THE DIVIDE-AND-COMBINE APPROACHES FOR MULTIVARIATE SURVIVAL ANALYSIS AND MULTISTATE SURVIVAL ANALYSIS IN BIG DATA

By

WEI WANG

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Public Health

Written under the direction of

Shou-En Lu

And approved by

_____

_____

_____

_____

_____

New Brunswick, New Jersey

October 2020

ABSTRACT OF THE DISSERTATION

The Divide-and-Combine Approaches for Multivariate Survival Analysis and Multistate Survival Analysis in Big Data

by Wei Wang

Dissertation Director: Prof. Shou-En Lu

Multivariate failure time data can be unordered or ordered, which can be analyzed using multivariate survival analysis and multistate survival analysis, respectively. When sample sizes are extraordinarily large, both analyses could face computational challenges. In this dissertation, we propose divide-and-combine approaches to analyze large-scale multivariate failure time data in both multivariate survival analysis and multistate survival analysis. Our approaches are motivated by the Myocardial Infarction Data Acquisition System (MIDAS), a New Jersey statewide database that includes 73,725,160 admissions to non-federal hospitals and emergency rooms (ERs) from 1995 to 2017. We propose to randomly divide the full data into multiple subsets and propose a weighted method to combine these estimators obtained from individual subsets. In divided subsets, estimated regression parameters and estimated cumulative hazards are calculated, respectively, for multivariate survival analysis and multistate survival analysis. Under mild conditions, we show that the combined estimators are asymptotically equivalent to the estimators obtained from the full data as if the data were analyzed all at once. In addition, to screen out risk factors with weak signals in multivariate survival analysis, we propose to perform the regularized estimation on the combined estimators using their combined confidence distributions. Theoretical properties of proposed approaches, such as asymptotic equivalence between divide-and-

combine analysis and full-data analysis, estimation consistency, selection consistency, and oracle properties are studied. Performances of proposed estimators are investigated using simulation studies. The MIDAS data are used to illustrate our proposed methodologies.

ACKNOWLEDGEMENTS

When writing this page of gratitude, I have been physically self-quarantined at home for four months due to the global pandemic of COVID-19, but I feel my heart is still with Rutgers and all the people there. I would like to take this opportunity to express my deepest gratitude to the people who have ever offered help or brought joy to me in the last six years.

First and foremost, my sincere appreciation and many thanks go to my advisor Dr. Shou-En Lu, who has been a tremendous mentor for me. I have benefited not only from her immense knowledge in statistics, but also from her enthusiasm in collaborative work with researchers in other fields. I believe they will have a far-reaching influence on my future career as an applied statistician. I am grateful for her patience and encouragement which motivates me to strive towards my goal and have made this journey much more enjoyable.

Next I would like to thank Dr. Yong Lin, Dr. Yaqun Wang, Dr. Sinae Kim, and Dr. Minge Xie for serving my dissertation proposal and defense committees. Their insightful comments and suggestions have perfected the ideas and the writing of this dissertation. I would also like to thank all the faculty members and staff at the Biostatistics and Epidemiology Department for their encouragement and support during all these years of my graduate studies. Moreover, I owe my thanks to Dr. Jerry Q. Cheng for providing valuable help which has led to significant improvements of this dissertation. I am also indebted to Huaibao Feng from Johnson & Johnson for being a good mentor during my summer internship.

Last but not least, I must express my very profound gratitude to my family and to my parents for providing me with unfailing support and continuous encouragement throughout my years of studies and throughout the process of researching and writing this dissertation. This accomplishment would not have been possible without them.

DEDICATION

To Linwei, Georgie, and Leo

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

1.1 Divide-and-Combine in Big Data

With the advancement of computing and storage technologies, data sets on a massive scale in terms of volume, intensity, and complexity ("*big data*") have become increasingly accessible. Big data are generated by a variety of sources, from internet search engines, social network tools, and internet of things, to electronic health records, medical imaging, and genomic sequencing, to name a few. Three characteristics distinguish them from traditional data: volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources), which is also known as the 3V definition of big data (Laney, 2001). High volume and high velocity may introduce scalability and storage bottleneck, noise accumulation, spurious correlation, incidental endogeneity, and measurement errors (Fan et al., 2014). High variety may bring non-traditional or even unstructured data types (Wang et al., 2016). As a result, researchers appreciate the great promises held by big data for discovering subtle population patterns and heterogeneities that are not possible with small-scale data, but also face the challenges of limited capacity of standard analytic tools.

Performing standard regression analysis in big data presents considerable computational challenges: 1) the entire data can be too big to be loaded and analyzed in one single processor; 2) the estimation algorithms of traditional software packages can be too complex and thus the computing tasks can take too long to wait for the results. To overcome these difficulties, sound statistical procedures with scalable computational capacities have been proposed. For example, bags of little bootstrap (Kleiner et al., 2014), divide and combine (Lin and Xi, 2011), and online updating (Schifano et al., 2016). We shall focus on the divide-and-combine approaches in this dissertation due to their simplicity and effec-

tiveness. Their principle is to divide a data set into multiple subsets, perform statistical inference in each subset separately and sometimes recursively, then combine these individual results into a final result. Divide-and-combine has a long history in computer science (Aho and Hopcroft, 1974), and has advanced rapidly in recent statistical development using the distributed computing architecture with a focus on the regularized estimation (Shamir et al., 2014; Chen and Xie, 2014; Lee et al., 2015; Tang et al., 2016; Battey et al., 2018; Wang et al., 2019; Jordan et al., 2019).

Most of the existing divide-and-combine approaches require one round communication only: estimates in subsets are aggregated to form a final result. For example, Chen and Xie (2014) combine LASSO estimators (Tibshirani, 1996) by a majority voting method; Battey et al. (2018) and Lee et al. (2015) simply average debiased LASSO estimators (Van de Geer et al., 2014); Tang et al. (2016) combine debiased LASSO estimators (Van de Geer et al., 2014) using the confidence distribution approach (Xie and Singh, 2013). While these "one-shot" approaches are highly efficient in communication, their statistical inference might not be available or may be sub-optimal in many occasions (Jordan et al., 2019). To achieve the best efficiency in statistical estimation, iterative algorithms are proposed. For instance, approximate Newton-type algorithms (Shamir et al., 2014) and one-step algorithms (Wang et al., 2019). However, these iterative algorithms require multiple-round communications to broadcast information back and forth between subsets and global aggregation, which makes them computationally intensive when compared to those "one-shot" approaches.

To date, all these divide-and-combine methods have been exclusively discussed and applied to univariate outcomes and mostly under the framework of generalized linear models (McCullagh, 1989) and the Cox proportional hazards model (Cox, 1972). None of them are applicable for dependent outcomes. This dissertation intends to fill this void and focuses on the multivariate failure time data.

1.2 Multivariate Failure Time Data

Time-to-event outcomes differ from lots of other measurements. There are chances that the prespecified events do not occur in a given study period and we say these outcomes are "censored". Because of censoring, appropriate statistical modelling approaches are needed to handle the possibly incomplete time-to-event data. Among them, the Cox proportional hazards model (Cox, 1972) and its associated statistical methods based on the partial likelihood (Cox, 1975) is the most popular one. Usually, we call the prespecified event a "failure", and the time period until a event occurs a "failure time".

Multivariate failure time data are commonly encountered in biomedical research where study subjects from the same cluster (e.g., patient or family) share common genetic and/or environmental factors such that the failure times within the same cluster are correlated. Multivariate failure times can be either unordered or ordered. Examples for correlated unordered failure times include times to blindness of the laser-treated eye and the other untreated one for an individual patient in a diabetic retinopathy study (Diabetic Retinopathy Study Research Group, 1981). Similarly, times to affective illness among relatives of the same proband in a genetic epidemiologic study of schizophrenia (Pulver et al., 1991) are also correlated and unordered. On the other hand, due to the biological process of a disease, it is possible that after experiencing the first failure, the risk of the next failure may change. For instance, times to tumor recurrences for one patient in a bladder cancer study (Byar, 1980) or times to multiple pyogenic infection episodes for another patient in a chronic granulomatous disease study (Fleming and Harrington, 2011) are correlated and ordered.

Analysis of multivariate unordered failure time data (commonly known as "multivariate survival anlaysis") has been extensively studied in the statistical literature (e.g., Hougaard (2000)). Extensions of the Cox proportional hazards model (Cox, 1972) for the multivariate survival analysis include marginal models and frailty models. The marginal model method estimates the marginal distributions of multivariate unorderd failure times. This method

either separately models the intra-cluster association or leaves the structure of the intra-cluster association unspecified but adjusts for it in the inference (Lin, 1994). The frailty model method, in contrast, formulates the dependence explicitly by using a frailty term which corresponds to a random block effect (Hougaard, 2000).

One key assumption for the above multivariate survival analysis is that the failure time distribution and the censoring time distribution for each of the multiple correlated failures are independent or at least conditionally independent given some covariates. However, there are situations in which the independence assumption may be violated. For example, the occurrences of multiple correlated failures may be mutually exclusive ("*competing risks*"), under which the Cox proportional hazards model (Cox, 1972) can still be used but the interpretation of the results is different (see details in Section 3 of Putter et al. (2007)). Another example is that multiple correlated failures may occur sequentially ("*event history*"). Rather than interested in time to the first failure, which may lose some information, researchers are more likely to be interested in what happens after the first non-fatal failure. The intermediate failures can provide more detailed information on the event history and allow for more precision in predicting future failures. Both scenarios (competing risks case and event history case) can be handled by the multistate models (Hougaard, 1999). When using multistate models, at any time in the time period we are considering, each subject is said to be in a state. The non-fatal failures are regarded as transitions from one state to another. Analysis of such multivariate ordered failure times is called multistate survival analysis.

