corresponding survival curve estimates using naïve approach by applying Cox PH model directly on the Observed-Data$_j$ in (3.2), comparing to the estimates using proposed approach under scenario (IV). The naïve curves appear different from the proposed estimates, especially in Figure 5.18 on the estimated survivor function of time to CVD, where the naïve estimates showed no treatment effect on time to CVD.

Table 5.4: Summary Statistics of BC-BRCAS Data $\mathcal{P}_{\text{final}}$ for Regression Analyses

		Total	$N(T_1^{\text{obs}})^{\dagger}$	$\overline{T_1^{\text{obs}}}^{\ddagger}$	$N(T_2^{\text{obs}})$	$\overline{T_2^{\text{obs}}}$	$N(D^{\text{obs}})$	$\overline{D^{\text{obs}}}$
Overall		36735	11025	5.30	5468	6.48	11330	7.56
Diagnosis Age Group	<40	2617	1019	4.61	109	7.15	721	5.89
	40+	34118	10006	5.37	5359	6.47	10609	7.68
Stage	Early (I and II)	33032	9421	5.61	5044	6.65	9685	8.00
	Late (III)	3703	1604	3.48	424	4.48	1645	4.96
Treatment	Chemo and Rad	11159	3516	4.94	899	6.40	2752	6.20
	Chemo	2866	945	4.27	250	6.19	734	5.80
	Rad	14470	4033	6.13	2562	6.97	4295	8.71
	No Chemo or Rad	8240	2531	4.86	1757	5.85	3549	7.59
Era[1]	I	7180	2772	5.27	2435	6.44	4876	8.42
	II	14166	4461	5.81	2285	6.73	3806	7.42
	III	15389	3792	4.71	748	5.86	2418	5.76

[1]Era I: Born 1900 - 1927; Era II: Born 1928 - 1945; Era III: Born 1946 - 1986

†,$N(T_1^{\text{obs}}), N(T_2^{\text{obs}}), N(D^{\text{obs}})$:,numbers of individuals with times to RSC, CVD, death (T_1, T_2, D) observed, respectively

‡,$\overline{T_1^{\text{obs}}}, \overline{T_2^{\text{obs}}}, , \overline{D^{\text{obs}}}$: sample means of the observed times to RSC„CVD, death (T_1, T_2, D), respectively

Table 5.5: Estimated Coefficients from Cox Models for $S_D(\cdot|Z)$, and $S_j^*(\cdot|Z)$ with BC-BRCAS Data $\mathcal{P}_{\text{final}}$

		$\hat{\beta}_D$	$\hat{se}$	sse	$\hat{\beta}_1^*$	$\hat{se}$	sse	$\hat{\beta}_2^*$	$\hat{se}$	sse
age at diagnosis		0.031	0.002	0.001	0.015	0.001	0.001	0.035	0.001	0.001
Stage	Early	-	-	-	-	-	-	-	-	-
	Late	0.879	0.029	0.026	0.744	0.026	0.025	0.704	0.027	0.029
Treatment	None	-	-	-	-	-	-	-	-	-
	Chemo Only	0.296	0.045	0.037	0.176	0.037	0.031	0.259	0.04	0.034
	Rad Only	-0.332	0.023	0.026	-0.250	0.020	0.023	-0.256	0.021	0.021
	Chemo + Rad	0.121	0.033	0.038	-0.034	0.028	0.033	0.122	0.029	0.039
Era[1]	I	-	-	-	-	-	-	-	-	-
	II	-0.641	0.030	0.032	-0.534	0.025	0.021	-0.445	0.026	0.023
	III	-0.715	0.050	0.044	-0.682	0.041	0.035	-0.613	0.044	0.036

[1]Era I: Born 1900 - 1927; Era II: Born 1928 - 1945; Era III: Born 1946 - 1986

Table 5.6: Estimates of τ and τ_{12} with BC-BRCAS Data $\mathcal{P}_{\text{final}}$ under Different Scenarios

Approach for θ_{12}, θ	Approach for S_j^*, S_D		Overall		None		Chemo		Rad		Both	
			τ_{12}	τ	τ_{12}	τ	τ_{12}	τ	τ_{12}	τ	τ_{12}	τ
(I) both constant	Marginal	est	0.446	0.752	0.41	0.719	0.588	0.834	0.361	0.722	0.641	0.852
		se	0.016	0.013	0.019	0.02	0.041	0.012	0.031	0.018	0.019	0.01
	Conditional	est	0.426	0.713	0.397	0.688	0.578	0.816	0.361	0.66	0.626	0.844
		se	0.014	0.017	0.02	0.027	0.036	0.01	0.023	0.017	0.125	0.013
(II) θ constant	Marginal	est	-	0.775	-	0.711	-	0.834	-	0.748	-	0.859
		se	-	0.009	-	0.012	-	0.008	-	0.013	-	0.004
	Conditional	est	-	0.719	-	0.662	-	0.816	-	0.659	-	0.837
		se	-	0.012	-	0.018	-	0.01	-	0.013	-	0.006
(III) θ_{12} constant	Marginal	est	0.437	-	0.634	-	0.423	-	0.451	-	0.472	-
		se	0.023	-	0.018	-	0.019	-	0.009	-	0.025	-
	Conditional	est	0.429	-	0.618	-	0.379	-	0.432	-	0.419	-
		se	0.007	-	0.008	-	0.008	-	0.012	-	0.023	-

Approach for θ_{12}, θ	Approach for S_j^*, S_D		Early		Late		Era 1		Era 2		Era 3	
			τ_{12}	τ	τ_{12}	τ	τ_{12}	τ	τ_{12}	τ	τ_{12}	τ
(I) both constant	Marginal	est	0.424	0.741	0.624	0.802	0.362	0.597	0.401	0.749	0.633	0.887
		se	0.015	0.013	0.017	0.016	0.029	0.045	0.026	0.011	0.026	0.005
	Conditional	est	0.405	0.722	0.551	0.76	0.354	0.657	0.385	0.765	0.683	0.875
		se	0.015	0.018	0.023	0.014	0.021	0.032	0.025	0.008	0.017	0.004
(II) θ constant	Marginal	est	-	0.753	-	0.818	-	0.638	-	0.764	-	0.881
		se	-	0.01	-	0.01	-	0.017	-	0.008	-	0.037
	Conditional	est	-	0.697	-	0.766	-	0.634	-	0.772	-	0.879
		se	-	0.012	-	0.013	-	0.02	-	0.007	-	0.011
(III) θ_{12} constant	Marginal	est	0.435	-	0.635	-	0.374	-	0.414	-	0.646	-
		se	0.011	-	0.012	-	0.023	-	0.027	-	0.014	-
	Conditional	est	0.409	-	0.552	-	0.412	-	0.409	-	0.676	-
		se	0.031	-	0.016	-	0.026	-	0.028	-	0.031	-

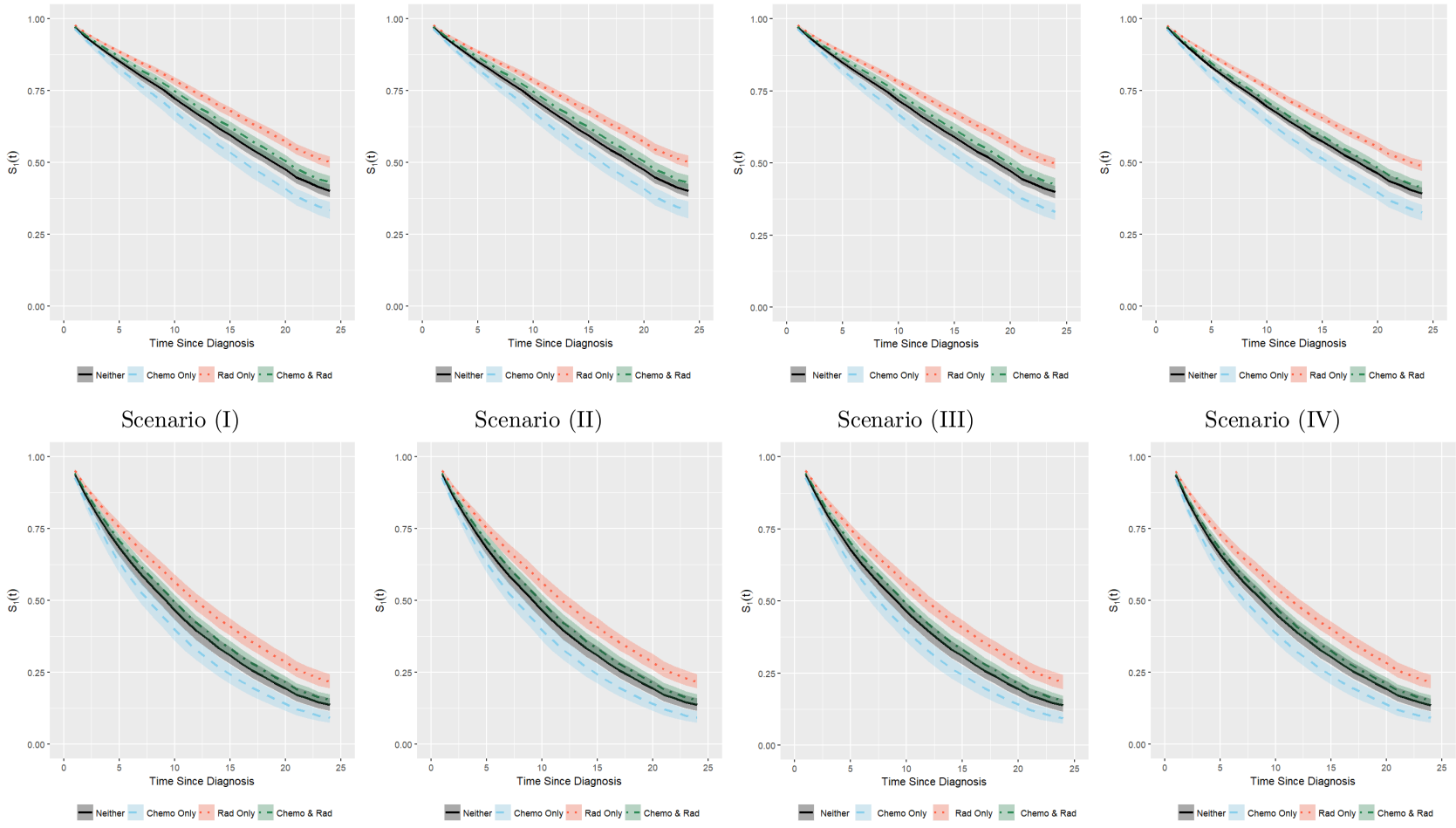
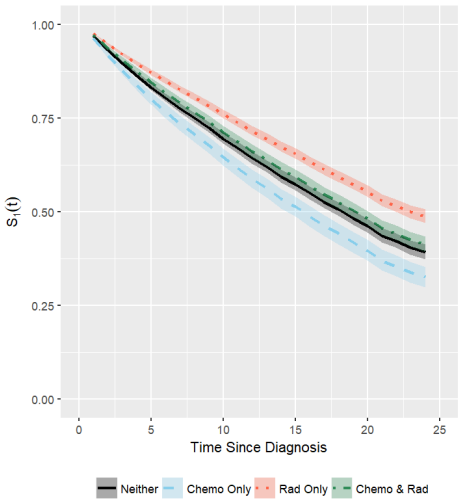
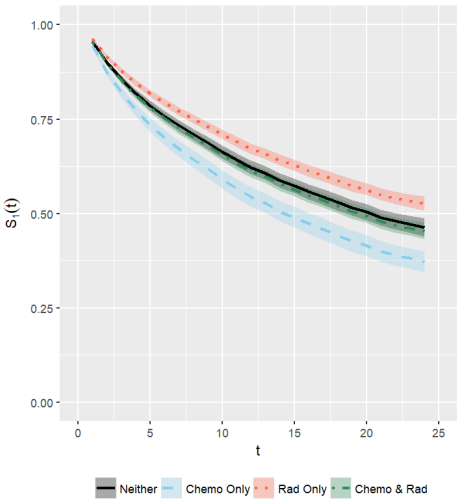


