Sample size calculation for competing risk models and design considerations for agricultural experiments

by

Rui Liu

M.A., University of Missouri, 2017

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2025

Abstract

This dissertation develops methods for the design and analysis of experiments seen in agricultural settings. First, we develop methods for determining requisite sample sizes for obtaining a pre-specified power in competing risk models, a multistate model for survival analysis. Typically, prior to conducting an "official" elaborate and expensive study, a small pilot study is performed to better understand the target population and the differences between treatments. For first-event classic survival analysis, a consistent hazard rate is often assumed to compute the sample size for the official experiment. However, for competing risk models, the hazards for each type of event need to be specified. Additionally, the number of occurrences for each type of event may be small, especially if we further stratify the data regarding the treatment or covariates. Thus, commonly-used approaches for modeling data from the pilot study—for example, using a parametric survival model assuming an exponential distribution—is unlikely to yield useful information from the pilot study. Therefore, we propose to use a flexible parametric survival model, the generalized log-gamma survival model, to extract information from the pilot study. After that, we simulate data related to the data generation mechanism of competing risks to fit a pre-specified statistical test intended for the official study. We repeat the previous step a large number of times to compute power. We adjust the size of the simulated dataset regarding the output power until the output power reaches a pre-specified level. The minimal sample size results in the pre-specified power is the one we recommend for the official experiment.

Next, cluster-randomized experiments, in which units are grouped together in some way and treatment is assigned to groups of units, are common in agricultural settings. For example, studies assessing the efficacy of a diet on the weight gain of pigs may assign the diet to pens, rather than individual pigs, as administration of the diet at the pen level is much less costly. However, in many settings, researchers may have some ability to form

the clusters of units prior to randomizing treatment. In this study, we determine best practices for forming clusters of units when such an option is available. Under the Neyman-Rubin Causal Model (NRCM), we derive expressions for the efficiency loss due to cluster randomization (relative to complete randomization) in a simplified setting where cluster sizes are equal. We show that efficiency loss is a function of the intra-cluster coefficient and the correlation between potential outcomes. The efficiency loss is minimized—and, in fact, cluster-randomized experiments can outperform completely randomized experiments—when the "discrepancy" between clusters is small—that is, when the distribution of units within each cluster is as similar as possible across clusters. We then introduce a heuristic algorithm for assigning units to clusters to ensure small discrepancy between clusters. We give a thorough simulation studies to verify our results. We conclude by applying our heurstic algorithm to improve divide and conquer methods for performing kernel ridge regression in large-to-massive data settings.

Sample size calculation for competing risk models and design considerations for agricultural experiments

by

Rui Liu

M.A., University of Missouri, 2017

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2025

Approved by:

Major Professor
Michael Higgins

Copyright

© Rui Liu 2025.

Abstract

This dissertation develops methods for the design and analysis of experiments seen in agricultural settings. First, we develop methods for determining requisite sample sizes for obtaining a pre-specified power in competing risk models, a multistate model for survival analysis. Typically, prior to conducting an "official" elaborate and expensive study, a small pilot study is performed to better understand the target population and the differences between treatments. For first-event classic survival analysis, a consistent hazard rate is often assumed to compute the sample size for the official experiment. However, for competing risk models, the hazards for each type of event need to be specified. Additionally, the number of occurrences for each type of event may be small, especially if we further stratify the data regarding the treatment or covariates. Thus, commonly-used approaches for modeling data from the pilot study—for example, using a parametric survival model assuming an exponential distribution—is unlikely to yield useful information from the pilot study. Therefore, we propose to use a flexible parametric survival model, the generalized log-gamma survival model, to extract information from the pilot study. After that, we simulate data related to the data generation mechanism of competing risks to fit a pre-specified statistical test intended for the official study. We repeat the previous step a large number of times to compute power. We adjust the size of the simulated dataset regarding the output power until the output power reaches a pre-specified level. The minimal sample size results in the pre-specified power is the one we recommend for the official experiment.

Next, cluster-randomized experiments, in which units are grouped together in some way and treatment is assigned to groups of units, are common in agricultural settings. For example, studies assessing the efficacy of a diet on the weight gain of pigs may assign the diet to pens, rather than individual pigs, as administration of the diet at the pen level is much less costly. However, in many settings, researchers may have some ability to form

the clusters of units prior to randomizing treatment. In this study, we determine best practices for forming clusters of units when such an option is available. Under the Neyman-Rubin Causal Model (NRCM), we derive expressions for the efficiency loss due to cluster randomization (relative to complete randomization) in a simplified setting where cluster sizes are equal. We show that efficiency loss is a function of the intra-cluster coefficient and the correlation between potential outcomes. The efficiency loss is minimized—and, in fact, cluster-randomized experiments can outperform completely randomized experiments—when the "discrepancy" between clusters is small—that is, when the distribution of units within each cluster is as similar as possible across clusters. We then introduce a heuristic algorithm for assigning units to clusters to ensure small discrepancy between clusters. We give a thorough simulation studies to verify our results. We conclude by applying our heurstic algorithm to improve divide and conquer methods for performing kernel ridge regression in large-to-massive data settings.

Table of Contents

List of Figures xi

List of Tables xii

1 Introduction 1

1.1 Introduction 1

1.2 Competing Risk Model 2

1.2.1 Traditional Survival Analysis 2

1.3 Cluster Randomized Trials 6

1.4 Discrepancy Theory 8

1.4.1 Notation and Preliminaries 8

1.4.2 Banaszczyk's Theorem and Relative Algorithm 10

1.4.3 Relationship to Vector Balancing 10

1.5 Kernel Ridge Regression 10

1.5.1 Ridge Regression 10

1.5.2 Kernel Functions and Hilbert Space 12

1.5.3 Kernel Ridge Regression 14

1.6 Conclusion 15

2 Sample Size Computation for Competing Risk Models Using Flexible Parametric Models 17

2.1 Introduction 17

2.2 Notation and Preliminiaries 19

2.2.1 Classic Survival Analysis 19

2.2.2 Parametric Survival Analysis Using the Generalized Log-Gamma Distribution . . . 22
2.3 Cause-specific Model for Competing Risk . . . 27
2.3.1 The Weighted Log-rank Test . . . 32
2.3.2 Suitability of the Generalized Log-gamma Model For Competing Risk Data . . . 33
2.4 Data Generation for Competing Risk . . . 39
2.5 Sample Size Computation For Competing Risk Data . . . 40
2.6 Simulation . . . 41
2.7 Conclusion . . . 68

3 On the design of cluster-randomized trials with user-specified clusters . . . 70
3.1 Introduction . . . 70
3.2 Cluster Randomized Trials . . . 72
3.2.1 Horvitz-Thompson Estimation of the Population Mean . . . 73
3.2.2 Estimates of the Average Treatment Effect . . . 79
3.3 Threshold Clustering for Cluster Heterogeneity . . . 81
3.4 Simulation Example . . . 83
3.5 Conclusion . . . 84

4 Improved Divide and Conquer Methods Using Discrepancy Theory for Kernel Ridge Regression . . . 88
4.1 Introduction . . . 88
4.2 Notation and Preliminaries . . . 90
4.2.1 Kernel Ridge Regression . . . 90
4.2.2 Regularized Regression and Kernel Ridge Regression . . . 92
4.2.3 Divide and Conquer . . . 95
4.3 A Discrepancy Method Approach for Divide and Conquer . . . 98

4.3.1 Discrepancy Theory 99
4.3.2 The Low Discrepancy Dividing Process 101
4.4 Simulation and Result 102
4.5 Results 103
4.6 Conclusion 105

5 Conclusion 106
5.1 Future Work 108

Bibliography 110

A Results for Chapter 3 115
A.1 Horvits-Thompson Estimator for Cluster Randomized Trials 115

B Results for Chapter 4 124

List of Figures

2.1 Cause-specific survival functions for event type 1 under control 35

2.2 Cause-specific survival functions for event type 1 under treatment 36

2.3 Cause-specific survival functions for event type 2 under control 37

2.4 Cause-specific survival functions for event type 2 under treatment 38

2.5 Cause-specific hazard for simulation 1 45

2.6 Power comparison for simulation 1 46

2.7 Cause-specific hazard for simulation 2 49

2.8 Power comparison for simulation 2 50

2.9 Cause-specific hazard for simulation 3 53

2.10 Power comparison for simulation 3 54

2.11 Cause-specific hazard for simulation 4 57

2.12 Power comparison for simulation 4 58

2.13 Cause-specific hazard for simulation 5 61

2.14 Power comparison for simulation 5 62

2.15 Cause-specific hazard for simulation 6 65

2.16 Power comparison for simulation 6 66

3.1 Comparison of difference from the true parameter 85

3.2 Variance comparison 86

4.1 Low discrepancy dividing 102

4.2 Average MSE of different dividing method 104

List of Tables

2.1 Summarization of power and sample size for simulation 1 47

2.2 Summarization of power and sample size of simulation 2 51

2.3 Summarization of power and sample size of simulation 3 55

2.4 Summarization of power and sample size of simulation 4 59

2.5 Summarization of power and sample size of simulation 5 63

2.6 Summarization of power and sample size of simulation 6 67

4.1 Comparison of average MSE 104

Chapter 1

Introduction

1.1 Introduction

This dissertation focuses on new methods for designing studies and analyzing data for agricultural applications. We have dedicated one chapter to developing methods for approximating power in competing risk models, which allow for multiple types of terminal events in a survival analysis. Even though competing risk models or multistate models are well studied, there is little guidance on deciding sample sizes based on a pilot study. Therefore, we propose our method of sample size computation in Chapter 2. We have an additional chapter focusing on narrowing the precision gap between cluster randomized trials and completely randomized experiments when clusters are allowed to be constructed by the researcher.. In this study, we use our knowledge of discrepancy theory to reduce between-cluster variability, aiming to decrease the design effect and narrow the precision gap between cluster randomized trials and simple random sampling. In addition, we also present a chapter that focuses on applying discrepancy to computation efficiency. In that chapter, we concentrate on discussing the method of model averaging of kernel ridge regression and how discrepancy theory inspires us to increase the efficiency of model fitting further.

The focus of this chapter is to provide the necessary background information for the

subsequent chapters. There are three sections. The section 1.2 discusses survival analysis and competing risk models. We focus on the introduction of cluster randomized trials in section 1.3. We discuss the concepts and basics theories of discrepancy theory and ϵ-approximation in section 1.4. We discuss the concepts of kernel and kernel ridge regression in section 1.5.

1.2 Competing Risk Model

For time-to-event responses, multiple types of events need to be considered. The competing risk model, or the multistate model, is an extension of traditional survival analysis that considers multiple types of events.

1.2.1 Traditional Survival Analysis

Before we give more details on the multi-state model, some notations and concepts of the survival model need to be clarified. Suppose there are n subjects under the study with a process $(X_t^{(i)})_{t>0}$, $(X_t^{(i)}) \in \{0, 1\}$, for $i = 1, 2, ..., n$. Here, 0 represents the initial state, and 1 represents the absorbing state. The end points for subject i are failure time, T_i, and failure cause, X_{T_i}. Additionally, we consider an independent right censoring C_i at this moment.

We are particularly interested in analyzing the *survival time* of subject i—length of time until subject i transitions from the initial state to the absorbing state. Rigorously, let $X_{it} \in \{0, 1\}$ denote the state of subject i at time t, $t \geq 0$. Then, the survival time T_i for subject i is

$$T_i = \inf\{t : X_{it} \neq 0\}. \tag{1.1}$$

We assume that $T_i, i = 1, \ldots, n$ are independent and identically distributed (iid) random variables; for ease of exposition, we may drop the subscript from T_i when we can do so unambiguously.

Note, the '0' state is often considered a "survival" state and the '1' state is often considered a "failure" state. Thus, if a subject does not transition from the initial state to the

absorbing state by time t, we may refer to that subject as "surviving" until time t. Similarly, we often refer to a subject transitioning from the initial state to the absorbing state at time t as subject "failing" at time t, and and the survival time may be referred to as the "failure time"

Let $f(t)$ and $F(t)$ denote the probability density function (pdf) and the cumulative density function (cdf) of the survival time T. The *survival function* $S(t)$ is the probability that a subject survives until time t:

$$S(t) = \mathbb{P}(T > t) = 1 - F(t) \tag{1.2}$$

The hazard function $h(t)$ is the instantaneous rate of failure for a subject given that the subject survives until time t:

$$h(t) = lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T < t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)} \tag{1.3}$$

The cumulative hazard function $H(t)$ is obtained by integrating the hazard function from time 0 to time t:

$$H(t) = \mathbb{P}(T \leq t) = \int_0^t h(u)du. \tag{1.4}$$

Note that the survival function can be expressed in terms of the cumulative hazard function:

$$S(t) = exp(-H(t)). \tag{1.5}$$

Furthermore, the pdf $f(t)$ can be rewritten in terms of $h(t)$ and $H(t)$:

$$f(t) = h(t)S(t) = h(t)exp(-H(t)). \tag{1.6}$$

The most common estimate in practice is the nonparametric Nelson-Aalen (Nelson, 1969) estimator of the cumulative hazard and the Kaplan-Meier (Kaplan and Meier, 1958) esti-

mator for the survival function. Giving some additional notations, that d_i is the number of transitions $0 \to 1$ at time t_i and n_i is the number of subjects at risk right before t_i (or at time t^-).

We can estimate these interested qualities in the framework of binomial likelihood. The Nelson-Aalen estimator of the cumulative hazard can be expressed as:

$$\hat{H}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i} \tag{1.7}$$

The Kaplan-Meier estimator of survival function is defined as:

$$\hat{S}(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i} \tag{1.8}$$

These concepts can be easily generalized to the multi-state model, where we investigate the time until a subject reaches one of several terminal states. In other words, we have multiple absorbing states, and T represents the time until the transition to the first observed absorbing state. By employing a competing risk model, specifically, a cause-specific model, we add specific information to the model that distinguishes between the potential event types. To further introduce the competing risk model, we introduce a new index j, where $j = 1, ..., J$ indicates the type of events. The stochastic process X_t has J values 0, 1,..., J.

We define the cause-specific hazards, α_{0j}, as the probability of a unit transition from the initial state to the absorbing state j within a time unit. The subscripts $0j$ stands for the transition from the initial state, 0, to the absorbing state j. Mathematically, we have

$$\alpha_{0j}(t) = lim_{\Delta t \to 0} \frac{\mathbb{P}(T \in [t, t + \Delta t), X_t = j | T > t)}{\Delta t} \tag{1.9}$$

The corresponding cumulative cause-specific is defined as

$$A_{0j}(t_i) = \int_0^{t_i} \alpha_{0j}(u) du \tag{1.10}$$

The failure time, T_i, is guided by the all-cause hazard function

$$\alpha_{0\cdot}(t) = \sum_{j=1}^{J} \alpha_{0j}(t) \tag{1.11}$$

with corresponding accumulated hazard:

$$A_{0\cdot} = \int_0^{t_i} \alpha_{0\cdot}(u)du = \sum_{j=1}^{J} \int_0^{t_i} \alpha_{0j}(u)du \tag{1.12}$$

The survival function can be computed directly by the all-cause accumulated hazard, that

$$S_{0\cdot}(t_i) = exp(-A_{0\cdot}(t_i)) = \mathbb{P}(T > t_i) \tag{1.13}$$

Therefore, the parallel definitions of process at risk, counts of transition, and the Nelson-Aalen estimator are specified as follows. We denote d_{ij} as the number of failures that the subjects transition from the initial state to absorbing state j at time t_i. The d_i denotes the total number of failures at time t_i or $d_i = \sum_{j=1}^{J} d_{ij}$. n_i is the number of subjects at risk at t_i^-. Therefore, the Nelson-Aalen estimator of the cause-specific accumulated hazard function can be expressed as:

$$\hat{A}_{0j}(t) = \sum_{t_i \leq t} \frac{d_{ij}}{n_i} \tag{1.14}$$

The all-cause accumulated hazard function is the sum of all J categories of cause-specific accumulated hazard:

$$\hat{A}_{0\cdot}(t) = \sum_{j=1}^{J} \hat{A}_{0j}(t) = \sum_{t_i \leq t} \frac{\sum_{j=1}^{J} d_{ij}}{n_i} \tag{1.15}$$

The survival function can be computed using the increment of the accumulated hazard function:

$$\hat{S}(t_i) = \prod_{t_i \leq t} (1 - \Delta \hat{A}_{0\cdot}(t)) = \prod_{t_i \leq t} (1 - \frac{d_i}{n_i}) \tag{1.16}$$

The competing risk model is a common tool to study time event data that has multiple

event types. However, we have not discovered any method that computes the sample size based on our pilot study. Therefore, we propose a method in Chapter 2 to discuss how we should compute the target sample size based on the pilot study.

1.3 Cluster Randomized Trials

In this section, we introduce cluster randomized trials to build a solid background knowledge. In general, there are restrictions, such as contamination of treatment or convenience of experimental design, that prohibit us from assigning an intervention to individual subjects. Instead, we consider clusters of subjects as the experimental units for randomly allocating the intervention to capture the treatment effect at the population level. In other studies, we want to understand the indirect effect of an intervention on subjects who do not directly receive the innervation. For these cases, CRT is an option. For example, we use CRT to study the impact of vaccination campaigns on the reduction in transmission of infectious diseases. This gives a more detailed picture of the prevalence of the vaccine. However, extra consideration should be given when using CRT because of the correlation among subjects in each cluster. This makes the CRT less precise compared to simple random sampling. To quantify the difference in precision, we introduce the concept of design effect (DEff). Before delving into this concept, we need to clarify some notations and concepts.

We frame the response to a continuous variable and assume the size of each cluster is equal. The generalization to discrete responses and different cluster sizes is easy. Suppose there are c clusters each with m units, x_{jk} is the response for the k-th unit in the j-th cluster, where $j = 1, ..., c$, $k = 1, ...m$. The true mean of the j-th cluster is μ_j, which is assumed to be sampled from a distribution with mean μ and variance σ_B^2. The within-cluster variance is defined as $\sigma_W^2 = Var(y_{jk}/\mu_j)$. The intracluster correlation coefficient, ρ, measures the correlation within a cluster.

$$\rho = \frac{\mathbb{E}(x_{jk} - \mu)(x_{jk'} - \mu)}{\mathbb{E}(x_{jk} - \mu)^2} \tag{1.17}$$

We often can approximate the intraclass correlation coefficient using

$$\rho = \frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2} \tag{1.18}$$

Note that between-cluster variability and within-cluster correlation are two ways of viewing the same phenomenon.

The variance of observed cluster means is the sum of the within-cluster variance and the between-cluster variance. We have $Var(\overline{x_j}|\mu_j) = \frac{\sigma_W^2}{m}$ which indicates $Var(\overline{x_j}) = \frac{\sigma_W^2}{m} + \sigma_B^2$.

The design effect measures the increase in variance caused by using CRT. It can be defined as the ratio between the variance for cluster sampling and the variance for simple random sampling.

$$DEff = 1 + (m-1)\rho = 1 + (m-1)\frac{\sigma_B^2}{\sigma_B^2 + \sigma_W^2} \tag{1.19}$$

Therefore, reducing between-cluster variability can improve precision and power.

There are design strategies to do this. Commonly, one is matching according to covariates at the cluster level. Another is using covariate adjustment (Hayes and Moulton, 2017). Matching ensures baseline comparability before any intervention. The strength of matching depends on the correlation between the covariate and response, and the closeness of matching on the covariate. This balancing is difficult to achieve via randomization for CRTs because the randomization unit is the cluster instead of the individual subject.

Typically, there are only a few clusters in a study. Increasing baseline comparability increases the efficiency and precision of estimation. The reason for this is that when the baseline is comparable, it reduces the variation between clusters. However, the comparison test is made within each matched pair. This decreases the degree of freedom of tests. Another disadvantage of matching is the lack of replication of matched pairs. This also enlarges the problem of dropout on matched design. If one cluster is removed from the trial, the entire matched pair is lost.

Adjustment models are a common alternative method to remove imbalance and improve precision. For CRTs, the adjustment models are two-stage models. In the first stage, regression models of the covariates without the treatment effect are fitted to the response. We obtain the covariate-adjusted residuals from this stage. These residuals are further compared regarding the interventions in the second stage to test whether treatment effects are similar on average. The disadvantage of the adjustment models is that they assume linearity of the covariates. In addition, the adjusted models do not work well for the matched model because of the involvement of additional parameters. There may be a violation of asymptotic assumptions due to many parameters approaching the number of clusters. Therefore, if the experimental condition allows, a threshold for Mahalanobis distance with individual covariates would be a better approach.

1.4 Discrepancy Theory

We begin by outlining the foundations of combinatorial discrepancy and describing key properties of "low-discrepancy" algorithms that facilitate their use across many problems in computer science and statistics.

1.4.1 Notation and Preliminaries

Let $(\boldsymbol{X}, \mathcal{S})$ be a set system, where $\boldsymbol{X}$ is the ground set of n elements, numbered 1 through n, and $\mathcal{S}$ is a collection of m sets, numbered 1 through m, where $S_j \subset \mathbf{X}$, for all $S_j \in \mathcal{S}$. Let $\boldsymbol{\Omega}$ denote all mappings $\chi : \boldsymbol{X} \to \{-1, 1\}$; note that $|\boldsymbol{\Omega}| = 2^n$. For any mapping $\chi \in \boldsymbol{\Omega}$ and for any set $S_j \in \mathcal{S}$ define

$$\chi(S_j) = \sum_{i \in S_j} \chi(i). \tag{1.20}$$

The discrepancy of the set system $(\boldsymbol{X}, \mathcal{S})$ with respect to the mapping χ is

$$disc(\chi, \mathcal{S}) = \max_{S_j \subset \mathcal{S}} |\chi(S_j)|, \tag{1.21}$$

and the discrepancy of the set system $(\boldsymbol{X}, \mathcal{S})$ is

$$disc(\mathcal{S}) = \min_{\chi \in \boldsymbol{\Omega}} disc(\chi, \mathcal{S}) \tag{1.22}$$

The set system can also be expressed in a matrix form. Let $\boldsymbol{A} \in \{0,1\}^{m \times n}$ denote an incidence matrix such that columns correspond to elements in $\boldsymbol{X}$ and rows correspond to sets in $\mathcal{S}$, and $\boldsymbol{A}_{ji} = 1$ is equal to 1 if and only if element $i \in \mathbf{X}$ is in set $S_j \in \mathcal{S}$.

$$\boldsymbol{A}_{ji} = \begin{cases} 1, & i \in S_j, \\ 0, & \text{otherwise.} \end{cases} \tag{1.23}$$

A mapping χ can be expressed as a vector $\mathbf{x} \in \{-1,1\}^n$ where the i^{th} component $\mathbf{x}_i = \chi(i)$. Thus, for any S_j,

$$\chi(\boldsymbol{S_j}) = (\boldsymbol{Ax})_j. \tag{1.24}$$

The discrepancy of the set system $(\boldsymbol{X}, \mathcal{S})$ with respect to the mapping χ is then

$$disc(\chi, \mathcal{S}) = max_{j \in [m]} (\boldsymbol{Ax})_j = ||\boldsymbol{Ax}||_\infty, \tag{1.25}$$

where $||\cdot||_\infty$ denotes the ℓ^∞ norm—the maximum value of the provided vector. Finally, the discrepancy of the set system $(\boldsymbol{X}, \mathcal{S})$ is

$$disc(\mathcal{S}) = min_{\boldsymbol{x} \in \{-1,1\}^n} ||\boldsymbol{Ax}||_\infty \tag{1.26}$$

1.4.2 Banaszczyk's Theorem and Relative Algorithm

Spencer (1985) discovered that the discrepancy of any set system with n elements and m sets is, at most, $O(\sqrt{nlog(\frac{m}{n+1})})$. If $m = n$, then the discrepancy is at most $6\sqrt{n}$. This degree of low discrepancy cannot be achieved by complete randomization.

A low discrepancy algorithm outputs an optimal mapping $\boldsymbol{x}^{\dagger}$ that minimizes the discrepancy of input $\boldsymbol{A}$. The latest low discrepancy algorithm is developed from Banaszczyk's Theorem. Banaszczyk (1998) and Dadush et al. (2016) show that a low discrepancy mapping can be obtained by generating a sub-Gaussian process based on the input $\boldsymbol{v_1}, ..., \boldsymbol{v_n}$.

1.4.3 Relationship to Vector Balancing

As it turns out, vector balancing problems can be viewed as a generalization of the discrepancy theory problem. We extend the definition in (1.26) to allow for $\mathbf{A}$ to be *any* $m \times n$ matrix:

$$disc(\boldsymbol{A}) = min_{\boldsymbol{x}\in\{-1,1\}^n}||\boldsymbol{A}\boldsymbol{x}||_{\infty} \tag{1.27}$$

1.5 Kernel Ridge Regression

1.5.1 Ridge Regression

Ridge Regression is proposed by Hoerl and Kennard (1970) for improving the instability of the Least Squares Estimator by adding a diagnose matrix, of $\lambda\boldsymbol{I}$, to the the $\boldsymbol{X^T X}$ to avoid singularity. The corresponding loss function is

$$\boldsymbol{L_{\lambda}}(\boldsymbol{\beta}, \boldsymbol{X}) = min_{\beta}\lambda||\beta||_2^2 + (\boldsymbol{y} - \boldsymbol{X^T\beta})^2 \tag{1.28}$$

which results the solution for $\boldsymbol{\beta}$,that

$$\boldsymbol{\beta} = (\boldsymbol{X^T X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X^T y} \tag{1.29}$$

The above solution is called the primal form. Alternatively, the dual form can be expressed by letting

$$\boldsymbol{\beta} = \boldsymbol{X^T \alpha} \tag{1.30}$$

by solving for $\boldsymbol{\alpha}$ we have

$$\boldsymbol{\alpha} = (\boldsymbol{X X^T} + \lambda \boldsymbol{I})^{-1} \boldsymbol{y} \tag{1.31}$$

The dual form implies the suitability of combining ridge regression with the kernel method via the $\boldsymbol{X X^T}$ term.

However, the ridge regression estimate is biased. Thus, the expected loss is a combination of squared bias and variance. Through cross-validation, the smaller λ is preferred, which introduces a smaller bias. An appropriate choice of λ can result smaller expected loss compared to the least square model, which gives better prediction accuracy.

From a geometric interpretation:

$$\begin{aligned} \hat{\boldsymbol{y}} &= \boldsymbol{X}(\boldsymbol{X^T X} + \boldsymbol{\lambda I})^{-1} \boldsymbol{X^T y} \\ &= \boldsymbol{U \Lambda}(\boldsymbol{\Lambda^2} + \boldsymbol{\lambda I})^{-1} \boldsymbol{\Lambda U^T y} \\ &= \sum_{i=1}^{p} \frac{\lambda_i^2}{\lambda_i^2 + \lambda} \boldsymbol{u_i u_i^T y} \end{aligned} \tag{1.32}$$

where $\boldsymbol{U}$ is the othonormal basis of $\boldsymbol{X}$ and λ_i is the corresponding eigenvalue to the eigenvector u_i.

Regarding the geometric interpretation, ridge regression shrinks the coordinates of the principal component coordinates, and $\frac{\lambda_i^2}{\lambda_i^2+\lambda}$ can be viewed as the shrinkage factor. Therefore, the smaller the variability is on the principal component coordinate, the greater the shrinkage is on the same coordinate. The larger λ is, the more the projection is shrunk in every coordinate. In addition, the $\frac{\lambda_i^2}{\lambda_i^2+\lambda}$ is also related to the effective degree of freedom associated

with $\boldsymbol{\beta}$

$$df(\lambda) = \sum_{i=1}^{p} \frac{\lambda_i^2}{\lambda_i^2 + \lambda} \tag{1.33}$$

As we can see, if the $\lambda = 0$ and the $\boldsymbol{X^T X}$ are not singular with the effective degree of freedom, $df(\lambda) = p$.

1.5.2 Kernel Functions and Hilbert Space

The advantage of using the kernel model is modeling a nonlinear relationship without explicitly working on these complicated functions. By expressing the data in the kernel function, we can work directly in terms of kernels that allow us implicitly to work on a feature space that is higher in dimension without explicitly introducing the feature vectors. In this section, we focus on the fundamentals of the kernel model, from the definition of kernel function to Mercer's theorem.

A kernel function κ for $\boldsymbol{x_i}$, $\boldsymbol{x_j} \in \boldsymbol{X}$ is defined as:

$$\kappa(\boldsymbol{x_i}, \boldsymbol{x_j}) = \langle \phi(\boldsymbol{x_i}), \phi(\boldsymbol{x_j}) \rangle \tag{1.34}$$

where ϕ is a mapping from $\boldsymbol{X}$ to an inner product feature space $\mathcal{F}$. An inner product feature space can be defined as a vector space of $\boldsymbol{X}$ that satisfies a real value symmetric dot product $\langle \cdot, \cdot \rangle$ such that $\langle \boldsymbol{x}, \boldsymbol{x} \rangle \geq 0$. Therefore, elements in the feature space can be expressed implicitly by linear functions of the dot product instead of explicitly constructing it.

The feature space can be viewed as a set of functions:

$$\mathcal{F} = \{ \sum_i \alpha_i \kappa(\boldsymbol{x_i},) : \boldsymbol{x_i} \in \boldsymbol{X}, \alpha_i \in \mathbb{R} \} \tag{1.35}$$

The space is closed under scalar multiplication and addition of functions such that

$$f, g \in \mathcal{F} \Rightarrow (f+g)(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x}) \tag{1.36}$$

where $f(\boldsymbol{x}) = \sum_i \alpha_i \kappa(\boldsymbol{x_i}, \boldsymbol{x})$, and $g(\boldsymbol{x}) = \sum_j \beta_i \kappa(\boldsymbol{z_j}, \boldsymbol{x})$.

The kernel matrix is a $N \times N$ Gram matrix that the i^{th} and the j^{th} entries is $\kappa_{ij} = \kappa(\boldsymbol{x_i}, \boldsymbol{x_j}) = \langle \phi(\boldsymbol{x_i}), \phi(\boldsymbol{x_j}) \rangle$. where ϕ is a feature mapping that is associated with a certain type of kernel function. The kernel matrix is a symmetric matrix that contains all the information about the pairwise relations within the dataset. However, this information is limited, and any coordinate rotation leaves the matrix unchanged.

The finitely positive semi-definite function plays its central role in defining the kernel matrix. A function $\kappa : \boldsymbol{X} \times \boldsymbol{X} \mapsto \mathbb{R}$ is a finite positive semi-definite if it is a symmetric function for which the matrices formed by restricting to a finite subset of the space X are positive semi-definite.

If a function $\kappa : \boldsymbol{X} \times \boldsymbol{X} \mapsto \mathbb{R}$, which is either continuous or has a finite domain, that satisfies the finitely positive semi-definite property, can be decomposed $\kappa(\boldsymbol{x_i}, \boldsymbol{x_j}) = \langle \boldsymbol{x_i}, \boldsymbol{x_j} \rangle$ into a feature space map ϕ into a Hilbert space $\mathcal{F}$ applied to both its arguments followed by the evaluation of the inner product in $\mathcal{F}$. That is called the characterization of the kernel.

The inner product

$$\langle f, g \rangle = \sum_i \sum_j \alpha_i \beta_j \kappa(\boldsymbol{x_i}, \boldsymbol{z_j}) \tag{1.37}$$

The representing property of the kernel can be given as:

$$\langle f, \kappa(\boldsymbol{x}, \cdot) \rangle = \langle f, \phi(\boldsymbol{x}) \rangle = \sum_i \alpha_i \kappa(\boldsymbol{x_i}, \boldsymbol{x}) = f(\boldsymbol{x}) \tag{1.38}$$

$$||f||^2 = \sum_i \sum_j \alpha_i \alpha_j \kappa(\boldsymbol{x_i}, \boldsymbol{x_j}) = \boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha} \tag{1.39}$$

Thus, the function f can be represented as a linear function that is defined by an inner product in the feature space $\mathcal{F}$

For complicity, Mercer's theorem is included. Under conditions of symmetry, continuity, and positive semi-definiteness, the kernel matrix can be decomposed as:

$$\boldsymbol{K} = \sum_{i=1}^{\infty} \lambda_i \boldsymbol{v_i} \boldsymbol{v_i}^T \tag{1.40}$$

where $\boldsymbol{v_i}$ are eigenvectors, λ_i are eigenvalues. Therefore, the eigenvectors v_i can be treated as functions of $\mathcal{F}$ and the eigenvalues as the distribution over these functions. Another construction for the feature space $\mathcal{F}$ can be expressed to the eigenfunctions of the integral:

$$f(\cdot) \to \int_{\boldsymbol{X}} \kappa(\boldsymbol{x}, \cdot) f(\boldsymbol{x}) d\boldsymbol{x} \tag{1.41}$$

the reproducing theorem guarantees that $\langle v_i, v_i \rangle_{\mathcal{H}} = \frac{1}{\lambda_i}$, $\langle v_i, v_{i'} \rangle_{\mathcal{H}} = 0$, for $i \neq i'$. For any function in space $\mathcal{H}$, the coefficient for basis v_i is defined as $\theta_i = \langle f, v_i \rangle_{L^2(\mathbb{P})}$. Therefore, the corresponding function can be expanded as $f = \sum_{i=1}^{\infty} \theta_i v_i$, and $||f||_2^2 = \sum_{i=1}^{\infty} \theta_i^2$, $||f||_{\mathcal{H}}^2 = \sum_{i=1}^{\infty} \frac{\theta_i^2}{\lambda_i}$.