1.3 Regularization Using Confidence Distribution

Variable selection is a classical but critically important problem. This is because in practice, the number of available covariates is typically large, but only a small subset of them are related to the response. Classical methods for variable selection, such as best-subset selection and forward/backward stepwise selection, are widely used as they have been successfully

integrated into many commonly-used statistical software packages. Despite their popularity, the sampling properties (e.g., unbiasedness and distributions) are mostly unknown (Fan and Li, 2002). In the past decades, various variable selection techniques through regularization have been studied in depth and extended into univariate survival analysis using the Cox proportional hazards model (Cox, 1972), such as the LASSO estimator (Tibshirani, 1997), the SCAD estimator (Fan and Li, 2002), the adaptive LASSO estimator (Zhang and Lu, 2007; Wang and Leng, 2007), and the MCP estimator (Zhang et al., 2010). Among them, only the SCAD estimator has been studied in multivariate survival analysis (Fan and Li, 2002; Cai et al., 2005), but the nonconvex penalty term makes its optimization hard to solve for globally. In this dissertation, we propose a confidence distribution approach to perform regularized estimation to achieve variable selection and statistical inference. Our proposed method applies to multivariate survival analysis and its convex penalty term makes it computationally appealing.

Confidence distribution (CD) (Xie and Singh, 2013) is a sample dependent distribution function that can be used to estimate and provide all aspects of statistical inference for a parameter of interest. It provides simple and interpretable summaries of what can be reasonably learned from data and an assumed model (Cox, 2013). CD has a long history (see, e.g, Cox (1958), Fisher (1956), and Efron (1993)) but recent development has redefined the CD concept and focused on providing inference tools for problems in modern applied statistics (Xie et al., 2011). The useful features can be seen in Xie et al. (2011), Tian et al. (2011) and Liu et al. (2015).

In this dissertation, as opposed to optimizing the objective function typically constructed from the original data in the sample(s), we propose to perform regularized estimation by optimizing the objective function based on the CD of regression parameters. This approach leads to substantial dimensionality reduction and savings in computation time when the sample size is much larger than the number of covariates. With a proper choice of regularization parameters, our proposed regularized estimators have some desired

statistical properties: estimation consistency, selection consistency, and oracle properties. Moreover, our proposed confidence distribution approach enables regularized estimation in multivariate survival analysis using existing software packages (e.g., the R package, `glmnet` (Friedman et al., 2010)), without the need for any new algorithm specific to multivariate survival analysis to perform regularized estimation.

1.4 A Motivating Example: MIDAS Study

Cardiovascular diseases (CVDs) are the number one cause of death globally, resulting in 17.7 million deaths annually (Thomas et al., 2018). CVDs prediction is one of the most effective measures for CVDs control. Existing predictive models for CVDs, such as Framingham Risk Score (D'Agostino et al., 2008) recommended by the American College of Cardiology/American Heart Association (ACC/AHA), Systematic Coronary Risk Evaluation (SCORE) algorithm (Piepoli et al., 2016) recommended by the European Society of Cardiology (ESC), and QRISK score (Hippisley-Cox et al., 2017) recommended by the National Institute for Health and Care Excellence (NICE) in the United Kingdom, are only applicable to either a combined CVD outcome consisting of myocardial infarction (MI), heart failure (HF), stroke, and cerebrovascular disease, or one single disease. Because the multivariate failure times recorded for one patient are usually clustered and correlated but those existing predictive models implicitly ignore the intra-cluster association, a new predictive model considering intra-cluster association is needed. Moreover, existing predictive models are typically developed using information on patient cohorts of relatively small sample size with limited number of risk factors due to the limited computing capacity. It is imperative to build a new predictive model using big data from a large population with a long follow-up to fully capture the important risk factors, especially those with weak to moderate effects. In developing predictive models, analyzing big data and identifying significant risk factors is the key step.

This dissertation employs the the Myocardial Infarction Data Acquisition System (MI-

DAS) database to identify significant risk factors. MIDAS has been published elsewhere (e.g., Kostis et al. (2007), Swerdel et al. (2016), Wellings et al. (2018)). It is a New Jersey statewide database that includes all admissions to non-federal hospitals and emergency rooms (ERs). It contains 73,725,160 records of hospital admissions and ER visits of 15,519,554 New Jersey patients from 1995 to 2017. MIDAS was originally dedicated for research on cardiovascular diseases and now it has been used in a broader fields of biomedical research. It records patient admission dates and a wide range of admission causes, including MI, HF, stroke, and other cardiovascular diseases. In addition, MIDAS contains patient clinical characteristics, for instance, age, gender (male and female), race (white, black, and other), length of stay, admission type (inpatient, ER outpatient, non-ER outpatient, same day surgery outpatient, and other outpatient), discharge type (discharge to hospice, discharge to home, discharge to long-term care, discharge to short-time care, and other), health insurance payer (medicare, medicaid, HMO, blue cross plans, commercial, and self pay), diagnosis year and month, comorbidity conditions and medical procedures received (diabetes, hypertension, chronic obstructive pulmonary disease (COPD), liver disease, renal disease, anemia, cannabidiol use, cancer, obesity, saccular aneurysms, heart valve disease, conduction disorder, cardiac catheterization, percutaneous coronary intervention, coronary artery bypass surgery, cardiac ablation, cardiac resynchronization therapy, and artificial cardiac pacemaker).

MIDAS can be more useful by linking with other databases. For example, patient county-level socioeconomic information obtained from New Jersey State Health Assessment Data (`https://www-doh.state.nj.us/doh-shad/home/Welcome.html`), including general health status (percentage of fair or poor condition), percentage of health care coverage, education attainment (percentage of high school attainment), poverty status (percentage of poverty), median household income, percentage of blood cholesterol screening history, percentage of high cholesterol diagnosis, percentage of high blood pressure diagnosis, percentage of angina diagnosis, percentage of stroke diagnosis, and percentage

of obesity diagnosis, can be linked to the MIDAS database through ZIP code. Hospital characteristics, such as teaching status (teaching, minor-teaching, and non-teaching), location (inner city, urban, suburban, and rural), and size (number of beds), can be linked to the MIDAS database using CMS certification number (CCN). Moreover, the new jersey death record can also be linked to the MIDAS database using probabilistic matching to obtain patient death dates.

The long follow-up in the general population not only lends the MIDAS database a major advantage in unveiling risk factors with weak to moderate effects that cannot be discovered in a short follow-up period or in a particular population with selected characteristics, but also helps understand and characterize the natural history of diseases over a long period. When taking into account the intrinsic order and failure-related dependence among multivariate cardiovascular-related failure times, the multistate stochastic processes used in multistate survival analysis provide a framework to longitudinally describe transitions (i.e., CVDs) of patients between their different health states, and to dynamically predict transition probabilities for patients with particular characteristics.

1.5 Research Questions and Objectives

Motivated by utilizing data sets with extraordinarily large sample size (n) and large number of covariates (d) ($n \gg d$), such as MIDAS, to better understand the natural courses of diseases and more accurately capture significant risk factors for diseases, our research interests are centered on relating multivariate unordered failure times to a collection of risk factors, and building a multistate stochastic model for multivariate ordered failure times to predict the prognosis of patients. With that in mind, our multivariate survival analysis is focused on estimating and identifying significant regression parameters; whereas our multistate survival analysis is mainly used for estimating cumulative hazards and predicting transition probabilities.

The objective of this dissertation is to develop divide-and-combine approaches for mul-

tivariate survival analysis and multistate survival analysis to analyze large-scale multivariate failure time data. Specifically,

1) develop a divide-and-combine approach for multivariate survival analysis to estimate regression parameters.

2) develop a confidence distribution approach to perform regularized estimation in multivariate survival analysis.

3) develop a divide-and-combine approach for multistate survival analysis to estimate cumulative hazards and predict transition probabilities.

The rest of this dissertation is organized as follows. Chapter 2 reviews the related literature on multivariate survival analysis and multistate survival analysis. Major topics for the former include both marginal models and frailty models, and their statistical inference procedures; those for the latter include structures of multistate stochastic processes, setups of multistate models, and statistical inference procedures. Regularization using the confidence distribution is also briefly reviewed in Chapter 2. The divide-and-combine approach to estimating regression parameters in multivariate survival analysis is proposed in Chapter 3, in which we show the asymptotic equivalence between the divide-and-combine estimators and the "analyzed-all-at-once" full-data estimators. By using the asymptotic distributions of the divide-and-combine estimators, a confidence distribution based regularized estimation approach in multivariate survival analysis is proposed in Chapter 4, with the establishment of estimation consistency, selection consistency, and oracle properties of the regularized estimators. In Chapter 5, we propose a divide-and-combine approach in multistate survival analysis to estimating cumulative hazards and predicting transition probabilities, in which we show that the predicted transition probabilities using both divide-and-combine analysis and full-data analysis are asymptotically equivalent. Numerical illustrations of the proposed methods, including simulation studies and real data examples, are presented in

Chapter 6. We conclude this dissertation with a discussion and possible future work in Chapter 7.

CHAPTER 2
LITERATURE REVIEW

2.1 Multivariate Survival Analysis

Multivariate failure times can be either unordered or ordered as described in Section 1.2. When they are unordered (or referred to as "parallel data" in Hougaard (2000)), their failure time distributions are considered independent of the censoring distributions, or at least conditionally independent given some covariates. Analysis of such data is called multivariate survival analysis. This section focuses on multivariate survival analysis and its related statistical models. Two distinct classes are reviewed: marginal models and frailty models. These two classes of models differ in interpretation and by procedures used to handle correlation within a cluster. The appropriateness of one model over the other depends on the research interest, the study design, and the nature of the data structure under study. These issues are discussed in detail in this section. In particular, extensions of the Cox proportional hazards model (Cox, 1972) in both classes are of interest in this dissertation.

In multivariate survival analysis, consider n independent clusters, with cluster i potentially experiencing K distinct types of failures, $i = 1, 2, \ldots, n$. Let T_{ik} and C_{ik} denote, respectively, the failure and censoring times for the k^{th} type of failure in the i^{th} cluster, and let $\boldsymbol{Z}_{ik}$ be a d-dimensional vector of possibly time-varying covariates, with $n \gg d$. We assume d can be large but still finite. Correspondingly, let $X_{ik} = min(T_{ik}, C_{ik})$ be the observed time and $\delta_{ik} = I(T_{ik} \leq C_{ik})$ be the censoring indicator, where $I(.)$ is the indicator function. It is assumed that T_{ik} and C_{ik} are conditionally independent given $\boldsymbol{Z}_{ik}$ and that the censoring mechanism is noninformative. Denote the observed data by $\mathbf{D}_i = \{X_{ik}, \delta_{ik}, \boldsymbol{Z}_{ik};\ k = 1, 2, \ldots, K\}$ for $i = 1, 2, \ldots, n$, and assume that $\mathbf{D}_i$'s are an independent and identically distributed random sample from a certain population $\{X, \delta, \boldsymbol{Z}\}$.