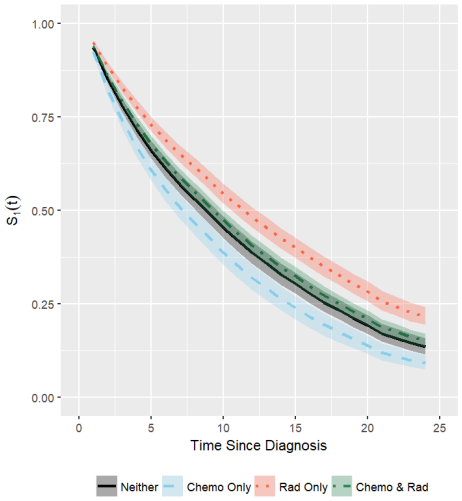
Figure 5.16: Marginal Survivor Function Estimates of $S_1(\cdot)$ under Four Scenarios with BC-BRCAS Data $\mathcal{P}_{\text{final}}$.First Column to Last Column: Scenarios (I) to (IV). Top Row and Bottom Row: Early and Late Stage.



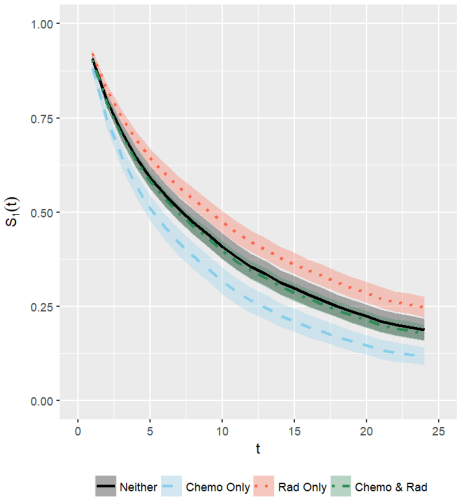
(a)
$S_1(t|X = 49, \text{early stage, each treatment})$, proposed approach



(b)
$S_1(t|X = 49, \text{early stage, each treatment})$, naïve approach

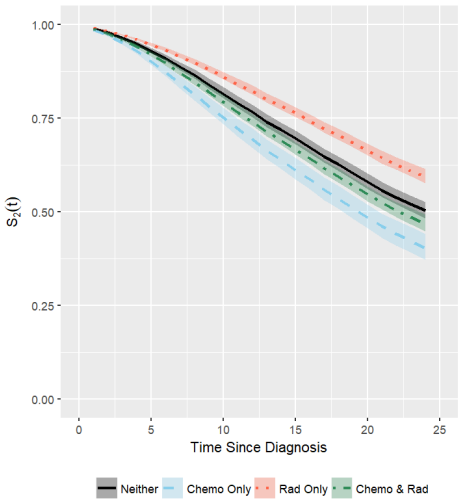


(c) $S_1(t|X = 49, \text{late stage, each treatment})$, proposed approach

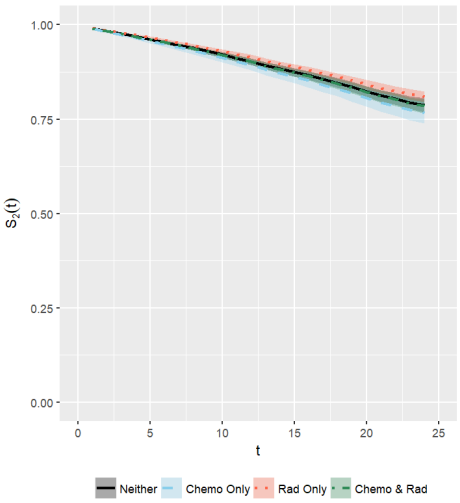


(d) $S_1(t|X = 49, \text{late stage, each treatment})$, naïve approach

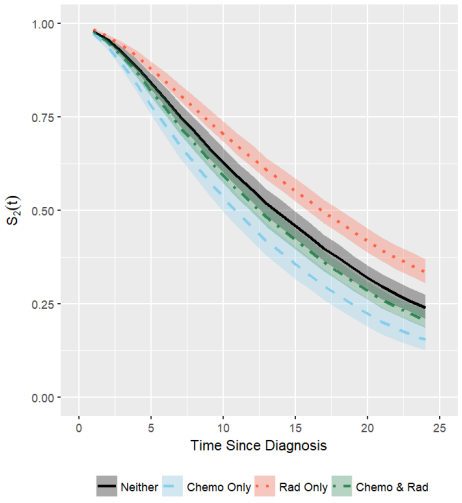
Figure 5.17: Marginal survivor function estimates for T_1 (time to RSC) for early and late stage at diagnosis, using proposed approach and naïve approach. Each plot shows curves for four treatment groups with diagnosis age at 49 and born in era II.



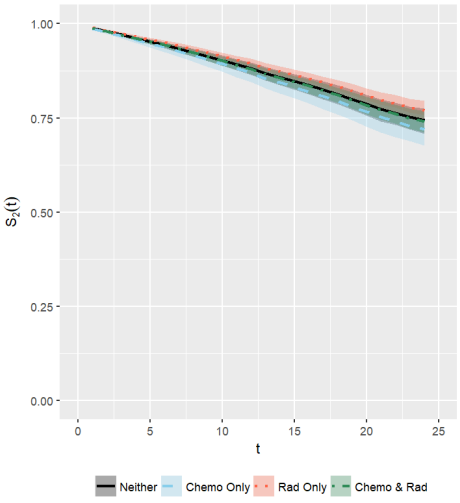
(a)
$S_2(t|X = 49, \text{early stage}, \text{each treatment})$, proposed approach



(b)
$S_2(t|X = 49, \text{early stage}, \text{each treatment})$, naïve approach



(c) $S_2(t|X = 49, \text{late stage}, \text{each treatment})$, proposed approach



(d) $S_2(t|X = 49, \text{late stage}, \text{each treatment})$, naïve approach

Figure 5.18: Marginal survivor function estimates for T_2 (time to CVD) for early and late stage at diagnosis, using proposed approach and naïve approach. Each plot shows curves for four treatment groups with diagnosis age at 49 and born in era II.