1.5.3 Kernel Ridge Regression

Compared with ridge regression, kernel ridge regression has comparably weaker assumptions. It is a flexible, non-parametric method that has a wide range of applications since it is smoother and less complicated compared with other non-linear models.

In the framework of kernel ridge regression, the optimization problem becomes

$$argmin_{f\in\mathcal{F}} \sum_i (f(\boldsymbol{x_i}) - y_i)^2 + \lambda||f||^2_{\mathcal{F}} \tag{1.42}$$

Therefore, the solution becomes:

$$\boldsymbol{\alpha} = (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}\boldsymbol{y} \tag{1.43}$$

or we can directly use the equation 1.31 of the dual form, by simply replacing XX^T with a suitable kernel matrix. This solution also matches the result of the optimization problem above.

Thus, the prediction becomes:

$$f(\boldsymbol{x}) = \sum_i \alpha_i \kappa(\boldsymbol{x_i}, \boldsymbol{x}) \tag{1.44}$$

From the above, we can easily see that as the sample size, N, increases, the kernel matrix's size increases, making the computation more complicated and time-consuming. Therefore, the model average method divides the full-size data into multiple subsets of data to increase the computation efficiency. We focus the discussion on the model average method and application of discrepancy theory to this procedure in chapter 4.

1.6 Conclusion

In the involvement processes of statistical methods, exciting discoveries continuously emerge. New extensions are required to complete the system in the application field, especially for

sample size and power computation. For this dissertation, we propose a sample size computation method for the competing risk model in Chapter 2. Our method can offer further guidance in experimental design on sample size for the multistate model. Other discoveries can be further improved by inspiration from the latest research in other disciplines. In this dissertation, we get our inspiration from the discrepancy theory. We apply the discrepancy theory to the model averaging method for cluster randomized trials and kernel ridge regression. We discuss how it can increase the precision of the cluster randomized trials in Chapter 3. Applying a low discrepancy algorithm during clustering can reduce between-cluster variance and design effect. Additionally, in Chapter 4, we discuss how a low discrepancy algorithm can enhance the computational efficiency of model averaging within the subset procedure.

Chapter 2

Sample Size Computation for Competing Risk Models Using Flexible Parametric Models

2.1 Introduction

Survival analysis is a branch of statistical analysis that focuses on analyzing the time before a target event is observed. In other words, survival models provide a statistical framework for analyzing time-to-event data. For example, survival models may be used in oncology studies to determine the efficacy of the latest chemotherapy treatment regimen. We may treat patient death as the event of interest and compare the survival times between patients that receive new treatment versus those that receive the current standard care (Cole et al., 1996; Fizazi et al., 2019; Solomon et al., 2018).

Competing risk models, or multi-state models, extend the survival analysis framework to accommodate studies with multiple types of terminal events (Beyersmann et al., 2011,?). For instance, an oncology study may be interested in the effect of a treatment on both the time-to-death and the time-to-remission.

Competing risk models not only present an average probability of survival over all events of interest, but also provides multiple specific analyses and comparisons according to the different event types, thereby providing a more detailed analysis than offered by traditional single-state event survival analysis models.

When designing studies utilizing competing-risk models, it is critical to accurately determine the requisite sample sizes to detect pre-specified differences in survival times; if sample sizes are too small, then studies will not have sufficient power to detect these differences, and if sample sizes are too large, then the cost of performing these studies may be unnecessarily high. Additionally, different studies may be interested in different types of effects. For example, studies may be interested in one type of event, and conduct sample size computation only concentrates on the interested type of event (Latouche and Porcher, 2007; Latouche et al., 2004; Schulgen et al., 2005). Others may be interested in detecting significance at least on one type of event based on a joint test Yang et al. (2018). There has been limited research on calculating sample sizes for detecting significance across multiple types of events. Additionally, much of this work has relied on stringent model assumptions (for example, "parallel trends") and may perform poorly when these assumptions are violated.

We consider a new framework for computing these sample sizes across a variety of different model specifications when pilot data is available. To be explicit, we begin by assuming, prior to the design of an elaborate experiment, that data from a much smaller pilot study is available. We then use a flexible parametric survival model to accurately estimate the cause-specific hazards. We then repeatedly simulate data from these fitted models to estimate the power of a specified test we intend to use for the elaborate experiment with different input sample sizes. The sample size that reaches the pre-specified power level is the one we recommend for a later official experiment.

The paper is organized as follows. In section 2.2.1, we review concepts of classic survival analysis. We concentrate on these concepts to facilitate our extensions for illustrating competing risk analysis in the following section. In section 2.3, we introduce the cause-specific

model for competing risk because it is highly related to the data generation. In section **??**, we discuss how to generate time-event data for multiple event types. Then, in section **??**, we review the parametric models we plan to use for extracting information from the pilot study. We describe the method we use for evaluating these parametric models in Section **??**. We demonstrate the versatility of the flexible parametric model using an example dataset from all those we have tested. Additionally, we discuss a newly discovered property of the generalized log-gamma model in Section **??**.

2.2 Notation and Preliminiaries

2.2.1 Classic Survival Analysis

We begin by reviewing classic survival analysis concepts that can be directly extended to competing risk models or that relate to some goodness-of-fit tests that will be discussed in future sections. Consider a study on n subjects, numbered 1 through n. Each unit i is observed from time $t = 0$ to time $c > 0$. We study the transition of subject i from an initial state, 0, to an absorbing state, 1. We are particularly interested in analyzing the *survival time* of subject i—length of time until subject i transitions from the initial state to the absorbing state. Rigorously, let $X_{it} \in \{0, 1\}$ denote the state of subject i at time t, $t \geq 0$. Then, the survival time T_i for subject i is

$$T_i = \inf\{t : X_{it} \neq 0\}. \tag{2.1}$$

We assume that $T_i, i = 1, \ldots, n$ are independent and identically distributed (iid) random variables; for ease of exposition, we may drop the subscript from T_i when we can do so unambiguously.

Note, the '0' state is often considered a "survival" state and the '1' state is often considered a "failure" state. Thus, if a subject does not transition from the initial state to the

absorbing state by time t, we may refer to that subject as "surviving" until time t. Similarly, we often refer to a subject transitioning from the initial state to the absorbing state at time t as subject "failing" at time t, and and the survival time may be referred to as the "failure time"

Let $f(t)$ and $F(t)$ denote the probability density function (pdf) and the cumulative density function (cdf) of the survival time T. The *survival function* $S(t)$ is the probability that a subject survives until time t:

$$S(t) = \mathbb{P}(T > t) = 1 - F(t) \tag{2.2}$$

The hazard function $h(t)$ is the instantaneous rate of failure for a subject given that the subject survives until time t:

$$h(t) = lim_{\Delta t \to 0} \frac{\mathbb{P}(t < T < t + \Delta t | T > t)}{\Delta t} = \frac{f(t)}{S(t)} \tag{2.3}$$

The cumulative hazard function $H(t)$ is obtained by integrating the hazard function from time 0 to time t:

$$H(t) = \mathbb{P}(T \leq t) = \int_0^t h(u)du. \tag{2.4}$$

Note that the survival function can be expressed in terms of the cumulative hazard function:

$$S(t) = exp(-H(t)). \tag{2.5}$$

Furthermore, the pdf $f(t)$ can be rewritten in terms of $h(t)$ and $H(t)$:

$$f(t) = h(t)S(t) = h(t)exp(-H(t)). \tag{2.6}$$

We assume that, given infinite time, all subjects will eventually fail. However, failure may occur during a period for which the unit is not observed. For example, a subject may

fail after the end point of the study c, or may drop out of the study prior to time c for reasons unrelated to the study. This results in *right-censoring* of the survival times T Censoring is considered an incomplete observation and provides partial information about the distribution of T. In other words, in scenarios where the subjects depart from the study or the study ends before an interesting event is observed, we, at least, know the subject stays in the initial state at the last observation and assume the transition takes place in (C_i, ∞). In our notation, we know that $T_i > t$ for subject i, if subject i is censored. We always assume the censoring mechanism is independent of the mechanism by which the subject transitions from the 0 state to the 1 state. In addition, we only consider the right censoring for simplicity. The response is recorded as a combination of the observation time $\tilde{t} = t_i \wedge c_i$ and the censor indicator $k_i = 1\{t_i \leq c_i\}$.

There are a variety of approaches for estimating survival curves and hazard functions, ranging from structured, parametric approaches to completely non-parametric approaches. Due to the nature of risk and the existence of outliers, non-parametric methods often outperform parametric methods, and hence, these approaches are more commonly used in practice. However, there are situations where it is advantageous to use a parametric approach. For example, when data is scarce, nonparametric or semi-parametric methods may be unable to model survival times well, and parametric methods may be preferred.

Because pilot studies traditionally are fairly small in scale, we will use a parametric approach to model the survival times from the pilot study. However, the follow-up study is often much larger, and hence, semi-parametric and nonparametric approaches may be effective for the follow-up study. Thus, when performing power calculations for the larger study, we will assume that data from the follow-up study will be modeled using a nonparametric approach.

2.2.2 Parametric Survival Analysis Using the Generalized Log-Gamma Distribution

Parametric methods assume that survival times T_i follow a well-known distribution. Once the distribution is specified, the pdf $f(t)$ and survival function $S(t)$ can be deduced, and likelihood-based methods can then be used to estimate unknown parameters of the model—for example, by maximizing the likelihood. Note, when right-censoring events are independent of the survival times, the likelihood of observing the survival times $t_1, t_2, \ldots, t_n$ is

$$L(\theta|\mathbf{t}) = \prod_i f(t_i)^{k_i} S(t_i)^{1-k_i}. \tag{2.7}$$

Common distributions for modeling survival times include the exponential, Weibull, log-normal, and log-logistic distributions. More recently, the generalized log-gamma distribution has been shown to be effective for modeling survival times. The aforementioned distributions can be shown to be special cases or transformations of the generalized log-gamma distribution. We now describe this distribution in detail.

Generalized Log-Gamma Distribution

The flexible parametric survival models can be constructed using two distributions: the generalized gamma distribution and the generalized F distribution. For the generalized gamma model, there are two parameterizations. One is the original, which was first proposed by Stacy (1962). We are focusing on the generalized log-gamma distribution, an extension of the generalized gamma distribution. The later extended the parameter space of q from $[0, \infty]$ to $[-\infty, \infty]$. This extension is believed to stabilize the fitting process by smoothly updating it via the Newton-Raphson algorithm. More details can be found at Prentice (1974).

For $T > 0$,$\beta > 0$, $d > 0$ and $k > 0$. The density of the generalized gamma distribution is expressed as:

$$f(t|\beta, d, k) = \frac{\frac{k}{\beta^d} t^{d-1} exp(-(\frac{t}{\beta})^k)}{\Gamma(\frac{d}{k})} \tag{2.8}$$

By re-write that $\gamma = \frac{d}{k}$, we have

$$f(t|\beta,\gamma,k) = \frac{\frac{k}{\beta^{k\gamma}} t^{k\gamma-1} exp(-(\frac{t}{\beta})^k)}{\Gamma(\gamma)} \tag{2.9}$$

By re-parameterize by letting:

$$q = \frac{1}{\sqrt{\gamma}} \rightarrow \gamma = q^{-2}$$
$$\sigma = \frac{1}{k\sqrt{\gamma}} \rightarrow k = \frac{q}{\sigma}$$
$$\mu = log(\beta) + \frac{log(\gamma)}{k} \rightarrow \beta = exp(-\mu)[(q^{-2})^{\frac{\sigma}{q}}]$$

Afterword, the density function can be expressed as:

$$f(t|\mu,\sigma,q) = \frac{|q|}{\sigma t \Gamma(q^{-2})} [q^{-2}(exp(-\mu)t)^{\frac{q}{\sigma}}]^{q^{-2}} exp[-q^{-2}(exp(-\mu)t)^{\frac{q}{\sigma}}] \tag{2.10}$$

This re-parameterization expands the parameter spaces that $\sigma > 0$, $\mu \in \mathbb{R}$ and $q \in \mathbb{R}$ and makes the process of updating maximum likelihood estimation smoother. We refer to μ as the location parameter, σ as the scale parameter, and q as the shape parameter. We are going to dive deeper on the meaning of each parameter later when we discuss the quantile function. Further, since $exp(q[\frac{log(t)-\mu}{\sigma}]) = [exp(-\mu)t]^{\frac{q}{\sigma}}$, and we can express the density function as:

$$f(t|\mu,\sigma,q) = \begin{cases} \frac{|q|(q^{-2})^{q^{-2}}}{t\sigma\Gamma(q^{-2})} exp\{q^{-2}[qw - exp(qw)]\}, & q \neq 0 \\ \frac{1}{(2\pi\sigma^2)^{\frac{1}{2}}} exp[-\frac{(log(t)-\mu)^2}{2\sigma^2}], & q = 0 \end{cases} \tag{2.11}$$

where $w = \frac{log(t)-\mu}{\sigma}$. We refer equation 2.11 as the density of generalized log gamma distribution. Many commonly-used distributions in survival analysis, including the Weibull, exponential, and log-normal distributions, are special cases of this generalized gamma dis-

tribution. For example, if we let $q = \sigma$, we have the two-parameter gamma distribution. We can have a Weibull distribution by letting $q = 1$, and obtaini an exponential distribution by letting $q = \sigma = 1$. Further, by observation of equation 2.11, if $q = 0$ we have a log-normal distribution.

In survival analysis, the survival function built by using the generalized log gamma distribution is expressed as:

$$S(t) = \begin{cases} 1 - \int_0^u \frac{x^{q^{-2}} exp(-x)}{\Gamma(q^{-2})} & \text{if} \quad q > 0 \quad \text{where} \quad u = q^{-2} exp(|q| \cdot \frac{log(t)-\mu}{\sigma}) \\ \int_0^u \frac{x^{q^{-2}} exp(-x)}{\Gamma(q^{-2})} & \text{if} \quad q < 0 \quad \text{where} \quad u = q^{-2} exp(-|q| \cdot \frac{log(t)-\mu}{\sigma}) \end{cases} \tag{2.12}$$

Furthermore, the cumulative hazard function can be found by $H(t) = -log(S(t))$ and the hazard function is $h(t) = \frac{f(t)}{S(t)}$.

The interpretation of each parameter is related to the quantile function of the generalized log gamma distribution. A quantile function indicates the percentage, p, of the subjects that experience the event at time t. Suppose $\Gamma^{-1}(p;\gamma)$ denotes the quantile function of the gamma distribution with mean and variance equal to γ. And $\Gamma(\Gamma^{-1}(p;\gamma);\gamma) = p$. Therefore, we can further define that the p-th percentile of the standard generalized gamma, such that $\mu = 0$ and $\sigma = 1$ can be defined as:

$$t_{GG(0,1,q)}(p) = [q^2 \Gamma^{-1}(p; q^{-2}]^{\frac{1}{q}}, \text{if} \quad q > 0 \tag{2.13}$$

$$[q^2 \Gamma^{-1}(1-p; q^{-2}]^{\frac{1}{q}}, \text{if} \quad q < 0 \tag{2.14}$$

The quantile function for a generalized log gamma can be expressed as:

$$log(t_{GG(\mu,\sigma,q)}(p)) = \mu + \sigma log(t_{GG(0,1,q)}(p)) \tag{2.15}$$

As we can see above, q is the parameter that determines the logs of the standard quantile.

The percentiles are then scaled by σ and shifted by μ. In other words, we can say, for fixed σ and q, $\mu = log(median) - \sigma log(t_{GG(0,1,q)}(0.5)$. For fixed q and independent μ, σ determines the ratio of the third quartile over the first quartile, IQR, or we can express σ as $log(IQR)/\{log(t_{GG(0,1,q)}(0.75)/t_{GG(0,1,q)}(0.25))\}$. In addition, the shape of the hazard function is also mainly determined by σ and q.

Nonparametric Classic Survival Analysis

We will begin by introducing the nonparametric method, the Kaplan-Meier model (Kaplan and Meier, 1958). However, before doing so, we need to define some additional notations. We define d_i as the number of transitions $0 \to 1$ at time t_i. We further define n_i as the number of subjects at risk right before t_i (or at time t^-). We can estimate these interested qualities in the framework of binomial likelihood. The Nelson-Aalen estimator of the cumulative hazard can be expressed as:

$$\hat{H}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i} \tag{2.16}$$

with the corresponding variance estimator in the framework of binomial likelihood:

$$Var(\hat{H}(t)) = \sum_{t_i \leq t} \frac{d_i(n_i - d_i)}{n_i^3}. \tag{2.17}$$

The Kaplan-Meier estimator of survival function is defined as:

$$\hat{S}(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i} \tag{2.18}$$

We can use the log-rank test to test whether two survival curves across different treatment groups are different. Here, we only present the log-rank test for comparing two survival functions because we can make a direct connection to the goodness of fit test we are going to discuss later. However, if more than two survival functions need to be compared, we can use a multiple-group log-rank test.

	Group 1	Group 2	Total
"Death"	d_{1i}	d_{2i}	d_i
"Survival"	$n_{1i} - d_{1i}$	$n_{2i} - d_{2i}$	$n_i - d_i$
Total	n_{1i}	n_{2i}	n_i

Consider two treatment groups with survival functions S_1 and S_2 respectively. We consider a test of the null hypothesis $H_0 : S_1 = S_2$. The corresponding number of failures at t_i are d_{1i} and d_{2i}, and the number of subjects at risk are denoted as n_{1i} and n_{2i}. Then we have a contingency table at t_i as: Under H_0 and conditional on the margins D_{1i} follows a hypergeometric distribution. $\mathbb{P}(D_{1i}) = \binom{d_i}{D_{1i}}\binom{n_i - d_i}{n_{1i} - D_{1i}} / \binom{n_i}{n_{1i}}$. Additionally, in large studies, by the Lyapunov central limit theorem, we have

$$\frac{\{\sum_i D_{1i} - \mathbb{E}(D_{1i})\}^2}{\sum_i Var(D_{1i})} \sim \chi_1^2 \tag{2.19}$$

The observed-expected form will be discussed again when we address the goodness of fit test for the survival function.

While the Kaplan-Meier model's strata method is useful, it falls short when handling multiple covariates and random effects. In such cases, the semi-parametric Cox model serves as a better alternative.

The Semi-parametric Classic Survival Model

The semi-parametric model for classic survival analysis is the Cox proportional hazard that is proposed by Cox (1972). it assume that

$$h_i(t) = h_0(t) exp(\boldsymbol{x_i^T \beta}) \tag{2.20}$$

where $h_0(t)$ is the baseline of the hazard, also known as the nonparametric part, which does not involve any assumption. The expression $\exp(\boldsymbol{x_i^T \beta})$ represents the parametric component. We do not include an intercept term in $\boldsymbol{x_i}$ to keep the baseline from further restrictions. We

call the model the proportional hazard model because the hazard ratio only includes the parametric part, and it is constant over time. The proportional hazard also implies that:

$$H_i(t) = H_0(t) \cdot exp(\boldsymbol{x_i^T \beta}) \tag{2.21}$$

$$S_i(t) = S_0(t)^{exp(\boldsymbol{x_i^T \beta})} \tag{2.22}$$

which will be useful when we explore the Nelson-Aalen-Breslow estimator of the nonparametric component.

Under the proportional hazard assumption, without considering censoring, we have:

$$\mathbb{P}(T_1 < T_2 < ... < T_i) = \prod_i \frac{exp(\boldsymbol{x_i \beta})}{\sum_{l=i}^{n} exp(\boldsymbol{x_l \beta})} \tag{2.23}$$

We can further derive a rank-based partial likelihood from this probability. It is called partial likelihood because it does not specify the probability of observing a given set of data. As we can see, the partial likelihood is free of $h_0(t)$. We use it to estimate the parametric part and develop asymptotic properties for $\boldsymbol{\beta}$. After estimating the parametric part, we can use the Nelson-Aalen-Breslow method to estimate the nonparametric part. By conditioning on $\hat{\boldsymbol{\beta}}$ we have

$$\hat{H}_0(t) = \sum_{t_i \leq t} \frac{d_i}{\sum_{R(t)} [exp(\boldsymbol{x_i^T \beta})]} \tag{2.24}$$

where $R(t)$ is the set of subject in risk at time t_i.

2.3 Cause-specific Model for Competing Risk

The competing risk model can be viewed as an extension of the classic survival model. We now have more than one absorbing state, denoted by j, $j = 1, ..., J$. We assume each transition from the initial state, 0, to the j-th absorbing state is guided by a specific rate. The survival time, T_i, is still our primary interest. However, the survival time is affected by

all the rates of transitions because if a subject survives, they have successfully navigated all these risks. Therefore, the process X_T becomes, for $t \geq 0$, $X_t \in \{0, 1, ..., J\}$ and $T_i = inf(t : X_t \neq 0)$. We still assume the censoring, C_i, is independent from T_i.

We define the cause-specific hazards, α_{0j}, as the probability of a unit transition from the initial state to the absorbing state j within a time unit. The subscripts $0j$ stands for the transition from the initial state, 0, to the absorbing state j. Mathematically, we have

$$\alpha_{0j}(t) = lim_{\Delta t \to 0} \frac{\mathbb{P}(T \in [t, t + \Delta t), X_t = j | T > t)}{\Delta t} \tag{2.25}$$

or

$$\alpha_{0j}(t)dt = \frac{\mathbb{P}(T \in dt, X_T = j | T > t)}{\Delta t} \tag{2.26}$$

The corresponding cumulative cause-specific is defined as

$$A_{0j}(t_i) = \int_0^{t_i} \alpha_{0j}(u)du \tag{2.27}$$

The failure time, T_i, is guided by the all-cause hazard function

$$\alpha_{0\cdot}(t) = \sum_{j=1}^{J} \alpha_{0j}(t) \tag{2.28}$$

with corresponding accumulated hazard:

$$A_{0\cdot} = \int_0^{t_i} \alpha_{0\cdot}(u)du = \sum_{j=1}^{J} \int_0^{t_i} \alpha_{0j}(u)du \tag{2.29}$$

The survival function can be computed directly by the all-cause accumulated hazard, that

$$S_{0\cdot}(t_i) = exp(-A_{0\cdot}(t_i)) = \mathbb{P}(T > t_i) \tag{2.30}$$

Additionally, we introduce the cumulative incident function, the joint probability of failure

for absorbing state j. The comparison of absorbing states decisively requires an analysis of these corresponding cumulative incident functions. It is clearer than the comparison of these corresponding accumulated hazard functions. The cumulative incident function can be computed by the survival function at t^- and the cause-specific hazard $\alpha_{0j}(t)$.

$$\mathbb{P}(T \leq t, X_T = j) = \int_0^t exp(-\int_0^t \alpha_{0\cdot}(u^-)du) \cdot \alpha_{0j}(u)du \tag{2.31}$$

Many of the same approaches for estimating and performing inferences on competing risk survival models follow similarly to those for traditional survival analysis. For completeness, we now detail these approaches.

The Nonparametric Cause-specific Model

Before diving into the nonparametric cause-specific model for competing risk. We need to define and clarify some notations. They are comparable to the notations from the nonparametric model for classic survival analysis. We denote d_{ij} as the number of failures that the subjects transition from the initial state to absorbing state j at time t_i. The d_i denotes the total number of failures at time t_i or $d_i = \sum_{j=1}^{J} d_{ij}$. n_i is the number of subjects at risk at t_i^-. Therefore, the Nelson-Aalen estimator of the cause-specific accumulated hazard function can be expressed as:

$$\hat{A}_{0j}(t) = \sum_{t_i \leq t} \frac{d_{ij}}{n_i} \tag{2.32}$$

The all-cause accumulated hazard function is the sum of all J categories of cause-specific accumulated hazard:

$$\hat{A}_{0\cdot}(t) = \sum_{j=1}^{J} \hat{A}_{0j}(t) = \sum_{t_i \leq t} \frac{\sum_{j=1}^{J} d_{ij}}{n_i} \tag{2.33}$$

The estimator of $Var(\hat{A}_{0j}(t))$, $Var(\hat{A}_{0\cdot}(t))$ and the confidence interval follow the same formulas of the Nelson-Aalen estimator of the classic survival analysis. The survival function can

be computed using the product integral of the increment of accumulated hazard function:

$$\hat{S}(t_i) = \prod_{t_i \leq t}(1 - \Delta \hat{A}_{0\cdot}(t)) = \prod_{t_i \leq t}(1 - \frac{d_i}{n_i}) \tag{2.34}$$

The estimator of $Var(\hat{S}(t))$ and the confidence interval also follow the same formulas as the Kaplan-Meier estimator of classic survival analysis. The nonparametric estimator of the cumulative incidence function can be computed by simply plugging in the Kaplan-Meier estimator of the survival function at t^- and the cause-specific accumulated hazard function at t.

$$\hat{\mathbb{P}}(T \leq t, X_T = j) = \sum_{t_i \leq t} \hat{S}(t_i^-)\frac{d_{ij}}{n_i} \tag{2.35}$$

The Semi-parametric Cause-specific Model

For the cause-specific model, the Cox regression structure is applied to the cause-specific hazard instead of the all-cause hazard. Therefore, we can write the cause-specific hazard as:

$$\alpha_{0j} = \alpha_{0j,0} exp(\boldsymbol{x_i^T \beta_{0j}}) \tag{2.36}$$

where the $\alpha_{0j,0}$ is the baseline hazard for transition $0 \to j$, and each transition has its own specific linear coefficients $\boldsymbol{\beta_{0j}}$. Therefore, if we assume no common effect among each transition, fitting Cox regression on each transition is common in practice. However, there are R packages that require us to input cause-specific covariates $\boldsymbol{x_{0j,i}}$ and output the coefficients of the Cox regression as a whole in case of additional common effect among transitions. This changes the Cox regression for cause-specific hazard as:

$$\alpha_{0j}(t) = \alpha_{0j,0}(t) exp(\boldsymbol{x_{0j,i}^T \beta}) \tag{2.37}$$

if we assume no common effect among transitions, we have $\boldsymbol{\beta} = (\boldsymbol{\beta_{01}, ..., \beta_{0J}})$ and $\boldsymbol{x_{01,i}} = [\boldsymbol{x_i^T, 0^T, .., 0^T}]^{\boldsymbol{T}}$, ..., $\boldsymbol{x_{0J,i}} = [\boldsymbol{0^T, ..., x_i^T}]^{\boldsymbol{T}}$.

Like the Cox regression of classic survival analysis, we use partial likelihood to estimate the parametric part and develop statistical tests for coefficients. We also use methods like Nelson-Aalen-Breslow to estimate the nonparametric part via the relation that $A_{0j}(t) = A_{0j,0}(t) \cdot exp(\boldsymbol{x}_{\boldsymbol{0j,i}}^{\boldsymbol{T}}\boldsymbol{\beta})$

$$\hat{A}_{0j,0}(t) = \sum_{t_i \leq t} \frac{d_{ij}}{\sum_{R(t)} exp(\boldsymbol{x}_{\boldsymbol{0j,i}}^{\boldsymbol{T}}\boldsymbol{\beta})} \tag{2.38}$$

and

$$\hat{A}_{0j}(t) = \hat{A}_{0j,0}(t) \cdot exp(\boldsymbol{x}_{\boldsymbol{0j}}^{\boldsymbol{T}}\hat{\boldsymbol{\beta}}) \tag{2.39}$$

The all-cause accumulated hazard function is the sum of all the cause-specific accumulated hazard functions.

$$\hat{A}_{0\cdot}(t) = \sum_{j=1}^{J} \hat{A}_{0j}(t) \tag{2.40}$$

The survival function can be derived from the product integral and the increments of the all-cause accumulated hazard.

$$\hat{S}(t) = \prod_{t_i \leq t} (1 - \Delta \hat{A}_{0\cdot}(t)) \tag{2.41}$$

The estimator of the cumulative incidence function can be derived by combining the survival function at t^- and the increment of the cause-specific accumulated hazard at t.

$$\hat{\mathbb{P}}(T \leq t, X_T = j) = \hat{S}(t-)\Delta \hat{A}_{0j}(t) \tag{2.42}$$

However, if we assume the cause-specific hazard functions follow the Cox regression, then the all-cause hazard $\alpha_{0\cdot}(t) = \sum_{j=1}^{J} \alpha_{0j}(t)$ does not follow the Cox model in general. We can justify the semi-parametric cause-specific model by showing that the misspecified models still provide consistent estimates, and the estimators of coefficients are the least false parameters.

The model diagnosis is a current research topic. However, it is common practice to check proportionality by graphing the all-cause accumulated hazard or the cause-specific

accumulated hazard against their corresponding baselines.

The nonparametric and semiparametric models are the most commonly used in practice. Parametric models are less frequently utilized, particularly in official experiments that require review by regulatory agencies. However, we combine a parametric cause-specific model with Bayesian statistics to conduct early-stage explorations with small sample sizes.

We can create more complex multistate models by incorporating intermediate states, allowing transitions to other states after moving to the current intermediate state, like the competing risk model with only absorbing states. We assume each transition has a special transition rate, α_{lj}. The subscribe lj indicates a transition from state l to state j. We can assume different dependencies on previous states. The most common one is assuming the future transition only depends on the current state. In other words, the future transition and the previous transition are independent, given the current state. Further, it is common to assume the transition rate is a function of time t. In other words, $\alpha_{lj}(0) \neq \alpha_{lj}(t)$. Officially, we call it time-inhomogeneous Markov multistate models. The transition probabilities from time s to t are an important component of this model. We denote it as $\boldsymbol{P}(s,t)$ with its lj element $P_{lj}(s,t)$ denotes the transition probability of $l \to j$ in the time interval (s,t). By the property of Markov chain, we know that for time $s < v < t$, $\boldsymbol{P}(s,t) = \boldsymbol{P}(s,v) \cdot \boldsymbol{P}(v,t)$. The transition probability matrix can be estimated by the Aalen-Johansen estimator that $\boldsymbol{P}(s,t) \approx \prod_{i=1}^{n}(\boldsymbol{I} + \Delta\boldsymbol{A}(t_i))$. However, the time-inhomogeneous Markov multistate models are less likely to be conducted in practice unless the time course of research is extended than usual.

2.3.1 The Weighted Log-rank Test

We discuss the weighted log-rank test in this section because the sample size computation methods we are going to compare with in section 2.6 are based on modifications of this test.

We begin by presenting it within the context of classic survival analysis, assuming there

are two treatment levels for simplicity. To test the null hypothesis that

$$H_0 : h_1(t) = h_2(t) \tag{2.43}$$

Following the notation in Section 2.2.1, d_i is the number of failures by time t_i, and d_{ik} is the number of failures in strata k by time t_i. n_i is the number of subjects at risk by time t_i, and n_{ik} is the number of subjects at risk in strata k by time t_i. The weighted log-rank test is defined as

$$U_k = \sum_{t_i=0}^{t} W_i n_k (\frac{d_{ik}}{n_{ik}} - \frac{d_i}{n_i}) \tag{2.44}$$

where W_i is an estimated weight function of t_i. The test quantity $n^{-\frac{1}{2}} \frac{U_k}{\hat{\sigma}}$ converge to a standard normal distribution where

$$\hat{\sigma}^2 = n^{-1} \sum_{t_i=0}^{t} W_i^2 \frac{n_{i1} n_{i2}}{n_i} \frac{d_i}{n_i} \tag{2.45}$$

One method we compare with is sample size computation for the supremum weighted log-rank test. The supremum weighted log-rank test is based on the limiting distribution of the $sup_{t_i} |n^{-\frac{1}{2}} \frac{U_k}{\sigma^2}|$. Another method we compare with is based on the chi-square joint test proposed by Li and Yang (2016), which can be viewed as a joint weighted log-rank test, and the joint statistics coverage of a multivariate normal distribution.