A full likelihood of the observed data is given by

$$\mathcal{L} = \prod_{i=1}^{n} \prod_{k=1}^{K} \{\lambda_{ik}(X_{ik}|\boldsymbol{Z}_{ik})\}^{\delta_{ik}} S_{ik}(X_{ik}|\boldsymbol{Z}_{ik}), \tag{2.1}$$

where $\lambda_{ik}(.|\boldsymbol{Z}_{ik})$ and $S_{ik}(.|\boldsymbol{Z}_{ik})$ are the conditional hazard function and the conditional survival function of T_{ik} given $\boldsymbol{Z}_{ik}$.

2.1.1 Marginal Models

Marginal models analyze multivariate unordered failure time data by modeling marginal distributions of multivariate failure times and estimating the intra-cluster association separately or even leave the structure of intra-cluster association unspecified when the regression parameters are of primary interest. The former approach includes a two-step estimation method (Mahé and Chevret, 1999) to take into account the correlation, while the latter approach finds the estimate under the (incorrect) assumption of independence within clusters. This yields directly the final estimate of regression coefficients. The uncertainty/variance of the regression coefficient estimate is evaluated by a sandwich/robust estimator. The latter approach is called the independence working model method, which is closely related to the generalized estimating equations (GEE). The independence working model method is of interest in this dissertation for its flexibility with model assumptions and its convenience for implementation.

Model Setups

By formulating the marginal distribution of each type of failure with a proportional hazard model (Cox, 1972), the hazard function for the k^{th} type of failure in the i^{th} cluster is

$$\lambda_{ik}(t|\boldsymbol{Z}_{ik}) = \lambda_{0k}(t) e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{ik}(t)}, \tag{2.2}$$

where $\lambda_{0k}(.),\ k = 1, 2, \ldots, K$, are unspecified baseline hazard functions for K distinct failure types, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_d)^T$ is a vector of unknown regression parameters having an interpretation of log hazard ratio at a population average level. When $K = 1$, model (2.2) reduces to the model with identical baseline hazard functions. Model (2.2) can accommodate type-specific regression parameters by appropriately specifying $\boldsymbol{Z}_{ik}$ for $k = 1, 2, \ldots, K$, such as introducing type-specific covariates (Lin, 1994) or including interactions of failure types and covariates (Therneau and Lumley, 2015). We used these approaches in the simulation studies and the real data example (Chapter 6) to allow varying regression parameters among failure types.

Note that the marginal survival function is fully determined by model (2.2) through $S_{ik}(t) = \exp(-\int_0^t \lambda_{ik}(s|\boldsymbol{Z}_{ik})ds)$. However, the specification of the joint survival function is not completed without providing the intra-cluster association. The joint survival function for the K-variate failure time $T_i = (T_{i1}, T_{i2}, \ldots, T_{iK})$ can be expressed as

$$\begin{aligned} S_i(t_{i1}, t_{i2}, \ldots, t_{iK}) &= Pr(T_{i1} > t_{i1}, T_{i2} > t_{i2}, \ldots, T_{iK} > t_{iK}|\boldsymbol{Z}_{i1}, \boldsymbol{Z}_{i2}, \ldots, \boldsymbol{Z}_{iK}) \\ &= C\left(S_{i1}(t_{i1}), S_{i2}(t_{i2}), \ldots, S_{iK}(t_{iK})\right), \end{aligned} \tag{2.3}$$

where $C(.)$ is a copula function, i.e., a distribution function with uniform (0, 1) marginals and parameterized by θ (possibly a vector). Clearly, different choices of C lead to different joint survival functions for T_i, whereas the same marginal distributions specified by model (2.2) for each of K types of failures are preserved. For example, a commonly used copula in multivariate survival analysis is the Clayton-Oakes copula (Clayton and Cuzick, 1985; Oakes, 1989), in which $C(u, v, \ldots, s) = L\{L^{-1}(u) + L^{-1}(v) + \cdots + L^{-1}(s)\}$, where

$$L(s) = (1 + s)^{1/(1-\theta)} \quad \text{and} \quad L^{-1}(s) = s^{1-\theta} - 1, \tag{2.4}$$

with $\theta \geq 0$. $L(.)$ is the Laplace transform of the one-parameter gamma density, whose

density function is given by

$$f_U(u_i|\theta) = \frac{u_i^{(2-\theta)/(\theta-1)} e^{-u_i}}{\Gamma(1/(\theta-1))}. \tag{2.5}$$

This gamma density corresponds to mean of $\frac{1}{\theta-1}$ and variance of $\frac{1}{\theta-1}$. By using the Clayton-Oakes copula, the joint survival function in (2.3) then becomes

$$S_i(t_{i1}, t_{i2}, \ldots, t_{iK}) = \left[\sum_{k=1}^{K} \{S_{ik}(t_{ik})\}^{1-\theta} - K + 1\right]^{1/(1-\theta)}. \tag{2.6}$$

In addition, the intra-cluster association can be measured by Kendall's tau[1], given by

$$\kappa = 4 \int_0^\infty s p(s) p^{''}(s) ds - 1, \tag{2.7}$$

where $p(.)$ is the Laplace transform of copula-related density function and $p^{''}(.)$ is the second derivative of $p(.)$. When multivariate failure times within clusters are independent, $\kappa = 0$. When $p(.) = L(.)$, $\kappa = \frac{\theta-1}{\theta+1}$.

Estimation and Inference

A number of researchers studied the estimation and inference procedures in the marginal proportional hazards model. Wei et al. (1989) and Lee et al. (1992) proposed a Cox-type semiparametric model for stratified and unstratified multivariate unorderd failure time data, respectively. They extended the marginal parametric model considered in Huster et al. (1989) and borrowed the independence working model assumption therein. Spiekerman and Lin (1998) proposed a general semiparametric regression model which has a nested structure allowing for different baseline hazard functions among distinct failure types and imposing a common baseline hazard function on the failure times of the same type. The models considered in Wei et al. (1989) and Lee et al. (1992) are special cases of this model.

[1]Kendall's tau is denoted by κ to avoid confusion with τ, the end of the follow-up period.

Spiekerman and Lin (1998) developed the rigorous asymptotic theory in the marginal proportional hazards model for regression parameter estimators via the elegant counting process martingale theory, which was firstly connected with the Cox proportional hazards model by Andersen and Gill (1982).

In the counting process notation, $N_{ik}(t) = \delta_{ik} I(X_{ik} \leq t)$ records the number of the k^{th} failures observed on the i^{th} cluster by time t, and $Y_{ik}(t) = I(X_{ik} \geq t)$ indicates whether the i^{th} cluster is at risk for the k^{th} type of failure at time t. Before stating the estimation and inference procedures in the marginal proportional hazards model, it is convenient to introduce the following notation. Denote the true value of $\boldsymbol{\beta}$ by $\boldsymbol{\beta}_0$, and the end of the follow-up period by τ. For the k^{th} $(k = 1, 2, \ldots, K)$ failure type, we define

$$\mathbf{S}_k^{(r)}(\boldsymbol{\beta}, t) = n^{-1} \sum_{i=1}^{n} Y_{ik}(t) e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{ik}(t)} \boldsymbol{Z}_{ik}(t)^{\otimes r}, \quad \mathbf{s}_k^{(r)}(\boldsymbol{\beta}, t) = \mathcal{E}\left\{ \mathbf{S}_k^{(r)}(\boldsymbol{\beta}, t) \right\},$$

$$\mathbf{E}_k(\boldsymbol{\beta}, t) = \frac{\mathbf{S}_k^{(1)}(\boldsymbol{\beta}, t)}{S_k^{(0)}(\boldsymbol{\beta}, t)}, \quad \mathbf{e}_k(\boldsymbol{\beta}, t) = \frac{\mathbf{s}_k^{(1)}(\boldsymbol{\beta}, t)}{s_k^{(0)}(\boldsymbol{\beta}, t)},$$

$$\mathbf{V}_k(\boldsymbol{\beta}, t) = \frac{\mathbf{S}_k^{(2)}(\boldsymbol{\beta}, t)}{S_k^{(0)}(\boldsymbol{\beta}, t)} - \mathbf{E}_k(\boldsymbol{\beta}, t)^{\otimes 2}, \quad \mathbf{v}_k(\boldsymbol{\beta}, t) = \frac{\mathbf{s}_k^{(2)}(\boldsymbol{\beta}, t)}{s_k^{(0)}(\boldsymbol{\beta}, t)} - \mathbf{e}_k(\boldsymbol{\beta}, t)^{\otimes 2},$$

where $\mathbf{a}^{\otimes 0} = 1$, $\mathbf{a}^{\otimes 1} = \mathbf{a}$, and $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$ and $\mathcal{E}$ denotes expectation. Throughout this dissertation, we assume that the following regularity conditions hold for the marginal proportional hazards model. For some constant $\tau > 0$:

(M1) $\Pr\{Y_{ik}(t) = 1, \text{ for all } t \in [0, \tau]\} > 0$ for all i and k.

(M2) $|Z_{ikj}(0)| + \int_0^\tau |dZ_{ikj}(u)| < B_Z$ a.s. for all i, k, j and some constant $B_Z < \infty$.

(M3) There exists a neighborhood $\mathcal{B}$ of $\boldsymbol{\beta}_0$ such that for $r = 0, 1, 2$ and $k = 1, 2, \ldots, K$,

$$\sup_{t \in [0,\tau], \boldsymbol{\beta} \in \mathcal{B}} \left\| \mathbf{S}_k^{(r)}(\boldsymbol{\beta}, t) - \mathbf{s}_k^{(r)}(\boldsymbol{\beta}, t) \right\|_{\max} \xrightarrow{P} \mathbf{0}, \text{ as } n \to \infty,$$

where $\|\mathbf{a}\|_{\max} = \sup_i |\mathrm{a}_i|$ for a column vector $\mathbf{a} = (\mathrm{a}_i)$, and $\|\mathbf{A}\|_{\max} = \sup_{i,j} |\mathrm{a}_{ij}|$ for a matrix $\mathbf{A} = (\mathrm{a}_{ij})$.