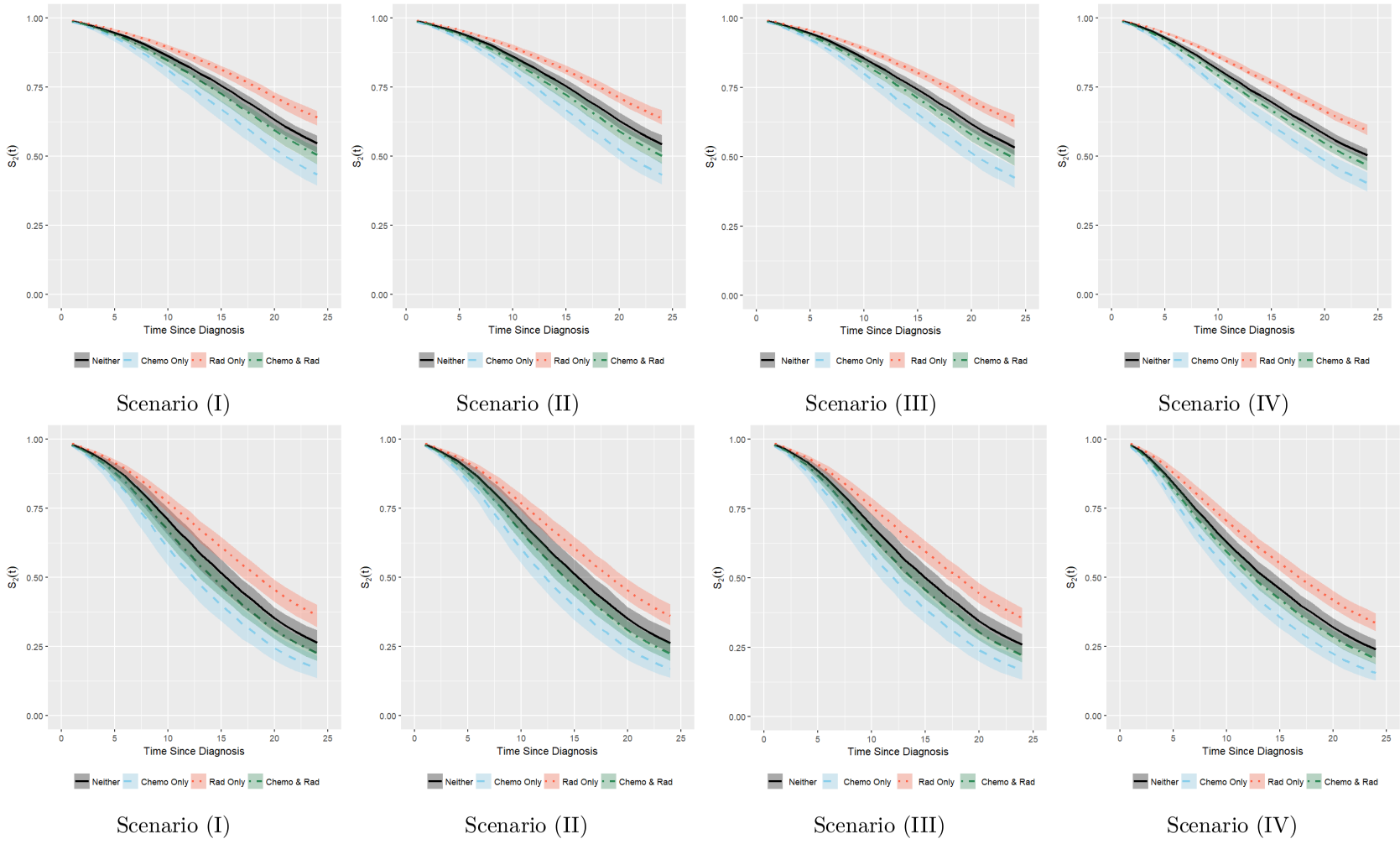


Figure 5.19: Marginal Survivor Function Estimates of $S_2(\cdot)$ under Four Scenarios with BC-BRCAS Data $\mathcal{P}_{\text{final}}$. First Column to Last Column: Scenarios (I) to (IV). Top Row and Bottom Row: Early and Late Stage.

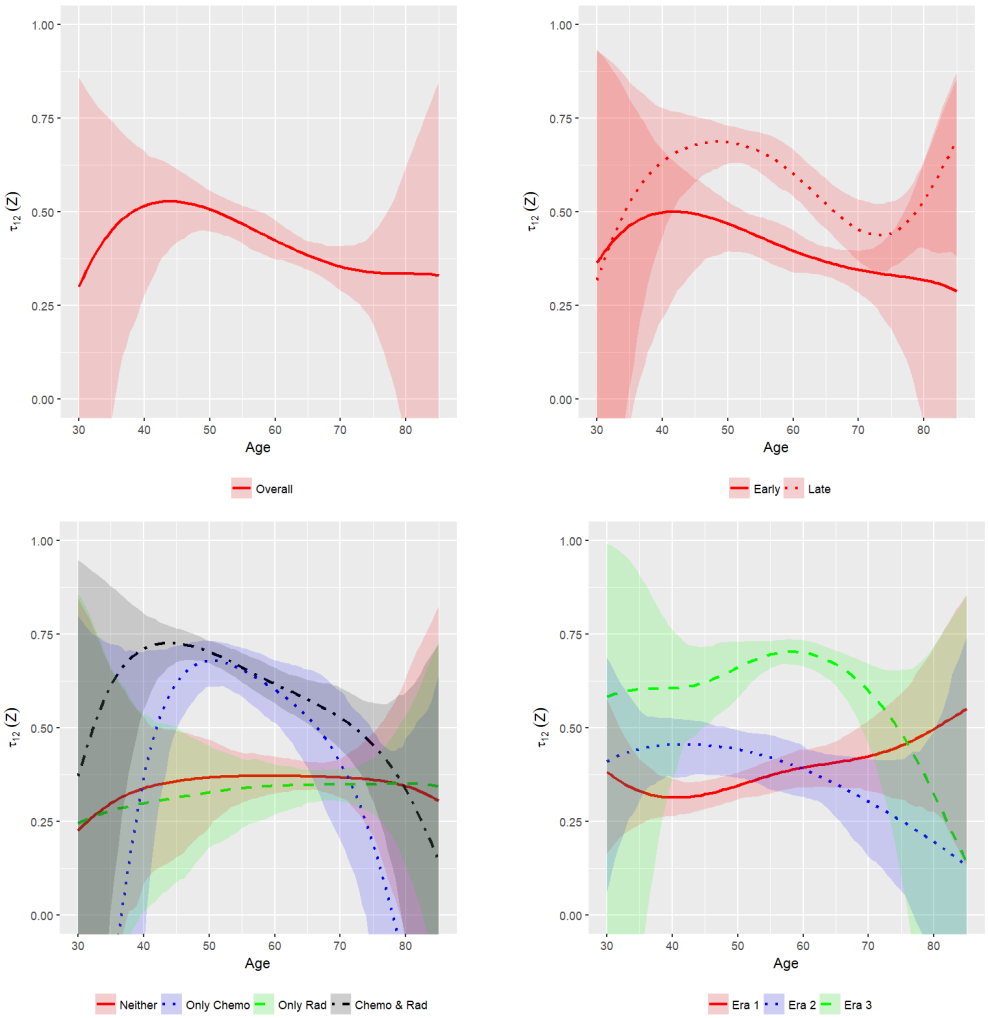


Figure 5.20: Estimates of $\tau_{12}(Z)$ under Scenario (IV) with BC-BRCAS Data $\mathcal{P}_{\text{final}}$ by Stage at Diagnosis.

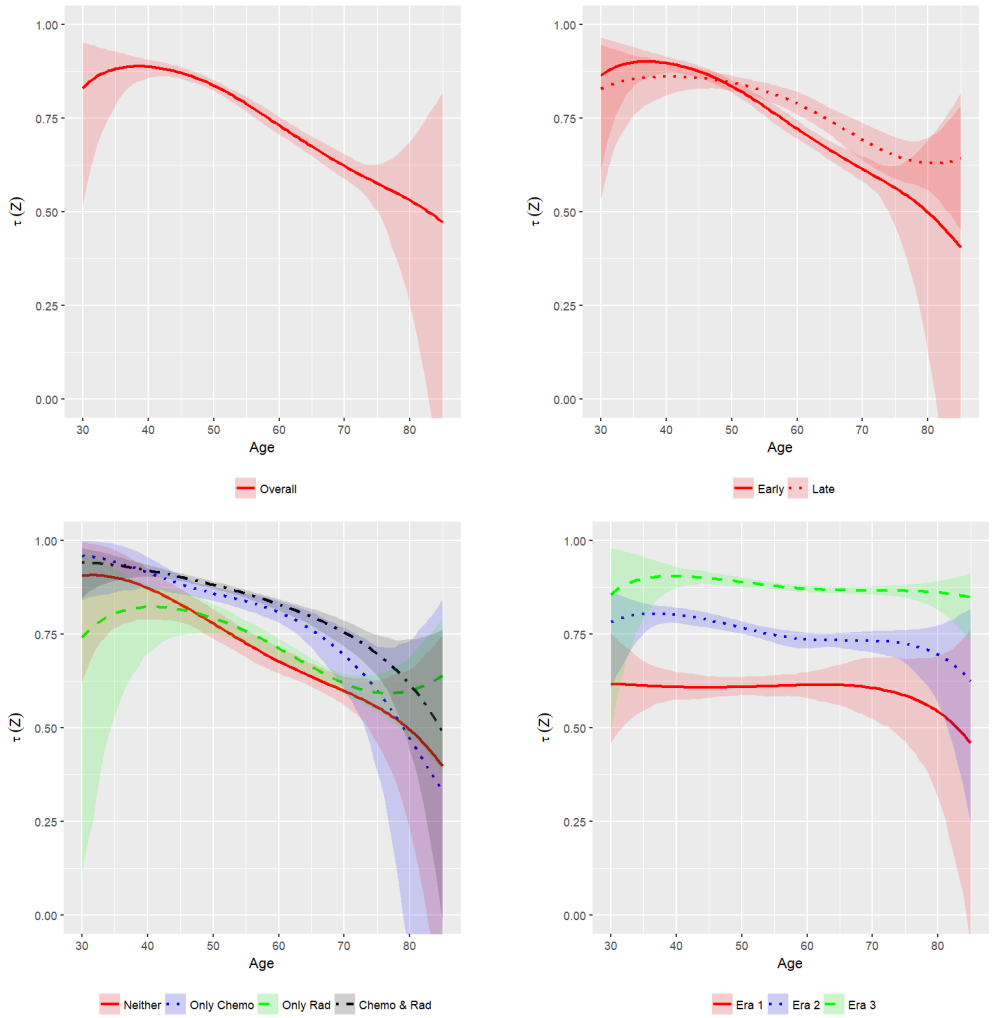


Figure 5.21: Estimates of $\tau(Z)$ for Scenario (IV) with BC-BRCAS Data $\mathcal{P}_{\text{final}}$ by Stage at Dignosis.