2.3.2 Suitability of the Generalized Log-gamma Model For Competing Risk Data

As we can see from previous sections, the inversion method for generating survival data is broadly accepted. Our focus in this section is to verify the flexibility of the generalized log gamma model and to test its capability of approximating the Kaplan-Meier survival curves. Among all these datasets we have tested, we present this interesting example here.

This competing risk dataset is the *sir.adm* dataset from the *R* package *mvna*. There are

six variables in this dataset: one identity indicator *id*, one categorical treatment variable *pneu* with two classes, one-time response *time*, one absorbing state indicator *status* with two failure types (1, 2) and censor (0), and two covariate variables *age* and *sex*. We exclude the covariates and use the strata method to handle the treatment variable. As we can see from figure 2.1 through 2.4. Most of the failures occurred at the beginning. We have an unbalanced treatment assignment, with most cases in the control group. In addition, events of type one dominated the dataset. This makes the approximation using parametric models more challenging for the treatment group, especially for those who end up with event two and are assigned to the treatment group. By approximating the Kaplan-Meier method using a generalized log-gamma model for these challenging cases, we can illustrate the advantages of employing a flexible parametric model for sample computation. We first separate the data by the treatment indicator because we use the strata method. Followed by fitting the individual cause-specific survival functions on the separated dataset. We assume independent censoring with an estimated censoring rate because of the small number of censoring cases in this dataset.

In our example, the generalized log gamma survival function performs better than other models for both absorbing states 1 and 2 in both the control and treatment groups. We can see this supremum of the generalized log gamma survival function in 2.1 through 2.4. In our example, most failure events happen in the early stage of the time course. Therefore, the likelihood function is dominated by the early failure events, and all the commonly used parametric models fail to approximate the tail behavior of the Kaplan-Meier survival curves, especially with a small number of failure cases. Last, in our case, only two percent of events are censored. In practice, we suggest using the independent censoring assumption and a consistent censor rate to generate censoring in these similar conditions. The figures of the survival curves are presented below.

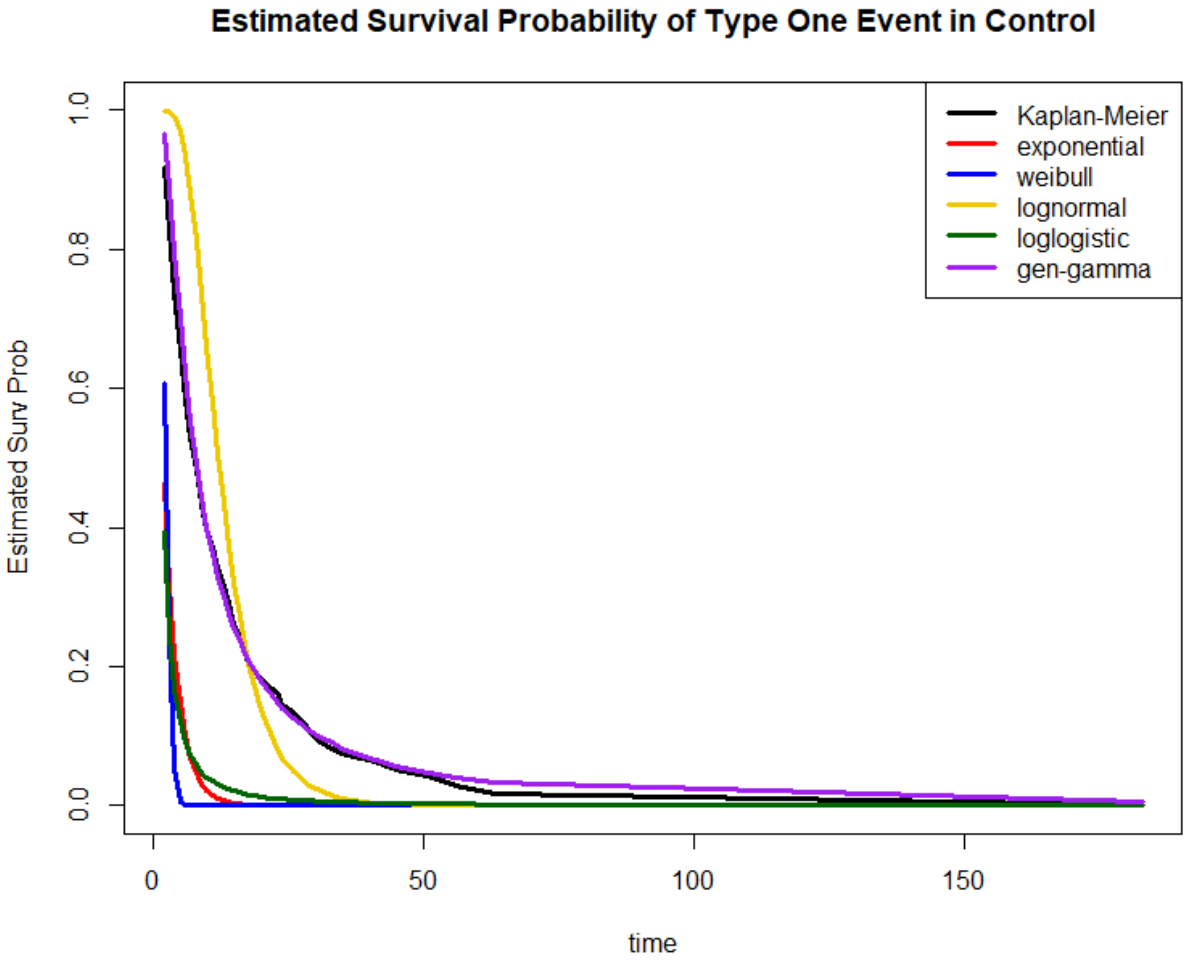

Figure 2.1: *Cause-specific survival functions for event type 1 under control*

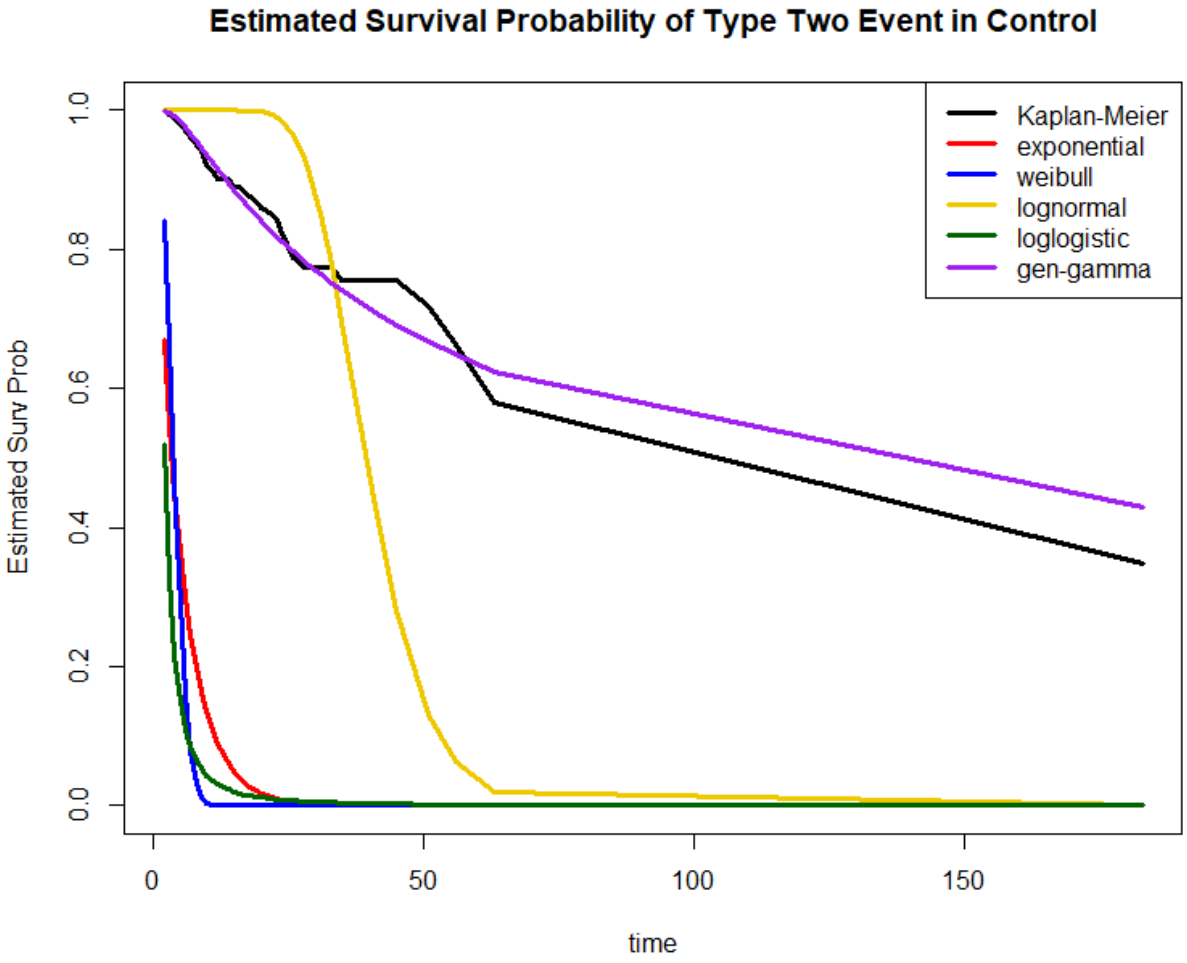

Figure 2.2: *Cause-specific survival functions for event type 1 under treatment*

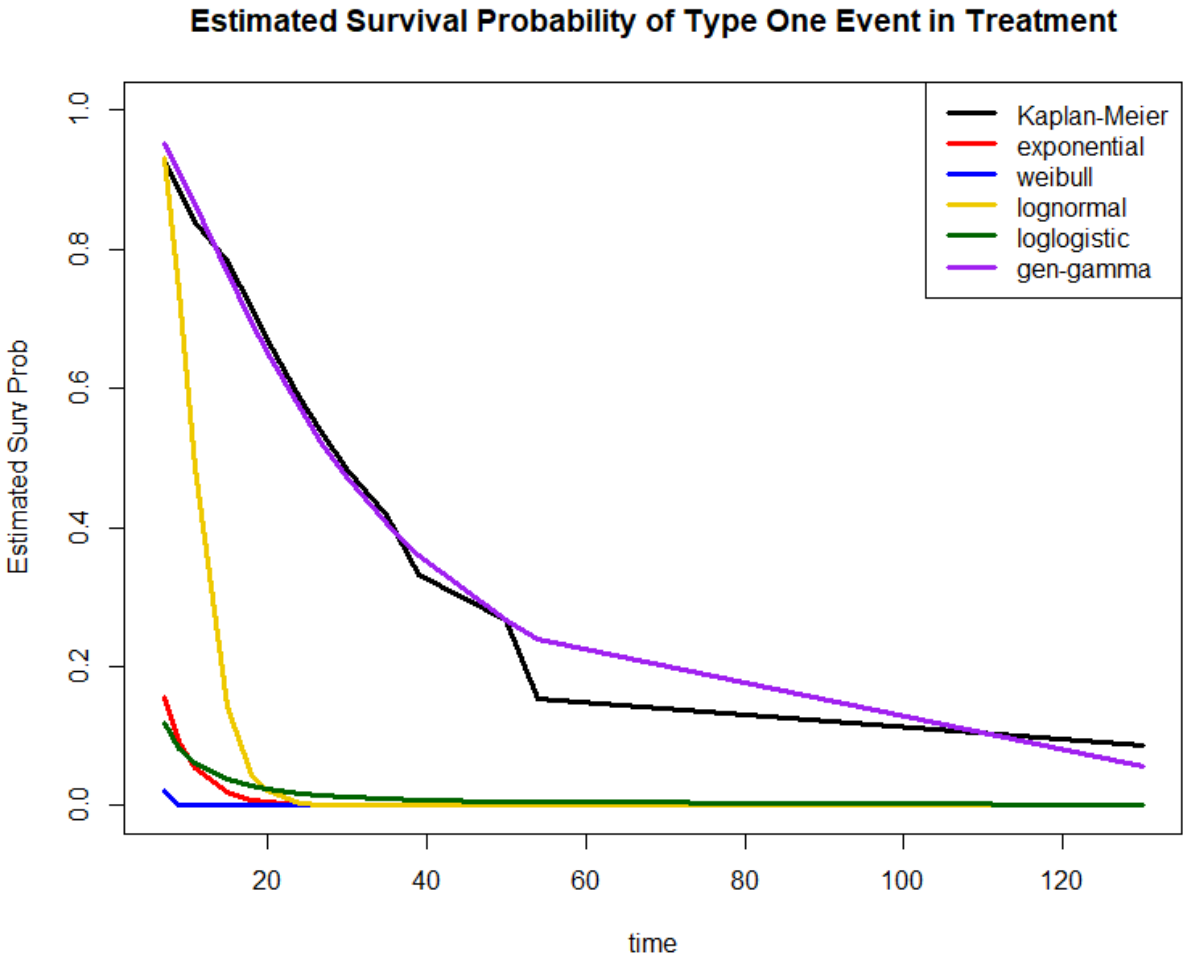

Figure 2.3: *Cause-specific survival functions for event type 2 under control*

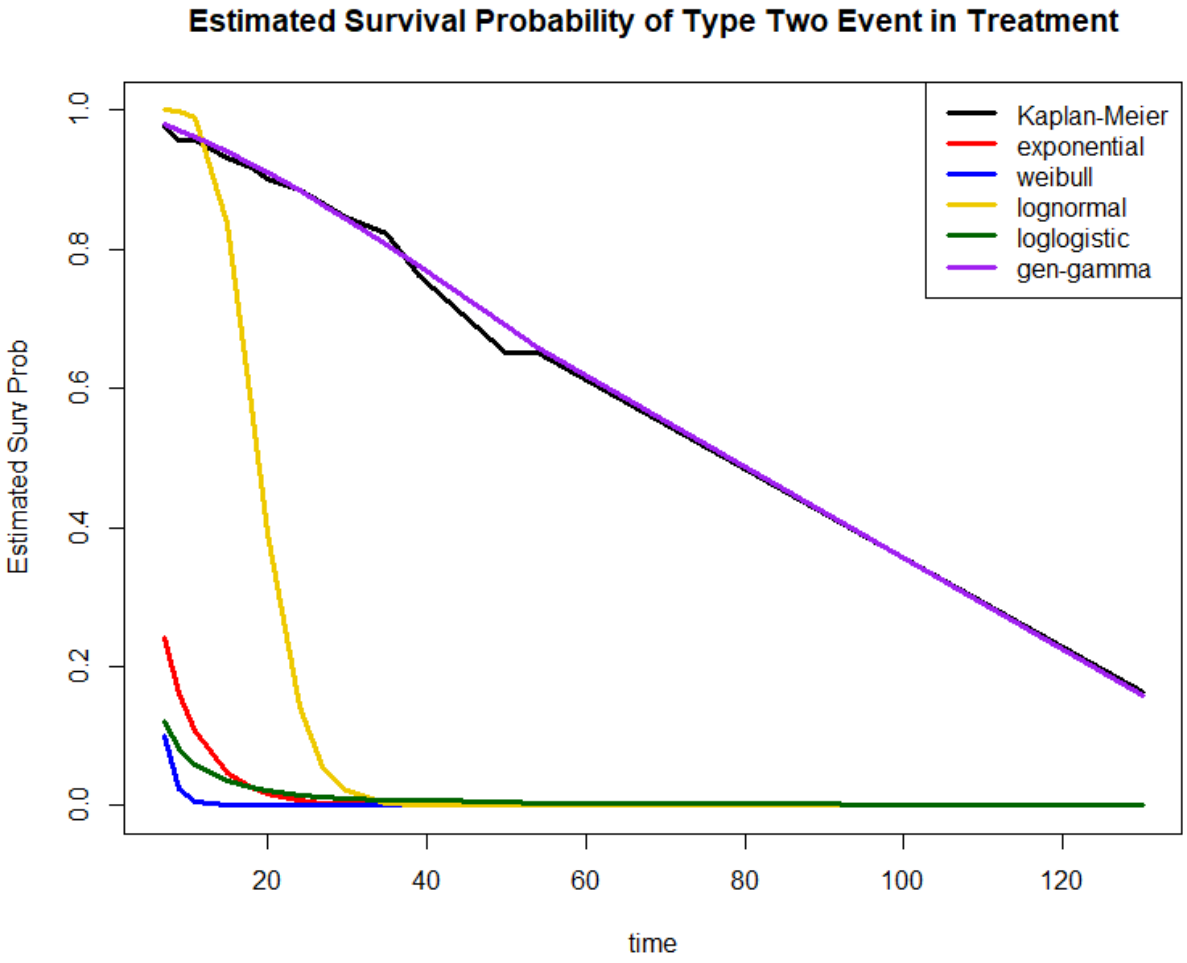

Figure 2.4: *Cause-specific survival functions for event type 2 under treatment*

2.4 Data Generation for Competing Risk

In this section, we focus on how to generate competing risk data regarding the cause-specific hazard functions. We recommend using the inversion method to generate survival time according to the all-cause hazard function, and the absorbing state depends on a multinomial trial with the probabilities of absorbing state determined by the cause-specific hazard functions. At this moment, we assume right censoring, and it is independent of the survival time. To end this section, we also explain why we suggest using parametric survival functions to approximate the non-parametric survival curve that fits the pilot data by enumerating the limitations of the data-generating procedure.

There are four steps to generate time-event data of multiple event types.

1. In the first step, we need to specify the cause-specific hazards,α_{0j}.
2. We simulate the failure times, T_i, with respect to the all-cause hazard function $\alpha_{0\cdot}(t) = \sum_{j=1}^{J} \alpha_{0j}(t)$.
3. We run a multinomial trial with the probability of failure on event type j is $\frac{\alpha_{0j}(T)}{\sum_{j=1}^{J} \alpha_{0j}(T)}$.
4. We generate the right censoring times, C_i.

One thing worth mentioning is that in Step 2, the commonly used method to generate survival time is the inversion method. Suppose we have the cumulative all-hazard function

$$A_{0\cdot}(t) = \int_0^t \alpha_{0\cdot}(u)du \tag{2.46}$$

which is a monotone-increasing function. the CDF of T is also a monotone-increasing function:

$$F(t) = \mathbb{P}(T \leq t) = 1 - exp(-A_{0\cdot}(t)) \tag{2.47}$$

We assume both $F(\cdot)$ and $A_{0\cdot}(\cdot)$ are invertible with inverse functions $F(\cdot)^{-1}$ and $A_{0\cdot}(\cdot)^{-1}$.

Then we can generate $F(T)$ from a uniform distribution $[0, 1]$. Given the relationship, that

$$\mathbb{P}(F(T) \leq u) = \mathbb{P}(T \leq F^{-1}(u)) = F(F^{-1}(u)) = u \tag{2.48}$$

where $u \in [0, 1]$. Therefore, if U is generated from a uniform distribution $[0, 1]$, then $F^{-1}(U)$ has the same distribution as T. We can further generate T regarding the cumulative hazard function because

$$F^{-1}(u) = A_{0\cdot}^{-1}(-log(1-u)) \tag{2.49}$$

where $u \in [0, 1]$. This inversion can be computed using statistical software, for example, by using the function `uniroot()` in `R`. We use this method in step 3 to generate time-event data.

As we can see above, we need an invertible function regarding the cumulative hazard functions to generate survival time. This poses a limitation on how we can use the information we extract from a pilot study. We propose fitting parametric survival models to the pilot data because the maximum likelihood optimization justifies the goodness of the parameter selection under certain distributional assumptions. We also fit the non-parametric Kaplan-Meier survival curve to select the most suitable parametric hazard function, further to generate time-event data for sample size or power computation. Before we propose our method for evaluating these parametric models regarding the nonparametric survival curve, we plan to introduce flexible parametric survival models that better approximate the nonparametric model than commonly used parametric survival models.

2.5 Sample Size Computation For Competing Risk Data

We now detail our approach for determining requisite sample sizes for testing differences in survival times using data from a pilot study with competing risks. The steps of our approach are as follows:

1. Use a generalized log-gamma model to extract information about each cause-specific hazard.

2. Investigate differences in survival times for each cause to gain preliminary insights on the effect of treatment.

3. Use the simulation method discussed in Section 2.4 to repeatedly generate data from our fitted models using different sample sizes.

4. For a given sample size, generate many hypothetical "large experiment" datasets. For each dataset, perform the desired hypothesis test, and estimate the power of the test by computing fraction of tests that reject the null hypothesis.

5. The input sample size that is closest to our pre-specified power is the one we recommend for the official experiment.

One benefit of our approach is that we do not require stringent assumptions on each of the cause-specific survival times, for example, proportional or constant hazards. While this prevents us from explicitly stating the specific alternative hypothesis for which we are computing power—for example, a statement about the hazard ratio—we believe that these alternative hypotheses are not very helpful when stringent assumptions required to test for these alternatives do not hold. Instead, we can gain intuition on the true effect of treatment, without stringent assumptions, by performing an exploratory data analysis to investigate differences in these fitted models across treatment groups.

2.6 Simulation

We conducted six simulation studies to compare our method with other methods. For each simulation study, we begin by defining the population distributions of the cause-specific hazards. We assume two treatment levels: control and treatment. The treatment groups are balanced, meaning that half of the subjects are assigned to the treatment group and

the other half are assigned to the control group. Afterward, we evaluate the power of the specified test by simulating data from the true distributions. This allows us to ensure that the effect size leads to a steadily increasing power as the sample size grows. Ultimately, we expect the power to reach 0.8 within 1000 units. We refer to this estimated power as the actual power.. To estimate the actual power, we generate data from the true distribution following the steps in section 2.4 with a pre-specified sample size, then we run the specified statistical test on this dataset. We repeated this process 500 times to estimate the actual power by dividing the number of significances by the total number of simulations. After computing the actual power, we generated 400 units from the true distributions, 200 units for each treatment group, as our "pilot" study. Then we fit the generalized log-gamma model to extract information on each cause-specific hazard. After obtaining the estimated $\hat{q}, \hat{\mu}$, and $\hat{\sigma}$ for the generalized log-gamma model, we can compute the $\hat{H}(t) = -log(\hat{S}(t))$ and $h(t) = \frac{\hat{f}(t)}{\hat{S}(t)}$ by the estimated pdf and survival functions according to section 2.2.2. We, then, follow the data generation steps from section 2.4 to compute the power of the specified test regarding a list of sample sizes (The number of subjects increases from 100 to 1,000, with each increase consisting of 50 additional subjects). We still set the treatment assignment to be balanced. For the comparison methods, we input the estimated actual power levels along with the information required by the formula from the "pilot" study to determine the corresponding sample size.

There are two comparison methods involved in our simulations. One is proposed by Latouche and Porcher (2007). Their method estimates the sample size based on one primary interest event. Their computation method is based on the supremum weight log-rank test, with the null hypothesis $H_0 : \alpha_{01,1} = \alpha_{01,0}$. Here we continue to use the notation we used in section 2.3 that α_{01} is the cause-specific hazard for transition $0 \to 1$, $\alpha_{01,1}$ indicates the cause-specific hazard for transition $0 \to 1$ of the treatment group, and $\alpha_{01,0}$ indicates the cause-specific hazard for transition $0 \to 1$ of the control group. Their computation method is built on *Schoenfeld's* formula and an additional adjustment for nonproportionality.

The *Schoenfeld's* formula is expressed as $\frac{(z(1-\frac{\alpha}{2})+z(1-\beta))^2}{(log(HR))^2\mathbb{P}_t(1-\mathbb{P}_t)\mathbb{P}_e}$ where α is the type I error, β is the type II error, $log(HR)$ is the log hazard ratio, $\mathbb{P}_t$ is the probability of treatment assignment, and $\mathbb{P}_e$ is the probability of the primary interest event. The *Schoenfeld's* is broadly used for computing sample size with the assumption of proportionality. For a non-proportional scenario, we time the output of *Schoenfeld's* with an additional adjustment for nonproportionality, $\mathcal{A}$, which does not depend on any weight of the test. The second method is proposed by Yang et al. (2018), which is based on the χ^2 joint test proposed by Li and Yang (2016) with the hypothesis that $H_0 : \alpha_{01,1} = \alpha_{01,0}$ and $\alpha_{0.,1} = \alpha_{0.,0}$. We use the *R* package *powerCompRisk* to compute the sample size for comparison. They assume consistent hazards and proportionality. We utilized Cox regression with the baseline defined as an exponential distribution to extract information from the pilot study to compute the sample size. We use the supremum weighted log-rank test and χ^2 joint test in our simulations without setting weights because both comparison methods compute sample size without considering the setting weight.

The first and second simulations assume the survival times follow exponential distributions; therefore, the assumptions of consistent hazards or proportionality are satisfied. Following the third and fourth simulations, we assume the survival times follow Weibull distributions; therefore, we can make comparisons among methods when the assumptions of consistent hazards or proportionality are violated. Last, we assume the survival times follow generalized log gamma distributions in our fifth and sixth simulations to compare different methods when the shapes of hazard functions are more complicated.

We conduct our first simulation using the setting from Latouche and Porcher (2007) for a better comparison and understanding of how our method works under the assumption of exponential distributions. They assume that the survival times follow exponential distributions, and the parameters are set as follows: $\alpha_{01,1} = 0.26$, $\alpha_{01,1} = 0.19$, and $\alpha_{02,0} = \alpha_{02,1} = 0.14$. In other words, the treatment only influences the type one events and decreases the hazard consistently for the treatment group over time. In other words, subjects in the control group

who transition to absorbing state one are more likely to fail. As the study continues, the probability of failure remains at the same level. The figures of hazard functions and survival functions are displayed in Figure 2.5.

This scenario could occur, for example, in an oncology study to compare the novel chemotherapy with a standard one. We may identify two events as failures: the occurrence of severe side effects and death. The 01 event 01 stands for developing severe side effects, and 02 stands for death. For the setting of the first simulation study, the novel treatment reduces severe side effects, but both treatments have the same death rates. By a ssuming the survival times follow exponential distributions, the rates of death and developing severe side effects are consistent over time. The outputs of the estimated actual power, our method, and the comparison method are summarized in figure 2.6 and table 2.1

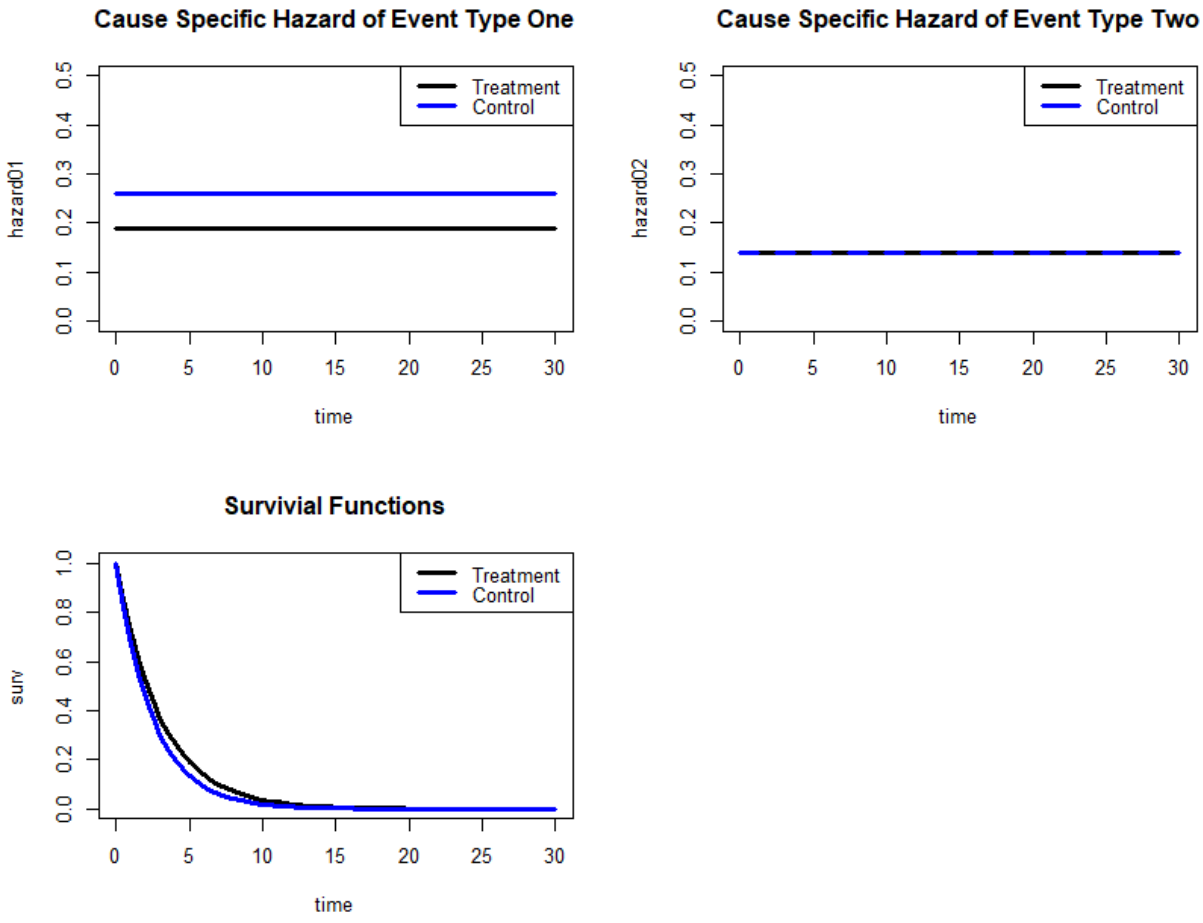

Figure 2.5: *Cause-specific hazard for simulation 1*

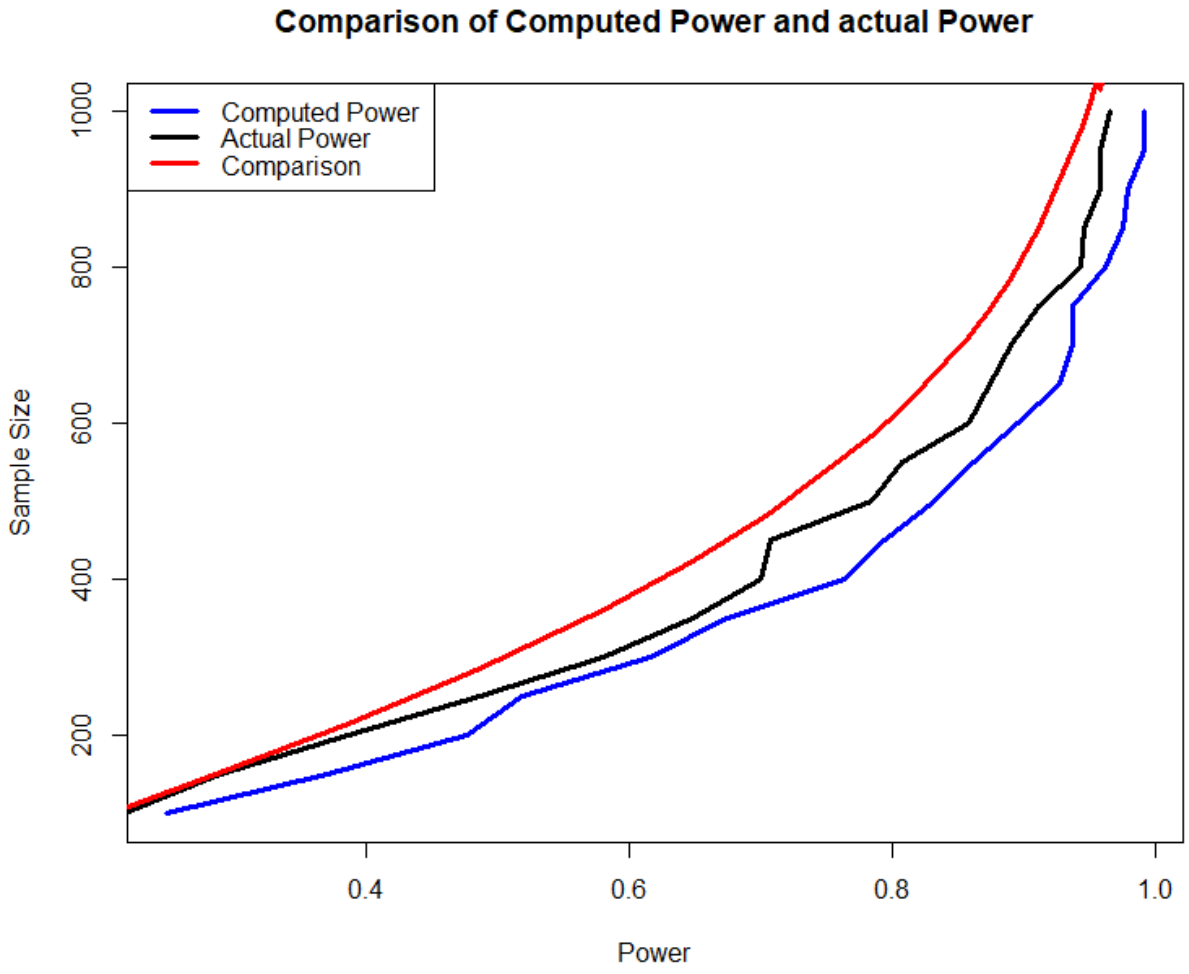

Figure 2.6: *Power comparison for simulation 1*

Summarization of Power and Sample Size of Simulation 1										
Computation	Sample size	100	150	200	250	300	350	400	450	500
	Power	0.248	0.370	0.476	0.518	0.616	0.674	0.764	0.794	0.832
Actual	Sample size	100	150	200	250	300	350	400	450	500
	Power	0.216	0.288	0.386	0.486	0.580	0.648	0.700	0.708	0.784
Comparison	Sample size	100	150	200	250	300	350	400	450	500
	Power	0.204	0.285	0.362	0.434	0.502	0.564	0.621	0.672	0.718
Computation	Sample size	550	600	650	700	750	800	850	900	950
	Power	0.862	0.896	0.928	0.938	0.938	0.962	0.976	0.980	0.992
Actual	Sample size	550	600	650	700	750	800	850	900	950
	Power	0.808	0.858	0.874	0.890	0.912	0.944	0.946	0.958	0.958
Comparison	Sample size	550	600	650	700	750	800	850	900	950
	Power	0.758	0.794	0.825	0.851	0.875	0.894	0.911	0.926	0.938