(M4) $\mathbf{A}(\boldsymbol{\beta}_0) = \sum_{k=1}^{K} \int_0^\tau \mathbf{v}_k(\boldsymbol{\beta}_0, u) s_k^{(0)}(\boldsymbol{\beta}_0, u) \lambda_{0k}(u) du$ is positive definite.

(M5) $\int_0^\tau \lambda_{0k}(u) du < \infty$ for each k.

(M6) $\mathbf{s}_k^{(r)}(\boldsymbol{\beta}, t)$ $(k = 1, 2, \ldots, K;\ r = 0, 1, 2)$ are continuous functions of $\boldsymbol{\beta} \in \mathcal{B}$ uniformly in $t \in [0, \tau]$ and are bounded on $\mathcal{B} \times [0, \tau]$, $s_k^{(0)}(\boldsymbol{\beta}, t)$ $(k = 1, 2, \ldots, K)$ are bounded away from 0 on $\mathcal{B} \times [0, \tau]$, and

$$\mathbf{s}_k^{(1)}(\boldsymbol{\beta}, t) = \frac{\partial}{\partial \boldsymbol{\beta}^T} s_k^{(0)}(\boldsymbol{\beta}, t), \quad \mathbf{s}_k^{(2)}(\boldsymbol{\beta}, t) = \frac{\partial^2}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} s_k^{(0)}(\boldsymbol{\beta}, t)$$

for $k = 1, 2, \ldots, K$, $\boldsymbol{\beta} \in \mathcal{B}$, and $t \in [0, \tau]$.

Following Breslow's idea (Breslow, 1972) and treating $\lambda_{0k}(.)$ in model (2.2) as piecewise constant between uncensored failure times, under the independence working model assumption, the full likelihood in (2.1) can be derived into a pseudo-partial likelihood (Cox, 1972, 1975; Spiekerman and Lin, 1998) upon which the statistical inference on $\boldsymbol{\beta}$ is usually based

$$\mathcal{PL}(\boldsymbol{\beta}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left\{ \frac{e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{ik}(X_{ik})}}{n S_k^{(0)}(\boldsymbol{\beta}, X_{ik})} \right\}^{\delta_{ik}}. \tag{2.8}$$

The detailed mathematical derivation from the full likelihood to the pseudo-partial likelihood can be found in Appendix A.1. The first and minus second derivatives of logarithm of $\mathcal{PL}(\boldsymbol{\beta})$ in (2.8) are given by

$$\mathbf{U}(\boldsymbol{\beta}) = \frac{\partial log \mathcal{PL}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} = \sum_{i=1}^{n} \sum_{k=1}^{K} \int_0^\tau \{ \boldsymbol{Z}_{ik}(t) - \mathbf{E}_k(\boldsymbol{\beta}, t) \} \, dN_{ik}(t), \tag{2.9}$$

and

$$\mathcal{I}(\boldsymbol{\beta}) = -\frac{\partial^2 log \mathcal{PL}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T \partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \sum_{k=1}^{K} \int_0^\tau \mathbf{V}_k(\boldsymbol{\beta}, t) dN_{ik}(t). \tag{2.10}$$

By solving $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}$, we can obtain the maximum pseudo-partial likelihood estimator $\hat{\boldsymbol{\beta}}$, which is unique if $\mathcal{I}(\boldsymbol{\beta})$ is nonsingular.

Theorem 2.1.1 (Lemma 1 and Corollary 1 in Spiekerman and Lin (1998))**.** Under regularity conditions (M1) to (M6) in Section 2.1.1, the maximum pseudo-partial likelihood estimator $\hat{\boldsymbol{\beta}}$ satisfies the following as $n \to \infty$:

(1) (Consistency) $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}_0$;

(2) (Asymptotic Normality) $n^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_1 = \{\mathbf{A}^{-1}(\boldsymbol{\beta}_0)\}\{\mathbf{B}(\boldsymbol{\beta}_0)\}\{\mathbf{A}^{-1}(\boldsymbol{\beta}_0)\}^T)$;

(3) $\boldsymbol{\Sigma}_1$ can be consistently estimated by the sandwich estimator $\{\hat{\mathbf{A}}^{-1}(\hat{\boldsymbol{\beta}})\}\{\hat{\mathbf{B}}(\hat{\boldsymbol{\beta}})\}\{\hat{\mathbf{A}}^{-1}(\hat{\boldsymbol{\beta}})\}^T$, where $\hat{\mathbf{A}}(\hat{\boldsymbol{\beta}}) = n^{-1}\mathcal{I}(\hat{\boldsymbol{\beta}})$, and $\hat{\mathbf{B}}(\hat{\boldsymbol{\beta}}) = n^{-1}\sum_{i=1}^{n} \hat{\mathbf{w}}_i^{\otimes 2}$, with

$$\mathcal{I}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\int_0^{\tau} \mathbf{V}_k(\hat{\boldsymbol{\beta}}, t)dN_{ik}(t), \tag{2.11}$$

and

$$\begin{aligned}\hat{\mathbf{w}}_i = \sum_{k=1}^{K}\Bigg[&\int_0^{\tau}\left\{\boldsymbol{Z}_{ik}(t) - \mathbf{E}_k(\hat{\boldsymbol{\beta}}, t)\right\} dN_{ik}(t) \\ &- \sum_{j=1}^{n}\int_0^{\tau}\left\{\boldsymbol{Z}_{ik} - \mathbf{E}_k(\hat{\boldsymbol{\beta}}, s)\right\}\frac{Y_{ik}(s)e^{\hat{\boldsymbol{\beta}}^T\boldsymbol{Z}_{ik}(s)}dN_{jk}(s)}{nS_k^{(0)}(\hat{\boldsymbol{\beta}}, s)}\Bigg].\end{aligned} \tag{2.12}$$

Proof. The detailed proof can be found in Spiekerman and Lin (1998). ■

2.1.2 Frailty Models

Frailty models assumes that the correlation within a cluster is induced by a set of unobserved random quantities, called frailties. Given the frailty, failures in the same cluster are assumed to be independent. Frailty models are useful in estimating correlations in multivariate survival analysis, in which, the correlation structure is specified by incorporating a random effect (frailty) that is common to failures within the same cluster. The covariate effect is then interpreted as being conditional on the frailties and is cluster specific. The simplest model is the shared frailty model. In this model, all the failures within each cluster

share a common frailty, each failure belongs to precisely one cluster, and frailties of different clusters are independent. More complex models are possible. Frailties can be nested; individuals within a family may share a common frailty, while families within communities share another common frailty. In this dissertation, the shared frailty model is of interest due to its simplicity and sound statistical properties.

Model Setups

Instead of modeling the marginal distribution of T_{ik}, we model the conditional distribution of T_{ik} given frailties. In particular, the conditional hazard function for the k^{th} type of failure in the i^{th} cluster, conditional on the cluster frailty u_i, is assumed to take the form

$$\lambda_{ik}(t|\boldsymbol{Z}_{ik}, u_i) = u_i \lambda_0(t) e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{ik}}, \tag{2.13}$$

where $\lambda_0(.)$ is the unspecified baseline hazard function and $\boldsymbol{\beta}$ represents a d-dimensional unknown regression parameter having an interpretation of log hazard ratio specific to the individual cluster. Of note, in our proportional hazards frailty model, $\boldsymbol{Z}_{ik}$ is time-independent. We assume that, given $\boldsymbol{Z}_{ik}$ and u_i, the censoring is noninformative, and that the frailty u_i is independent of $\boldsymbol{Z}_{ik}$ and has a density $f_U(u_i|\theta)$, where θ is an unknown parameter. Similar to the marginal proportional hazards model in (2.2), the proportional hazards frailty model also can accommodate type-specific regression parameters by introducing type-specific covariates.

Note that in the marginal proportional hazards model, it is usually assumed that the number of failure types are the same across clusters; whereas in the proportional hazards frailty model, clusters rather than K distinct failure types are of primary interest, such that the number of failure types can vary in different clusters. Our proportional hazards frailty model in (2.13) can handle such situations by letting $K = K_i$ and the statistical properties below remain the same.

In the frailty model, under the conditional independence assumption, the joint conditional survival function for the i^{th} cluster is $S_{i,u}(t_{i1}, t_{i2}, \ldots, t_{iK}) = \exp[-u_i\{\Lambda_{i1}(t_{i1}|\boldsymbol{Z}_{i1}) + \Lambda_{i2}(t_{i2}|\boldsymbol{Z}_{i2}) + \cdots + \Lambda_{iK}(t_{iK}|\boldsymbol{Z}_{iK})\}]$, where $\Lambda_{ik}(t|\boldsymbol{Z}_{ik}) = \int_0^t \lambda_0(s)e^{\boldsymbol{\beta}^T\boldsymbol{Z}_{ik}}ds$ is the conditional cumulative baseline hazard for the k^{th} failure type in the i^{th} cluster. The joint survival function can be obtained by integrating out these frailties with respect to their density function

$$\begin{aligned} S_i(t_{i1}, t_{i2}, \ldots, t_{iK}) &= Pr(T_{i1} > t_{i1}, T_{i2} > t_{i2}, \ldots, T_{iK} > t_{iK}|\boldsymbol{Z}_{i1}, \boldsymbol{Z}_{i2}, \ldots, \boldsymbol{Z}_{iK}) \\ &= \int_0^\infty S_{i,u}(t_{i1}, t_{i2}, \ldots, t_{iK}) f_U(u_i|\theta)du_i. \end{aligned} \tag{2.14}$$

The joint survival function is determined by the frailty density. For example, in the commonly used gamma shared frailty model, by using the same gamma density function as in (2.5), it can be shown that the joint survival function in (2.14) is the same as the one in (2.6).