5.6 Discussion

As an extension of Chapter 3 and Chapter 4, this chapter applied the proposed approaches to regression setting to examine covariate effects. We explored real data analyses (IV) which addressed the research questions that motivated this thesis work. One might attempt to estimate the location of the knots for the splines. However, it will substantially increase the complexity of the analysis, and it is not developed here for cubic splines. As an alternative, an automatic knot selection procedure can be developed in attempts to compromise between flexibility of the model and complexity of the analysis (Rosenberg 1995).

Chapter 6

Comparison in CVD age Between BC Breast Cancer Cohort and Age–Matched Controls

In this chapter, we apply the modeling and inference procedures proposed in previous chapters to deal with informative censoring caused by one single terminating event, in the setting of regression analyses. Conventional approaches such as the Cox proportional hazards model in survival analyses require the assumption of noninformative censoring. Thus, this chapter proposes an approach to deal with informative censoring caused by a terminating event. We model the event time jointly with the terminating event by an Archimedean copula function. This allows one to account for informative censoring, and it yields a consistent estimator of the marginal survivor function in the semicompeting risks data setting. We propose an easy-to-implement inference procedure using a pseudolikelihood approach. Simulation studies were conducted to verify the consistency and efficiency of the proposed approach, as well as robustness against model misspecification. We applied the proposed approach to a case control study in an attempt to evaluate the difference in age at first cardiovascular disease between breast cancer survivors and the general population in presence of a terminating event. This is more practical and useful in real-life examples when informative censoring is present.

6.1 Notation and Modeling

6.1.1 Notation

We aim to estimate the marginal survivor function $S(\cdot|Z)$ with the study's right-censored event time when T is potentially correlated with D. Adopting the conventional notation, let Δ_D be the indicator $I\{D \leq C_A\}$, and $U = T \wedge C$ with $\Delta = I\{T \leq C\}$. Suppose that

the study data are n independent realizations of $\{(U,\Delta),(C,\Delta_D);Z\}$, denoted by

$$\text{Observed-Data} = \bigcup_{i=1}^{n}\Big\{\Big[\{(u_i,\delta_i)\}\bigcup\{(c_i,\delta_{Di})\}\Big];z_i : i=1,\ldots,n\Big\}. \tag{6.1}$$

This is one set of semicompeting-risks data on T associated with $C = D \wedge C_A$, together with the observed covariate z_i,

$$\text{Observed-Data} = \{(u_i,\delta_i,c_i,\delta_{Di};z_i) : i=1,\ldots,n\}. \tag{6.2}$$

We perform inference on the distributions of the event times T over the intervals $[0,v]$ with v chosen to be slightly smaller than $\max_i\{u_i\}$.

6.1.2 Model Specification

We assume that the administrative censoring time C_A is independent of the event time T and the time to the terminating event D, and assume the joint survivor function of T with D conditional on Z is equal to

$$Pr(T\geq t, D\geq d|Z) = \mathcal{A}_{[2]}(S_T(t|Z),S_D(d|Z);\theta(Z)). \tag{6.3}$$

The association parameter function $\theta(Z)$ is an unknown function of Z which characterizes the correlation between $S_T(t|Z)$ and $S_D(d|Z)$. When Z is categorical one can estimate θ nonparametrically for each level of category. When Z is continuous, we model $\theta(Z)$ through a linear combination of B-splines basis function (see for example Rosenberg 1995), indexed by a set of dimensional parameters $\boldsymbol{\alpha}$: $\theta(Z) = \sum_{i=1}^{K}\boldsymbol{\alpha}'\mathcal{B}$ where $\mathcal{B}$ is the vector of known B-spline basis function of degree 3. Now the likelihood function may be represented as:

$$Pr(T\geq t, D\geq d|Z) = \mathcal{A}_{[2]}(S_T(t|Z),S_D(d|Z);\boldsymbol{\alpha}). \tag{6.4}$$

Let $\dot{h}(r)$ be $dh(r)/dr$ for a function $h(r)$ and $h^{(a_1,a_2)}(r_1,r_2;\phi)$ be $\partial h^{(a_1+a_2)}(r_1,r_2;\phi)/\partial r_1^{a_1}\partial r_2^{a_2}$ for a function $h(r_1,r_2;\phi)$ with well-defined partial derivatives. The likelihood function with the available data under the copula model (6.4) is

$$\begin{aligned} & L(S_T(\cdot|Z),S_D(\cdot|Z),\boldsymbol{\alpha}|\text{Observed-Data}) \\ = & \prod_{i=1}^{n}\left\{(-1)^{\delta_i+\delta_{Di}}\frac{\partial^{\delta_i}\mathcal{A}_{[2]}^{(0,\delta_{Di})}(S_T(u_i|Z),S_D(c_i|Z);\boldsymbol{\alpha})}{\partial u^{\delta_i}}\dot{S}_D(c_i|Z)^{\delta_{Di}}\right\}. \end{aligned} \tag{6.5}$$

The current observations on D are right-censored with the noninformative censoring time C_A. There is a readily available consistent estimator for $S_D(\cdot)$, e.g., the Kaplan–Meier estimator, denoted as $\tilde{S}_D(\cdot)$. In regression setting, there is also readily available models and estimating procedures to obtain consistent estimator of $S_D(\cdot|Z)$, denoted as $\tilde{S}_D(\cdot|Z)$. For

example, in the analysis of real data in section (6.3), the Cox model was applied:

$$S_{D|Z}(t) = \exp\{H_{0D}(t)e^{\beta_D Z}\} \tag{6.6}$$

Following the idea of the pseudolikelihood estimation procedure under a copula model (e.g., Lawless & Yilmaz 2011), we may consider a pseudo-MLE of $\boldsymbol{\alpha}$ by maximizing $L(S_T(\cdot|Z), \tilde{S}_D(\cdot|Z), \boldsymbol{\alpha}|\text{Observed-Data})$, which is proportional to

$$\prod_{i=1}^{n} \Big\{(-1)^{\delta_i+\delta_{Di}} \frac{\partial^{\delta_i} \mathcal{A}_{[2]}^{(0,\delta_{Di})}\big(S_T(u_i|Z), \tilde{S}_D(c_i|Z); \boldsymbol{\alpha}\big)}{\partial u^{\delta_i}}\Big\}, \tag{6.7}$$

with respect to $\boldsymbol{\alpha}$ only. The resulting estimator, with the trade-off of some efficiency loss, can be much easier to implement than its MLE counterpart.

In principle, one may maximize (6.5) under model (6.3) with respect to $\boldsymbol{\alpha}, S_T(\cdot|Z)$, and $S_D(\cdot|Z)$ to obtain their MLE. However, similar to the remarks in previous chapters, this requires quite intensive-computing. Furthermore, there is no readily available consistent estimator for $S_T(\cdot|Z)$ with the current semicompeting-risks data on T. These considerations motivate the two procedures in Section 6.2 for estimating the survivor function $S_T(\cdot|Z)$ under model (6.3).

6.1.3 More on Modeling

Estimating the marginal survivor function of the event times T with the semicompeting-risks data is of interest in many situations, and it is the goal of this chapter. When the copula function $\mathcal{A}_{[2]}(\cdot;\theta(Z))$ in (6.3) is an Archimedean copula with its generator $\psi(\cdot;\theta(Z))$, it yields

$$S_T(t|Z) = g(S_T^*(t|Z), S_D(t|Z); \theta(Z)) = \psi^{-1}\{\psi(S_T^*(t|Z);\theta(Z)) - \psi(S_D(t|Z);\theta(Z));\theta(Z)\}, \tag{6.8}$$

where $S_T^*(t) = P(T_j^* \geq t|Z)$ is the survivor function of $T_j^* = T_j \wedge D$ conditional on Z. T^* is subject to noninformative censoring time C_A only, and therefore can be estimated through the Kaplan–Meier estimator for unconditional distribution, or it can be estimated through the Cox model:

$$S_{j|Z}^*(t) = \exp\{H_{0j}^*(t)e^{\beta_j^* Z}\} \tag{6.9}$$

6.2 Pseudolikelihood Based Estimation Procedure

Using the idea underlying two-stage estimation procedures with a copula model (e.g., Oakes 1994, Genest et al. 1995), we estimate $S_T(\cdot|Z)$ under the model (6.3). The estimation procedure yields a consistent estimator for the marginal survivor function. We also present the asymptotic properties of the estimators.