Table 2.1: *Summarization of power and sample size for simulation 1*

We conduct the second simulation to compare with the method proposed by Yang et al. (2018). We follow their setting that the cause-specific hazard ratios are equal and consistent. The survival times follow exponential distributions. However, they only reported the hazard ratios in their article. Therefore, we continuously use the setting of hazard functions for type one events in simulation one and change the hazard ratios of the second type of events, Therefore, we set both events have the same hazard ratio with $\alpha_{01,1} = 0.26$, $\alpha_{01,1} = 0.19$, $\alpha_{02,0} = 0.19$, and$\alpha_{02,1} = 0.14$. Thus, the treatment is effective for both types of events, reducing the risks associated with them consistently over time. In other words, subjects in the control group are more likely to make transitions to any absorbing state. The probability of failure remains the same for the whole course of study. The figures of hazard functions are displayed in figure 2.7. To illustrate the setting using the oncology example, the novel treatment not only decreases the death rate but also decreases the probability of developing severe side effects, and the rates of death and developing severe side effects are consistent over time. The outputs of the estimated actual power, our method, and the comparison method are summarized in figure 2.8 and table 2.2

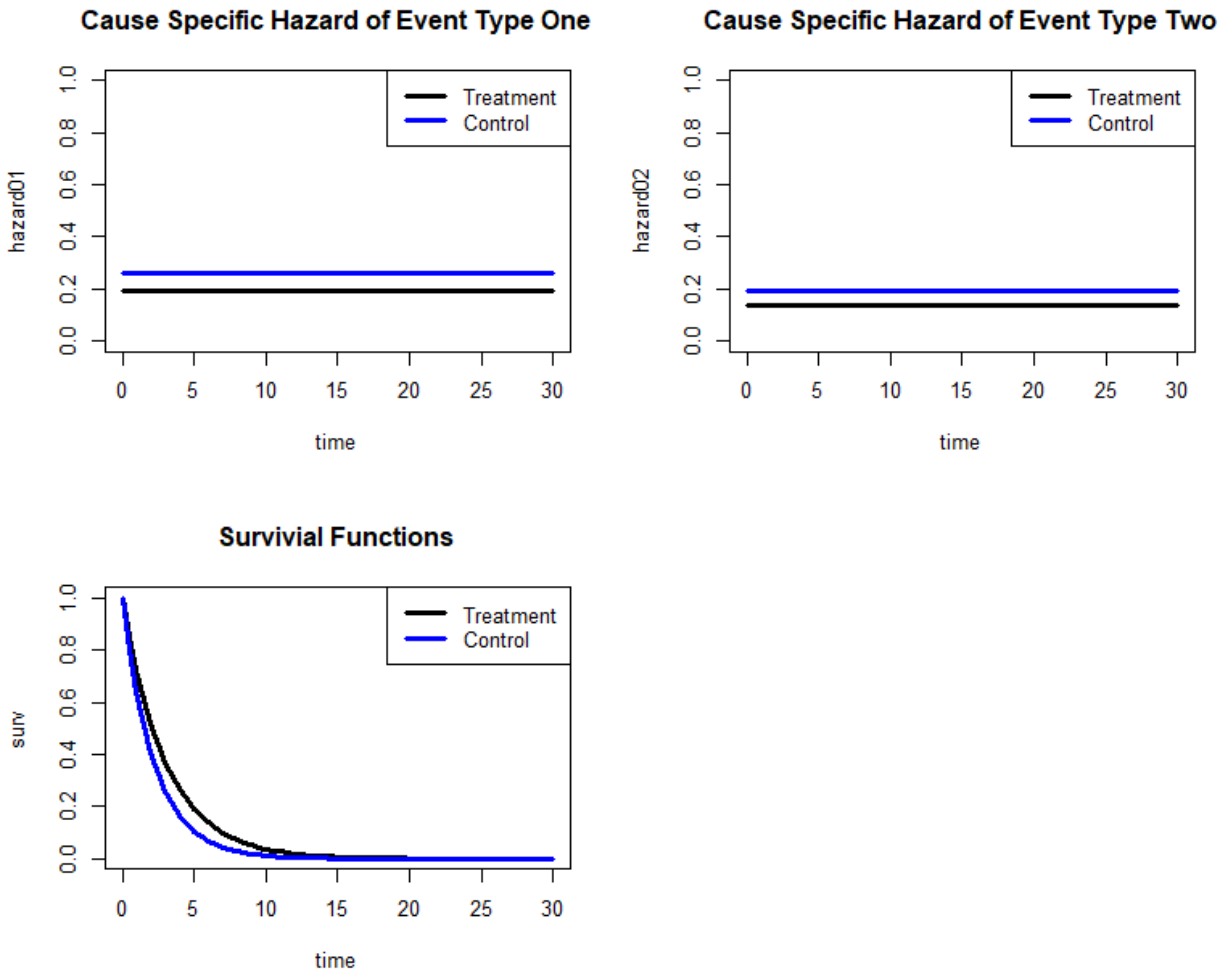

Figure 2.7: *Cause-specific hazard for simulation 2*

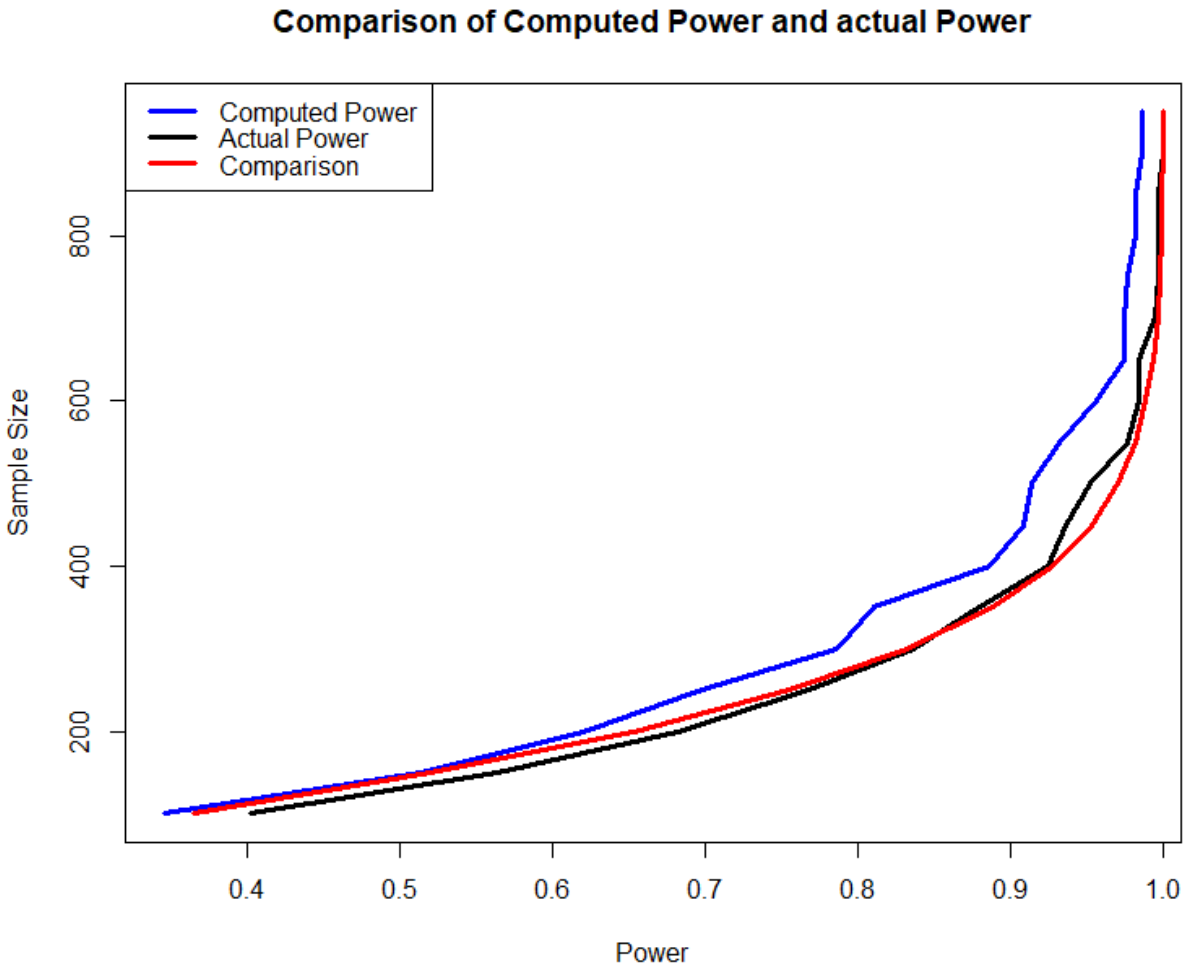

Figure 2.8: *Power comparison for simulation 2*

Summarization of Power and Sample Size of Simulation 2										
Computation	Sample size	100	150	200	250	300	350	400	450	500
	Power	0.346	0.514	0.620	0.698	0.786	0.810	0.886	0.908	0.914
Actual	Sample size	100	150	200	250	300	350	400	450	500
	Power	0.402	0.564	0.682	0.768	0.836	0.880	0.924	0.936	0.952
Comparison	Sample size	100	150	200	250	300	350	400	450	500
	Power	0.348	0.497	0.629	0.732	0.811	0.870	0.913	0.942	0.962
Computation	Sample size	550	600	650	700	750	800	850	900	950
	Power	0.932	0.956	0.974	0.974	0.976	0.982	0.982	0.986	0.986
Actual	Sample size	550	600	650	700	750	800	850	900	950
	Power	0.976	0.984	0.984	0.994	0.996	0.996	0.996	1.000	1.000
Comparison	Sample size	550	600	650	700	750	800	850	900	950
	Power	0.976	0.985	0.990	0.994	0.996	0.998	0.9986	0.9992	0.9995

Table 2.2: *Summarization of power and sample size of simulation 2*

The third simulation aims to compare our method to Latouche and Porcher (2007) without assuming proportionality. The simulated data is generated using the Weibull distributions. The hazard function of the Weibull distribution is given by $h(t) = p\lambda^p t^{p-1}$, where p is a shape parameter and λ is a scale parameter. We aim to create a scenario in which the treatment effect is only effective for the first type of event and results in a reduced hazard for the treatment group. Therefore, we set $p_{01,0} = 1.4$ and $\lambda_{01,0} = 0.17$ for $\alpha_{01,0}$. $p_{01,1} = 1.2$ and $\lambda_{01,0} = 0.12$ for $\alpha_{01,1}$. For the hazards of the type two events, we set $p_{02,0} = p_{02,1} = 0.95$ and $\lambda_{02,0} = \lambda_{02,1} = 0.175$ for $\alpha_{02,0} = \alpha_{02,1}$. Subjects in the control group and those transitioning to type one events are more likely to fail. As the study continues, the longer these subjects survive, the more likely they are to transition to the absorbing state one. The population hazard functions are presented in 2.9. To illustrate the setting using the oncology example, the novel treatment only reduces the rate of developing severe side effects, and both treatments influence the death rate in the same way. For the subjects under the standard treatment, the probability of developing side effects increases over time. It escalates at the beginning of the study, and the increasing rate gradually reduces throughout the study. The subjects receiving the novel treatment have a reduced likelihood of experiencing side effects, and this rate remains more consistent over the course of the study. The result is summarized in the figure 2.10 and the table 2.3.

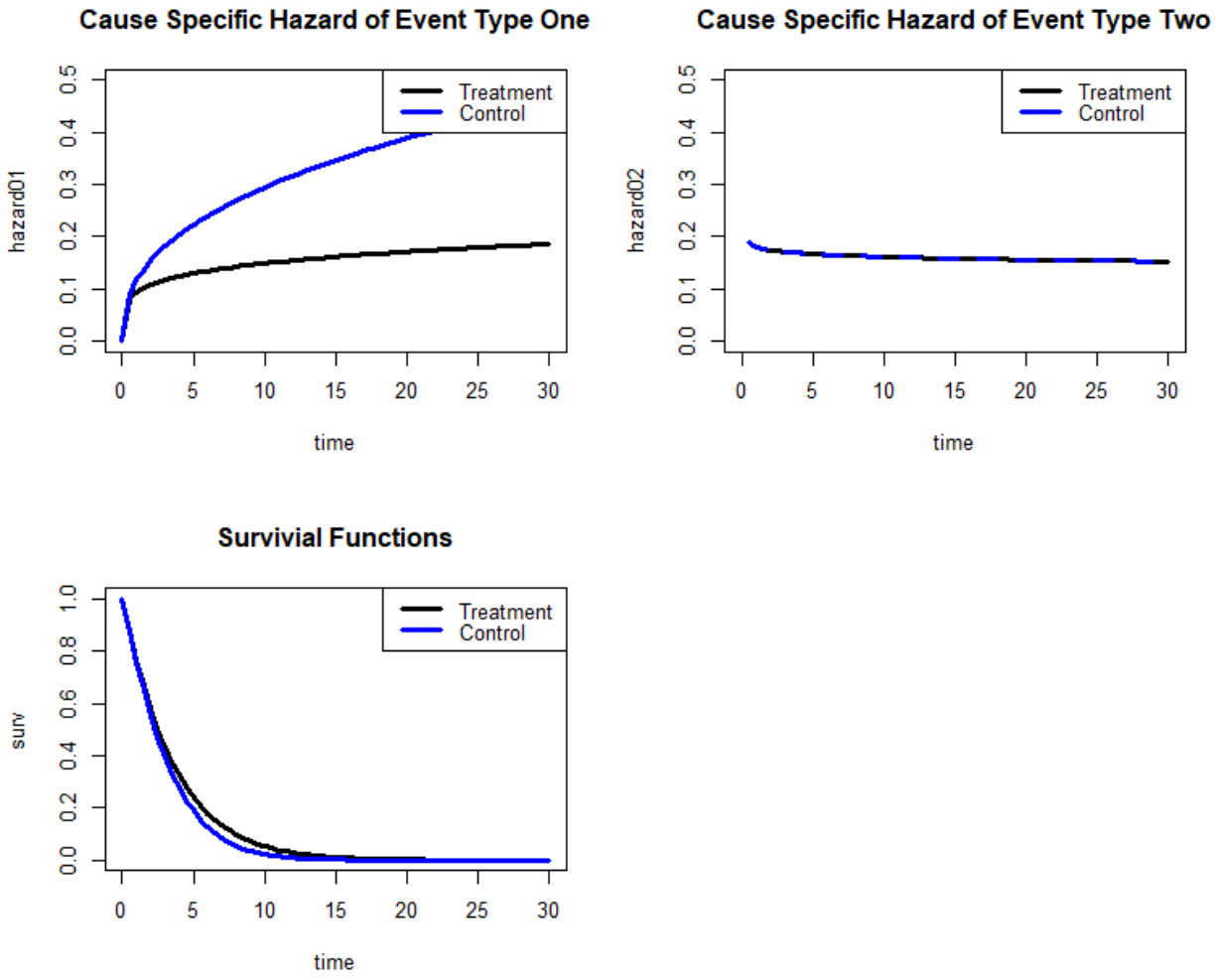

Figure 2.9: *Cause-specific hazard for simulation 3*

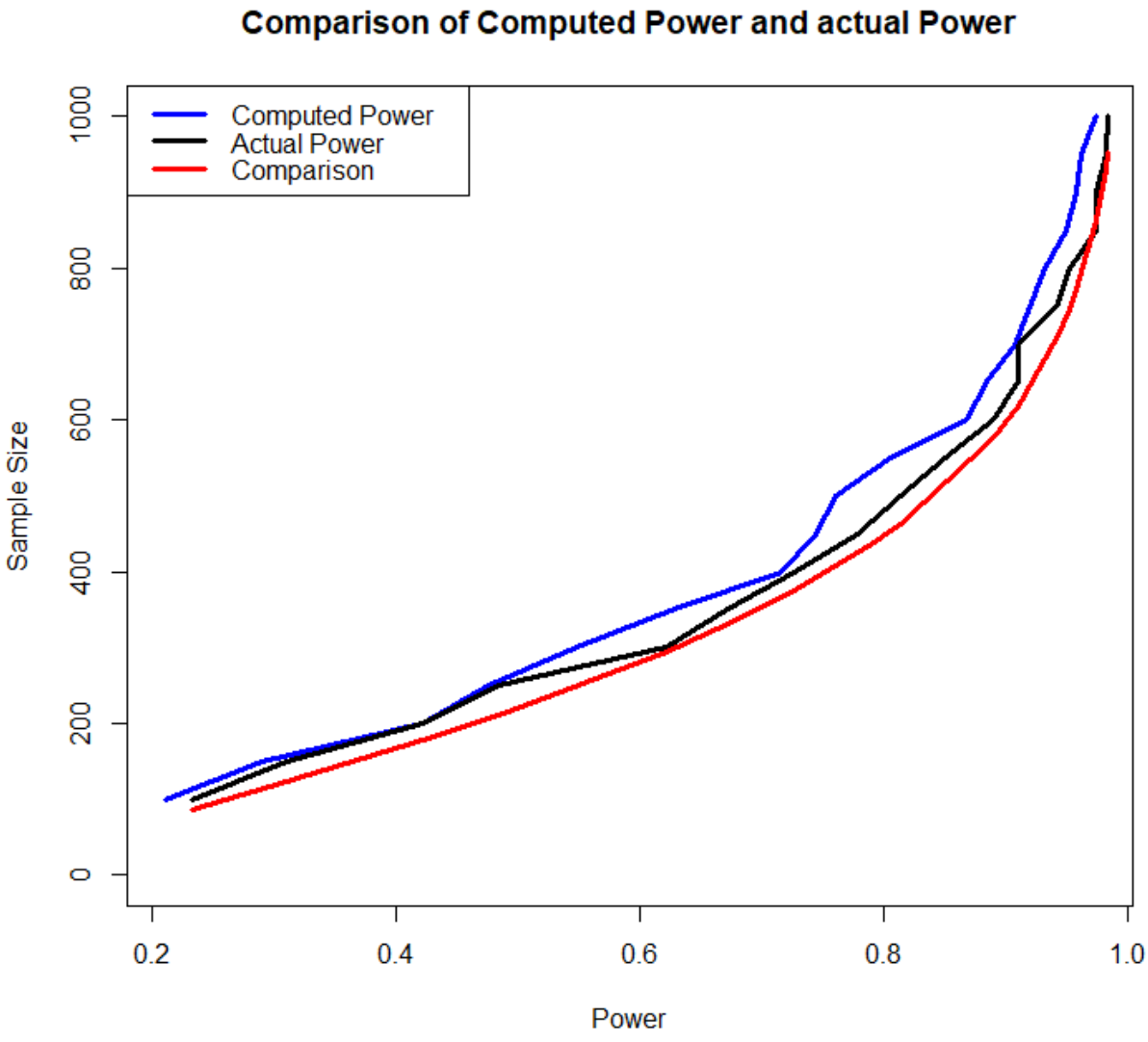

Figure 2.10: *Power comparison for simulation 3*

Summarization of Power and Sample Size of Simulation 3										
Computation	Sample size	100	150	200	250	300	350	400	450	500
	Power	0.210	0.290	0.422	0.476	0.548	0.628	0.716	0.744	0.760
Actual	Sample size	100	150	200	250	300	350	400	450	500
	Power	0.232	0.310	0.422	0.484	0.622	0.672	0.726	0.778	0.814
Comparison	Sample size	100	150	200	250	300	350	400	450	500
	Power	0.260	0.365	0.463	0.552	0.629	0.697	0.755	0.803	0.842
Computation	Sample size	550	600	650	700	750	800	850	900	950
	Power	0.806	0.868	0.884	0.908	0.920	0.932	0.950	0.958	0.962
Actual	Sample size	550	600	650	700	750	800	850	900	950
	Power	0.850	0.890	0.910	0.910	0.942	0.952	0.974	0.974	0.982
Comparison	Sample size	550	600	650	700	750	800	850	900	950
	Power	0.874	0.901	0.923	0.940	0.953	0.964	0.972	0.979	0.984

Table 2.3: *Summarization of power and sample size of simulation 3*

The fourth simulation is designed to compare our method with Yang et al. (2018). We use a setting similar to the second simulation. We have the same setting on event type one as we did for simulation three. we set $\alpha_{02,0} = 0.95 \cdot 0.175^{0.95} \cdot t^{0.95-1}$ and $\alpha_{02,1} = 0.75 \cdot 0.125^{0.75} \cdot t^{0.75-1}$. The setting is present in figure 2.11. To illustrate this setting using the oncology example, the novel treatment decreases both the rates of developing severe side effects and death. The patterns of developing severe side effects are the same as those in the third simulation. For the subjects under the standard treatment, the death rate remains almost consistent over time. On the contrary, the subjects receiving the novel treatment exhibit lower death rates, with a significant decline observed at the beginning of the study. The decline in the rate continues, but decelerates throughout the study. The comparison is presented in figure 2.12 and table 2.4.

.

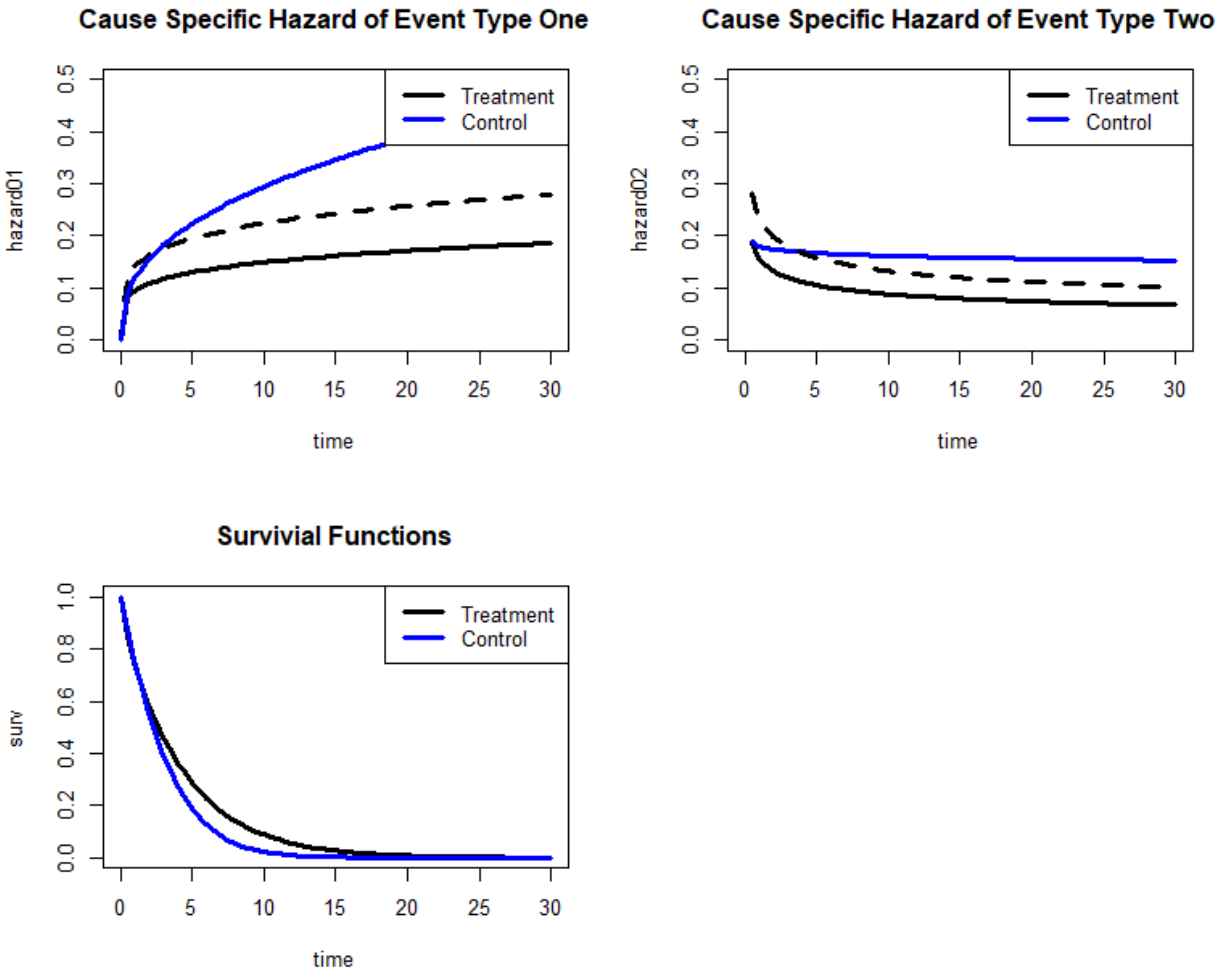

Figure 2.11: *Cause-specific hazard for simulation 4*

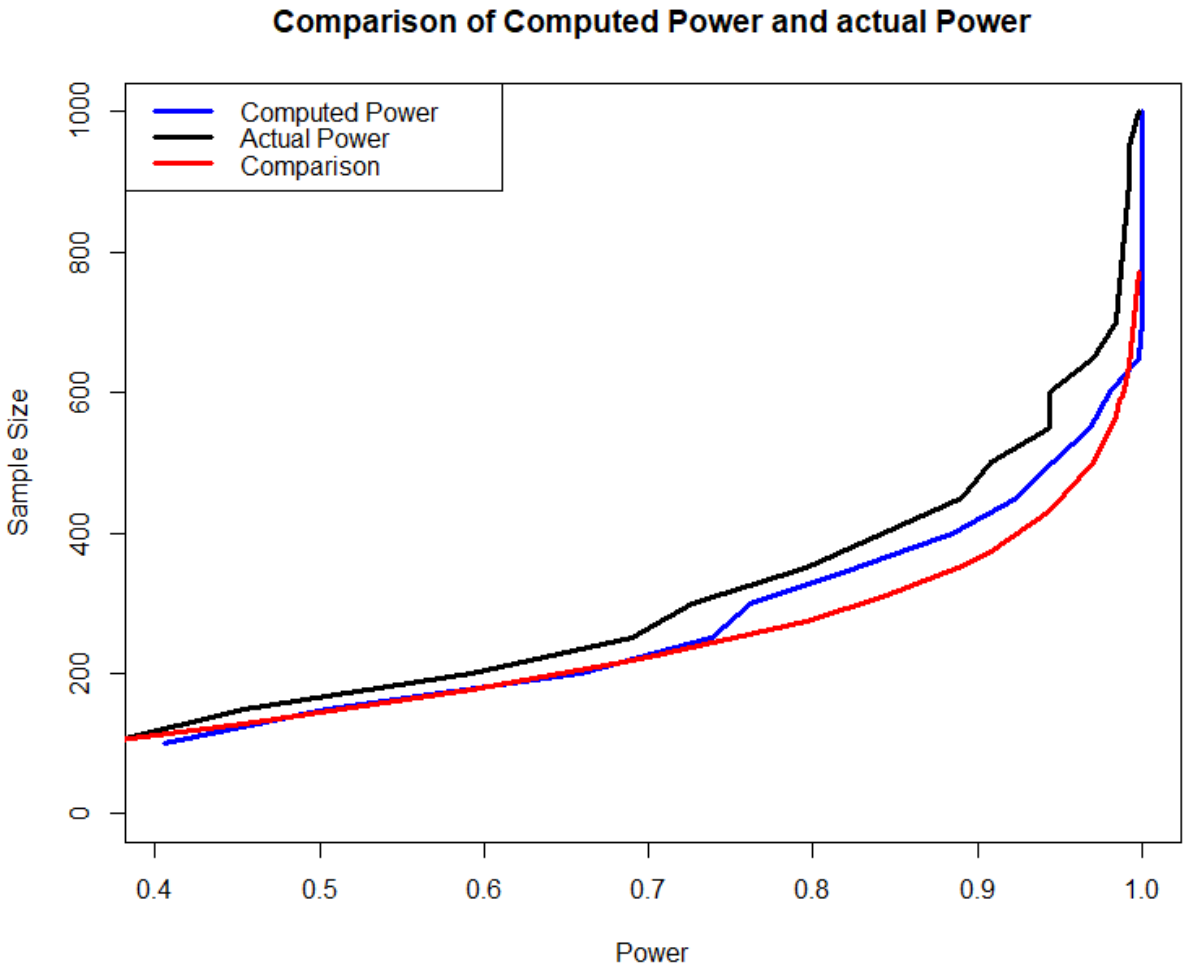

Figure 2.12: *Power comparison for simulation 4*

Summarization of Power and Sample Size of Simulation 4										
Computation	Sample size	100	150	200	250	300	350	400	450	500
	Power	0.406	0.502	0.658	0.738	0.762	0.826	0.884	0.924	0.946
Actual	Sample size	100	150	200	250	300	350	400	450	500
	Power	0.370	0.454	0.592	0.690	0.726	0.796	0.842	0.890	0.908
Comparison	Sample size	100	150	200	250	300	350	400	450	500
	Power	0.365	0.520	0.653	0.755	0.832	0.888	0.927	0.953	0.970
Computation	Sample size	550	600	650	700	750	800	850	900	950
	Power	0.946	0.968	0.980	0.998	1.000	1.000	1.000	1.000	1.000
Actual	Sample size	550	600	650	700	750	800	850	900	950
	Power	0.944	0.944	0.970	0.984	0.986	0.988	0.990	0.992	0.992
Comparison	Sample size	550	600	650	700	750	800	850	900	950
	Power	0.982	0.988	0.993	0.996	0.9975	0.9985	0.9991	0.9995	0.9997

Table 2.4: *Summarization of power and sample size of simulation 4*

For the fifth and sixth simulations, we aim to compare our method with the comparison methods when the shapes of the hazards are more complex and the assumptions of consistent hazards or proportionality are violated. We use generalized log-gamma distributions as our true distributions. The hazard functions for both events under treatment and control effects are presented in figure 2.13. For $\alpha_{01,0}$, we have $q_{01,0} = -1$, $\sigma_{01,0} = 0.5$ and $\mu_{01,0} = 1.35$. For $\alpha_{01,1}$, we have $q_{01,1} = -1.5$, $\sigma_{01,1} = 0.5$ and $\mu_{01,1} = 1.5$. We have $\alpha_{02,0} = \alpha_{02,1}$ with $q_{02,0} = q_{02,1} = -1.5$, $\sigma_{02,0} = \sigma_{02,1} = 0.5$ and $\mu_{02,0} = \mu_{02,1} = 1.3$. As we can see, the hazard functions are no longer monotone increasing or decreasing like the Weibull cases. The treatment only influences the type one event, and decreases the hazard. Specifically, the majority of subjects who transition to type one event in the control group are more likely to have shorter survival times than the treatment group, especially at the early stage of the study. However, in due course, subjects who survive tend to be more likely to live, and even subjects from the control group are still more likely to fail. To illustrate this setting using the oncology example, the novel treatment only reduces the probability of developing severe side effects. For the subjects under the standard treatment, the probability of developing side effects escalates at the beginning of the study. However, the rate decreases once it reaches a climax, and the decreasing rate gradually stagnates. The subjects receiving the novel treatment have a reduced likelihood of experiencing side effects, especially at the beginning of the study. The rate of developing side effects changes over time in a similar pattern to subjects under the standard treatment. The results of sample size computation and power estimation are present in figure 2.14 and table 2.5.