Here, to be consistent with the formulation in the R package, `frailtySurv` (Monaco et al., 2018), used in this dissertation, we assume $u_1, u_2, \ldots, u_n$ are independent realizations of a one parameter gamma density with mean of one and variance of θ, whose density is given by

$$f_U(u_i|\theta) = \frac{u_i^{1/\theta-1}e^{-u_i/\theta}}{\theta^{1/\theta}\Gamma(1/\theta)}. \tag{2.15}$$

Correspondingly, the joint survival function in (2.14) becomes

$$S_i(t_{i1}, t_{i2}, \ldots, t_{iK}) = \left[\sum_{k=1}^{K}\{S_{ik}(t_{ik})\}^{-\theta} - K + 1\right]^{-1/\theta}, \tag{2.16}$$

where $S_{ik}(t_{ik})$ is the marginal survival function obtained by having $t_{il} = 0$ $(l = 1, 2, \ldots, K, l \neq k)$ in $S_{i,u}(t_{i1}, t_{i2}, \ldots, t_{iK})$ and integrating out frailties. It follows that

$$S_{ik}(t_{ik}) = \int_0^\infty \exp\left[-u_i\left\{\Lambda_{ik}(t_{ik}|\boldsymbol{Z}_{ik})\right\}\right] f_U(u_i|\theta)du_i. \tag{2.17}$$

Of note, these integrals in (2.14) and (2.17) are nothing but the Laplace transforms of the frailty density function, thus they can be easily solved. In the proportional hazards frailty model, the intra-cluster dependence can also be measured by Kendall's tau using (2.7). When $p(.)$ is the Laplace transform of the density function in (2.15), $\kappa = \frac{\theta}{\theta+2}$.

Estimation and Inference

In frailty models, the estimation of the regression coefficients ($\boldsymbol{\beta}$), the cumulative baseline hazard function ($\Lambda_0(.)$), and the dependence parameter (θ) in the frailty density function, has been extensively studied in the statistical literature. For example, under the semi-parametric and shared frailty setting adopted in this dissertation, Klein (1992) proposed estimators using the expectation-maximization (EM) algorithm in the partial likelihood, which is a natural estimation tool since the frailties are latent parameters. There are also other estimators in this setting, such as the hierarchical likelihood (HL) based estimators (Ha et al., 2001), the maximum penalized partial likelihood (PPL) estimators (Therneau et al., 2003), and the maximum penalized likelihood estimators (MPLE) (Rondeau et al., 2006). Among them, the EM algorithm method perhaps is one of the most frequently used approaches with a notable challenge in estimating the variance of the estimated parameters and establishing the asymptotic distributions of the estimators (Parner, 1998; Zeng et al., 2008). In this dissertation, we build up our proposed approach based on the estimators proposed by Gorfine et al. (2006) because they established the asymptotic properties of the estimators and implemented their estimation and inference procedures in the R package, `frailtySurv`.

Before stating the estimation and inference procedures in the proportional hazards frailty model, we introduce the following notation and regularity conditions. Denote the true value of $\boldsymbol{\gamma} = (\theta, \boldsymbol{\beta}^T)^T$ by $\boldsymbol{\gamma}_0 = (\theta_0, \boldsymbol{\beta}_0^T)^T$ and the end of the follow-up period by τ. Throughout this dissertation, we assume that the following regularity conditions hold for the proportional hazards frailty model. For some constant $\tau > 0$:

(F1) $\mathcal{E}\{\sum_{k=1}^{K}(\tau)\} > 0$ for all i, where $\mathcal{E}$ denotes expectation.

(F2) The frailty u_i has finite moments up to order (K+2).

(F3) $|Z_{ikj}(0)| + \int_0^\tau |dZ_{ikj}(u)| < B_Z$ a.s. for all i, k, j and some constant $B_Z < \infty$.

(F4) The parameter $\boldsymbol{\gamma}$ lies in a compact subset of $\mathbb{R}^{1+d}$ containing an open neighbourhood of $\boldsymbol{\gamma}_0$.

(F5) $\int_0^\tau \lambda_0(u)du < \infty$.

(F6) The function $f'_U(u_i|\theta) = df_U(u_i;\theta)/d\theta$ is absolutely integrable.

(F7) The censoring distribution has at most finitely many jumps on $[0, \tau]$.

(F8) Pr($\delta_{ik} > 0$ for at least two k's, $k = 1, 2, \ldots, K) > 0$ for all i.

(F9) Pr$[\{\partial \mathbf{U}(\boldsymbol{\gamma}, \hat{\Lambda}_0(.))/\partial\boldsymbol{\gamma}\}|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_0}$ is invertible$] \xrightarrow{P} 1$ as $n \to \infty$.

(F10) Either of the following two conditions holds:

(a) There exist $b(\theta) > 0$ and $C(\theta) > 0$ such that

$$\sup_\theta \left| \frac{f_U(u_i|\theta)}{C(\theta)u_i^{b(\theta)-1}} - 1 \right| \xrightarrow{P} 0, \text{ as } h \to 0,$$

with h is the cumulative hazard and $b(\theta)$ bounded from below over θ (see Lemma 1 in Zucker et al. (2008) for more details);

(b) There exists $a > 0$ independent of θ such that $f_U(u_i|\theta)$ is increasing in u_i over $u_i \in [0, a]$ and we have

$$\lim_{u_i \to 0} \left[\sup_\theta f_U(u_i|\theta) \right] = 0.$$

Under the conditional independence assumption, the full likelihood in (2.1) can be rewritten as

$$\mathcal{L}(\boldsymbol{\beta}, \theta, \Lambda_0(.)) = \prod_{i=1}^{n} \left[\prod_{k=1}^{K} \left\{ \lambda_0(X_{ik}) e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{ik}} \right\}^{\delta_{ik}} \right] \int_0^{\infty} u_i^{A_i(\tau)} \exp\left\{ -u_i \sum_{k=1}^{K} \Lambda_0(X_{ik}) e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{ik}} \right\} f_U(u_i|\theta) du_i, \tag{2.18}$$

where $A_i(t) = \sum_{k=1}^{K} \delta_{ik} I(X_{ik} \leq t)$, $\Lambda_0(.)$ is the cumulative baseline hazard function. The corresponding first derivative of logarithm of $\mathcal{L}(\boldsymbol{\beta}, \theta, \Lambda_0(.))$ in (2.18) with respect to θ and $\boldsymbol{\beta}$ are given by

$$\mathbf{U}(\theta) = \sum_{i=1}^{n} \frac{\int_0^{\infty} u_i^{A_i(\tau)} H_i f_U'(u_i|\theta) du_i}{\int_0^{\infty} u_i^{A_i(\tau)} H_i f_U(u_i|\theta) du_i}, \tag{2.19}$$

and

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \delta_{ik} \boldsymbol{Z}_{ik} - \sum_{i=1}^{n} \frac{\left\{ \sum_{k=1}^{K} \Lambda_0(X_{ik}) e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{ik}} \boldsymbol{Z}_{ik} \right\} \int_0^{\infty} u_i^{A_i(\tau)+1} H_i f_U(u_i|\theta) du_i}{\int_0^{\infty} u_i^{A_i(\tau)} H_i f(u_i|\theta) du_i}, \tag{2.20}$$

with $f_U'(u_i|\theta) = df_U(u_i|\theta)/d\theta$ and $H_i = \exp\left\{ -u_i \sum_{k=1}^{K} \Lambda_0(X_{ik}) e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{ik}} \right\}$. The maximum pseudo-likelihood estimator $\hat{\boldsymbol{\gamma}} = \left(\hat{\theta}, \hat{\boldsymbol{\beta}}^T \right)^T$ can be obtained by solving $\mathbf{U}(\theta, \boldsymbol{\beta}, \Lambda_0(.)) = \mathbf{0}$, where $\mathbf{U}(\theta, \boldsymbol{\beta}, \Lambda_0(.)) = \left(\mathbf{U}(\theta), \mathbf{U}(\boldsymbol{\beta})^T \right)^T$ with $\Lambda_0(.)$ substituted by a Breslow-type estimator $\hat{\Lambda}_0(.)$. This pseudo-likelihood approach proposed Gorfine et al. (2006) avoids complicated iterative optimization process in the EM algorithm by using a simplified Breslow-type plug-in estimator for $\Lambda_0(.)$, which is not computationally intensive compared to the EM algorithm method.

Theorem 2.1.2 (Section 3 in Gorfine et al. (2006) and Section 3 in Zucker et al. (2008))**.** Under regularity conditions (F1) to (F10) in Section 2.1.2, the maximum pseudo-likelihood estimator $\hat{\boldsymbol{\gamma}}$ satisfies the following as $n \to \infty$:

(1) (Consistency) $\hat{\boldsymbol{\gamma}} \xrightarrow{P} \boldsymbol{\gamma}_0$;

(2) (Asymptotic Normality) $n^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_2 = \{\mathbf{A}^{-1}(\boldsymbol{\gamma}_0)\}\{\mathbf{B}(\boldsymbol{\gamma}_0)\}\{\mathbf{A}^{-1}(\boldsymbol{\gamma}_0)\}^T)$;

(3) $\boldsymbol{\Sigma}_2$ can be consistently estimated by the sandwich estimator $\{\hat{\mathbf{A}}^{-1}(\hat{\boldsymbol{\gamma}})\}\{\hat{\mathbf{B}}(\hat{\boldsymbol{\gamma}})\}\{\hat{\mathbf{A}}^{-1}(\hat{\boldsymbol{\gamma}})\}^T$, where $\hat{\mathbf{A}}(\hat{\boldsymbol{\gamma}}) = n^{-1}\mathcal{I}(\hat{\boldsymbol{\gamma}})$, $\hat{\mathbf{B}}(\hat{\boldsymbol{\gamma}}) = \hat{\mathbf{V}}(\hat{\boldsymbol{\gamma}}) + \hat{\mathbf{G}}(\hat{\boldsymbol{\gamma}}) + \hat{\mathbf{C}}(\hat{\boldsymbol{\gamma}})$. Of note, $\mathcal{I}(.)$ is the minus second derivative of logarithm of $\mathcal{L}(\boldsymbol{\beta}, \theta, \Lambda_0(.))$ in (2.18), and $\hat{\mathbf{V}}(\hat{\boldsymbol{\gamma}})$, $\hat{\mathbf{G}}(\hat{\boldsymbol{\gamma}})$, and $\hat{\mathbf{C}}(\hat{\boldsymbol{\gamma}})$ can be found in Appendix A.2.