6.2.1 Estimating the Association Parameter with the Observed-Data

Under model (6.3), as it is given in (6.8), the marginal survivor function $S_T(t|Z) = g(S_T^*(t|Z), S_D(t|Z); \theta(Z))$, a known function of the marginal survivor function of $T_j^* = T_j \wedge D$ and the marginal survivor function of D upon $\theta(Z)$ for $j = 1, 2$. With known $S_T^*(t|Z)$ and $S_D(t|Z)$, $S_T(t|Z)$ is known only upon the parameter $\theta(Z)$. When Z is categorical, $\theta(\cdot)$ can be estimated nonparametrically, and when Z is continuous we specify it through $\theta(Z) = \sum_{i=1}^{K} \boldsymbol{\alpha}' \mathcal{B}$. In other words, the estimation of an unknown function $\theta(\cdot)$ can be reduced to estimating finite dimensional parameters, denoted $\boldsymbol{\alpha}$.

In addition, note that the likelihood function in (6.5) becomes

$$\prod_{i=1}^{n} \left\{ (-1)^{\delta_i + \delta_{Di}} \frac{\partial^{\delta_i} \mathcal{A}_{[2]}^{(0,\delta_{Di})}(S_T(u_i|Z), S_D(c_i|Z); \boldsymbol{\alpha})}{\partial S_T(u|Z)^{\delta_i}} \dot{S}_T(u_i|Z)^{\delta_i} \dot{S}_D(c_i|Z)^{\delta_{Di}} \right\},$$

which is proportional to

$$\begin{aligned} & L(\boldsymbol{\alpha}; S_T(\cdot|Z), S_D(\cdot|Z) | \text{Observed-Data}) \\ = & \prod_{i=1}^{n} \left\{ (-1)^{\delta_i + \delta_{Di}} \frac{\partial^{\delta_i} \mathcal{A}_{[2]}^{(0,\delta_{Di})}(S_T(u_i|Z), S_D(c_i|Z); \boldsymbol{\alpha})}{\partial S_T(u|Z)^{\delta_i}} \right\} \end{aligned} \tag{6.10}$$

when $S_T(\cdot|Z)$ and $S_D(\cdot|Z)$ are known. This leads to the following estimation procedure.

Provided with consistent estimators for $S_T(\cdot|Z)$ and $S_D(\cdot|Z)$, we maximize the resulting pseudolikelihood function of $\boldsymbol{\alpha}$ or, equivalently, its log-transformation with respect to the parameters $\boldsymbol{\alpha}$, to derive a pseudo-MLE:

$$\hat{\boldsymbol{\alpha}}_n = \text{argmax}_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha} | \tilde{S}_T(\cdot|Z; \boldsymbol{\alpha}), \tilde{S}_D(\cdot|Z); \text{Observed-Data}). \tag{6.11}$$

This pseudo-MLE procedure is computationally easy to implement. We present below an iterative algorithm to calculate $\hat{\boldsymbol{\alpha}}_n$.

ALGORITHM. Using the estimated $\tilde{S}_T^*(\cdot|Z)$ and $\tilde{S}_D(\cdot|Z)$ together with the current estimate $\boldsymbol{\alpha}^{(k-1)}$ and $S_T^{(k-1)}(\cdot|Z)$ and with $k \geq 1$,

Step 1. obtain the updated estimate for $\boldsymbol{\alpha}$ as

$$\boldsymbol{\alpha}^{(k)} = \text{argmax}_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha} | S_T^{(k-1)}(\cdot|Z), \tilde{S}_D(\cdot|Z); \text{Observed-Data});$$

Step 2. obtain the updated estimates for $S_T(\cdot|Z)$ as $S_T^{(k)}(t|Z) = \tilde{S}_T(t|Z; \boldsymbol{\alpha}^{(k)}) = g(\tilde{S}_T^*(t|Z), \tilde{S}_D(t|Z); \boldsymbol{\alpha}^{(k)})$.

Repeat steps 1 and 2 until the sequence $\{\boldsymbol{\alpha}^{(k)} : k = 0, 1, \ldots\}$ converges. The limit is $\hat{\boldsymbol{\alpha}}_n$ defined in (6.11).

6.3 Analysis of BC-BRCAS Data (V)

This subsection applies the proposed approach to the cohort and controls $\mathcal{P}_0 \cup \mathcal{Q}_0$, as described in Chapter 2, to compare the age at first CVD between cases and controls, to achieve *goal 3* in Chapter 1.

6.3.1 Study Description

The study group includes a breast cancer survivor cohort which includes with all women diagnosed with breast cancer in BC from 1989 to 2010, and a gender and age-matched control group, as defined in in Chapter 2. The time scale we consider in this chapter is age because the subjects in the control group do not have diagnosis dates. The covariates vector $Z = (Z_1, Z_2)$, where Z_1 is the indicator for case or controls, and Z_2 is the calendar year of diagnosis. Z_2 is first categorized into three eras as a discrete variable in the first set of analysis, i.e. era I: 1900-1927, era II: 1928-1945, and era III: 1946-1989, and then is treated as a continuous covariate in the second set of analysis. Subgroup analysis is conducted stratified by stage at diagnosis. Table 6.1 is a summary of the breast cancer cohort and the controls with their observed event-times.

6.3.2 Estimates of Conditional Survivor Functions

Z_2 is discrete

Figure 6.1 includes six plots, for the six subgroups with different stages at diagnosis and birth eras. Each plot shows two sets of estimated marginals for case and control respectively. Solid curves are the naïve estimates using the Cox model, and dashed curves are the estimates using the proposed approach. Figure 6.2 shows the difference $S_1(t|Z) - S_0(t|Z)$ in estimated marginal survivor function for CVD between breast cancer survivors ($S_1(t|Z)$) and controls ($S_0(t|Z)$), as well as the log difference $\log S_1(t|Z)/S_0(t|Z) = \log\left(S_1(t|Z)\right) - \log\left(S_2(t|Z)\right)$.

The conclusions are different using the proposed approach and the naïve approach where we directly apply the Cox model without considering informative censoring. After adjustment of informative censoring, the distribution of T seems different between the breast cancer survivors and the controls and the survivors are at a lower survival rate to CVD, especially those diagnosed at a late stage.

Z_2 is continuous

Then we treated the birth year as a continuous variable, defined as time since the year 1900 to birth year , and approximated $\theta(Z)$ through B-spline functions. Figure 6.3 shows the conditional survivor curve given $Z_2 = 14$, $Z_2 = 37$ and $Z_2 = 68$, respectively. (i.e. those born in 1914, 1937 and 1968 from three different birth eras), and figure 6.4 presents the difference in estimated survival between breast cancer survivors and the controls. The

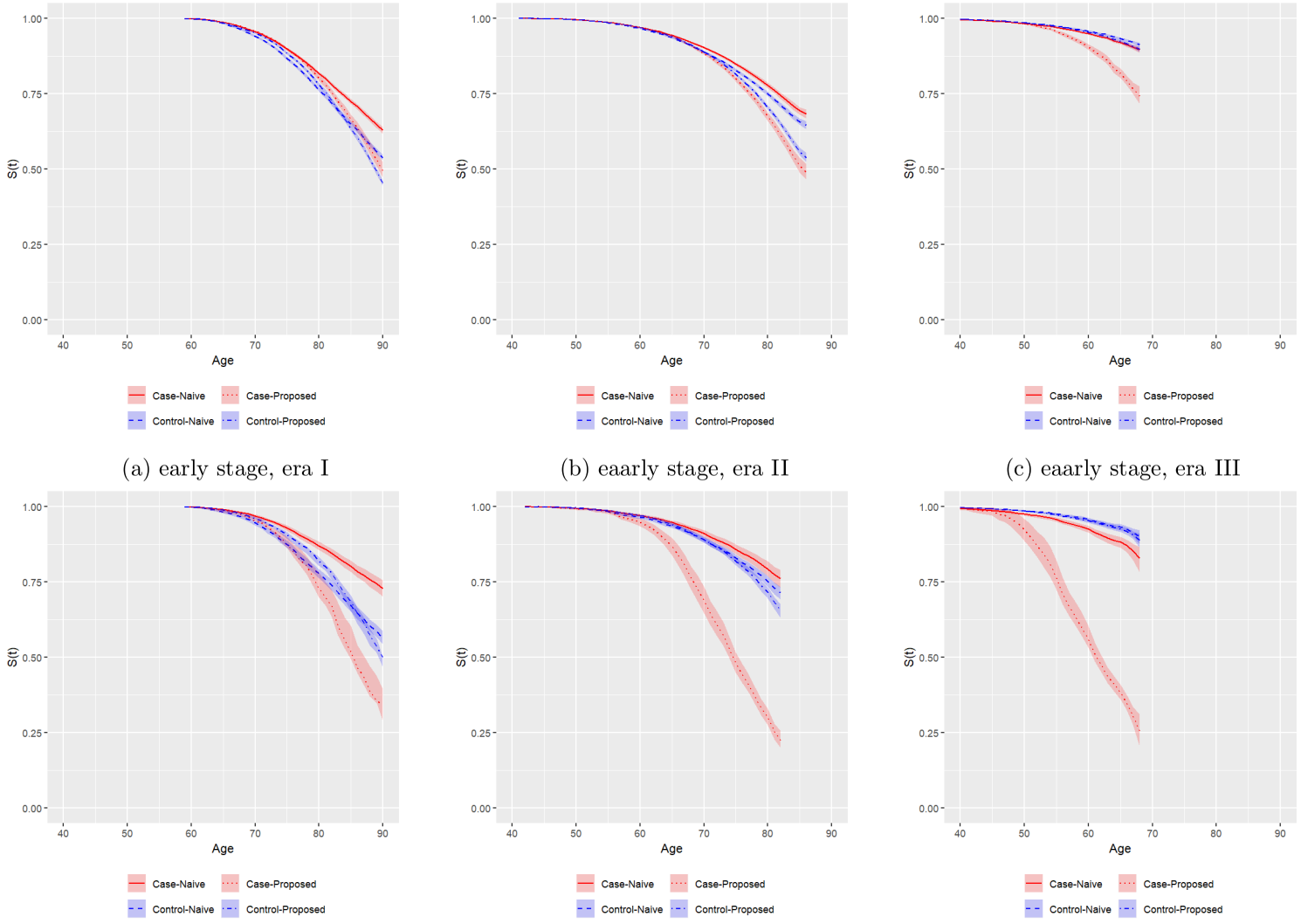
differences are significant, and the conclusions with and without adjusting for informative censoring are opposite.