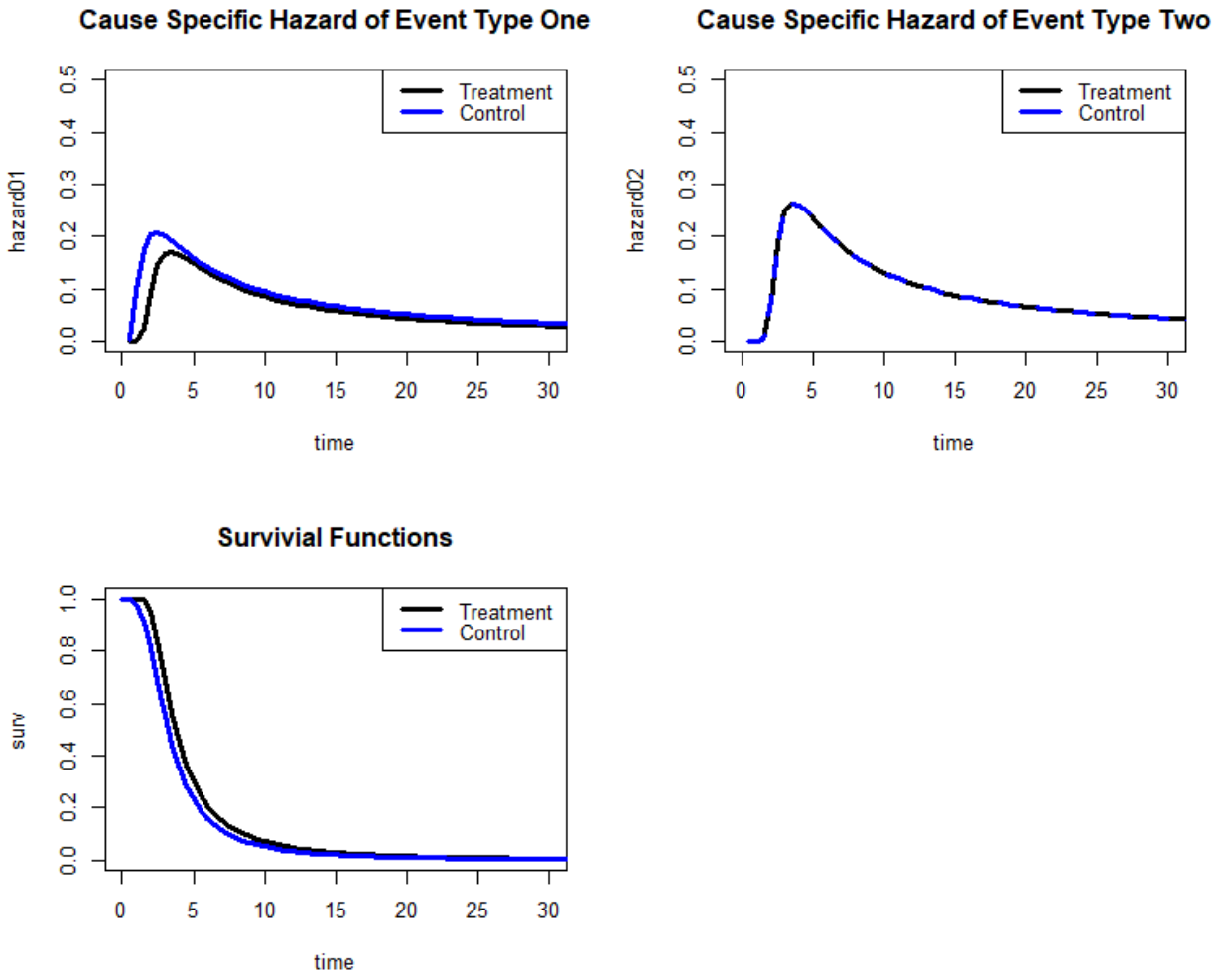

Figure 2.13: *Cause-specific hazard for simulation 5*

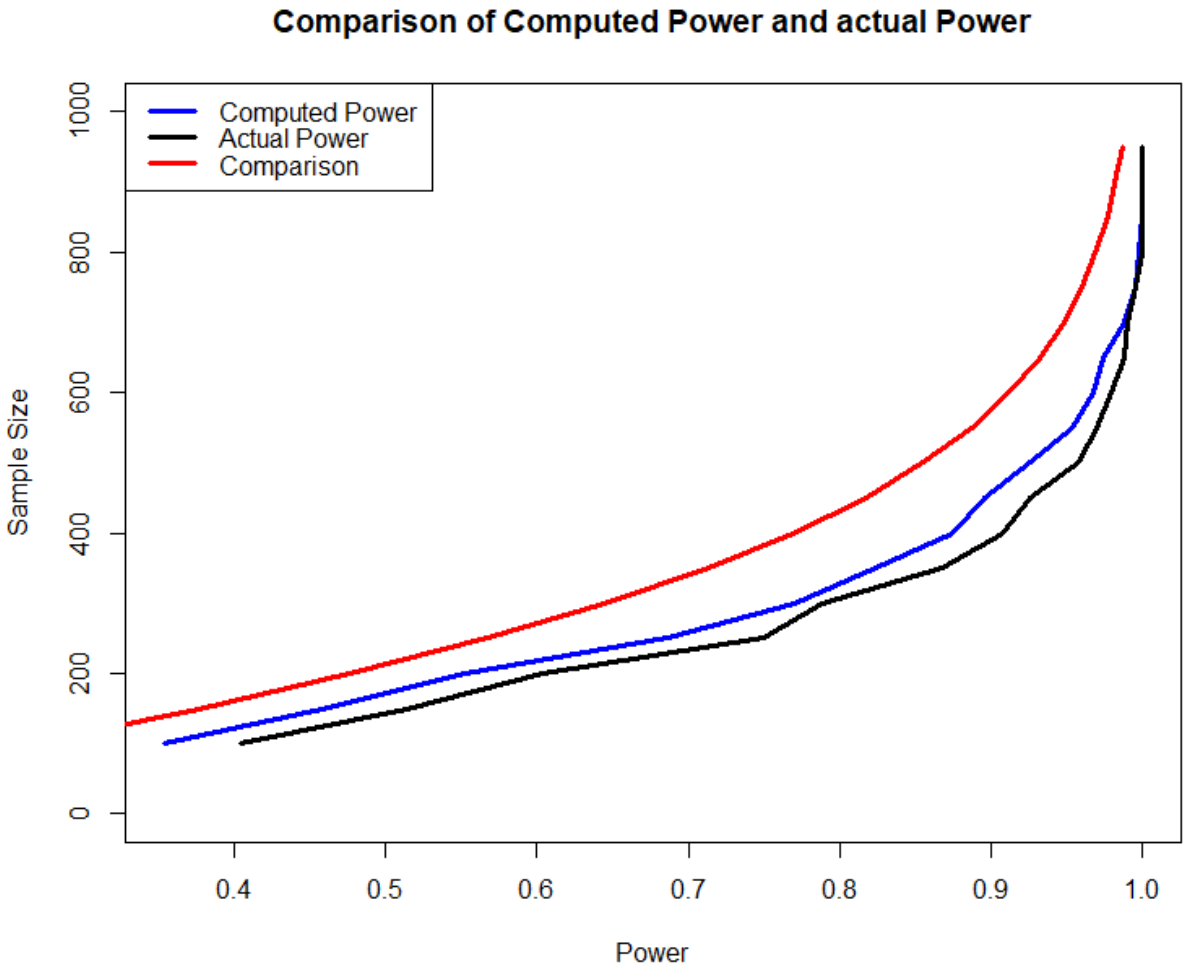

Figure 2.14: *Power comparison for simulation 5*

Summarization of Power and Sample Size of Simulation 5										
Computation	Sample size	100	150	200	250	300	350	400	450	500
	Power	0.354	0.458	0.522	0.686	0.770	0.824	0.874	0.896	0.926
Actual	Sample size	100	150	200	250	300	350	400	450	500
	Power	0.404	0.514	0.604	0.750	0.788	0.868	0.908	0.926	0.958
Comparison	Sample size	100	150	200	250	300	350	400	450	500
	Power	0.267	0.376	0.476	0.567	0.645	0.713	0.769	0.817	0.855
Computation	Sample size	550	600	650	700	750	800	850	900	950
	Power	0.954	0.968	0.974	0.988	0.996	0.998	1.000	1.000	1.000
Actual	Sample size	550	600	650	700	750	800	850	900	950
	Power	0.970	0.980	0.988	0.990	0.996	1.000	1.000	1.000	1.000
Comparison	Sample size	550	600	650	700	750	800	850	900	950
	Power	0.887	0.912	0.932	0.948	0.960	0.969	0.977	0.982	0.9866

Table 2.5: *Summarization of power and sample size of simulation 5*

We continue to use a similar setting to that used in the second and fourth simulations. The treatment reduces the risks associated with both type one and type two events. All hazards are in arch shapes to ensure that the assumption of consistent hazard and proportionality is violated considerably. For $\alpha_{01,0}$, we set $q_{01,0} = -1$, $\sigma_{01,0} = 0.5$, and $\mu_{01,0} = 1.42$. For $\alpha_{01,1}$, we set $q_{01,1} = -1.5$, $\sigma_{01,1} = 0.5$, and $\mu_{01,1} = 1.5$. For $\alpha_{02,0}$, we set $q_{02,0} = -1.5$, $\sigma_{02,0} = 0.5$, and $\mu_{02,0} = 1.42$. For $\alpha_{02,1}$, we set $q_{02,1} = -2$, $\sigma_{02,1} = 0.5$, and $\mu_{02,1} = 1.5$. In other words, the majority of subjects who transition to both types of events in the control group are more likely to have shorter survival times than the treatment group, especially at the early stage of the study. Even in the latter course of the study, subjects tend to be more likely to survive, but the subjects from the control group are still more likely to fail. The figure of hazards and survival functions of the true population is presented in figure 2.15. In the framework of the oncology example, the novel treatment not only reduces the probability of developing severe side effects but also the death rate. For the subjects under the standard treatment, the probability of developing side effects escalates at the beginning of the study. However, the rate decreases once it reaches a climax, and the decrease gradually stagnates. The subjects receiving the novel treatment have a reduced likelihood of experiencing side effects, especially after the rate reaches the climax. The rate of developing side effects changes over time in a similar pattern to subjects under the standard treatment. The patterns of death rates resemble the likelihood of developing side effects. The summaries of sample size computation and estimated powers are presented in figure 2.16 and table 2.6

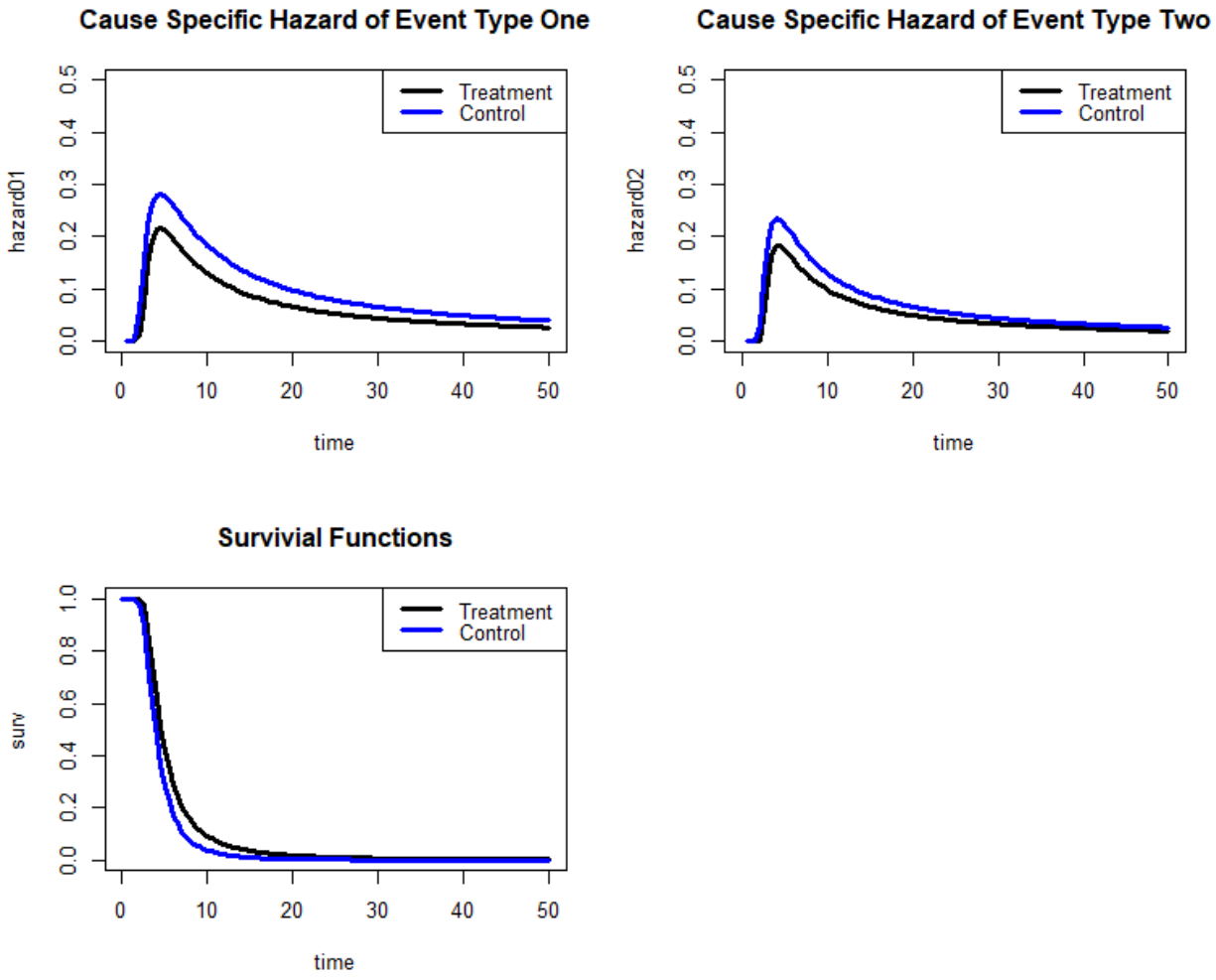

Figure 2.15: *Cause-specific hazard for simulation 6*

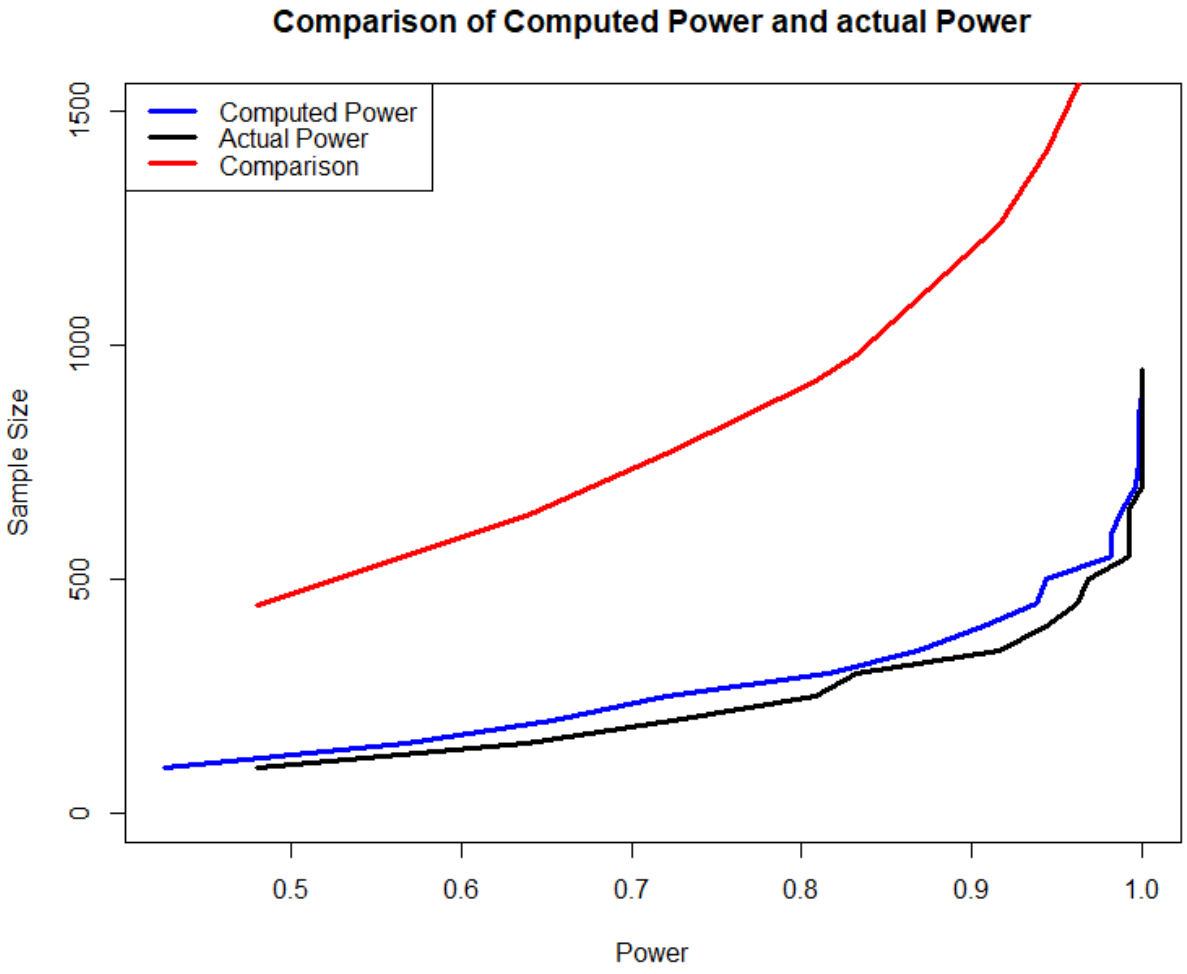

Figure 2.16: *Power comparison for simulation 6*

Summarization of Power and Sample Size of Simulation 6										
Computation	Sample size	100	150	200	250	300	350	400	450	500
	Power	0.426	0.568	0.654	0.720	0.816	0.868	0.906	0.938	0.944
Actual	Sample size	100	150	200	250	300	350	400	450	500
	Power	0.480	0.638	0.724	0.808	0.832	0.916	0.944	0.962	0.968
Comparison	Sample size	100	150	200	250	300	350	400	450	500
	Power	0.136	0.185	0.236	0.287	0.339	0.387	0.436	0.483	0.528
Computation	Sample size	550	600	650	700	750	800	850	900	950
	Power	0.982	0.976	0.988	0.996	0.998	0.994	0.996	1.000	1.000
Actual	Sample size	550	600	650	700	750	800	850	900	950
	Power	0.992	0.990	0.992	1.000	0.998	0.998	1.000	1.000	1.000
Comparison	Sample size	550	600	650	700	750	800	850	900	950
	Power	0.571	0.611	0.648	0.683	0.714	0.744	0.771	0.796	0.8185

Table 2.6: *Summarization of power and sample size of simulation 6*

As we can see, both our method and the comparison methods yield similar results, with the estimated sample sizes closely aligning with the estimated actual sample size when the consistent hazards and proportionality assumptions are satisfied. Comparison methods relying on proportionality or consistent hazards even perform acceptably when the true hazards are in comparably simple shapes (monotonously increasing or decreasing) without being consistent or proportional. However, when the hazards are in arch shapes and the treatment flattens the hazards, assuming consistent hazards or proportionality leads to underestimating the effect sizes. This, in turn, is more likely to result in an overestimate of the required sample size.

2.7 Conclusion

Today, we can achieve more precise and specific measurements, providing a deeper understanding of the study subjects. More sophisticated statistical methods are required to handle this data. Competing risk models are one of them. How to model competing risk data is a well-studied topic. However, more studies are needed on the sample size/power computation of this method to guide statisticians and research scientists in planning their studies after the pilot. Therefore, we propose our method of providing a scientific reference for sample size. We use flexible parametric survival models to approximate the non-parametric cause-specific survival curves to extract useful information from the pilot study. We further generate competing risk data based on the best approximation of parametric models. These generated data can be used for further model fitting and test computation for reference on sample size. Our method does not depend on the assumptions of proportionality or consistent hazards. In scenarios where the hazards are in arch shapes and the treatment flattens the arches, our method outperforms methods that rely on assumptions of proportionality or consistent hazards. We are planning to extend our method to more complex multistate competing risk data. However, there are limitations to using this method. There are conditions we want

to meet in the pilot studies. We want each type of competing event to be well-represented in the pilot studies. A heavily uneven distribution of event types poses more challenges in fitting cause-specific models on minority events. In other words, if there are only a few cases of one of the competing events when we fit cause-specific models, the approximation for that typical event becomes less accurate.

Chapter 3

On the design of cluster-randomized trials with user-specified clusters

3.1 Introduction

A cluster randomized trial (CRT) is an experiment in which observational units—the units on which responses are collected and which comprise the population of interest—are clustered together into prior to randomizing treatment, and treatment is assigned to these clusters of units rather than the observational units themselves. The reasons for randomizing treatment across clusters rather than individual units are almost always logistical rather than to increase the power of statistical inferences. For many CRTs, it may be extremely costly, logistically challenging, infeasible, or even illegal, to adjust the design to accommodate randomization to individual units.

For example, in a study assessing the efficacy of a diet regimen on the weight gain of pigs, it may require substantial effort to allow pigs within the same enclosure to be assigned to different diets; instead, it may be much easier to perform an experiment in which all pigs within the same enclosure are assigned to the same diet. In this case, a CRT may be preferred by the experimenter—in this case, the experimental units would be the pens

housing the pigs, and the observational units would be the pigs themselves.

The logistical efficiency provided by a CRT comes with a cost—almost always, parameter estimates will be less precise and statistical inferences will be less powerful in a CRT than those obtained from, say, an experiment where treatment is assigned completely at random to individual units. This loss in precision can be quantified by the *intraclass correlation coefficient* (ICC)—a measure of how correlated responses are within each cluster. Larger values of ICC correspond to cases in which responses within each cluster are similar, which in turn reduces the precision of estimates and the power of statistical inferences with respect to completely randomized designs.

However, there are settings in which the experimenter will have the ability to specify clusters. For instance, an experimenter may be able to assign particular pigs to pens prior to randomizing treatment to pens. In which case, it may be possible to improve the design of the experiment through careful assignment of observational units to clusters.

To shed light on best practices for specifying clusters, we consider a simplified experimental setting in which there is two treatment categories—treatment and control—and exactly half of the units receive each treatment condition. We compare the precision of estimates for the average treatment effect (ATE) under two randomization schemes: Complete randomization of treatment to units and cluster randomization in which each cluster contains the same number of units.

We show, under the Neyman-Rubin potential outcome model of response, that cluster randomization will perform comparable to complete randomization—and, in special cases, will outperform complete randomization—when the variance of responses within each cluster is large relative to the variance across clusters. Thus, in the design of such a CRT, the experimenter should aim to create clusters so that the units within each cluster are as varied as possible, or equivalently, so that that the distribution of units within a cluster are the same across clusters. Note, this is essentially the opposite goal of statistical blocking, in which the experimenter aims to form groups of similar units prior to randomizing treatment

within each group.

We then develop a method utilizing threshold clustering (TC) (Higgins et al., 2016; Luo et al., 2019) for forming clusters with large within-cluster variability. We confirm through a simulation study that our approach provides significant reduction in the variance of ATE estimators relative to the case where units within each cluster are selected to be similar—which is often done in practice. However, in many cases, our TC approach does not perform significantly better than simply completely randomizing units to clusters.

3.2 Cluster Randomized Trials

Suppose there are N clusters, numbered 1 to N, with M units assigned to each cluster, numbered 1 to M. Each unit is assigned to one of two treatment conditions–treatment and control. Let T_{ij} denote a treatment indicator for the jth unit in cluster i; $T_{ij} = 1$ if this unit receives the treatment condition and $T_{ij} = 0$ otherwise. Let y_{ij} denote the observed response of the jth unit in cluster i.

We assume that these responses follow the Neyman-Rubin model of response (Holland, 1986; Imbens and Rubin, 2015; Rubin, 1974; Splawa-Neyman et al., 1990):

$$y_{ij} = y_{1ij}T_{ij} + y_{0ij}(1 - T_{ij}). \tag{3.1}$$

The y_{tij} terms are *potential outcomes*–the hypothetical response of the jth unit in cluster i if that unit receive treatment condition t. Potential outcomes y_{tij} are assumed to be non-random and only observed if the corresponding unit receives treatment condition t. Thus, under this model of response, the randomness in response is solely attributable to the randomness of the treatment assignment.

We assume that treatment assignment is balanced–that is, $MN/2$ units receive each treatment condition. Under cluster randomization, this can be ensured by sampling $n = N/2$ clusters and assigning all units belonging to these sampled clusters to the treatment condition;

all other units receive the control condition. Under complete randomization, this is equivalent to sampling Mn units completely at random–irrespective of the cluster assignment–and assigning these units to treatment; all other units are assigned to the control condition. Hence, many of our calculations follow straightforwardly from sampling theory (Lohr, 2021).

We are interested in estimating the average treatment effect (ATE)

$$\begin{aligned}\Delta &= \frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{1ij} - y_{0ij}) = \frac{1}{MN}\sum_{i}^{N}\sum_{j}^{M} y_{1ij} - \frac{1}{MN}\sum_{i}^{N}\sum_{j}^{M} y_{0ij} \\ &= \bar{y}_{1U} - \bar{y}_{0U} \end{aligned} \tag{3.2}$$

where

$$\bar{y}_{tU} = \frac{1}{MN}\sum_{i}^{N}\sum_{j}^{M} y_{tij}. \tag{3.3}$$

denotes the population mean under treatment condition t. We compare the variance of estimates of the ATE under complete randomization and cluster randomization. As usual, we estimate the ATE by estimating these population means separately and taking the difference of these estimates.

3.2.1 Horvitz-Thompson Estimation of the Population Mean

Horvitz-Thompson (HT) estimation (Horvitz and Thompson, 1952) provides a pathway for unbiased estimation of a population total regardless of the randomization scheme used to assign treatment.

The population total under the treatment condition t is

$$t_t = \sum_{i}^{N}\sum_{j}^{M} y_{tij}. \tag{3.4}$$

The population mean under treatment condition t is obtained by dividing the total by the

number of units under study:

$$\bar{y}_t = \frac{1}{MN} t_t \tag{3.5}$$

The HT estimator provides an unbiased estimate for t_t by dividing the response of each treatment given t by the probability that the unit received treatment. In particular, given the treatment assignment probabilities $\pi_{1ij} = P(T_{ij} = 1)$, the HT estimator for t_1 is:

$$\hat{t}_{1,HT} = \sum_{i}^{N} \sum_{j}^{M} \frac{y_{1ij} T_{ij}}{\pi_{1ij}} \tag{3.6}$$

and for t_0 is

$$\hat{t}_{0,HT} = \sum_{i}^{N} \sum_{j}^{M} \frac{y_{0ij}(1 - T_{ij})}{1 - \pi_{1ij}}. \tag{3.7}$$

An unbiased estimator of $\bar{y}_{tU}$ can then be obtained by dividing the corresponding estimate by MN, the number of experimental units in our study:

$$\hat{\mu}_t = \frac{1}{MN} \hat{t}_{t,HT} \tag{3.8}$$

Note, in all experimental settings we consider in this paper, $\pi_{1ij} = 1/2$ for all i, j. Hence, the estimator $\hat{\mu}_{t,HT}$ is equivalent to the sample mean of all units given treatment condition t.

While the form of the point estimates are the same, the two randomization schemes will lead to different variances between the two estimators. Following Lohr (2021), we now discuss differences the variances of $\hat{\mu}_{t,HT}$ under complete and cluster randomization. To differentiate between the two randomization schemes, we will let $\hat{\mu}_{t,CRE}$ denote the HT estimator under complete randomization and $\hat{\mu}_{t,CRT}$ denote the HT estimator under cluster randomization. Again, the following calculations are assuming that cluster sizes are equal.

It is a well known result for finite experiments that, under complete randomization of an

experiment with Mn out of MN units receiving treatment condition t,

$$Var(\hat{\mu}_{t,CRE}) = \frac{1}{Mn}\left(1 - \frac{Mn}{MN}\right) S_t^2 \tag{3.9}$$

where S_t^2 is the variance of the potential outcomes under treatment condition t:

$$S_t^2 = \frac{1}{MN-1}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(y_{tij} - \bar{y}_{tU}\right)^2. \tag{3.10}$$

The variance under cluster randomization can be derived similarly, except we treat the clusters as the individual units, and use the cluster totals as the responses. We can borrow the previous results if we begin by considering estimates of the average cluster total

$$\bar{t}_{tU} = \frac{1}{N}\sum_{i=1}^{N} t_{tiU} \tag{3.11}$$

where

$$t_{tiU} = \sum_{j=1}^{N} y_{tij}. \tag{3.12}$$

denotes the total of the potential outcomes within cluster i under treatment condition t. Since $\bar{y}_{tU} = (1/M)\bar{t}_{tU}$, an estimator for $\bar{y}_{tU}$ can then be obtained by taking the estimator for $\bar{t}_{tU}$ and dividing by M. Finally, to better make comparisons between the two randomization schemes, let

$$\bar{y}_{tiU} = \frac{1}{M} t_{tiU} = \frac{1}{M}\sum_{j=1}^{M} y_{tij} \tag{3.13}$$

denote the mean of the potential outcomes in cluster i under treatment condition t.

The HT estimator of $\bar{t}_{1U}$ is

$$\hat{\bar{t}}_{1,HT} = \frac{1}{N}\hat{t}_{1,HT} = \frac{1}{N}\sum_{i=1}^{N}\frac{T_i}{\pi_{1i}}\sum_{i=1}^{M} y_{1ij} \tag{3.14}$$

where π_{1i} and T_i are moved outside of the second sum to explicitly denote that, under cluster

randomization, the treatment assignment given to a unit only depends on the treatment assignment given to the cluster. For completeness, we can also define $\hat{\bar{t}}_{0,HT} = \frac{1}{N}\hat{t}_{0,HT}$.