Proof. The detailed proof can be found in Zucker et al. (2008). ■

2.2 Multistate Survival Analysis

Multivariate ordered failure time data collected in many clinical or epidemiological longitudinal studies can be used to gain insight into the disease processes of patients. It is particularly useful in studying the prognosis of patients (e.g., prediction of survival for cardiovascular disease patients) in the course of time by incorporating intermediate events. Multistate models are a natural choice to analyze such data. In this dissertation, we adopt a finite-state Markov multistate model where the hazard for each possible transition in the multistate stochastic processes is estimated by a separate Cox proportional hazards model (Cox, 1972). In multistate survival analysis, this model is also known as the "Andersen-type Cox Markov model" (Andersen et al., 1991).

2.2.1 Examples of Multistate Processes

Multistate models are models for multivariate ordered failure time data in which all subjects start in one or possibly more states (e.g., post transplantation for patients with liver cirrhosis or index HF hospitalization for patients with cardiovascular diseases) and eventually may end up in one (or more) absorbing state(s) (e.g., death or relapse). In between, intermediate states can be visited. Some subjects are censored before they reach an absorbing state.

The classical statistical model for univariate survival analysis may be considered as a special case of multistate models with only two states, namely alive state and dead state. The force of transition from the alive state to the dead state is the hazard function $\lambda(t)$ of the failure time distribution. A typical example of multistate models is illustrated in Figure 2.1. This example is often referred to as the "illness-death model" and the simplest true multistate model. In Figure 2.1, health states (i.e., healthy, illness, and death) are represented by boxes. Transitions (i.e., failures) are represented by arrows going from one state to another. Although, as suggested by the name, the typical application of this illness-death model is one where the illness state is an unfavorable intermediate state, this

is not necessarily the case. It could correspond to a favorable development during the disease processes. In this dissertation, we shall restrict to uni-directional models though bi-directional is possible. Our interests are probabilities of transitions between these states, especially those probabilities given some intermediate states. For example, the probabilities of state 1 to state 3 are different depending on experiencing state 2 or not.

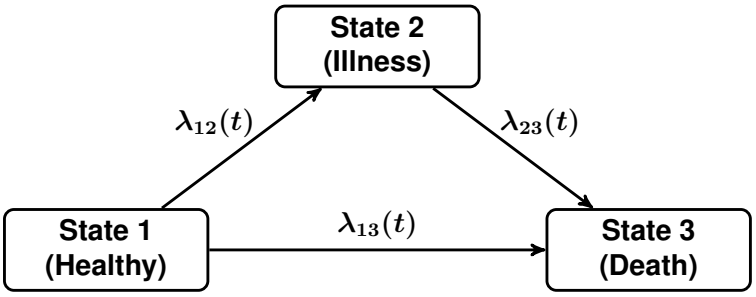

Figure 2.1: The illness-death model.

2.2.2 Multistate Models

In multistate survival analysis, a multistate model is modeling a stochastic process $X(t)$ with a finite state space $\mathcal{Q} = \{1, 2, \ldots, Q\}$. Consider n independent subjects, with subject i potentially visiting Q states, $i = 1, 2, \ldots, n$. The value of the process $X_i(t)$ denotes the state being occupied by subject i at time t. The transition hazard $\lambda_{i,hj}(t)$, which expresses the instantaneous risk of a transition from state h to state j at time t for subject i, is defined as

$$\lambda_{i,hj}(t) = \lim_{\Delta t \to 0} \frac{Pr(X_i(t + \Delta t) = j | X_i(t) = h)}{\Delta t}. \tag{2.21}$$

The Markov assumption adopted in this dissertation is implicitly present in definition (2.21). It says only the current state and time govern the future trajectory of the process. Formally, $Pr(X_i(t + \Delta t) = j | X_i(t) = h, \{X_i(s), s < t\}) = Pr(X_i(t + \Delta t) = j | X_i(t) = h)$. Correspondingly, the cumulative transition hazard is defined as $\Lambda_{i,hj}(t) = \int_0^t \lambda_{i,hj}(s)ds$. Cumulative transition hazards between states can be gathered into a $Q \times Q$ matrix $\mathbf{\Lambda}_i(t)$, with diagonal elements $\Lambda_{i,hh}(t) = -\sum_{j \neq h} \Lambda_{i,hj}(t)$. If a direct transition from state h to

state j is impossible, $\Lambda_{i,hj}(t) = 0$. Hereinafter, a transition from state h to state j refers to the direct transition, unless otherwise specified. The cumulative transition hazard matrix of the illness-death model in Figure 2.1 is illustrated below in (2.22):

$$\mathbf{\Lambda}_i(t) = \begin{pmatrix} -\{\Lambda_{i,12}(t) + \Lambda_{i,13}(t)\} & \Lambda_{i,12}(t) & \Lambda_{i,13}(t) \\ 0 & -\Lambda_{i,23}(t) & \Lambda_{i,23}(t) \\ 0 & 0 & 0 \end{pmatrix}. \tag{2.22}$$

The transition probability matrix $\mathbf{P}_i(s,t)$ is of primary interest in this dissertation. Its element $P_{i,hj}(s,t) = Pr(X_i(t) = j | X_i(s) = h)$ denotes the transition probability from state h to state j in the time interval $(s,t]$. $P_{i,hj}(s,t)$ combines both direct and indirect transitions from state h to state j. In Markov multistate models, we have

$$P_{i,hj}(s,t) = \sum_{q=1}^{Q} P_{i,hq}(s,u) P_{i,qj}(u,t). \tag{2.23}$$

The corresponding matrix form of transition probabilities can also be found

$$\mathbf{P}_i(s,t) = \prod_{u \in (s,t]} \{\mathbf{I} + d\mathbf{\Lambda}_i(u)\}, \tag{2.24}$$

where $\mathbf{I}$ is the identity matrix and $\prod$ is the sign of product integral, which has the same relation to a product as the well-known integral has to a sum. The detailed derivation for $\mathbf{P}_i(s,t)$ and a heuristic explanation of the product integral can be found in Appendix A.3.

In Markov multistate models with finite state spaces, the estimation and inference procedures are two steps: 1) estimation of cumulative transition hazards using nonparametric or semiparametric approaches; 2) estimation of transition probabilities using estimated cumulative transition hazards. This dissertation is focused on the semiparametric approach and assume a Cox proportional hazards model (Cox, 1972) for each possible transition in the multistate stochastic processes.

The Markov assumption adopted in this dissertation implies that the time t in our multistate processes refers to the time since the subject entered the initial state (this category is also called "clock-forward" models in Putter et al. (2007)). If the time scale is modified to represent the time since entry of the current state (this category is also called "clock-reset" models in Putter et al. (2007)), the Markov assumption cannot hold. The resulting multistate model is called a Markov renewal model or a semi-Markov model. In both Markov models and semi-Markov models, the sojourn time in a current state or the current state itself as a time-dependent covariate can also be incorporated. In this dissertation we only focus on Markov multistate models and shall not distinguish Markov models, semi-Markov models, or non-Markov models, unless necessary.

2.2.3 Estimation of Transition Hazards in Cox Models

Model Setups

We use Cox proportional hazards model (Cox, 1972) for each transition separately. The hazard function in transition from state h to state j for subject i is

$$\lambda_{i,hj}(t|\boldsymbol{Z}_{i,hj}) = \lambda_{0,hj}(t)e^{\boldsymbol{\beta}^T\boldsymbol{Z}_{i,hj}}, \tag{2.25}$$

where $\lambda_{0,hj}(t)$, $h, j = 1, 2, \ldots, Q$ and $h \neq j$, are unspecified baseline hazard functions for different transitions, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_d)^T$ is a vector of regression parameters that describe the effect of covariates, and $\boldsymbol{Z}_{i,hj}$ is a time-independent covariate vector for transition from h to j. Similar to the marginal proportional hazards model in (2.2), the Markov proportional hazards multistate model in (2.25) can also accommodate transition-specific regression coefficients by introducing appropriate transition-specific covariates, as elaborated in the simulation studies and the real data example (Chapter 6).

Yet time-dependent covariates are possible in estimating transition hazards using the Cox proportional hazards model (Cox, 1972), prediction of transition probabilities based

on most types of time-dependent covariates, in practical applications, gives problems of estimation and interpretation (e.g., Cortese and Andersen (2010)). Thus, generally we need the covariate $\boldsymbol{Z}_{i,hj}$ to be time-independent and measured at the baseline to enable prediction of transition probabilities. However, there is one exception: a time-dependent covariate $\tilde{Z}(t)$ used to distinguish between different transitions into the same state, can be used in model (2.25) and for further prediction of transition probabilities. This exceptional time-dependent covariate $\tilde{Z}(t)$ is useful in modeling the proportionality among transitions into the same state. For instance, in the illness-death model example (illustrated in Figure 2.1), if we assume $\lambda_{0,23}(t) = \tilde{\omega}\lambda_{0,13}(t)$, model (2.25) can accommodate this proportionality between transitions (i.e., $2 \to 3$ and $1 \to 3$) into the same state 3 by denoting $e^{\tilde{\beta}\tilde{Z}(t)} = \tilde{\omega}$ and letting $\tilde{Z}(t) = 1$ if the subject has already visited state 2 or $\tilde{Z}(t) = 0$ otherwise. Of note, although prediction is in general no longer possible when endogenous covariates are used (see Section 6.3 in Kalbfleisch and Prentice (2002)), this time-dependent covariate $\tilde{Z}(t)$ is a special kind of endogenous covariate that serves only to distinguish between transitions sharing the same baseline hazard. Thus the use of $\tilde{Z}(t)$ does not cause problems in prediction of transition probabilities.

Estimation and Inference

The estimation of cumulative transition hazards in model (2.25) can be achieved in the same way as in the marginal proportional hazards model, except that K distinct types of failures in the marginal proportional hazards model are replaced by the transitions from state h to state j for $h, j = 1, 2, \ldots, Q$ and $h \neq j$. We adopt the method discussed in Andersen et al. (1991) and Andersen et al. (1993).