Table 6.1: Summary Statistics of BC-BRCAS Data $\mathcal{P}_0$ and $\mathcal{Q}_0$

	N	$N(T^{\text{obs}})^{\dagger}$	$\overline{T^{\text{obs}}}^{\ddagger}$	$N(D^{\text{obs}})$	$\overline{D^{\text{obs}}}$
BC-BRCA Data $\mathcal{P}_0$	51,612	7,952	74.9	19,212	74.0
Controls $\mathcal{Q}_0$	103,224	19,578	72.2	25,597	78.4

†: number of subjects who has observed T

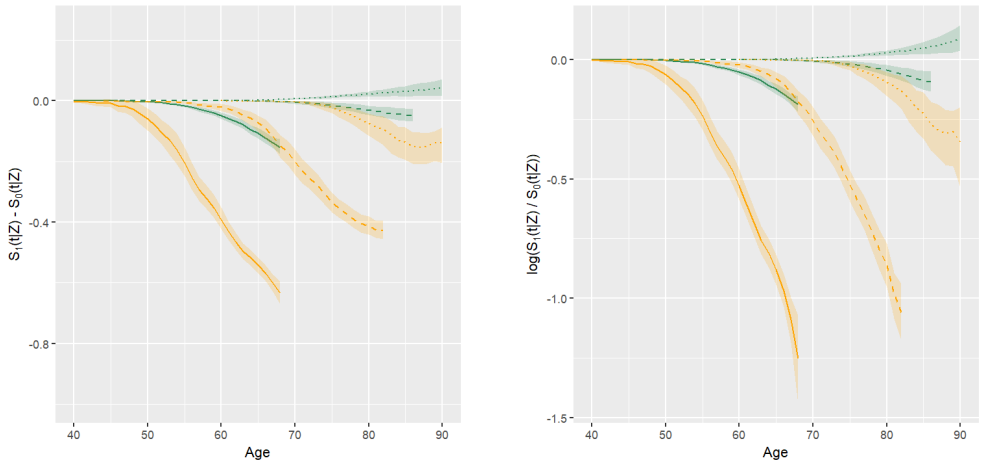
‡: mean observed T_2



(a) early stage, era I (b) eaarly stage, era II (c) eaarly stage, era III

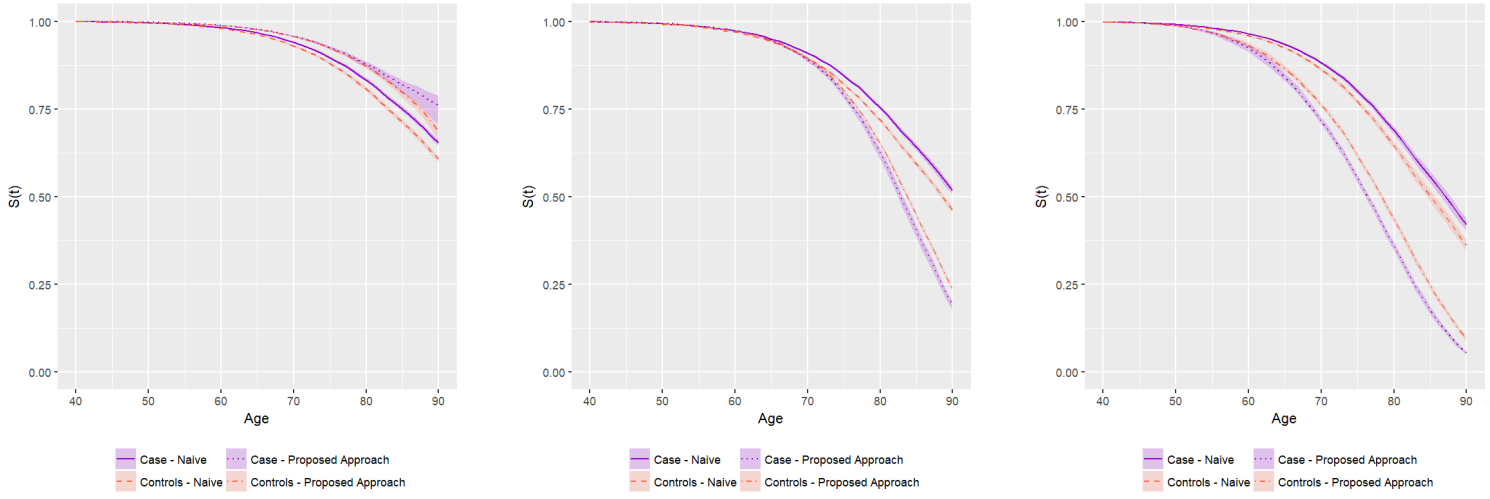
(d) late stage, era I (e) late stage, era II (f) late stage, era III

Figure 6.1: Estimates of marginal survivor functions $S(\cdot)$ of age at first CVD, for 6 subgroups using proposed approach with Clayton copula and Cox PH model. Dashed curves: estimates using proposed approach. Solid curves: estimates using naïve approach. Red curves: breast cancer cohort. Blue curves: controls.

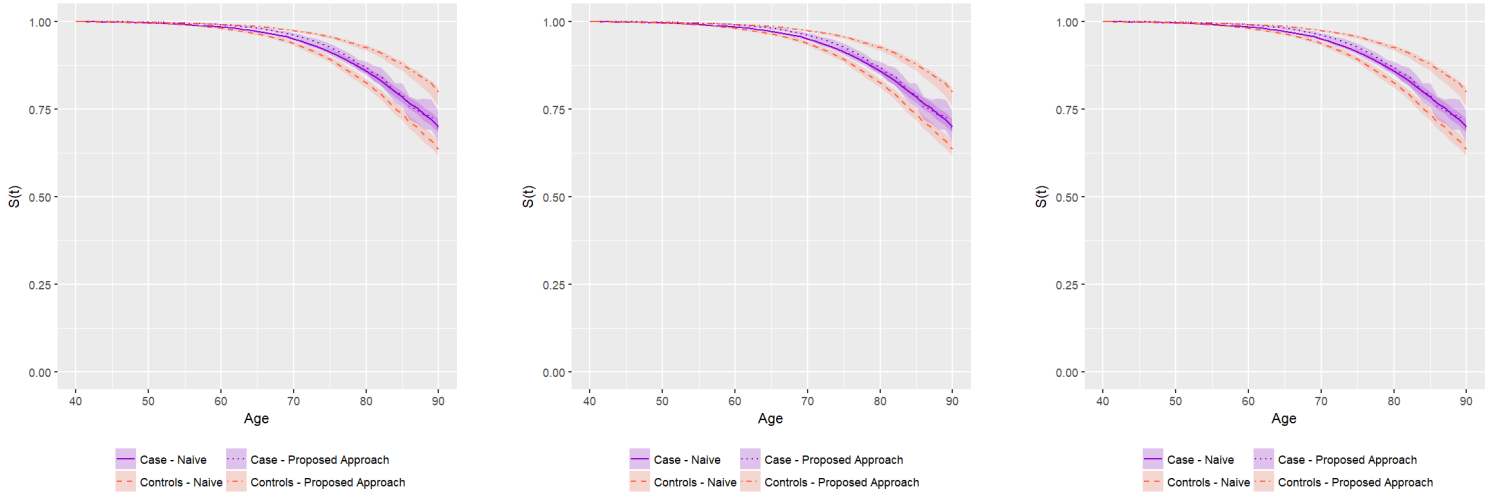


(a) Difference Between Marginal Estimates (b) Log Ratio Between Marginal Estimates

Figure 6.2: Difference in the estimates of marginal survivor functions $S(\cdot)$ of age at first CVD between breast cancer survivors and the controls, with Z_2 treated as continuous variables, for 6 subgroups using proposed approach and Cox PH model. Orange curves: late stage. Green curves: early stage. Dotted: era I. Dashed: era II. Solid: era III.