Note that, under our current experimental setup, $\hat{t}_{t,HT}$ is simply the sample mean of the cluster totals under treatment condition t. Hence, we apply our previous result for simple random sampling, and obtain

$$Var(\hat{\bar{t}}_{t,HT}) = \frac{1}{n}\left(1 - \frac{n}{N}\right) S_{c,t}^2 \tag{3.15}$$

where

$$S_{c,t}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(t_{tiU} - \bar{t}_{tU})^2 = \frac{M^2}{N-1}\sum_{i=1}^{N}(\bar{y}_{tiU} - \bar{y}_{tU})^2. \tag{3.16}$$

Finally, we obtain

$$Var(\hat{\mu}_{t,CRT}) = \frac{1}{M^2}Var(\hat{\bar{t}}_{t,HT}) = \frac{1}{M^2 n}\left(1 - \frac{n}{N}\right) S_{c,t}^2 \tag{3.17}$$

We can interpret the terms S_t^2 and $S_{c,t}^2$ in terms of a sums-of-squares decomposition. Let

$$SSW_t = \sum_{i=1}^{N}\sum_{j=1}^{M}(\bar{y}_{tij} - \bar{y}_{tU})^2 \tag{3.18}$$

$$SSB_t = \sum_{i=1}^{N}\sum_{j=1}^{M}(\bar{y}_{tiU} - \bar{y}_{tU})^2 = M\sum_{i=1}^{N}(\bar{y}_{tiU} - \bar{y}_{tU})^2 \tag{3.19}$$

$$SSTO_t = \sum_{i=1}^{N}\sum_{j=1}^{M}(\bar{y}_{tijU} - \bar{y}_{tU})^2 \tag{3.20}$$

denote the within, between, and total sums of squares of the potential outcomes under treatment condition t respective. As usual, we have $SSTO_t = SSW_t + SSB_t$. Additionally,

we can express S_t^2 and $S_{c,t}^2$ in terms of these quantities. Note that

$$S_t^2 = \frac{1}{MN-1} SSTO_t = MSTO_t \tag{3.21}$$

$$S_{c,t}^2 = \frac{M}{N-1} SSB_t = M(MSB_t) \tag{3.22}$$

where MSB_t and $MSTO_t$ denote the mean squares between and mean squares total. We then can express the variance of $Var(\hat{\mu}_{t,CRT})$ as a function of MSB_t:

$$Var(\hat{\mu}_{t,CRT}) = \frac{1}{Mn}\left(1 - \frac{n}{N}\right) MSB_t. \tag{3.23}$$

Thus, for cluster randomized experiments, reduced variability in the cluster means across clusters leads to reduced variability in the sample mean. Additionally, since S_t^2 is a function of the sum-of-squares total–which remains the same regardless of the cluster formation–we confirm that the sample mean is unaffected by cluster formation in completely randomized experiments.

This idea can be made more precise via the introduction of the intra-cluster correlation (ICC), which is a measure of the homogeneity within clusters. The ICC is defined as the Pearson correlation coefficient across the $NM(M-1)$ pairs of potential outcomes within each cluster. For a given treatment t, this is

$$ICC_t = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{j\neq j'}(y_{tij} - \bar{y_{tU}})(y_{tij'} - \bar{y_{tU}})}{(MN-1)(M-1)S_t^2} \tag{3.24}$$

Given that $(MN-1)S^2 = SSTO$ and $SSTO = SSB + SSW$ the numerator of the ICC can

be re-written as:

$$\begin{aligned}
&\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{j\neq j'}(y_{ij}-\bar{y_U})(y_{ij'}-\bar{y_U}) \\
&=\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{j'=1}^{M}(y_{ij}-\bar{y_U})(y_{ij'}-\bar{y_U})-\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y_U})^2 \\
&=\sum_{i=1}^{N}M^2(\bar{y_{iU}}-\bar{y_U})^2-\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y_U})^2 \\
&=M(SSB)-SSTO \\
&=(M-1)SSTO-M(SSW)
\end{aligned}$$

Therefore, we can rewrite the ICC in terms of the SSW and SSTO:

$$ICC_t=1-\frac{M}{M-1}\frac{SSW_t}{SSTO_t} \tag{3.25}$$

and we make a connection between ICC, S_t^2, and the MSB:

$$MSB_t=\frac{MN-1}{M(N-1)}S_t^2\left(1+(M-1)ICC_t\right). \tag{3.26}$$

Note that, while ICC_t tends to be positive in CRTs, it can be negative when almost all variability in potential outcomes occurs within clusters, and there is very little variability in potential outcomes across clusters–or to express this in perhaps a more intuitive way, when clusters are formed "as heterogeneously as possible." When clusters are formed completely at random, we expect ICC_t to be close to 0.

We then can show that the relative variances in the sample means can be expressed as a function of the ICC:

$$\frac{Var(\hat{\mu}_{t,CRT})}{Var(\hat{\mu}_{t,CRE})}=\frac{MSB_t}{S_t^2}=\frac{MN-1}{M(N-1)}[1+(M-1)ICC_t] \tag{3.27}$$

Again, while this ratio will almost always be greater than 1, thus implying that the variance for the sample mean is greater for CRTs, we may have ratio less than 1 in the special case where

$$1 + (M-1)ICC_t \leq \frac{M(N-1)}{MN-1} \implies$$
$$ICC_t \leq -\frac{1}{MN-1}. \tag{3.28}$$

In this special case, CRTs will provide a more precise estimate of the population mean of potential outcomes. When conducting CRTs, considering the ICC is crucial. The ICC plays a vital role in both the experimental design and data analysis of CRTs (Ouyang et al., 2024).

3.2.2 Estimates of the Average Treatment Effect

We now extend our results obtained for the sample mean to investigate how estimators of the average treatment effect (ATE) Δ compare under complete randomization and cluster randomization.

Since we have already obtained unbiased estimators of $\bar{y}_{1U}$ and $\bar{y}_{0U}$, we can obtain an unbiased estimator for $\Delta = \bar{y}_{1U} - \bar{y}_{0U}$ by taking the difference of these estimators:

$$\hat{\Delta}_{HT} = \hat{\mu}_1 - \hat{\mu}_0 \tag{3.29}$$

The variance of the Horvitz-Thompson estimator is then

$$Var(\hat{\Delta}) = Var(\hat{\mu}_1) + Var(\hat{\mu}_0) - 2Cov(\hat{\mu}_1,\ \hat{\mu}_0). \tag{3.30}$$

We have already computed $Var(\hat{\mu}_1)$ and $Var(\hat{\mu}_0)$ under complete randomization and cluster randomization. We now turn our attention to $Cov(\hat{\mu}_1, \hat{\mu}_0)$.

We show in Appendix A that, under complete randomization, $Cov(\hat{\mu}_{1,CRE}, \hat{\mu}_{0,CRE})$ is

$$Cov(\hat{\mu}_{1,CRE}, \hat{\mu}_{0,CRE}) = -\frac{1}{Mn}\left(1 - \frac{n}{N}\right)\gamma_{10} \tag{3.31}$$

where γ_{10} is the covariance of the potential outcomes under treatment and control:

$$\gamma_{10} = \frac{1}{MN-1}\sum_{k=1}^{MN}(y_{1ij} - \bar{y_{1U}})(y_{0ij} - \bar{y_{0U}}) \tag{3.32}$$

To compute $Cov(\hat{\mu}_{1,CRT}, \hat{\mu}_{0,CRT})$, we can do the same trick as when analyzing variances and apply results for complete randomization while treating cluster totals as the units of interest. It is then immediate that

$$Cov(\hat{\mu}_{1,CRT}, \hat{\mu}_{0,CRT}) = -\frac{1}{n}\left(1 - \frac{n}{N}\right)\gamma_{c10} \tag{3.33}$$

where

$$\gamma_{c10} = \frac{1}{N-1}\sum_{k=1}^{MN}(t_{1iU} - \bar{t_{1U}})(t_{0iU} - \bar{t_{0U}}) \tag{3.34}$$

In particular, we note that while $Cov(\hat{\mu}_{1,CRE}, \hat{\mu}_{0,CRE})$ depends on γ_{10}, the correlation between treated and control potential outcomes, $Cov(\hat{\mu}_{1,CRT}, \hat{\mu}_{0,CRT})$ only depends on γ_{c10}, the covariance between the potential outcome totals. These covariances may be very different; for example, if each cluster contained an identical set of M units and there are no unit-level treatment effects nor effects due to cluster assignment, then $\gamma_{10} = S_1 S_0$ and $\gamma_{c10} = 0$.

As a quick aside, the covariances γ_{10} and γ_{c10} are not estimable as each term in these covariances requires knowledge of the potential outcomes under treatment and control for a given unit; we only observe the potential outcome for one treatment status for each unit. Instead, the covariance can be bounded via Young's inequality (Young, 1912). In particular:

$$-\frac{S_1^2 + S_0^2}{2} \leq \gamma_{10} \leq \frac{S_1^2 + S_0^2}{2}. \tag{3.35}$$

and

$$-\frac{S_{c,1}^2 + S_{c,0}^2}{2} \leq \gamma_{c10} \leq \frac{S_{c,1}^2 + S_{c,0}^2}{2}. \tag{3.36}$$

This inequality may be helpful in analyzing differences between variances between complete randomization and cluster randomization in special extreme cases.

Combining these results, we obtain

$$\frac{Var(\hat{\Delta}_{CRT})}{Var(\hat{\Delta}_{CRD})} = \frac{1}{M}\left(\frac{S_{c,1}^2 + S_{c,0}^2 + 2\gamma_{c10}}{S_0^2 + S_1^2 + 2\gamma_{10}}\right) \tag{3.37}$$

Note that the numerator is entirely dependent on the variability of within-cluster totals and the covariance between these totals across treatment groups. The denominator is dependent only on the variability of potential outcomes and the covariance of these outcomes across treatment groups. Hence, we gain further clarity that, when designing a CRT and clusters are able to be specified by the experimenter, the experimenter should assign units to clusters in a way that reduces between-cluster variability.

3.3 Threshold Clustering for Cluster Heterogeneity

While we cannot observe the potential outcomes prior to performing an experiment, we often have access to pre-treatment covariates that are often correlated with responses. Depending on the strength of association, we may be able to incorporate these covariates to improve the design of cluster-randomized experiments–in particular, by ensuring that the distribution of these covariates within a cluster are similar across clusters.

To ensure similar distributions in covariates across clusters, we propose using threshold clustering (TC) (Higgins et al., 2016; Luo et al., 2019). Given a dissimilarity measure that is small when two unit's covariate values are similar, TC forms groups of units so that each group contains a pre-specified number of units and so that the within-group dissimilarity is small. In particular, if a CRT has N clusters, we can perform TC so that each group

contains at least N units, and assign each unit within a group to a different cluster. If a group contains more than N units, we may assign the remaining units to clusters at random in a way to preserve equality in the cluster sizes.

TC is easiest to describe in the context of graph theory. We view dataset as a complete graph $G = (V, E)$ where each vertex $i \in V$ corresponds to a unit under study and E is an edge set where an edge ij is drawn between every unique pair of units i, j. Every edge ij has a cost c_{ij}; we assume that this cost is some kind of dissimilarity measure that is small if the pre-treatment covariates between i and j are similar. Common choices for the dissimilarity measure include the Euclidean distance and the Mahalanobis distance between the pre-treatment covariates.

For a pre-specified group size k, TC finds a partition $\mathbf{b} = \{V_1, \ldots, V_k\}$ such that each block of the partition V_i contains at least k units and so that the maximum dissimilarity within a block is small—if the dissimilarity measure satisfies the triangle inequality, this will produce a maximum dissimilarity within a factor of 4 of optimal across all possible partitions where each block contains at least k units. In the context of this paper, we will be setting $k = N - 1$ and setting $|V| = MN$. TC proceeds as follows:

1. Construct a $k-1$ nearest neighbor subgraph, $G_{nn} = (V, E_{nn})$—the subgraph of G where an edge $ij \in E_{nn}$ if and only if c_{ij} is one of the $k-1$ smallest values of $c_{i\ell}, \ell \in \{1, \ldots, MN\}$ or is one of the $k-1$ smallest values of $c_{\ell j}, \ell \in \{1, \ldots, N\}$.

2. Select *seed units* $S \subset V$ such that

   (a) There is no path of two edges or fewer in G_{nn} connecting any two seeds.

   (b) For any vertex $i \notin V$, there is a path of two edges or fewer connecting i to a node $s \in S$

3. Form clusters comprised of a cluster seed and all units adjacent to that seed in G_{nn}.

4. Assign any remaining unassigned vertices j to the group containing the seed $s \in S$ with the smallest edge weight c_{sj}.

Once the partition $\mathbf{b}$ is formed, we place one unit from each block $V_i \in V$ in each cluster. Any leftover units can be placed into clusters completely at random.

3.4 Simulation Example

In this paper, we conducted a simulation with one treatment effect and three covariates. The coefficients for each covariate are fixed at specific values, ensuring that the effects of the covariates on the response are on a comparable scale to the treatment effect. There are 50 subjects and 10 clusters. Each cluster contains 5 subjects. This sample size is common in studies on animals. Treatment is randomly assigned to each cluster. The goal of this simulation is to demonstrate that a well-controlled between-cluster variance can produce a more precise estimate of the treatment effect. Therefore, there are three ways to form the clusters. One is to sample five subjects randomly. One is using the threshold clustering algorithm to identify the most similar ten subjects regarding the covariates at first, then each subject is equally likely assigned to each cluster. The last way is to assign the five subjects most similar to each other to one cluster. The second way is the clustering method which ensures the between-cluster variance is as large as possible. We promote the second method. The third method represents the worst-case scenario, where the between-cluster variance is maximized. The random clustering method is included as a standard. Compared with random clustering, our method ensures that we avoid the worst scenario. The population-level treatment effect is, then, estimated by the linear mixed model. The point estimate and standard error are recorded for further model comparison. The whole simulation is repeated 500 times. Here, we will also explore the advantages and disadvantages of using adjusted models. The main concern of using covariates adjustment is adding more bias to the model especially when the covariates influence the response is not linearly. even it helps

with acceptance for trials close to the boundary of significance if the insignificance is due to a baseline imbalance. The other disadvantage of covariates adjustment is that if the covariates are not associated with the response there is a loss of degrees of freedom. In addition, it affects the effect size if the covariates are correlated with treatment assignment.

We present the results regarding the bias and variance of the comparison. As we can see from 3.1, the bias of the trials that the between-cluster variance is minimized is more concentrated around zero than the trials that the subjects are grouped by similar covariates. This bias is caused by the imbalance of covariates. This also explains why the method of minimizing between-cluster variance has a comparably more concentrated bias than the random clustering method. The result regarding the variability of estimates also presents the supreme of our method. As figure 3.2 shows, the variances of the estimates are less and more concentrated if we minimize the between-cluster variance. Even there is no obvious difference from threshold clustering to random clustering. However, the threshold clustering produces less extreme values on variance. This simulation supposes our idea.

3.5 Conclusion

The CRT is a versatile experimental design that works well when individual subjects are naturally connected, or when it is not feasible to assign treatments to individual subjects. However, CRT leads to less accurate statistical tests compared to simple random sampling because of the correlation within clusters. By quantifying this inaccuracy as the design effect (DEff), it provides insight into improving the CRT, or at least preventing the worst scenario during the design process. matching, stratification, and covariate adjustment are commonly used methods to achieve this improvement. Assigning individual subjects to each cluster based on a threshold of Mahalanobis distance with individual covariates is the trend in CRT. It leads to a decrease in between-cluster variability and an improvement in imbalance when the experimental condition is allowed. Therefore, we propose to use a low discrepancy

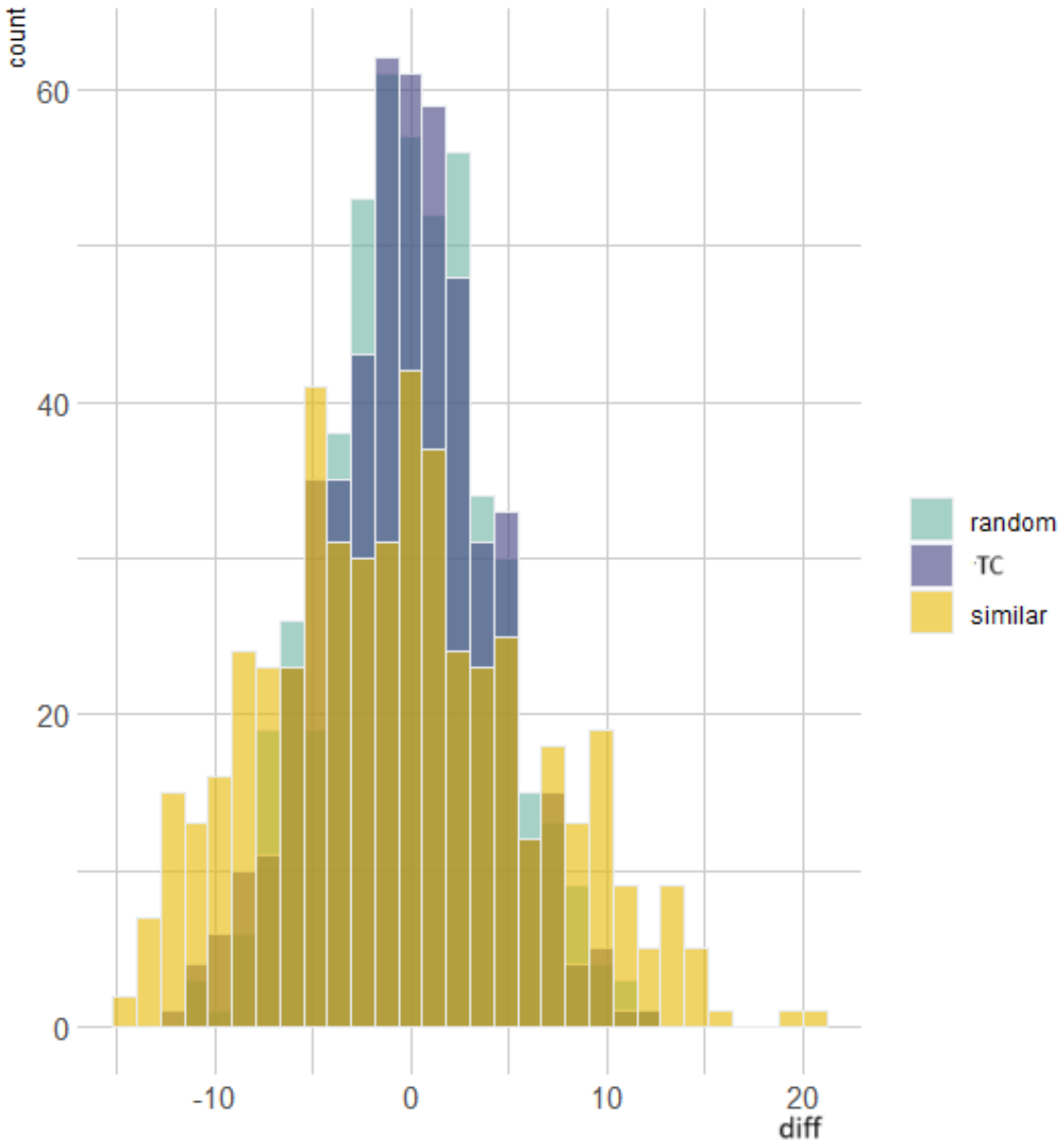

Figure 3.1: *Comparison of difference from the true parameter*

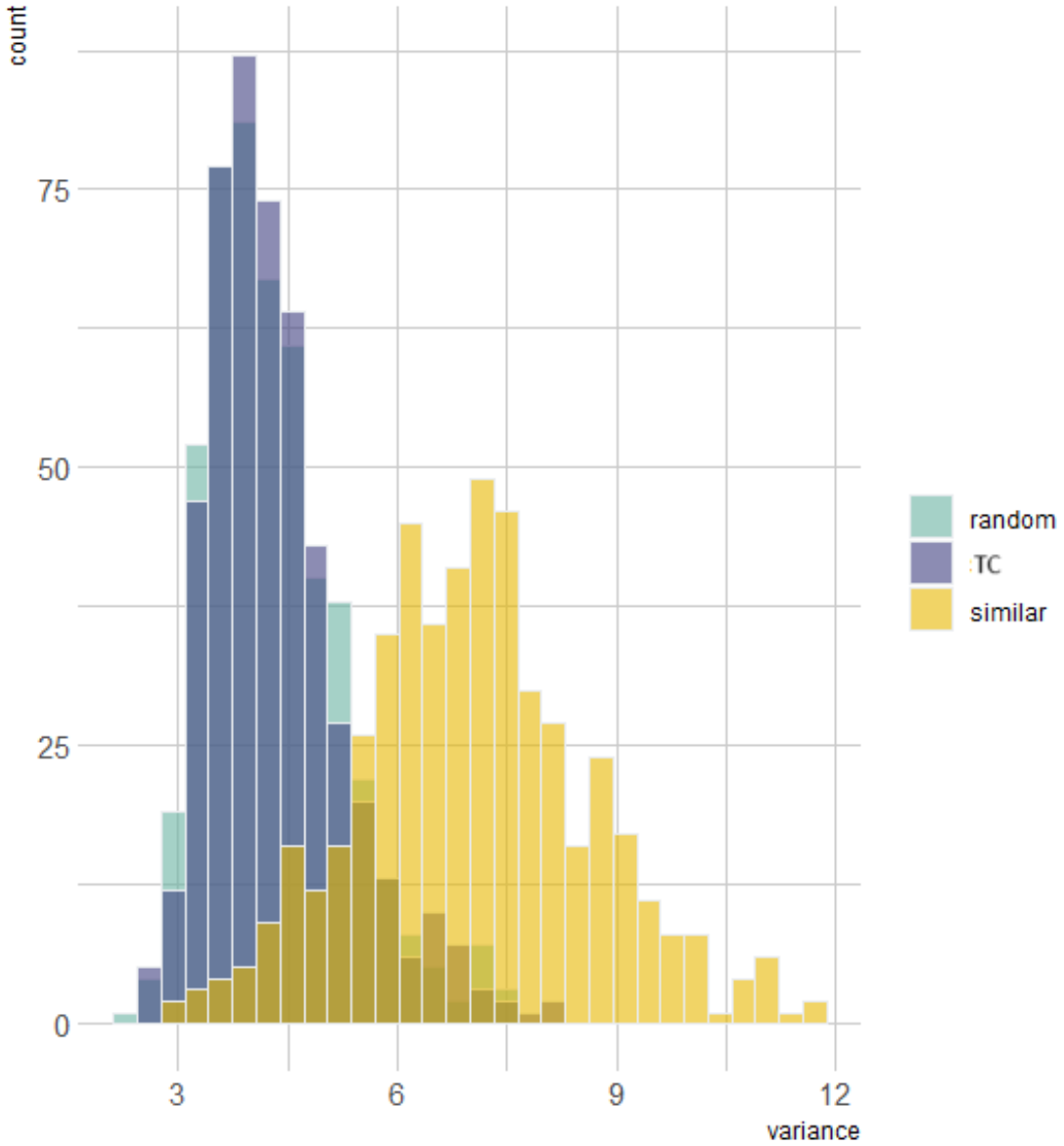

Figure 3.2: *Variance comparison*

algorithm based on the discrepancy theorem and ϵ-approximation. By doing so, we can randomly assign the most similar subjects with their covariates to each cluster to control the correlation within each cluster and improve the accuracy of CRT. This method can be further combined with covariates adjustment. These claims are confirmed by our simulation study. from section 3.4, our method outputs a more concentrated bias around zero and less variability of the estimate compared to the worst scenario. We believe that if there are fewer subjects in each cluster with fewer clusters, the method of minimizing between-cluster variance and balancing covariates becomes more important.

Chapter 4

Improved Divide and Conquer Methods Using Discrepancy Theory for Kernel Ridge Regression

4.1 Introduction

Kernel Ridge Regression (KRR) is a widely used machine learning method ideal for situations when the dimensionality of the covariate space is large and responses follow a non-linear relationship with respect to the covariates (Shawe-Taylor and Cristianini, 2004). Common applications include studies on gene expression, long-term clinical trials, classification of disease type, and prediction of treatment progress (Peng et al., 2020), (Singh, 2022) (Wang et al., 2022). However, while KRR often performs well in practice, it is a computationally intensive method and implementation may be still a challenge when the sample size, N, is large.

Many methods have been developed to reduce the computational time and memory required for KRR (Raskutti et al., 2013; Schölkopf et al., 1998; Williams and Seeger, 2000; Yao et al., 2007). One of the most common methods is *divide and conquer*, which involves

dividing the whole sample to multiple subsets, fitting KRR on each subset, and averaging KRRs from all the subsets. Zhang et al. (2015) suggest forming subsets of the full dataset through simple random sampling (SRS). They prove that this method not only ensures good convergence rates to the true data-generating function, but it also provides a substantial reduction in computational time—from $\mathcal{O}(N^3)$ to $\mathcal{O}(\frac{N^3}{m^2})$, where m is the number of subsets. Additionally, Zhang et al. (2015) clarifies the relationship between the KRR tuning parameter λ and the bias and the variance of the estimator; generally, increasing λ corresponds to an increase in bias and a decrease in variance.

However, Zhang et al. (2015) and Xu et al. (2018) both mention that the SRS approach for forming subsets often lead to under-regularized KRR models that tend to select a smaller regulating tuning parameters λ when fitting KRR within each subset than when fitting KRR to the entire dataset. This causes the variability of the models fit within each subset to be quite large, though the aggregate model will have small bias overall.

A variety of approaches have been proposed to overcome this issue of under-regularization. Some methods have focused on improved methods for forming the subsets. For example, Tandon et al. (2016) propose to divide the full data set to different independent continuous segments/portions. Other methods work to improve the tuning parameter used in the fit of the models within each subset. For example, Xu et al. (2018) suggest to using distribution generalized cross-validation to select the subset tuning parameter. The resulting subset tuning parameter is closer to the one obtained from fitting the model to the full dataset, thereby reducing variability after aggregating models.

We propose applying a sampling approach from discrepancy theory—the Gram-Schmidt Random Walk—to implement divide and conquer for KRR. The Gram-Schmidt Random Walk can produce subsets that resemble the full set of units in distribution. To fully reflect the dynamic of balancing bias and variance, another tuning parameter, ϕ, is introduced to control the proportion of low discrepancy sampling and random sampling. The low discrepancy algorithm is executed when ϕ is set to zero. Otherwise, the algorithm runs as an SRS

when ϕ is set to one. We can then find an optimal tuning parameter that minimizes the MSE within each subset by setting ϕ across a range of values.

We show via simulation that our sampling approach results in subsamples that are closer in distribution to the full dataset than expected under SRS. Additionally, this approach helps solve the issue of under-regularization typical with other divide and conquer methods—tuning parameter estimates within each subset will be closer to those for the full dataset under our approach. We also show that the optimal balancing of bias and variance can be obtained by adjusting the algorithm tuning parameter ϕ. Differences between our approach and SRS are largest for smaller datasets. Our conclusions corroborate those in Karnin and Liberty (2019).

The rest of the chapter is organized as follows. In 4.2.1, we first review the concepts of kernel ridge regression (KRR), introduce Zhang et al. (2015)'s divide and conquer method, and detail bias and variance tradeoffs related to the KRR tuning parameter. method and these important results regarding the dynamics of bias and variance in 4.2. We describe concepts in discrepancy theory and describe the Gram-Schmidt random walk in 4.3.1. We implement our divide and conquer approach with Gram-Schmidt for KRR in **??**. We perform a simulation study and discuss the results from the simulation in Section 4.4. Section 4.6 concludes.

4.2 Notation and Preliminaries

4.2.1 Kernel Ridge Regression

Kernel ridge regression (KRR) is a useful method for making predictions that combines the benefits of kernel methods with those from regularized regression. In general, kernel methods allow for effective modeling of non-linear relationships between covariates and responses (Shawe-Taylor and Cristianini, 2004). Additionally, regularized regression methods permit regression-type prediction and estimation under high-dimensional settings. We now

briefly summarize these components of KRR in detail.

Kernel Methods

Suppose we have N units, numbered 1 through N. Let $\boldsymbol{N} = \{1, 2, \ldots, N\}$ denote the *full sample* of units. Each unit $i \in \boldsymbol{N}$ has a response y_i and a d-dimensional vector of covariates $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \ldots, x_{id}) \in \mathbb{R}^d$. Define $\boldsymbol{X} = \{\boldsymbol{x}_i\}_{i=1}^N$.

Kernel methods aim to fit non-linear models to data by first transforming the data to a higher-dimensional space and fitting a linear model to these transformed data. Notably, for many settings, this transformation ϕ does not need to be explicitly defined. Rather, we only need to specify the inner product of $\langle \phi(\boldsymbol{x_i})\phi(\boldsymbol{x_j}) \rangle$ for all $\boldsymbol{x_i}$, $\boldsymbol{x_j} \in \boldsymbol{X}$ using a kernel function κ.

Rigorously, let $\boldsymbol{\phi}$ denote a mapping from $\mathbb{R}^d$ to an inner product feature space $\mathcal{F} \subset \mathbb{R}^N$:

$$\mathcal{F} = \left\{ \sum_i \alpha_i \phi(\boldsymbol{x_i}) : \boldsymbol{x_i} \in \mathbb{R}^d, \alpha_i \in \mathbb{R} \right\}. \tag{4.1}$$

Generally, we only express this transformation implicitly via kernel transformation. A kernel function κ with respect to ϕ is a mapping from $\boldsymbol{x_i}$, $\boldsymbol{x_j} \in \boldsymbol{X}$ to $\mathbb{R}$ satisfying

$$\kappa(\boldsymbol{x_i}, \boldsymbol{x_j}) = \langle \phi(\boldsymbol{x_i}), \phi(\boldsymbol{x_j}) \rangle. \tag{4.2}$$

Intuitively, κ measures the similarity of the two input vectors–larger similarities correspond to more-similar points. For example, a Gaussian kernel is given as: $\kappa(\boldsymbol{x_i}, \boldsymbol{x_j}) = \exp\left(-\frac{||\boldsymbol{x_i} - \boldsymbol{x_j}||^2}{2b}\right)$, where b is a bandwidth.

Note, we are making an assumption that the kernel function can be expressed as an inner-product of two unknown functions. However, this will hold under certain conditions. In particular, Mercer's theorem guarantees such a decomposition if the kernel function is symmetric, continuous, and positive semi-definite (Mercer, 1909).

Note, if κ is such a function, then for any two functions f, $g \in \mathcal{F}$, where $f = \sum_i \alpha_i \phi(\boldsymbol{x_i})$

and $g = \sum_{j=1}^{N} \beta_j \phi(\boldsymbol{x_j})$, the inner product of f and g is:

$$\langle f, g \rangle = \sum_i \sum_j \alpha_i \beta_j \phi(\boldsymbol{x_i}) \phi(\boldsymbol{x_j}) = \sum_i \sum_j \alpha_i \beta_j \kappa(\boldsymbol{x_i}, \boldsymbol{x_j}). \tag{4.3}$$

The representing property of the kernel can be given as

$$\langle f, \phi(\boldsymbol{x}) \rangle = \sum_i \alpha_i \kappa(\boldsymbol{x_i}, \boldsymbol{x}) = f(\boldsymbol{x}) \tag{4.4}$$

The last equation implies that a function f can be represented as a linear function in the feature space, even the function f is a non-linear complex function in the original coordinate space. By computing the kernel matrix, the original data is transformed to the Reproducing Kernel Hilbert Space to make the pattern of data linear in the higher dimension inner product space. Therefore, a linear model can be applied on these transformed data. The corresponding norm in the feature space is given as:

$$||f||_{\mathcal{F}}^2 = \langle f, f \rangle = \sum_i \sum_j \alpha_i \alpha_j \kappa(\boldsymbol{x_i}, \boldsymbol{x_j}) = \boldsymbol{\alpha}^T \boldsymbol{K} \boldsymbol{\alpha} \tag{4.5}$$

4.2.2 Regularized Regression and Kernel Ridge Regression

Ridge Regression (RR) was proposed by Hoerl and Kennard (1970) to overcome the challenge that post by the singularity of $\boldsymbol{X^T X}$ when applying ordinary least squares regression. It is a common method used in genome regression modeling when the data scale is large and multicollinearity is common. Intuitively, RR works by adding a small value λ to the diagonal entries of $X^T X$ to avoid singularity.

To be precise, the loss function of RR is

$$\boldsymbol{L}_\lambda(\boldsymbol{\beta}, \boldsymbol{X}) = min_{\boldsymbol{\beta}}(\boldsymbol{y} - \boldsymbol{X^T \beta})^2 + \lambda ||\boldsymbol{\beta}||_2^2. \tag{4.6}$$

By taking the derivative of the loss function with respect to $\boldsymbol{\beta}$ and setting it equal to zero,

we obtain

$$\boldsymbol{X^T y} - \boldsymbol{X^T X\beta} - \lambda\boldsymbol{\beta} = \boldsymbol{X^T y} - (\boldsymbol{X^T X} + \lambda\boldsymbol{I})\boldsymbol{\beta} = 0. \tag{4.7}$$

Solving for $\boldsymbol{\beta}$, we obtain

$$\boldsymbol{\beta} = (\boldsymbol{X^T X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X^T y}. \tag{4.8}$$

This solution is known as the primal form.