Before stating the asymptotic properties of the cumulative estimated transition hazards, we need to define some notation. Adopting the notation in De Wreede et al. (2010) and Andersen and Keiding (2002), let $N_{i,hj}(t)$ be the number of direct transitions from state h to state j for subject i in the time interval $[0, t]$, and $Y_{i,h}(t) = I\{X_i(t-) = h\}$. For the

transition from state h to state j $(h, j = 1, 2, \ldots, Q, h \neq j)$, we define

$$\mathbf{S}_{hj}^{(r)}(\boldsymbol{\beta}, t) = n^{-1} \sum_{i=1}^{n} Y_{i,h}(t) e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{i,hj}} \boldsymbol{Z}_{i,hj}^{\otimes r}, \quad \mathbf{s}_{hj}^{(r)}(\boldsymbol{\beta}, t) = \mathcal{E}\left\{\mathbf{S}_{hj}^{(r)}(\boldsymbol{\beta}, t)\right\},$$

$$\mathbf{E}_{hj}(\boldsymbol{\beta}, t) = \frac{\mathbf{S}_{hj}^{(1)}(\boldsymbol{\beta}, t)}{S_{hj}^{(0)}(\boldsymbol{\beta}, t)}, \quad \mathbf{e}_{hj}(\boldsymbol{\beta}, t) = \frac{\mathbf{s}_{hj}^{(1)}(\boldsymbol{\beta}, t)}{s_{hj}^{(0)}(\boldsymbol{\beta}, t)},$$

$$\mathbf{V}_{hj}(\boldsymbol{\beta}, t) = \frac{\mathbf{S}_{hj}^{(2)}(\boldsymbol{\beta}, t)}{S_{hj}^{(0)}(\boldsymbol{\beta}, t)} - \mathbf{E}_{hj}(\boldsymbol{\beta}, t)^{\otimes 2}, \quad \mathbf{v}_{hj}(\boldsymbol{\beta}, t) = \frac{\mathbf{s}_{hj}^{(2)}(\boldsymbol{\beta}, t)}{s_{hj}^{(0)}(\boldsymbol{\beta}, t)} - \mathbf{e}_{hj}(\boldsymbol{\beta}, t)^{\otimes 2},$$

where $\mathbf{a}^{\otimes 0} = 1$, $\mathbf{a}^{\otimes 1} = \mathbf{a}$, and $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^T$ and $\mathcal{E}$ denotes expectation. Throughout this dissertation, we assume that the same regularity conditions (M1) to (M6) as in Section 2.1.1, hold for the Markov proportional hazards multistate model — except that the failure type k should be replaced by transition from state h to state j.

The regression coefficient $\boldsymbol{\beta}$ is estimated by $\hat{\boldsymbol{\beta}}$ from solving $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}$, where

$$\mathbf{U}(\boldsymbol{\beta}) = \sum_{h \neq j} \sum_{i=1}^{n} \int_0^{\tau} \{\boldsymbol{Z}_{i,hj} - \mathbf{E}_{hj}(\boldsymbol{\beta}, t)\} \, dN_{i,hj}(t), \tag{2.26}$$

with τ is the end of the follow-up period. The use of (2.26) in estimating regression coefficients is justified in Appendix A.4. The cumulative baseline transition hazard function, $\Lambda_{0,hj}(t) = \int_0^t \lambda_{0,hj}(s) ds$, is estimated by the Breslow estimator (Breslow, 1972)

$$\hat{\Lambda}_{0,hj}(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_{i,hj}(s)}{nS_{hj}^{(0)}(\hat{\boldsymbol{\beta}}, s)}. \tag{2.27}$$

Denote the true value of $\boldsymbol{\beta}$ by $\boldsymbol{\beta}_0$. For a future subject with a covariate $\boldsymbol{z}_{0,hj}$ having the same structure of $\boldsymbol{Z}_{i,hj}$ as considered in model (2.25), the cumulative transition hazard function, $\Lambda_{hj}(t|\boldsymbol{z}_{0,hj}) = e^{\boldsymbol{\beta}_0^T \boldsymbol{z}_{0,hj}} \Lambda_{0,hj}(t)$ is estimated by

$$\hat{\Lambda}_{hj}(t|\boldsymbol{z}_{0,hj}) = \sum_{i=1}^{n} \int_0^t \frac{e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{z}_{0,hj}} dN_{i,hj}(s)}{nS_{hj}^{(0)}(\hat{\boldsymbol{\beta}}, s)}. \tag{2.28}$$

Theorem 2.2.1 (Section 3 in Andersen et al. (1991) and Theorem VII.2.3 (Page 503) in An-

dersen et al. (1993))**.** Under regularity conditions in Section 2.2.3, $\hat{\Lambda}_{hj}(t|\boldsymbol{z}_{0,hj})$ converges in probability to $\Lambda_{hj}(t|\boldsymbol{z}_{0,hj})$ uniformly in $t \in [0, \tau]$, i.e., as $n \to \infty$,

$$\sup_{t\in[0,\tau]} \left|\hat{\Lambda}_{hj}(t|\boldsymbol{z}_{0,hj}) - \Lambda_{hj}(t|\boldsymbol{z}_{0,hj})\right| \xrightarrow{P} 0. \tag{2.29}$$

Proof. The proof for Theorem 2.2.1 can be found in Andersen et al. (1991) and Andersen et al. (1993). A similar proof also can be found in Spiekerman and Lin (1998). ■

Theorem 2.2.2 (Section 3 in Andersen et al. (1991) and Corollary VII.2.6 (Page 505) in Andersen et al. (1993))**.** Under regularity conditions in Section 2.2.3, the stochastic process $n^{1/2}\left\{\hat{\Lambda}_{hj}(t|\boldsymbol{z}_{0,hj}) - \Lambda_{hj}(t|\boldsymbol{z}_{0,hj})\right\}$ converges weakly to a zero-mean Gaussian process whose variance function is given by

$$b^2(t|\boldsymbol{z}_{0,hj}) + \boldsymbol{a}^T(t|\boldsymbol{z}_{0,hj})\boldsymbol{\Sigma}_3\boldsymbol{a}(t|\boldsymbol{z}_{0,hj}), \tag{2.30}$$

where $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_3)$ and $\hat{\boldsymbol{\beta}}$ is obtained by solving $\mathbf{U}(\boldsymbol{\beta}) = \mathbf{0}$ in (2.26), and

$$\begin{aligned} b^2(t|\boldsymbol{z}_{0,hj}) &= \int_0^t \frac{e^{2\cdot\boldsymbol{\beta}_0^T\boldsymbol{z}_{0,hj}}\lambda_{0,hj}(s)ds}{s_{hj}^{(0)}(\boldsymbol{\beta}_0, s)}, \\ \boldsymbol{a}(t|\boldsymbol{z}_{0,hj}) &= \int_0^t \left\{\boldsymbol{z}_{0,hj} - \mathbf{e}_{hj}(\boldsymbol{\beta}_0, s)\right\} e^{\boldsymbol{\beta}_0^T\boldsymbol{z}_{0,hj}}\lambda_{0,hj}(s)ds. \end{aligned} \tag{2.31}$$

Corollary 2.2.2.1. Under conditions in Theorem 2.2.2, the variance function of the random process $n^{1/2}\left\{\hat{\Lambda}_{hj}(t|\boldsymbol{z}_{0,hj}) - \Lambda_{hj}(t|\boldsymbol{z}_{0,hj})\right\}$ can be uniformly consistently estimated by

$$\hat{b}^2(t|\boldsymbol{z}_{0,hj}) + \hat{\boldsymbol{a}}^T(t|\boldsymbol{z}_{0,hj})\hat{\boldsymbol{\Sigma}}_3\hat{\boldsymbol{a}}(t|\boldsymbol{z}_{0,hj}), \tag{2.32}$$

where

$$\begin{aligned}
\hat{b}^2(t|\boldsymbol{z}_{0,hj}) &= \sum_{i=1}^{n} \int_0^t \frac{e^{2\cdot\hat{\boldsymbol{\beta}}^T \boldsymbol{z}_{0,hj}} dN_{i,hj}(s)}{n\left\{S_{hj}^{(0)}(\hat{\boldsymbol{\beta}}, s)\right\}^2}, \\
\hat{\boldsymbol{a}}(t|\boldsymbol{z}_{0,hj}) &= \sum_{i=1}^{n} \int_0^t \frac{\left\{\boldsymbol{z}_{0,hj} - \mathbf{E}_{hj}(\hat{\boldsymbol{\beta}}, s)\right\} e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{z}_{0,hj}} dN_{i,hj}(s)}{nS_{hj}^{(0)}(\hat{\boldsymbol{\beta}}, s)}, \\
\hat{\boldsymbol{\Sigma}}_3^{-1} &= n^{-1} \sum_{h\neq j} \sum_{i=1}^{n} \int_0^\tau \mathbf{V}_{hj}(\hat{\boldsymbol{\beta}}, t) dN_{i,hj}(t).
\end{aligned} \tag{2.33}$$

The asymptotic normality of $\hat{\Lambda}_{hj}(t|\boldsymbol{z}_{0,hj})$ presented in Theorem 2.2.2 and Corollary 2.2.2.1 implicitly assume the cumulative baseline hazards of different transitions are uncorrelated, which is the general model setup adopted in this dissertation. However, as discussed in the part of model setups of Section 2.2.3, by incorporating a time-dependent covariate, model in (2.25) can be adjusted to accommodate possibly correlated transition hazards. We provide the asymptotic results for the cumulative transition hazard functions taking into account the possible correlation in the following Corollary 2.2.2.2. For the ease of notation, in the following corollary only, denote the type-specific cumulative baseline hazards by $\Lambda_{0k}(t)$ and $\Lambda_{0k'}(t)$ for types k and $k'(k, k' = 1, 2, \ldots, K)$. Note that multiple transition-specific baseline hazards may share the same type of baseline hazard, i.e., $\lambda_{0,hj}(t) = \lambda_{0k}(t)$ and $\lambda_{0,lm}(t) = \lambda_{0k}(t)$ for $h, j, l, m = 1, 2, \ldots, Q$, $h \neq j$, $l \neq m$, and $K \leq Q$. Other notation in the following corollary should be modified accordingly.