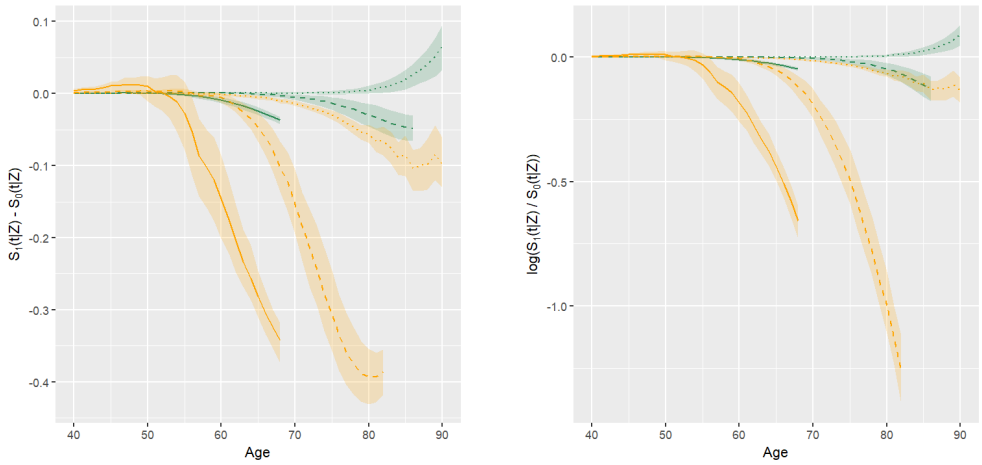


(a) early stage, diagnosis year 1914 (b) early stage, diagnosis year 1937 (c) early stage, diagnosis year 1968



(d) late stage, diagnosis year 1914 (e) late stage, diagnosis year 1937 (f) late stage, diagnosis year 1968

Figure 6.3: Estimates of marginal survivor functions $S(\cdot)$ of age at first CVD, with Z_2 treated as continuous variable, for 6 subgroups using proposed approach and Cox PH model.



(a) Difference Between Marginal Estimates (b) Log Ratio Between Marginal Estimates

Figure 6.4: Difference in the estimates of marginal survivor functions $S(\cdot)$ of age at first CVD between breast cancer survivors and the controls, with Z_2 treated as continuous variables, for 6 subgroups using proposed approach with clayton copula and Cox PH model. Orange curves: late stage. Green curves: early stage. Dotted: era I. Dashed: era II. Solid: era III.

6.4 Discussion

We applied the approach proposed in the dissertation to the semi-competing risk data setting and ran regression analysis to compare case control difference in time to CVD, to achieve *goal 1*. This is a direct application of the proposed methods to a single event time with observations subject to informative censoring. The analysis showed that ignoring informative censoring one could conclude that the general population will get CVD sooner than their age-matched breast cancer survivors, but the sets of estimates using the proposed approach showed a different conclusion.

It should be noted that the definition of CVD in this paper is based on hospitalization record, and may not reveal the real disease onset time. In addition, the data collection window is from 1986 to 2011, so left censoring exists and the event time T is time to the first CVD after 1986, and not necessarily the first CVD in the subjects' lifetime. However, the general comparison group, or the controls, are age-matched thus the comparison is still practically meaningful. Using age as time scale, one can reasonably assume that both left and right censoring are noninformative.

Chapter 7

Final Discussion

This dissertation was motivated and illustrated by the breast cancer survivorship program, but the methodology can be applied broadly for dealing with multiple event times with observations subject to informative censoring. We formulated the joint models of multiple event times, and developed inference procedures and associated applications. We justified our approach both theoretically through derivation of asymptotic properties, and numerically through simulation studies to study their finite sample performance. Moreover, we analyzed the motivating BC-BRCAS throughout the dissertation for each of the proposed methodologies.

7.1 Summary of Contributions

We started with a cross-sectional preliminary analyses which is a conventional approach in epidemiological research and presented the preliminary findings. Then we proposed in Chapter 3 a modeling approach by the Archimedean copula family and the developed associated pseudolikelihood-based procedure for the analysis of multiple event times in the presence of informative censoring due to a terminating event. The approach allows us to account for the informative censoring and to estimate validly the joint distribution of the multiple event times. It has the inference convenience associated with a copula model. One somewhat strong assumption is that the proposed modeling requires the same association between the event times, and between them jointly and the time to the terminating event. But it is still informative in that the association parameter may be viewed as an average of the associations with varying magnitudes between different pairs of event times. As shown in the real data analysis (II), there were strongly positive associations between the two event times, and each of them with death time across different subgroups. This supports the hypothesis that breast cancer patients are more likely to suffer CVD. On the other hand, the three individual association parameters do not necessarily appear the same. This

indicates that an alternative modeling would be desirable to allow different event time pairs to have different association parameters, or different dependence structures.

To mitigate the assumption of the same dependence structure, we formulated in Chapter 4 the correlation of the bivariate event time with the censoring time by embedding the bivariate distribution in a bivariate copula model. This allows us the convenience of inference under the conventional copula model. At the same time, the proposed model is more flexible, and thus potentially more appropriate in many practical situations, than modeling the event times and the associated censoring time jointly by a single multivariate copula. In addition, the joint survivor function can also be modeled through other bivariate functions such as (4.7), which leads to additional flexibility. Our approach can be extended to multiple (≥ 3) event times. We verified the consistency, efficiency and robustness through intensive simulation studies. The real-data analysis showed that the association between the bivariate event times is quite different from their dependence on the informative censoring time. This confirms the usefulness of our flexible modeling approach. Besides, the subgroup analyses revealed different dependence strengths amongst subgroups, as was also observed in Chapter 3. This leaded us to extend the methodology to regression setting which is of more interest in a practical setting.

Chapters 5 and 6 extended the proposed approach in Chapter 4 in regression setting. Chapter 5 focussed on the survivor cohort, to assess the effect of clinical factors such as age at diagnosis, stage at diagnosis, and treatment. Chapter 6 adapted the proposed methods with one single event time of interest, and real-data analysis was focussed on the case control comparison in an attempt to verify the research *hypothesis 1*: whether or not breast cancer patients suffer CVD earlier than the general population.

7.2 Future Investigations

In this dissertation, we proposed models to deal with multiple event times with informatively censored observations and developed a pseudo-MLE procedure, and comprehensively analyzed the breast cancer data using conventional and proposed approaches. However, there are other interesting possibilities for further investigation.

For example, to avoid a strong model assumption, one may consider the empirical copula instead of model (3.3) in Chapter 3. However, it is not straightforward to use the available approaches under that model with the current data. Besides one could also consider the density estimation of the marginal survivor function through smoothing techniques, and thus the pseudo-MLE could be obtained directly from (4.4).

The construction given in section 5.3.2, for example, for an approximate CB for the marginal survivor function $S_j(\cdot)$ can yield a coverage lower than the target level since it ignores the variation of the Kaplan–Meier estimates involved. An alternative approach to improve the construction is another interesting future project. One may consider adopting

the re-sampling procedure for constructing the CB of a survivor function; see, for example, Hu & Lagakos (1999) and Zhao et al. (2009).

We note here that the definition of CVD in this dissertation is based on hospitalization record, and may not reveal the real disease onset time. In addition, due to the data extraction window, the observations on T_2 are left censored; thus, the observed T_2 might not be the real time at 'first' CVD disease. In addition, only those diagnosed in or after the year 1989 were included; therefore, patients who entered the study were those who were still alive by 1989. Left truncation is potentially informative and is worthwhile to take into account in future research.

This thesis mainly considers nested copula modeling. However, another interesting area of research is to model the dependence through vine copula (Joe 1997), by linking the conditional distribution of $T_1|D$ and $T_2|D$, although the concept of $T|D$ may not be meaningful in a practical sense. Furthermore, for simulation purposes, it would be useful to generate samples from the nested Archimedean copula (Joe 1997) with a larger parameter in the outer copula than in the inner copula. This could be challenging because, given the mathematical properties of the copula-generating function, it is less convenient to sample from in this situation (see, e.g., Jaworski et al. 2010, Hofert 2012).

As one future investigation, one could consider the multistate model, (Xu et al. 2010, Farewell & Tom 2014, see, e.g.), given the following considerations addressed. First, the goal is to study the association between T_1 and T_2, which do not necessarily occur in a specific order. In addition, two 'states' (e.g. RSC and CVD) might happen at the same time, which cannot be directly addressed by the multi-state model. Furthermore, we note here that although the 'disease-free survival' (Andersen & Keiding 2012), formulated in our setting as $T^* = T \wedge D$, is not our main goal, our procedure uses the estimator of the distribution of T^* by applying readily available consistent estimator.

Lastly, the association between event times is different from causation. The occurrence of RSC does not necessarily cause a CVD. Also, as the results show in the case control comparison, the survivors have sooner CVD after a certain age. Thus, it would be interesting to explore causal inference to understand the reason behind this.

Bibliography

Andersen, P., Borgan, O., Gill, R. & Keiding, N. (1993), *Statistical models based on counting processes*, Springer series in statistics, Springer-Verlag New York.

Andersen, P. K. & Keiding, N. (2012), 'Interpretability and importance of functionals in competing risks and multistate models', *Statistics in medicine* **31**(11-12).

Bandeen-Roche, K. & Liang, K.-Y. (2002), 'Modelling multivariate failure time associations in the presence of a competing risk', *Biometrika* **89**(2), 299–314.

Bandeen-Roche, K. & Ning, J. (2008), 'Nonparametric estimation of bivariate failure time associations in the presence of a competing risk', *Biometrika* **95**(1), 221–232.

Bardia, A., Arieas, E., Zhang, Z., DeFilippis, A., Tarpinian, K., Jeter, S., Nguyen, A., Henry, N., Flockhart, D., Hayes, D., Hayden, J., Storniolo, A., Armstrong, D., Davidson, N., Fetting, J., Ouyang, P., Wolff, A., Blumenthal, R., Ashen, M. & Stearns, V. (2012), 'Comparison of breast cancer recurrence risk and cardiovascular disease incidence risk among postmenopausal women with breast cancer', *Breast Cancer Research and Treatment* **131**(3), 907–914.

Cheng, Y. & Fine, J. P. (2012), 'Cumulative incidence association models for bivariate competing risks data', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **74**(2), 183–202.