Alternatively, the dual form can be expressed by re-writing equation $(\boldsymbol{X^T X} + \lambda\boldsymbol{I})\boldsymbol{\beta} = \boldsymbol{X^T y}$ as which results

$$\boldsymbol{\beta} = \lambda^{-1}\boldsymbol{X^T}(\boldsymbol{y} - \boldsymbol{X\beta}) = \boldsymbol{X^T\alpha} \tag{4.9}$$

showing that, $\boldsymbol{\beta} = \sum \alpha_i \boldsymbol{x_i}$ by solving for $\boldsymbol{\alpha}$ we have

$$\boldsymbol{\alpha} = (\boldsymbol{X X^T} + \lambda\boldsymbol{I})^{-1}\boldsymbol{y}. \tag{4.10}$$

The dual form offers a different way of fitting a ridge regression on a symmetric semi-positive definite input, which implies the suitability of combining ridge regression with the kernel method. In the primal form, $\hat{y} = \langle\boldsymbol{\beta}, \boldsymbol{x}\rangle = \boldsymbol{y^T X}(\boldsymbol{X^T X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{x}$. In the dual form $\hat{y} = \langle\boldsymbol{\beta}, \boldsymbol{x}\rangle = \langle\sum \alpha_i \boldsymbol{x_i}, \boldsymbol{x}\rangle = \sum \alpha_i \langle\boldsymbol{x_i}, \boldsymbol{x}\rangle = \boldsymbol{y^T}(\boldsymbol{G}+\lambda\boldsymbol{I})^{-1}\boldsymbol{k}$, where the Gram matrix $\boldsymbol{G} = \boldsymbol{X X^T}$, and $k_i = \langle\boldsymbol{x_i}, \boldsymbol{x}\rangle$. As we can see, the dual solution contains the information from the training units in the form of an inner product. Similarly, the information of the new unit $\boldsymbol{x}$ is also contained in the Therefore, an extension of ridge regression can be made by simply replacing the Gram matrix with other symmetric semi-positive matrices.

In the framework of the kernel model, the optimization problem becomes

$$argmin_{f\in\mathcal{F}} \sum_i (f(\boldsymbol{x_i}) - y_i)^2 + \lambda||f||^2_{\mathcal{F}} \tag{4.11}$$

Which can be rewritten as:

$$min_{\boldsymbol{\alpha}}(\boldsymbol{y}-\boldsymbol{K\alpha})^{\boldsymbol{T}}(\boldsymbol{y}-\boldsymbol{K\alpha})+\lambda\boldsymbol{\alpha}^{\boldsymbol{T}}\boldsymbol{K\alpha} \tag{4.12}$$

Therefore, the solution becomes:

$$\hat{\boldsymbol{\alpha}}=(\boldsymbol{K}+\frac{1}{n}\lambda\boldsymbol{I})^{-1}\boldsymbol{y} \tag{4.13}$$

The prediction becomes:

$$f(\boldsymbol{x})=\sum_i \hat{\alpha}_i\kappa(\boldsymbol{x_i},\boldsymbol{x}) \tag{4.14}$$

Just like the dual outcomes we derived from the previous part.

To further understand the role of λ in the general ridge regression, a geometrical interpretation is presented below:

$$\begin{aligned}\hat{\boldsymbol{y}}&=\boldsymbol{X}(\boldsymbol{X^TX}+\boldsymbol{\lambda I})^{-1}\boldsymbol{X^Ty}\\&=\boldsymbol{V\Lambda}(\boldsymbol{\Lambda^2}+\boldsymbol{\lambda I})^{-1}\boldsymbol{\Lambda V^Ty}\\&=\sum_{i=1}^{p}\boldsymbol{v_i}\frac{\omega_i^2}{\omega_i^2+\lambda}\boldsymbol{v_i^Ty}\end{aligned} \tag{4.15}$$

where $\boldsymbol{V}$ is the othonormal basis of $\boldsymbol{X}$ and ω_i is the corresponding eigenvalue. Ridge regression actually shrinks the coordinates of the principal component coordinates, and $\frac{\omega_i^2}{\omega_i^2+\lambda}$ can be viewed as the shrinkage factor. Therefore, the smaller the variability is on the principal component coordinate, the greater the shrinkage is on the same coordinate. The larger λ is, the more the projection is shrunk in every coordinate.

A similar geometrical decomposition can be performed on kernel ridge regression:

$$\begin{aligned}\hat{\boldsymbol{y}} &= \boldsymbol{K}\boldsymbol{\alpha} \\ &= \sum_i \frac{\omega_i}{\omega_i + \lambda_i} \boldsymbol{v_i} \boldsymbol{v_i^T} \boldsymbol{y}\end{aligned} \tag{4.16}$$

where $\boldsymbol{v_i}$ are eigenvectors, ω_i are eigenvalues. In addition, $\lambda \boldsymbol{I}$ can be project on the eigenvector space such that $\frac{1}{n}\lambda \boldsymbol{I} = \sum_i \lambda_i \boldsymbol{v_i} \boldsymbol{v_i^T}$.

4.2.3 Divide and Conquer

While KRR can be an effective way of making predictions, a major disadvantage is that the computing time becomes prohibitive as the number of units N increases. The computational complexity for solving α is $\mathcal{O}(N^3 + N^2 d)$ with additional time of $\mathcal{O}(Nd)$ for evaluating a new data point. To overcome these computational limitations, Zhang et al. (2015) proposed a divide and conquer method will allows for application of KRR when N is large. In this method, the full sample $\boldsymbol{N}$ is randomly partitioned to m subsets $\boldsymbol{S_1}, \boldsymbol{S_2}, \ldots, \boldsymbol{S_m}$. A kernel ridge regression $\hat{f}_j$ is then fitted within each subset:

$$\hat{f}_j = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \left\{ \frac{1}{|\boldsymbol{S_j}|} \sum_{(\boldsymbol{x}, y) \in \boldsymbol{S_i}} (f(\boldsymbol{x}) - y)^2 + \lambda_j ||f||_{\mathcal{F}}^2 \right\}. \tag{4.17}$$

The final model is the average of each subset model:

$$\bar{f} = \frac{1}{m} \sum_{j=1}^{m} \hat{f}_j. \tag{4.18}$$

Zhang et al. (2015) further decomposes the expected MSE into more detailed parts as:

$$\begin{aligned}
\mathbb{E}[\| \bar{f} - f^* \|_2^2] &= ||\mathbb{E}[\hat{f}_1] - f^*||_2^2 + \frac{1}{m}\mathbb{E}[||\hat{f}_1 - \mathbb{E}[\hat{f}_1]||_2^2] \\
&\leq \| \mathbb{E}[\hat{f}_1] - f^* \|_2^2 + \frac{1}{m}\mathbb{E}[\| \hat{f}_1 - f^* \|_2^2] \\
&= \| \mathbb{E}[\hat{f}_1] - f^* \|_2^2 + \frac{1}{m}(\| f_\lambda^* - f^* \|_2^2 + \mathbb{E}[\| \hat{f}_1 - f^* \|_2^2] - \| f_\lambda^* - f^* \|_2^2) \\
&= T_1 + \frac{1}{m}(T_2 + T_3)
\end{aligned} \tag{4.19}$$

where

$$\begin{aligned}
f_\lambda^* &= \underset{f \in \mathcal{F}}{\operatorname{argmin}}\{\mathbb{E}[(f(X) - Y)^2] + \lambda \| f \|_{\mathcal{F}}^2\} \\
T_1 &= \| \mathbb{E}[\hat{f}] - f^* \|_2^2 \\
T_2 &= \| f_\lambda^* - f^* \|_2^2 \\
T_3 &= \mathbb{E}[\| \hat{f}_1 - f^* \|_2^2] - \| f_\lambda^* - f^* \|_2^2
\end{aligned}$$

The f^* is the true model, and the f_λ^* is the regularized population minimizer. The less and equal sign is due to the fact $\mathbb{E}[\hat{f}_j]$ minimizes $\mathbb{E}[||\hat{f}_j - f||_2^2]$ for $f \in \mathcal{F}$. T_1 and T_2 are bias and approximation error by using the ridge regression method. T_3 is the variance term. Equation (3.20) expresses the three trade-offs in this divide-and-conquer procedure. There is a trade-off between T_2 and T_3 that is controlled by λ. As λ increases, the bias, T_2, increases while the variance, T_3 decreases. The same relation can be discovered with T_1 and T_3. The last trade-off is between T_1 and computational time. The bias, T_1, increases when the number of subsets, m, increases, and computation time, $\frac{N^3}{m^2}$ decreases.

As Zhang et al. (2015) mentioned that the subset models are under regularized that the tuning parameter, λ, are smaller comparing to the full model, and the variances are enlarged in this scenario. Xu et al. (2018) also illustrated that the locally optimal choice of λ leads a balance of the trade-off only for the sub model. Xu et al. (2018) propose to use distributed generalized cross-validation to determine the tuning parameter in stead of using the regular

cross-validation. The enlarged variance may also affect the convergence rate.

Giving the assumptions that the population true model is $\boldsymbol{f^* = K^*\alpha^*}$, $\boldsymbol{y = f^* + \epsilon}$, and the points that have similar position relations with other points output the similar f^*'s, since $f^* = \sum \alpha_i^* \boldsymbol{k(x_i, x)}$ and α_i's are parameters. According to the geometrical interpretation of the kernel ridge regression. We propose a subset method that preserves the size and direction of the variability of the full set can result a local tuning parameter that is as close as the global tuning parameter. The low discrepancy algorithm that splits the most similar two points into two subsets could output subsets that select the local λ's that are similar to the global λ.

To better understand how this could work. An extreme scenario is studied. Suppose there are two subsets $\boldsymbol{S_1}$ and $\boldsymbol{S_2}$ are identical such that $\boldsymbol{x_{11} = x_{21}, x_{12} = x_{22}, ..., x_{1n} = x_{2n}}$, and $y_{11} = y_{21}, y_{12} = y_{22}, ..., y_{1n} = y_{2n}$. The full set ,$\boldsymbol{S}$, is the combination of S_1 and S_2. Giving the idea of splitting the two most similar points to each subset, S_1 and S_2 are the outputs. Fitting kernel ridge regressions on both S and S_1, we have the object to find a λ that gives the minimal prediction error. Giving that loss of one point should not change the eigenvalues and eigenvectors (Allen (1974)). The objective:

$$min_\lambda \frac{1}{n} \sum_i \parallel y_i - \boldsymbol{k_{-i}(x, x_i)}(\boldsymbol{K_{-i}} + \frac{1}{n-1}\lambda \boldsymbol{I})^{-1} \boldsymbol{y_{-i}} \parallel^2 \tag{4.20}$$

becomes to

$$min_\lambda \frac{1}{n} \parallel \boldsymbol{y} - \sum_j \frac{\omega_j}{\omega_j + \lambda_j} \boldsymbol{v_j v_j^T Y_-} \parallel^2 \tag{4.21}$$

For a giving y_i, the $\boldsymbol{y_{1,-i}}$ of the subset is one fold of two corresponding parts $\boldsymbol{y_{1,-i}}$ and $\boldsymbol{y_{2,-i}}$. In other words, $\boldsymbol{y_{-i}}$ are combination of $\boldsymbol{y_{1,-i}}$, $\boldsymbol{y_{2,-i}}$ and y_i. Both kernel matrices of

the subset and full set have the same variability on each $\boldsymbol{v_j}$. $\boldsymbol{y_{-1,i}}$ and $\boldsymbol{y_{2,-i}}$ have the same projection on each $boldsymbolv_j$. The projection of y_i on each $\boldsymbol{v_j}$ can be ignored with the set size is big. Therefore, the projection of $\boldsymbol{y_{-i}}$ on each $\boldsymbol{v_j}$ is twice as bigger as $\boldsymbol{y_{1,-i}}$, by average, these projections become equal. With each ω_j of the subset equals the full set. The shrinkage λ_j in each direction needs to be the same size to reach the minimum. When those two folds of data are not identical, to split the most similar kernel, $\boldsymbol{k(x, x_i)}$, into two sets ensures that the subsets kernel matrices have the closest form of eigen-decomposition. Since we assume that $\boldsymbol{y} = \boldsymbol{K^*\beta^*}$, which implies similar kernels should have similar y's, that indicates these two subsets have the most similar projections on the eigen-decomposition. We can conclude that the subset $\boldsymbol{y_1}$ has the most similar average projection on each eigenvector as the full set $\boldsymbol{y}$ when split into the most similar kernel to different subsets.

4.3 A Discrepancy Method Approach for Divide and Conquer

We propose to use low discrepancy algorithm to divide the full data set. This algorithm assigns the two most similar units in the full set to two different subsets. This gives the subsets that are highly similar to the full set. This results in a local λ that is similar to the global λ. The variance is smaller compared to the random dividing due to the similarity between each subset.

In addition, a new tuning parameter, ϕ is introduced to the low discrepancy algorithm. By generally increasing ϕ, we can generally increase the variance, decrease the bias, and find an optimal balance of the variance and bias trade-off. The new tuning parameter is introduced to the low discrepancy algorithm by modifying the input matrix. Harshaw et al. (2019) used this method to achieve an optimization of covariate balancing and randomness in experimental design. Generally, increasing the value of ϕ, the low discrepancy algorithm results in a division that is close to a random division. Therefore, the variance is generally

enlarged, and the bias generally decreases. An optimal bias and variance balance could be found during this procedure.

4.3.1 Discrepancy Theory

This subsection focuses on the systematic introduction of discrepancy theory, low discrepancy algorithm, the ϵ-approximation bound of the kernel density model, and how this bound gives insight into the divide and conquer method.

The combinatorial discrepancy, For any $n \times p$ matrix $\boldsymbol{X}$, is defined as

$$disc(\boldsymbol{X^T}) = min_{\boldsymbol{z}\in\{-1,1\}^n} ||\boldsymbol{X^T z}||_\infty \tag{4.22}$$

which is minimizing the l_∞ norm of the vector $\boldsymbol{X^T z}$ over all mapping $\mathcal{Z} : z \to \{-1, 1\}$. In this case, the discrepancy is also a vector balance problem.

In this section, the Gram-Schmidt random walk is the low discrepancy algorithm we are focusing on. This algorithm is based on Dadush et al. (2016)'s proportion that there exists a $O(1)$ sub-Gaussian distribution supported on $\{\sum_{i=1}^n z_i \boldsymbol{x_i} : z_i \in \{-1, 1\}\}$ regarding Banaszczyk (1998)'s theory. Therefore, a low discrepancy output can be obtained by generating sub-Gaussian process based on the input $\boldsymbol{x_1^T}, ..., \boldsymbol{x_n^T}$. Generally, generating sub-Gaussian process is much more effective than other algorithms. There are applications of this algorithm in different statistical studies. For example, Harshaw et al. (2019) applies Gram-Schmidt random walk on randomized experimental design and shows that this algorithm has better results on balancing covariates. Phillips (2016) applies Gram-Schmidt algorithm to model approximation for kernel density estimation, and shows the random walk requires a smaller subset, $\mathcal{O}(\frac{\sqrt{d}}{\epsilon})$, compared with other low discrepancy algorithms.

The Gram-Schmidt Walk algorithm proceeds as follows. The algorithm will produce an assignment of units into one of two groups $\boldsymbol{z_t} = (z_{t1}, z_{t2}, \ldots, z_{tn}) \in \{-1, 1\}$, where units i satisfying $z_{ti} = 1$ belong to one group, and all other units belong to the other group.

1. **(Initialize)**: Set $\boldsymbol{z_1} = (0, \ldots, 0)$, $\mathcal{A} = \{1, \ldots, n\}$, and $t = 1$.

2. **(Select pivot)**: Select a unit $p \in \mathcal{A}$ completely at random.

3. **(Compute the step direction)**: Define

$$\mathbb{U}_p = \{(\mu_1, \cdots, \mu_n) \in \mathbb{R}^n : \mu_p = 1, \mu_j = 0 \ \textit{if} \ j \notin \mathcal{A}\} \tag{4.23}$$

   and let $\boldsymbol{\mu_t} = \text{argmin}_{\mu \in \mathbb{U}_p} \parallel \boldsymbol{B\mu} \parallel$.

4. **(Choose size of step)**: Let

$$\Delta = \{\delta \in \mathbb{R} : \boldsymbol{z_t} + \delta \boldsymbol{\mu_t} \in [-1, 1]^n\} \tag{4.24}$$

   denote all possible step sizes δ, and let $\delta^+ = |\max \Delta|$ and $\delta^- = |\min \Delta|$. Set

$$\delta_t = \begin{cases} \delta^+, & \text{with probability } \frac{\delta^-}{\delta^- + \delta^+}, \\ -\delta^-, & \text{with probability } \frac{\delta^+}{\delta^- + \delta^+} \end{cases} \tag{4.25}$$

5. **(Make step)**: Set

$$\boldsymbol{z_{t+1}} \leftarrow \boldsymbol{z_t} + \delta_t \mu_t. \tag{4.26}$$

   and set

$$\mathcal{A}^* = \{i \in \mathcal{A} : |z_{(t+1)i}| < 1\} \tag{4.27}$$

6. **(Terminate or repeat)** If $|\mathcal{A}^*| = 0$, terminate. We have $\boldsymbol{z_{t+1}} \in \{-1, 1\}^n$. Otherwise, set $\mathcal{A} \leftarrow \mathcal{A}^*$, $t \leftarrow t + 1$, and return to Step 2.

It relaxes the assignment task from the integral $\{\pm 1\}$ to the interval $[-1, 1]$. The update step $\delta \boldsymbol{\mu_t}$ maintains $\boldsymbol{Bz_t}$ close to 0 by its direction $\boldsymbol{\mu_t}$ and enlarge the step size δ so $\boldsymbol{z_t}$ is close to $\{\pm 1\}^n$.

Harshaw et al. (2019) further propose a new form of input matrix $\boldsymbol{B}$ that can be view as a combination of identity matrix $\boldsymbol{I}$ and a covariates matrix $\boldsymbol{X}$ that the trade off is controlled by a parameter ϕ:

$$\boldsymbol{B} = \begin{bmatrix} \sqrt{\phi}\boldsymbol{I} \\ \xi^{-1}\sqrt{1-\phi}\boldsymbol{X}^{\boldsymbol{T}} \end{bmatrix} \tag{4.28}$$

where ξ is the maximum row norm of the matrix $\boldsymbol{X}$. Therefore, the bottom part and the top part are on the same scale. Harshaw et al. (2019) use this new form of input matrix to balance the trade-off of covariance pairing and randomness in experimental design. As we can see, the algorithm places greater weight on balancing the matrix $\boldsymbol{X}$ by setting ϕ as close to zero. On the other hand, the algorithm ignores the rows of $\boldsymbol{X}$ and randomly divides them by setting ϕ as close to 1.

4.3.2 The Low Discrepancy Dividing Process

We propose to divide the full set of data into a manner of binary tree. We first run the algorithm to receive the first layer of subsets, say S_{11} and S_{12}. the low discrepancy algorithm is applied to the first lay subsets separately, each subset gives two second lay subsets, say S_{11} is divided to S_{21} and S_{22}, S_{12} is divided to S_{23} and S_{24}. Generally speaking, we can repeatedly apply the low discrepancy algorithm to a set until the size of the resulting subset reaches a certain amount. However, we suggest dividing the data for only several layers, due to the additivity of ϵ-approximation. The more the full set is divided, the less accurate the ϵ-approximation becomes. We also suggest reversing the dividing process if the sample sizes of the resulting two subsets are highly different, especially at the bottom of the hierarchy. We want to receive two subsets that are similar in cardinality because the bound of the approximation is related to the cardinality of the subsets.

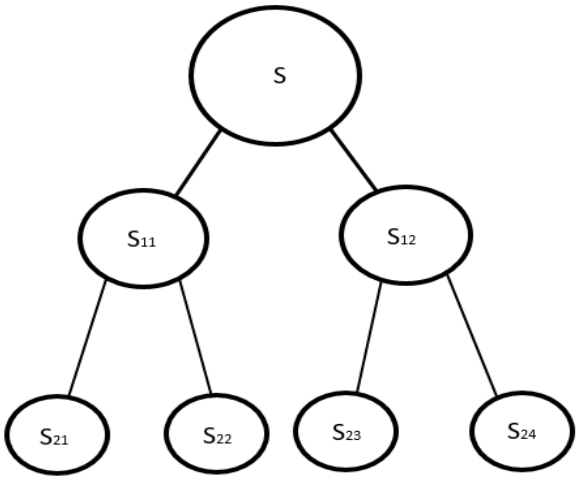

Figure 4.1: *Low discrepancy dividing*

4.4 Simulation and Result

We now conduct a simulation study to evaluate the efficacy of our divide-and-conquer algorithm. We evaluate our method on a few measures, including bias, mean-square error, and similarity in distribution of the subsets created by our divide and conquer method.

To conduct our simulations, we generate n independent and identically distributed p-dimensional covariate vectors from multivariate normal distribution $\mathcal{N}(\mathbf{0_p}, \boldsymbol{I_p})$, where $\mathbf{0_p}$ denotes a 0-vector of length p and We consider sample sizes $n = 400$, 800, 1200, 1600, 2000, 2400, 2800, 3200, and 3600 number of covariates $p = 4$.

We generate responses as

$$y = \boldsymbol{f}^*(\boldsymbol{x}) + \boldsymbol{\epsilon} = \boldsymbol{K}^*\boldsymbol{\alpha}^* + \boldsymbol{\epsilon} \tag{4.29}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, 0.5\boldsymbol{I})$ and the kernel matrix $\boldsymbol{K}^*$ takes one of the following forms:

- **(Gaussian)**: $\boldsymbol{K}^*_{ik} = -exp[(\boldsymbol{x}^*_{(k,)} - \boldsymbol{x}^*_{i,})^T(\boldsymbol{x}^*_{k,} - \boldsymbol{x}^*_{i,})]$.

The true model, f^* is generated by ensuring the design matrix, $\boldsymbol{X}^*$. $\boldsymbol{X}^*$ contains four columns, each one is generated by a standard normal distribution. the kernel matrix, $\boldsymbol{K}^*$, is computed as, $[K^*]_{ik} = -exp[(\boldsymbol{x}^*_{(k,)} - \boldsymbol{x}^*_{i,})^T(\boldsymbol{x}^*_{(k,)} - \boldsymbol{x}^*_{i,})]$. Further the $\boldsymbol{\alpha}^*$ is generated

according to the kernel density,$\boldsymbol{d}$, computed by $\boldsymbol{K^*}$. $\boldsymbol{\alpha^*} = \pi \cdot \boldsymbol{d} + 2.16$. $\boldsymbol{y}$ is generated as $\boldsymbol{y} = \boldsymbol{K^*\alpha^*} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ follows a $\mathcal{N}(\mathbf{0, 0.5})$. To test the effectiveness of the model, the full model with size $n = 400, 800, 1200, 1600, 2000, 2400, 2800, 3200, 3600$ is divided by random sampling and Gram Schmidt walk algorithm with $\phi = 0, 0.2, 0.4, 0.6$ to four groups. The full set kernel, $\boldsymbol{K}$, is divided for four subset kernels, $\boldsymbol{K_{11}}$, $\boldsymbol{K_{22}}$, $\boldsymbol{K_{33}}$ and $\boldsymbol{K_{44}}$. In other words:

$$\boldsymbol{K} = \begin{bmatrix} \boldsymbol{K_{11}} & \boldsymbol{K_{12}} & \boldsymbol{K_{13}} & \boldsymbol{K_{14}} \\ \boldsymbol{K_{21}} & \boldsymbol{K_{22}} & \boldsymbol{K_{23}} & \boldsymbol{K_{24}} \\ \boldsymbol{K_{31}} & \boldsymbol{K_{32}} & \boldsymbol{K_{33}} & \boldsymbol{K_{34}} \\ \boldsymbol{K_{41}} & \boldsymbol{K_{42}} & \boldsymbol{K_{43}} & \boldsymbol{K_{44}} \end{bmatrix} \tag{4.30}$$

The corresponding responses follow the same sub-scrabbles. For each subset, the tuning parameter, λ, is selected by 10-fold cross-validation. The estimate values, $\hat{\boldsymbol{y_{11}}}$, is estimate by $\boldsymbol{K_{11}} * \hat{\boldsymbol{\alpha_1}}$, where $\hat{\boldsymbol{\alpha}}_1 = [\boldsymbol{K_{11}}] + \frac{\lambda_1}{n_1}\boldsymbol{I}]^{-1}\boldsymbol{y_1}\mathbf{1}$. The predict value, $\hat{\boldsymbol{y_{12}}}$, is computed by $\boldsymbol{K_{12}} * \hat{\boldsymbol{\alpha_1}}$. $\hat{\boldsymbol{y_{13}}}$ and $\hat{\boldsymbol{y_{14}}}$ are computed similarly. The predict values, $\hat{\boldsymbol{y_{21}}}$ is $\boldsymbol{K_{21}} * \hat{\boldsymbol{\alpha}}_{\mathbf{2}}$, where $\hat{\boldsymbol{\alpha}}_2 = [\boldsymbol{K_{22}}] + \frac{\lambda_2}{n_2}\boldsymbol{I}]^{-1}\boldsymbol{y_{22}}$. The estimate values, $\hat{\boldsymbol{y_{21}}}$ is computed by $\boldsymbol{K_{21}} * \hat{\boldsymbol{\alpha}}_{\mathbf{2}}$. The values of $\hat{\boldsymbol{y_{23}}}$, $\hat{\boldsymbol{y_{24}}}$, $\hat{\boldsymbol{y_{31}}}$, $\hat{\boldsymbol{y_{32}}}$, $\hat{\boldsymbol{y_{33}}}$, $\hat{\boldsymbol{y_{34}}}$, $\hat{\boldsymbol{y_{41}}}$, $\hat{\boldsymbol{y_{42}}}$, $\hat{\boldsymbol{y_{43}}}$, $\hat{\boldsymbol{y_{44}}}$are computed similarly. The final model, $\bar{y} = \frac{1}{4}[\sum_{i=1}^{4} \hat{\boldsymbol{y_{21,2i}}}, \sum_{i=1}^{4} \hat{\boldsymbol{y_{22,2i}}}, \sum_{i=1}^{4} \hat{\boldsymbol{y_{23,2i}}}, \sum_{i=1}^{4} \hat{\boldsymbol{y_{24,2i}}}]^T$.

The MSE of each dividing method is computed following the final model. For each n, the simulation is repeated 100 times. The MSE of each divide method is summed and divided by 100. The averaged MSEs are further plotted as the size, n, on the horizontal axis and the averaged MSE on the vertical axis.

4.5 Results

As we can see, as the size increases, the MSE of random sampling decreases. However, the low discrepancy method with different *phi* increases. The reason behind that could be that the dimensions increase when the sample size increases. The average MSE of different

Comparison of Average MSE					
N	$\phi = 0$	$\phi = 0.2$	$\phi = 0.4$	$\phi = 0.6$	$\phi = 1$
400	0.1624	0.1629	0.1632	0.2095	0.2591
800	0.1737	0.1760	0.1815	0.1921	0.2480
1200	0.1818	0.1863	0.1827	0.1991	0.2412
1600	0.1889	0.1897	0.1935	0.2005	0.2367
2000	0.1951	0.1956	0.1967	0.2062	0.2369
2400	0.1978	0.1983	0.1964	0.2088	0.2406
2800	0.1913	0.1959	0.1943	0.2071	0.2380
3200	0.1986	0.2046	0.2045	0.2121	0.2389

Table 4.1: *Comparison of average MSE*

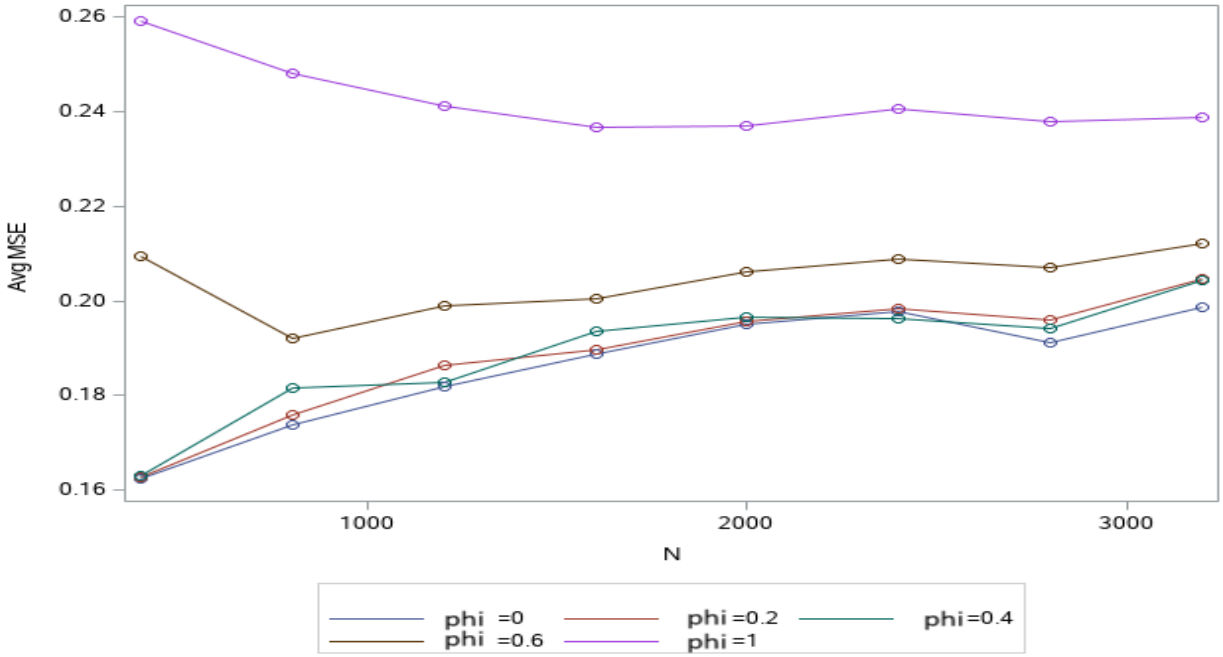

Figure 4.2: *Average MSE of different dividing method*

dividing methods with $\phi = 0$, $\phi = 0.2$, and $\phi = 0.4$ cross over each other. This demonstrates the optimal balance of bias and variance when the sample size changes.

In addition, with the general increases of ϕ, the MSE of low discrepancy dividing generally close to the random sampling. This confirms the capability of the tuning parameter, ϕ, on adjusting the proportion of low discrepancy dividing and random sampling. For those cases that the method with $\phi = 0$ produces the smallest MSE, the dividing process could add more variability than we expect. Therefore, the complete low-discrepancy dividing method that

results in the biggest subset model bias gives the optimal balancing of bias and variability, when the subset number is kept the same. We also expect our method to perform much better than random sampling when the distribution of data is skewed on multiple dimensions.

4.6 Conclusion

We have proposed a new low-discrepancy approach for performing divide and conquer in kernel ridge regression. Our approach employs the Gram-Schmidt random walk algorithm during the dividing process to ensure that each local tuning parameter is closely aligned with the global tuning parameter. By generally increasing the proportion of random division in the algorithm, we can optimize bias and variance. Through simulation, we have some preliminary results that, using a low-discrepancy approach in the dividing process, results in a model with smaller MSE compared to the dividing process that only employs random dividing. We can further improve the decrease in the MSE by finding an optimal combination of a random dividing and a low-discrepancy dividing. In future work, we intend to flesh out this approach to obtain more information about the stability of our results, the reduction in runtime, and the precision of our estimates.

Chapter 5

Conclusion

In this dissertation, we develop new methods for designing and analyzing agricultural studies. In Chapter 2, we explore a new method of sample computation for competing risk models when data from a pilot study are available. Competing risk models are more sophisticated statistical methods for more complicated time-event data that include multiple types of events as the absorbing event.

To compute sample size for competing risk models, we propose to use a flexible parametric model, the three-parameter generalized log-gamma model, to fit the pilot data to extract useful information on the cause-specific hazards. In this way, we do not need to assume constant hazards or proportionality. Assumptions of constant hazards or proportionality may lead to incorrect estimates of the effect size and miscalculations of the requisite sample size to achieve a pre-specified power. We discover that these assumptions may be severely violated when hazards are in arch shapes or U shapes and the treatment smooth and flat the hazards. We further proposed to use the Greenwood-Nam-D'Agostino statistics to evaluate the goodness of generalized log-gamma models by comparing against the nonparametric model. Alternatively, to graph the generalized log-gamma model along with the nonparametric model regarding the cause-specific hazards or survival functions of the pilot can also be a good method for evaluation.

After pinpointing each cause-specific hazard function, we follow the four-step data generation procedure. We generate the failure time using the inversion method indicated by the cumulative all-cause hazard function. Then, we determine the type of event by running multinomial trials with the probabilities equal to the ratio of each cause-specific hazard over the all-cause hazard. Last, we generate the censor time according to the underlying assumptions. The observed time is the minimum of the censor time and the failure time. Next, we use the simulation method to approximate the power of the statistical test we intend to use for the official experiment. To do so, we repeatedly generate data regarding the fitting of the generalized log-gamma model on the pilot data, then run the statistical test on each generated dataset. By adjusting the input sample size in the power approximation method, we can clearly observe how the sample size affects the power of the statistical test we plan to use for the official experiment. The sample size that achieves the pre-determined power level is the one we recommend for the official experiment.

In Chapter 3, we investigate best practices for designing cluster randomized trials when clusters are under the control of the experimenter. We verify that, under a simplified experimental setting and assuming the responses follow the Neyman-Rubin potential outcomes models, the precision of average treatment effect estimates under a CRT decreases as the intracluster correlation coefficient (the correlation of responses within each cluster) increases. However, when clusters can be specified by the experimenter, CRTs are made more precise when clusters are created "as heterogeneously as possible." This corresponds to the case where the distribution of units within a cluster is the same across clusters. In other words, we can enhance the precision of CRTs by assigning the most similar units to each distinct cluster. We develop a new method incorporating threshold clustering for assigning units to clusters to ensure similarly distributed clusters. Even though our method does not significantly outperform assigning units to clusters completely at random in our simulation study, we still show that both approaches vastly outperform CRTs in which units within each cluster are similar.

In chapter 4, we present our method to improve the computation efficiency for kernel ridge regression. Kernel ridge regression is an extension of ridge regression from linear modeling to nonlinear modeling by introducing a kernel transformation. According to the representing property, we only need to model the linear combination of the kernels instead of using complicated functions. However, as the sample size increases, the computational complexity increases. A broadly accepted method, called divide and conquer, is to divide the full dataset into multiple subsets and fit the kernel ridge regression on each subset. The final model is an average of these subset models. However, randomly divided subsets result in local tuning parameters that are smaller compared to the global tuning parameter, and the variances are enlarged in this scenario. Inspired by the application of low-discrepancy algorithms in instance selection, we propose using the Gram-Schmidt walk algorithm to divide the full dataset into subsets that are similar to each other. This approach ensures that the local tuning parameters are closer to the global tuning parameter, which helps reduce variance. Additionally, by adjusting the proportion of low-discrepancy dividing and random dividing in the algorithm, we can achieve a better balance between bias and variance, resulting in a smaller mean squared error (MSE). Our preliminary results from the simulation study also align with the theory.

5.1 Future Work

We are planning to extend our sample size computation method for competing risk to more complex multi-state models, such as illness-death models with recoveries, recurrent events, and the sequential nested model, which is based on time-inhomogeneous Markov processes with a finite state space. For the method of specifying clusters for CRTs, we plan to extend this method to more complex cluster-randomized trials, where the sizes of the clusters are unequal and the estimates of the average treatment effects are more nuanced. Last, for the improved divide and conquer methods, we intend to flesh out this approach to obtain more

information about the stability of our results, the reduction in runtime, and the precision of our estimates.

Bibliography

David M Allen. The relationship between variable selection and data agumentation and a method for prediction. *technometrics*, 16(1):125–127, 1974.

Wojciech Banaszczyk. Balancing vectors and gaussian measures of n-dimensional convex bodies. *Random Structures & Algorithms*, 12(4):351–360, 1998.

Jan Beyersmann, Arthur Allignol, and Martin Schumacher. *Competing risks and multistate models with R*. Springer Science & Business Media, 2011.

Bernard F Cole, Richard D Gelber, John M Kirkwood, Aron Goldhirsch, Edward Barylak, and Ernest Borden. Quality-of-life-adjusted survival analysis of interferon alfa-2b adjuvant treatment of high-risk resected cutaneous melanoma: an eastern cooperative oncology group study. *Journal of Clinical Oncology*, 14(10):2666–2673, 1996.

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

Daniel Dadush, Shashwat Garg, Shachar Lovett, and Aleksandar Nikolov. Towards a constructive version of banaszczyk's vector balancing theorem. *arXiv preprint arXiv:1612.04304*, 2016.

Karim Fizazi, NamPhuong Tran, Luis Fein, Nobuaki Matsubara, Alfredo Rodriguez-Antolin, Boris Y Alekseev, Mustafa Özgüroğlu, Dingwei Ye, Susan Feyerabend, Andrew Protheroe, et al. Abiraterone acetate plus prednisone in patients with newly diagnosed high-risk metastatic castration-sensitive prostate cancer (latitude): final overall survival analysis of a randomised, double-blind, phase 3 trial. *The Lancet Oncology*, 20(5):686–700, 2019.

Christopher Harshaw, Fredrik Sävje, Daniel Spielman, and Peng Zhang. Balancing covariates in randomized experiments with the gram–schmidt walk design. *arXiv preprint arXiv:1911.03071*, 2019.

Richard J Hayes and Lawrence H Moulton. *Cluster randomised trials*. Chapman and Hall/CRC, 2017.

Michael J Higgins, Fredrik Sävje, and Jasjeet S Sekhon. Improving massive experiments with threshold blocking. *Proceedings of the National Academy of Sciences*, 113(27):7369–7376, 2016.

Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

Daniel G Horvitz and Donovan J Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, 47(260): 663–685, 1952.

Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge university press, 2015.

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

Zohar Karnin and Edo Liberty. Discrepancy, coresets, and sketches in machine learning. In *Conference on Learning Theory*, pages 1975–1993. PMLR, 2019.

A Latouche and R Porcher. Sample size calculations in the presence of competing risks. *Statistics in medicine*, 26(30):5370–5380, 2007.

Aurélien Latouche, Raphaël Porcher, and Sylvie Chevret. Sample size formula for proportional hazards modelling of competing risks. *Statistics in medicine*, 23(21):3263–3274, 2004.

Gang Li and Qing Yang. Joint inference for competing risks survival data. *Journal of the American Statistical Association*, 111(515):1289–1300, 2016.

Sharon L Lohr. *Sampling: design and analysis*. Chapman and Hall/CRC, 2021.

Jianmei Luo, ChandraVyas Annakula, Aruna Sai Kannamareddy, Jasjeet S Sekhon, William Henry Hsu, and Michael Higgins. Hybridized threshold clustering for massive data. *arXiv preprint arXiv:1907.02907*, 2019.

James Mercer. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209(441-458):415–446, 1909.

Wayne Nelson. Hazard plotting for incomplete failure data. *Journal of Quality Technology*, 1(1):27–52, 1969.

Yongdong Ouyang, Fan Li, Xiaojuan Li, Julie Bynum, Vincent Mor, and Monica Taljaard. Estimates of intra-cluster correlation coefficients from 2018 usa medicare data to inform the design of cluster randomized trials in alzheimer's and related dementias. *Trials*, 25(1): 732, 2024.

Li-Hong Peng, Li-Qian Zhou, Xing Chen, and Xue Piao. A computational study of potential mirna-disease association inference based on ensemble learning and kernel ridge regression. *Frontiers in bioengineering and biotechnology*, 8:40, 2020.

Jeff M Phillips. Coresets and sketches. *arXiv preprint arXiv:1601.00617*, 2016.

Ross L Prentice. A log gamma model and its maximum likelihood estimation. *Biometrika*, 61(3):539–544, 1974.

G Raskutti, MJ Wainwright, and B Yu. Early stopping for non-parametric regression: An optimal data-dependent stopping rule. communication, control, and computing, 2013.

Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural computation*, 10(5):1299–1319, 1998.

Gabi Schulgen, Manfred Olschewski, Vera Krane, Christoph Wanner, Günther Ruf, and Martin Schumacher. Sample sizes for clinical trials with time-to-event endpoints and competing risks. *Contemporary clinical trials*, 26(3):386–396, 2005.

John Shawe-Taylor and Nello Cristianini. *Kernel methods for pattern analysis*. Cambridge university press, 2004.

Rahul Singh. Generalized kernel ridge regression for long term causal inference: Treatment effects, dose responses, and counterfactual distributions. *arXiv preprint arXiv:2201.05139*, 2022.

Benjamin J Solomon, Dong-Wan Kim, Yi-Long Wu, Kazuhiko Nakagawa, Tarek Mekhail, Enriqueta Felip, Federico Cappuzzo, Jolanda Paolini, Tiziana Usari, Yiyun Tang, et al. Final overall survival analysis from a study comparing first-line crizotinib versus chemotherapy in alk-mutation-positive non–small-cell lung cancer. *Journal of Clinical Oncology*, 36 (22):2251–2258, 2018.

Joel Spencer. Six standard deviations suffice. *Transactions of the American mathematical society*, 289(2):679–706, 1985.

Jerzy Splawa-Neyman, Dorota M Dabrowska, and Terrence P Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990.

Edney W Stacy. A generalization of the gamma distribution. *The Annals of mathematical statistics*, pages 1187–1192, 1962.

Rashish Tandon, Si Si, Pradeep Ravikumar, and Inderjit Dhillon. Kernel ridge regression via partitioning. *arXiv preprint arXiv:1608.01976*, 2016.

Chun-Chun Wang, Chi-Chi Zhu, and Xing Chen. Ensemble of kernel ridge regression-based small molecule–mirna association prediction in human disease. *Briefings in Bioinformatics*, 23(1), 2022.

Christopher Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. *Advances in neural information processing systems*, 13, 2000.

Ganggang Xu, Zuofeng Shang, and Guang Cheng. Optimal tuning for divide-and-conquer kernel ridge regression with massive data. In *International Conference on Machine Learning*, pages 5483–5491. PMLR, 2018.

Qing Yang, Wing K Fung, and Gang Li. Sample size determination for jointly testing a cause-specific hazard and the all-cause hazard in the presence of competing risks. *Statistics in medicine*, 37(8):1389–1401, 2018.

Yuan Yao, Lorenzo Rosasco, and Andrea Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

William Henry Young. On classes of summable functions and their fourier series. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 87(594):225–229, 1912.

Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *The Journal of Machine Learning Research*, 16(1):3299–3340, 2015.

Appendix A

Results for Chapter 3

A.1 Horvits-Thompson Estimator for Cluster Randomized Trials

To discuss the importance of controlling ICC for cluster randomized trials in a more detailed and specific manner. We are focusing on the Horvits-Thompson estimator at this moment, and comparing the estimator for cluster randomized trials on completely randomized design. For clarity, We now considering the special case of a one-stage CRT where there are N clusters, all clusters contain M units, and half of the clusters are assigned treatment under simple random sampling. We compare results to those from a completely randomized design, where half of the units (not clusters) are assigned treatment and the other half are given control.

Under the design of Cluster Randomized Trials (CRT), suppose we have N clusters with indication i, where $i = 1, ...N$. In addition, each cluster contains M units with indication j, where $j = 1, ..., M$. Control, denoted as 0, and treatment, denoted as 1, are randomly allocated evenly. In other words, $\frac{N}{2} = n$ cluster receive control and $\frac{N}{2} = n$ cluster receive treatment. The potential responses are y_{0ij}'s for the unit j that receives the control in the cluster i. Similarly, y_{1ij}'s are the potential responses for unit j that receives the treatment

in the cluster i. We also define I_0 as the set of clusters assigned to control and I_1 as the set of clusters assigned to treatment. For a comparable complete randomized design. There are MN units, $\frac{MN}{2}$ are assigned to control group and $\frac{MN}{2}$ are assigned to treatment group. We use k to indicate the units, where $k = 1, ..., K$, and $K = MN$. We further define the set of units that receive control as K_0, and the set of units that receive treatment as K_1. Further, to simplify equations, we defined $\sum_{j=1}^{M} y_{ij} = t_i$ to stand for the total response of cluster i. Similarly, $\sum_{j=1}^{M} y_{0ij} = t_{0i}$ stands for the total potential response of the control of cluster i. $\sum_{j=1}^{M} y_{1ij} = t_{1i}$ stands for the total potential response of the treatment of cluster i. In addition, $\bar{y_{0\cdot}}$ is the mean of the potential outcomes of control. $\bar{y_{1\cdot}}$ is the mean of the potential outcomes of treatment. The formulas of these means depend on the experimental designs. For a cluster randomized design, $\bar{y_{0\cdot}} = \frac{1}{MN}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{0ij}$. For a completely randomized design $\bar{y_{0\cdot}} = \frac{1}{MN}\sum_{k=1}^{MN} y_{0k}$. Similarly, $t_{0\cdot} = \frac{1}{N}\sum_{i=1}^{N} t_{0i}$, and $t_{1\cdot} = \frac{1}{N}\sum_{i=1}^{N} t_{1i}$. These definitions of cluster total only work on cluster randomized design. To compare the precision of a completely randomized design and a cluster randomized trial. We also define SSTO as $SSTO = \sum_{k=1}^{MN}(y_k - \bar{y_{\cdot}})^2$ for completely randomized design and $SSTO = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \bar{y_{\cdot}})^2$. There is no SSW or SSB for a completely randomized design. We define the SSW for a cluster randomized trial as $SSW = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \bar{y_{i\cdot}})^2$ and $SSB = \sum_{i=1}^{N}\sum_{j=1}^{M}(\bar{y_{i\cdot}} - \bar{y_{\cdot}})^2$.

The Horvitz-Thompson estimator for the complete randomized design is defined as:

$$\hat{\Delta_{HT,SRS}} = \frac{1}{MN}[\sum_{k=1}^{K}\frac{1}{\pi_{1k}}T_k y_k - \sum_{k=1}^{K}\frac{1}{\pi_{0k}}(1 - T_k)y_k] \tag{A.1}$$

where π_{0k} is the probability that the unit k is assigned to control. Similar as π_{1k}. In our case, $\pi_{0k} = \pi_{1k} = \frac{Mn}{MN}$. Therefore, the Horvitz-Thompson estimator can be re-write as:

$$\hat{\Delta_{HT,SRS}} = \frac{1}{MN}[\sum_{k\in K_1}\frac{MN}{Mn}y_k - \sum_{k\in K_0}\frac{MN}{Mn}y_k] \tag{A.2}$$

We further define $\sum_{k\in K_0}\frac{MN}{Mn}y_k = \hat{Y_{0HT,SRS}}$. We can see $\hat{Y_{HT,SRS}}$ is the unbiased estimate of

the population total of y_{0k} regarding the finite population principle, if we sample Mn units from a finite population of MN. Similar as $\hat{Y_{1HT,SRS}}$. The Horvitz-Thompson estimator for the cluster randomized trial is defined as

$$\hat{\Delta_{HT,CRT}} = \frac{1}{MN}[\sum_{i=1}^{N}\sum_{j=1}^{M}\frac{1}{\pi_{1ij}}y_{ij}T_{ij} - \sum_{i=1}^{N}\sum_{j=1}^{M}\frac{1}{\pi_{0ij}}y_{ij}(1-T_{ij})] \tag{A.3}$$

Where π_{0ij} is the probability that the unit j from cluster i is assigned to the control. Similar as π_{1ij}. In our case, $\pi_{0ij} = \pi_{1ij} = \frac{n}{N}$. In addition, $\pi_{0ij} = \pi_{0ij'}$, for $j \neq j'$. Therefore, we can simplify the notation and let $\pi_{0ij} = \pi_{0i} = \frac{n}{N}$. Similar as π_{1i}. By defining $\sum_{j=1}^{M} y_{ij} = t_i$, $\sum_{j=1}^{M} y_{0ij} = t_{0i}$, and $\sum_{j=1}^{M} y_{1ij} = t_{1i}$. We can re-write the Horvitz-Thompson estimator as:

$$\hat{\Delta_{HT,CRT}} = \frac{1}{MN}[\frac{N}{n}\sum_{i\in I_1} t_i T_i - \frac{N}{n}\sum_{i\in I_0} t_i(1-T_i)] \tag{A.4}$$

We further define $\frac{N}{n}\sum_{i\in I_0} t_i = \hat{Y_{0HT,CRT}}$. It is the unbiased estimate of the population total of t_{0i}, if we sample n clusters from a finite population of N clusters. We define $\hat{Y_{1HT,CRT}}$ in the similar way.

The variance of the Horvitz-Thompson estimator is defined as

$$Var(\hat{\Delta_{HT}}) = \frac{1}{(MN)^2}[Var(\hat{Y_{0HT}}) + Var(\hat{Y_{1HT}}) - 2Cov(\hat{Y_{0HT}} \quad \hat{Y_{1HT}})] \tag{A.5}$$

For the complete randomized design, the $\hat{Y_{0HT,SRS}}$ is the unbiased estimator of the total of the potential control outcomes. Therefore, regarding the finite population principle, the $Var(\hat{Y_{0HT,SRS}})$ is defined as:

$$Var(\hat{Y_{0HT,SRS}}) = (MN)^2\frac{MN-Mn}{MN}\frac{S_0^2}{Mn} \tag{A.6}$$

where $S_0^2 = \frac{1}{MN-1}\sum_{k=1}^{MN}(y_{0k} - \bar{y_{0\cdot}})^2$ and $\bar{y_{0\cdot}}$ is the mean of the control potential outcomes. Similarly, the $Var(\hat{Y_{1HT,SRS}})$ can be defined in the same manner.

The $Cov(\hat{Y_{0HT,SRS}} \quad \hat{Y_{1HT,SRS}})$ is defined as:

$$Cov(\hat{Y_{0HT,SRS}} \quad \hat{Y_{1HT,SRS}}) = -MN \cdot Cov(y_{1k} \quad y_{0k}) \tag{A.7}$$

where $Cov(y_{1k} \quad y_{0k}) = \frac{1}{MN-1}\sum_{k=1}^{MN}(y_{0k} - \bar{y_{0\cdot}})(y_{1k} - \bar{y_{1\cdot}})$. For the cluster randomized trial, the $\hat{Y_{0HT,CRT}}$ is the unbiased estimator of the total of potential control outcomes. Therefore, the $Var(\hat{Y_{0HT,CRT}})$ can be defined according to the finite population principle.

$$Var(\hat{Y_{0HT,CRT}}) = N^2\frac{N-n}{N}\frac{1}{n}S_{t_0}^2 \tag{A.8}$$

where $S_{t_0}^2 = \sum_{i=1}^{N}\frac{1}{N-1}(t_{0i} - t_{0\cdot})^2$ and $t_{0\cdot} = \frac{1}{N}\sum_{i=1}^{N} t_{0i}$. We can see that the $Var(\hat{Y_{0HT,CRT}})$ only involves the total of each cluster, therefore, only depends on the between-cluster variability. Similarly, $Var(\hat{Y_{1HT,CRT}})$ can be defined in the same manner. The $Cov(\hat{Y_{0HT,CRT}}\, \hat{Y_{1HT,CRT}})$ is defined as:

$$Cov(\hat{Y_{0HT,CRT}}\, \hat{Y_{1HT,CRT}}) = -N \cdot Cov(t_{0i}\, t_{1i}) \tag{A.9}$$

where $Cov(t_{0i}\, t_{1i}) = \frac{1}{N-1}\sum_{i=1}^{N}(t_{0i} - t_{0\cdot})(t_{1i} - t_{1\cdot})$.

To compare the complete randomized design to the cluster randomized trial. We define SSTO as $SSTO = \sum_{k=1}^{MN}(y_k - \bar{y_{\cdot}})$ for completely randomized design and $SSTO = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \bar{y_{\cdot}})$. There is no SSW or SSB for a completely randomized design. We define the SSW for a cluster randomized trial as $SSW = \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij} - \bar{y_{i\cdot}})$ and the SSB as $SSB = \sum_{i=1}^{N}\sum_{j=1}^{M}(\bar{y_{i\cdot}} - \bar{y_{\cdot}})$. We add a subscript 0 or 1 to indicate that it relates to the potential outcome of control or treatment. We first introduce the intracluster correlation coefficient (ICC). The ICC is a measure of homogeneity within the clusters and can be defined as the Pearson correlation coefficient for the $NM(M-1)$ pair of responses for each cluster. It can be defined as:

$$ICC = \frac{\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{j\neq j'}(y_{ij} - \bar{y_{\cdot}})(y_{ij'} - \bar{y_{\cdot}})}{(MN-1)(M-1)S^2} \tag{A.10}$$

Note,

$$
\begin{aligned}
(MN-1)S^2 &= SSTO \\
SSTO &= SSB + SSW \\
\sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{j\neq j'}(y_{ij}-\bar{y}_{\cdot})(y_{ij'}-\bar{y}_{\cdot}) &= \sum_{i=1}^{N}\sum_{j=1}^{M}\sum_{j'=1}^{M}(y_{ij}-\bar{y}_{\cdot})(y_{ij'}-\bar{y}_{\cdot}) - \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{\cdot})^2 \\
&= \sum_{i=1}^{N} M^2(\bar{y_{i\cdot}}-\bar{y}_{\cdot})^2 - \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\bar{y}_{\cdot})^2 \\
&= M(SSB) - SSTO \\
&= (M-1)SSTO - MSSW
\end{aligned}
$$

Therefore, we can relate the ICC to the SSTO, SSW, and SSB, then written as:

$$ICC = 1 - \frac{M}{M-1}\frac{SSW}{SSTO} \tag{A.11}$$

The expression of ICC can further imply that

$$MSB = \frac{MN-1}{M(N-1)}S^2[1+(M-1)ICC] \tag{A.12}$$

To demonstrate the comparison, we focus on the ratio of $Var(\hat{Y_{0HT,CRT}})$ and $Var(\hat{Y_{0HT,SRS}})$ first. To better express the ratio regarding the ICC, we can re-write the terms as:

$$Var(\hat{Y_{0HT,CRT}}) = N^2\frac{N-n}{N}\frac{M}{n}MSB_0 \tag{A.13}$$

This is due to that we can re-write $S_{t_0}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(t_{0i}-\bar{t_{0\cdot}})^2 = \frac{1}{N-1}\sum_{i=1}^{N}M^2(\bar{y_{0i\cdot}}-\bar{y_{0\cdot\cdot}})^2 = M(MSB_0)$, where $\bar{y_{0i\cdot}}$ stands for the average of potential control outcomes from cluster i, and $\bar{y_{0\cdot\cdot}}$ stands for the grand average of potential control outcomes.

$$Var(\hat{Y_{0HT,SRS}}) = N^2\frac{N-n}{N}\frac{M}{n}S_0^2 \tag{A.14}$$

Therefore, we can write the ratio as:

$$\frac{Var(\hat{Y_{0HT,CRT}})}{Var(\hat{Y_{0HT,SRS}})} = \frac{MSB_0}{S_0^2} = \frac{MN-1}{M(N-1)}[1+(M-1)ICC_0] \tag{A.15}$$

Now, we are planning to show the ratio of $Var(\hat{\Delta_{HT,CRT}})$ and $Var(\hat{\Delta_{HT,SRS}})$ in the similar manner.

$$\frac{Var(\hat{\Delta_{HT,CRT}})}{Var(\hat{\Delta_{HT,SRS}})} = \frac{N^2\frac{N-n}{N}\frac{1}{n}Var(t_{0i}) + N^2\frac{N-n}{N}\frac{1}{n}Var(t_{1i}) + 2NCov(t_{0i}\, t_{1i})}{(MN)^2\frac{MN-Mn}{MN}\frac{1}{Mn}Var(y_{0k}) + (MN)^2\frac{MN-Mn}{MN}\frac{1}{Mn}Var(y_{1k}) + 2MNCov(y_{0k}\, y_{1k})} \tag{A.16}$$

To further simplify this ratio, we introduce Young's inequality, which is commonly used in causal inference. There are simulation studies to justify the use of this inequality, that the bound is generally close to the covariance.

The young's inequality states:

$$Cov(t_{0i}, t_{1i}) \leq \frac{Var(t_{0i}) + Var(t_{1i})}{2} \tag{A.17}$$

Therefore, we can claim that,

$$-(Var(t_{0i}) + Var(t_{1i})) \leq 2Cov(t_{0i}\, t_{1i}) \leq Var(t_{0i}) + Var(t_{1i}) \tag{A.18}$$

similar for the term , $2Cov(y_{0k}\, y_{1k})$. Therefore, we can further write bounds for $Cov(\hat{Y_{0HT,CRT}}\, \hat{Y_{1HT,CRT}})$ and $Cov(\hat{Y_{0HT,SRS}}\, \hat{Y_{1HT,SRS}})$.

$$-N[Var(t_{0i}) + Var(t_{1i})] \leq Cov(\hat{Y_{0HT,CRT}}\, \hat{Y_{1HT,CRT}}) \leq N[Var(t_{0i}) + Var(t_{1i})] \tag{A.19}$$

$$-N[M(MSB_0 + M(MSB_1))] \leq Cov(\hat{Y_{0HT,CRT}}\, \hat{Y_{1HT,CRT}}) \leq N[M(MSB_0 + M(MSB_1))] \tag{A.20}$$

$$-MN[Var(y_{0k}) + Var(y_{1k})] \leq Cov(Y_{0\hat{HT,SRS}}\, Y_{1\hat{HT,SRS}}) \leq MN[Var(y_{0k}) + Var(y_{1k})] \tag{A.21}$$

$$-MN[S_0^2 + S_1^2] \leq Cov(Y_{0\hat{HT,SRS}}\, Y_{1\hat{HT,SRS}}) \leq MN[S_0^2 + S_1^2] \tag{A.22}$$

Therefore, we can compute the lower bound of the ratio by using the lower bound of $Var(\Delta_{\hat{HT,CRT}})$ versus the upper bound of the $Var(\Delta_{\hat{HT,SRS}})$, and compute the upper bound of the ratio by using the upper bound of $Var(\Delta_{\hat{HT,CRT}})$ versus the lower bound of the $Var(\Delta_{\hat{HT,SRS}})$. therefore, the lower bound of the ratio can be expressed as:

$$\frac{N^2\frac{N-n}{N}\frac{M}{n}MSB_0 + N^2\frac{N-n}{N}\frac{M}{n}MSB_1 - MN[MSB_0 + MSB_1]}{N^2\frac{N-n}{N}\frac{M}{n}S_0^2 + N^2\frac{N-n}{N}\frac{M}{n}S_1^2 + MN[S_0^2 + S_1^2]} \tag{A.23}$$

By further assuming the consistency of the treatment effect, we have $MSB_0 = MSB_1$, $S_0^2 = S_1^2$, and the lower bound of the ratio can be simplified as:

$$\frac{N-n-1}{N-n+1}\frac{MSB}{S^2} \tag{A.24}$$

$$\frac{N-n-1}{N-n+1}\frac{MN-1}{M(N-1)}[1 + (M-1)ICC] \tag{A.25}$$

The upper bound of the ratio can be expressed as:

$$\frac{N^2\frac{N-n}{N}\frac{M}{n}MSB_0 + N^2\frac{N-n}{N}\frac{M}{n}MSB_1 + MN[MSB_0 + MSB_1]}{N^2\frac{N-n}{N}\frac{M}{n}S_0^2 + N^2\frac{N-n}{N}\frac{M}{n}S_1^2 - MN[S_0^2 + S_1^2]} \tag{A.26}$$

By further assuming the consistency of the treatment effect, the upper bound of the ratio can be simplified as:

$$\frac{N-n+1}{N-n-1}\frac{MSB}{S^2} \tag{A.27}$$

$$\frac{N-n+1}{N-n-1}\frac{MN-1}{M(N-1)}[1 + (M-1)ICC] \tag{A.28}$$

Both the upper and the lower bounds of the ratio of the variance of the estimator for

cluster randomized trial and the variance of the estimator for the completely randomized design indicate a positive intracluster correlation decreases the precision of a cluster randomized trial. Therefore, balancing the covariates between clusters is important for cluster randomized trials.

In addition, the covariance of potential outcomes can also decompose to a within-cluster covariance and between-cluster covariance. This will elaborate on our simulation of how the covariance of potential outcomes influences the precision of cluster randomization trials.

The decomposition can be expressed as:

$$
\begin{aligned}
&\sum_{i=1}^{N}\sum_{j=1}^{M} y_{1ij}y_{0ij} - MN\bar{y}_1\bar{y}_0 \\
=& \sum_{i=1}^{N}\sum_{j=1}^{M} y_{1ij}y_{0ij} - M\sum_{i=1}^{N}\bar{y}_{1i}\bar{y}_{0i} + \sum_{i=1}^{N} M\bar{y}_{1i}\bar{y}_{0i} - MN\bar{y}_1\bar{y}_0 \\
=& \sum_{i=1}^{N}\left(\sum_{j=1}^{M} y_{1ij}y_{0ij} - M\bar{y}_{1i\cdot}\bar{y}_{0i\cdot}\right) + \sum_{i=1}^{N} M\bar{y}_{1i\cdot}\bar{y}_{0i\cdot} - MN\bar{y}_{1\cdot\cdot}\bar{y}_{0\cdot\cdot}
\end{aligned}
$$

Which decomposes the covariance between potential treatments to the summation of the covariance within cluster and M times the covariance across clusters.

Without loss of generality, we generated our simulation that the variability between-cluster and within-cluster dominate both potential outcomes. We have four settings regarding the variances between-cluster and within-cluster. In the first case, the variability is dominated by within-cluster variance. In the second case, the variability is primarily influenced by the variance between-clusters. The within-cluster and the between-cluster variances are equal in the last two cases. However, in the third case, their sizes are small, and in the fourth case, their sizes are large. In each case, we have three scenarios regarding the correlation between the potential outcomes: positive, zero, and negative. We discover that, if the within-cluster variance dominates, the cluster randomized trials are more precise

than the completely randomized designs. On the contrary, when the between-cluster variance dominates, or the between-cluster and within-cluster variances are equal, the cluster randomized trials are less precise than the completely randomized design. This matches our discussion on the ICC and our suggestions on minimizing the between-cluster variability. For the covariance between potential outcomes, when we have a positive correlation between the treated and control, we generally have larger variance of the estimator regardless of the designs. The variance of the estimator decreases as the correlation decreases. In other words, the negative correlation between potential outcomes provides the smallest variance of estimators. Even in the situations that the cluster randomized trials should be less precise than the completely randomized design, a negative correlated potential outcomes can result a cluster randomized trial with a smaller variance of estimator.

Standard Deviation of Estimator						
	Positive	Positive	None	None	Negative	Negative
	CRD	CRT	CRD	CRT	CRD	CRT
Within	3.29	2.21	2.24	1.93	2.74×10^{-16}	1.69×10^{-16}
Between	3.95	14.16	2.78	7.03	5.24×10^{-16}	5.20×10^{-16}
Equal(Small)	0.51	1.44	0.34	0.68	6.47×10^{-17}	4.98×10^{-17}
Equal(Lager)	5.15	14.38	3.44	6.87	6.00×10^{-16}	6.02×10^{-16}

Appendix B

Results for Chapter 4

Chernoff's Bound

let $\boldsymbol{X_1, X_2, ..., X_n}$ be independent uniform random variables. For any $a > 0$,

$$Pr(|X_1 + X_2 + \cdots + X_n| > a) < 2exp(-\frac{a^2}{2n}) \tag{B.1}$$

Discrepancy Bound of Random mapping for a set S_j, by Chernoff's Bound

$$Pr(|\chi(S_j)| > \sqrt{2|S_j|ln(2m)}) = Pr(|\sum_{i \in S_j} \chi(i)| > \sqrt{2|S_j|ln(2m)}) \leq 2exp(-\frac{\sqrt{2|S_j|ln(2m)}}{2||S_j||}) = \frac{1}{m} \tag{B.2}$$

by additivity of probability,

$$Pr(\exists\ S_j \subset \mathcal{S}, |\chi(S_j)| > \sqrt{2|S_j|ln(2m)}) < m \cdot \frac{1}{m} = 1 \tag{B.3}$$

Thus there exists a mapping that $\chi(\mathcal{S}) \leq \sqrt{2|S_j|ln(2m)} \leq \sqrt{2nlm(2m)}$