Cheng, Y., Fine, J. P. & Kosorok, M. R. (2007), 'Nonparametric association analysis of bivariate competing-risks data', *Journal of the American Statistical Association* **102**(480), 1407–1415.

Clarke, M., Collins, R., Darby, S., Davies, C., Elphinstone, P., Evans, V., Godwin, J., Gray, R., Hicks, C., James, S., MacKinnon, E., McGale, P., McHugh, T., Peto, R., Taylor, C. & Wang, Y. (2005), 'Effects of radiotherapy and of differences in the extent of surgery for early breast cancer on local recurrence and 15-year survival: An overview of the randomised trials', *The Lancet* **366**(9503), 2087–2106.

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

Cuzick, J., Stewart, H., Rutqvist, L., Houghton, J., Edwards, R., Redmond, C., Peto, R., Baum, M., Fisher, B. & Host, H. (1994), 'Cause-specific mortality in long-term survivors of breast cancer who participated in trials of radiotherapy', *Journal of Clinical Oncology* **12**(3), 447–453. PMID: 8120544.

Davis, M., Li, D., Wai, E., Tyldesley, S., Simmons, C., Baliski, C. & McBride, M. (2014), 'Hospital-related cardiac morbidity among survivors of breast cancer: Long-term risks and predictors', *Canadian Journal of Cardiology* **30**(10), S122–S123. Canadian Cardiovascular Congress 2014.

de Boor, C. (1978), *A Practical Guide to Splines*, Springer-Verlag.

Diao, L. & Cook, R. J. (2014), 'Composite likelihood for joint analysis of multiple multistate processes via copulas', *Biostatistics* **15**(4), 690–705.

Fan, J., Prentice, R. & Hsu, L. (2000), 'A class of weighted dependence measures for bivariate failure time data', *Journal of the Royal Statistical Society* **62**(1), 181–190.

Farewell, V. & Tom, B. (2014), 'The versatility of multi-state models for the analysis of longitudinal data with unobservable features', *Lifetime Data Analysis* **20**(1), 51–75.

Fine, J. & Jiang, H. (2000), 'On association in a copula with time transformations', *Biometrika* **87**(3), 559–571.

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

Genest, C., Ghoudi, K. & Rivest, L.-P. (1995), 'A semiparametric estimation procedure of dependence parameters in multivariate families of distributions', *Biometrika* **82**(3), 543–552.

Goethals, K., Janssen, P. & Duchateau, L. (2008), 'Frailty models and copulas: similarities and differences', *Journal of Applied Statistics* **35**(9), 1071–1079.

Hamilton, S. N., Tyldesley, S., Li, D., Olson, R. & McBride, M. (2015), 'Second malignancies after adjuvant radiation therapy for early stage breast cancer: Is there increased risk with addition of regional radiation to local radiation?', *International Journal of Radiation Oncology, Biology, Physics* **91**(5), 977–985.

Hofert, M. (2012), 'A stochastic representation and sampling algorithm for nested Archimedean copulas', *Journal of Statistical Computation and Simulation* **82**(9), 1239–1255.

Hofert, M., Kojadinovic, I., Maechler, M. & Yan, J. (2017), *Copula: Multivariate Dependence with Copulas*. R package version 0.999-18.

Hofert, M. & Mächler, M. (2011), 'Nested Archimedean copulas meet R: The nacopula package', *Journal of Statistical Software* **39**(9), 1–20.

Hougaard, P. (1984), 'Life table methods for heterogeneous populations: Distributions describing the heterogeneity', *Biometrika* **71**(1), 75–83.

Hougaard, P. (2012), *Analysis of Multivariate Survival Data*, Statistics for Biology and Health, Springer New York.

Hu, X. J. & Lagakos, S. W. (1999), 'Interim analyses using repeated confidence bands', *Biometrika* **86**, 517–529.

Huber, P. J. (1967), 'The behavior of maximum likelihood estimates under nonstandard conditions', *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **I**, 221–233.

Jaworski, P., Durante, E., Härdle, W. & Rychlik, T. (2010), *Copula Theory and Its Applications: Proceedings of the Workshop Held in Warsaw, 25–26 September 2009*, Vol. 198 of *Lecture Notes in Statistics*, Springer, Berlin, Heidelberg.

Jiang, H., Fine, J. P., Kosorok, M. R. & Chappell, R. (2005), 'Pseudo self-consistent estimation of a copula model with informative censoring', *Scandinavian Journal of Statistics* **32**(1), 1–20.

Joe, H. (1997), *Multivariate Models and Multivariate Dependence Concepts*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, Taylor & Francis.

Kojadinovic, I. & Yan, J. (2010), 'Modeling multivariate distributions with continuous margins using the copula R package', *Journal of Statistical Software* **34**(9), 1–20.

Lawless, J. F. & Yilmaz, Y. E. (2011), 'Comparison of semiparametric maximum likelihood estimation and two-stage semiparametric estimation in copula models', *Computational Statistics and Data Analysis* **55**, 2446–2455.

Li, D., Hu, X. J., McBride, M. L. & Spinelli, J. J. (2018), 'Multiple event times in the presence of informative censoring: Modeling and analysis by copulas'. Submitted for publication.

Li, Q. H. & Lagakos, S. W. (1997), 'Use of the Wei-Lin-Weissfeld method for the analysis of a recurring and a terminating event', *Statistics in Medicine* **16**, 925–940.

Li, Y., Tiwari, R. C. & Guha, S. (2007), 'Mixture cure survival models with dependent censoring', *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **69**(3), 285–306.

Liang, K.-Y., Self, S. G., Bandeen-Roche, K. J. & Zeger, S. L. (1995), 'Some recent developments for regression analysis of multivariate failure time data', *Lifetime Data Analysis* **1**(4), 403–415.

McBride, M. L., Groome, P., Turner, D., Jorgensen, M., Kendell, C., Porter, G., Jiang, L., Krzyzanowska, M., Lofters, A., Moineddin, R., Grunfeld, E. & Winget, M. (2016), 'Using canadian administrative data to evaluate primary and oncology care of breast cancer patients post-treatment: Subset of the canimpact study.', *Journal of Clinical Oncology* **34**, 5–6.

Mehta, S., L., Watson, E., K., Barac, M., A., Beckie, L., T., Bittner, Santos, V., Cruz-Flores, Santos, S., Dent, Santos, S., Kondapalli, Santos, L., Ky, Santos, B., Okwuosa, Santos, T., Piña, Santos, I. & Volgman, Santos, A. (2018), 'Cardiovascular disease and breast cancer: Where these entities intersect: A scientific statement from the american heart association', *Circulation* **137**(8), e30–e66.

Nelsen, R. (2006), *An Introduction to Copulas, Second Edition*, New York: Springer Sceience+Business Media Inc.

Ning, J. & Bandeen-Roche, K. (2014), 'Estimation of time-dependent association for bivariate failure times in the presence of a competing risk', *Biometrics* **70**(1), 10–20.

Ning, J., Chen, Y., Cai, C., Huang, X. & Wang, M.-C. (2015), 'On the dependence structure of bivariate recurrent event processes: inference and estimation', *Biometrika* **102**(2), 345–358.

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

Oakes, D. (1994), 'Multivariate survival distributions', *Journal of Nonparametric Statistics* **3**(3-4), 343–354.

Park, N.-J., Chang, Y., Bender, C., Conley, Y., Chlebowski, R. T., van Londen, G. J., Foraker, R., Wassertheil-Smoller, S., Stefanick, M. L. & Kuller, L. H. (2017), 'Cardiovascular disease and mortality after breast cancer in postmenopausal women: Results from the women's health initiative', *PLoS ONE* **12**(9).

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *http://www.R-project.org/*

Rosenberg, P. S. (1995), 'Hazard function estimation using b-splines', *Biometrics* **51**(3).

Serfling, R. (1980), *Approximation Theorems of Mathematical Statistics*, Wiley Series in Probability and Statistics - Applied Probability and Statistics Section Series, Wiley.

Shih, J. H. & Louis, T. A. (1995), 'Inferences on the association parameter in copula models for bivariate survival data', *Biometrics* **51**(4), 1384–1399.

Shorack, G. & Wellner, J. (2009), *Empirical Processes with Applications to Statistics*, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104).

Sklar, A. (1959), 'Fonctions de répartition à n dimensions et leurs marges', *Institut Statistique de l'Université de Paris* **8**, 229–231.

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.

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

Yan, J. (2007), 'Enjoy the joy of copulas: With a package copula', *Journal of Statistical Software* **21**(4), 1–21.

Zhang, S., Zhang, Y., Chaloner, K. & Stapleton, J. T. (2010), 'A copula model for bivariate hybrid censored survival data with application to the MACS study', *Lifetime Data Analysis* **16**(2), 231–249.

Zhao, L., Hu, X. J. & Lagakos, S. W. (2009), 'Statistical monitoring of clinical trials with multivariate response and/or multiple arms: A flexible approach', *Biostatistics* **10**, 310–323.

Zheng, M. & Klein, J. P. (1995), 'Estimates of marginal survival for dependent competing risks based on an assumed copula', *Biometrika* **82**(1), 127–138.

Zhong, Y. & Cook, R. J. (2016), 'Augmented composite likelihood for copula modeling in family studies under biased sampling', *Biostatistics* **17**(3), 437–452.

Canadian Institute for Health Information (2011): Discharge Abstract Database (Hospital Separations). V2. Population Data BC. Data Extract. MOH (2011). `http://www.popdata.bc.ca/data`