Corollary 2.2.2.2 (Proposition 1 in De Wreede et al. (2010))**.** Under conditions in Theorem 2.2.2, the process

$$n^{1/2}\left\{e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{z}_{0k}} \hat{\Lambda}_{0k}(t) - e^{\boldsymbol{\beta}_0^T \boldsymbol{z}_{0k}} \Lambda_{0k}(t),\ e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{z}_{0k'}} \hat{\Lambda}_{0k'}(t) - e^{\boldsymbol{\beta}_0^T \boldsymbol{z}_{0k'}} \Lambda_{0k'}(t)\right\} \tag{2.34}$$

converges weakly to a bivariate Gaussian process with mean zero and a covariance function

which can be estimated uniformly consistently by

$$\begin{aligned}
&\omega_{kk'} \sum_{i=1}^{n} \int_0^t \frac{e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{z}_{0k}} e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{z}_{0k'}} dN_{ik}(s)}{n \left\{ S_k^{(0)}(\hat{\boldsymbol{\beta}}, s) \right\}^2} \\
&+ \sum_{i=1}^{n} \int_0^t \frac{\left\{ \boldsymbol{z}_{0k} - \mathbf{E}_k(\hat{\boldsymbol{\beta}}, s) \right\}^T e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{z}_{0k}} dN_{ik}(s)}{n S_k^{(0)}(\hat{\boldsymbol{\beta}}, s)} \times \hat{\boldsymbol{\Sigma}}_3 \\
&\times \sum_{i=1}^{n} \int_0^t \frac{\left\{ \boldsymbol{z}_{0k'} - \mathbf{E}_{k'}(\hat{\boldsymbol{\beta}}, s) \right\} e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{z}_{0k'}} dN_{ik'}(s)}{n S_{k'}^{(0)}(\hat{\boldsymbol{\beta}}, s)},
\end{aligned} \tag{2.35}$$

in which $\omega_{kk'} = 1$ if $k = k'$ and 0 otherwise.

Proof. The detailed proofs for Theorem 2.2.2, Corollary 2.2.2.1, and Corollary 2.2.2.2 can be found in Andersen et al. (1991), Andersen et al. (1993), and De Wreede et al. (2010). ■

2.2.4 Estimation of Transition Probabilities

When there are no covariates, Aalen and Johansen (1978) suggested to estimate the transition probability matrix in (2.24) by the nonparametric Aalen-Johansen estimator. Andersen et al. (1991) extended this method to the case where each transition hazard is specified via a Cox proportional hazards model (Cox, 1972) and derived the asymptotic properties of the Aalen-Johansen estimator in the semiparametric approach. Indeed, the estimation of transition probabilities depends only on the estimation of the transition hazards. This relation does not depend on the presence or absence of covariates, nor how hazards are related to covariates — nonparametrically, semiparametrically, or parametrically (De Wreede et al., 2010). In this dissertation, we are interested in estimating transition probabilities for a given subject with a covariate $\boldsymbol{z}_0$, thus in the remainder of this section, we suppress the subscript i in the notation of $\mathbf{P}$ and $\boldsymbol{\Lambda}$ for simplicity. That has been said, we estimate the transition probability matrix $\mathbf{P}(s,t)$ in (2.24) by the Aalen-Johansen estimator

$$\hat{\mathbf{P}}(s,t|\boldsymbol{z}_0) = \prod_{u \in (s,t]} \left\{ \mathbf{I} + d\hat{\boldsymbol{\Lambda}}(u|\boldsymbol{z}_0) \right\}, \tag{2.36}$$

where $\hat{\mathbf{\Lambda}}(u|\boldsymbol{z}_0)$ is the semiparametric estimator of $\mathbf{\Lambda}(u)$ for a given subject with a covariate $\boldsymbol{z}_0$. Note that $\boldsymbol{z}_0$ is a basic covariate and can be extended into the transition-specific structure $\boldsymbol{z}_{0,hj}$ for $h, j = 1, 2, \ldots, Q$ and $h \neq j$.

Transition Probability Matrix

Since $\mathbf{\Lambda}(t|\boldsymbol{z}_0)$ is a $Q \times Q$ matrix, the dimension of the transition probability matrix $\mathbf{P}(s, t|\boldsymbol{z}_0)$ is also $Q \times Q$. Here, we illustrate the structure of the transition probability matrix using the illness-death model as shown in Figure 2.1 and demonstrate its relation with univariate survival analysis.

Given $\boldsymbol{\lambda}(t|\boldsymbol{z}_0)$ of the illness-death model in (2.37),

$$\boldsymbol{\lambda}(t|\boldsymbol{z}_0) = \begin{pmatrix} -\{\lambda_{12}(t|\boldsymbol{z}_0) + \lambda_{13}(t|\boldsymbol{z}_0)\} & \lambda_{12}(t|\boldsymbol{z}_0) & \lambda_{13}(t|\boldsymbol{z}_0) \\ 0 & -\lambda_{23}(t|\boldsymbol{z}_0) & \lambda_{23}(t|\boldsymbol{z}_0) \\ 0 & 0 & 0 \end{pmatrix}, \tag{2.37}$$

under the Markov model framework adopted in this dissertation, we find the transition probability matrix $\mathbf{P}(s, t|\boldsymbol{z}_0)$ in (2.38)

$$\mathbf{P}(s, t|\boldsymbol{z}_0) = \begin{pmatrix} P_{11}(s, t|\boldsymbol{z}_0) & P_{12}(s, t|\boldsymbol{z}_0) & P_{13}(s, t|\boldsymbol{z}_0) \\ 0 & P_{22}(s, t|\boldsymbol{z}_0) & P_{23}(s, t|\boldsymbol{z}_0) \\ 0 & 0 & P_{33}(s, t|\boldsymbol{z}_0) \end{pmatrix}, \tag{2.38}$$

as the solution of the Kolmogorov forward equation

$$\frac{\partial}{\partial t}\mathbf{P}(s, t|\boldsymbol{z}_0) = \mathbf{P}(s, t|\boldsymbol{z}_0)\boldsymbol{\lambda}(t|\boldsymbol{z}_0). \tag{2.39}$$

The derivation of (2.39) can be found in Appendix A.3. Apparently, state 3 is the absorbing

state and thus $P_{33}(s,t|\boldsymbol{z}_0) = 1$. And by (2.39), we have

$$\frac{\partial}{\partial t}P_{11}(s,t|\boldsymbol{z}_0) = -\left\{\lambda_{12}(t|\boldsymbol{z}_0) + \lambda_{13}(t|\boldsymbol{z}_0)\right\} P_{11}(s,t|\boldsymbol{z}_0), \tag{2.40}$$

thus we get the solution

$$P_{11}(s,t|\boldsymbol{z}_0) = \exp\left[-\int_s^t \left\{\lambda_{12}(u|\boldsymbol{z}_0) + \lambda_{13}(u|\boldsymbol{z}_0)\right\} du\right]. \tag{2.41}$$

Similarly, $P_{22}(s,t|\boldsymbol{z}_0) = \exp\left\{-\int_s^t \lambda_{23}(u|\boldsymbol{z}_0)du\right\}$. Further, we have $P_{23}(s,t|\boldsymbol{z}_0) = 1 - P_{22}(s,t|\boldsymbol{z}_0)$, and $P_{13}(s,t|\boldsymbol{z}_0) = 1 - P_{11}(s,t|\boldsymbol{z}_0) - P_{12}(s,t|\boldsymbol{z}_0)$. Note that $P_{13}(s,t|\boldsymbol{z}_0)$ includes not only the direct transition $1 \to 3$ but also the indirect transition $1 \to 2 \to 3$. The last probability $P_{12}(s,t|\boldsymbol{z}_0)$ is the solution to

$$\frac{\partial}{\partial t}P_{12}(s,t|\boldsymbol{z}_0) = \lambda_{12}(t|\boldsymbol{z}_0)P_{11}(s,t|\boldsymbol{z}_0) - \lambda_{23}(t|\boldsymbol{z}_0)P_{12}(s,t|\boldsymbol{z}_0). \tag{2.42}$$

After solving the differential equation in (2.42), we find the solution

$$P_{12}(s,t|\boldsymbol{z}_0) = \int_s^t P_{11}(s,u|\boldsymbol{z}_0)\lambda_{12}(u|\boldsymbol{z}_0)P_{22}(u,t|\boldsymbol{z}_0)du. \tag{2.43}$$

The detailed step-by-step derivation leading to (2.43) is given in Appendix A.5.

In the illness-death model, $P_{13}(0,t|\boldsymbol{z}_0)$ gives the conditional probability of death prior to time t, given healthy at time 0, which is closely related to $\{1 - S(t)\}$ in univariate survival analysis, where $S(t)$ is the overall survival probability. When compared to univariate survival analysis, using multistate models enables more accurate prediction in the probability of death by incorporating intermediate events. For instance, $P_{23}(s,t|\boldsymbol{z}_0)$ gives the conditional probability of death prior to time t, given ill at time s, which is more accurate than $\{1 - S(t)\}$ when the information of being ill is available.

Estimation and Inference

By incorporating the estimate of cumulative hazard matrix $\hat{\mathbf{\Lambda}}(t|\boldsymbol{z}_0)$ whose elements can be obtained in (2.28), into the Aalen-Johansen estimator $\hat{\mathbf{P}}(s,t|\boldsymbol{z}_0)$ in (2.36), we can obtain the estimate for transition probabilities. As $\hat{\mathbf{P}}$ itself is a $Q\times Q$ matrix, its estimated variance matrix is defined as $\hat{\mathbf{var}}(\mathbf{vec}(\hat{\mathbf{P}}))$, where $\mathbf{vec}(\hat{\mathbf{P}})$ is defined as the vectorized $Q^2\times 1$ matrix where the columns are stacked on top of each other. Thus the dimension of $\hat{\mathbf{var}}(\hat{\mathbf{P}})$ is $Q^2\times Q^2$.

Theorem 2.2.3 (Section 3 in Andersen et al. (1991) and Theorem IV.4.1 (Page 317) in Andersen et al. (1993))**.** Under regularity conditions in Section 2.2.3, $\hat{\mathbf{P}}(s,t|\boldsymbol{z}_0)$ converges in probability to $\mathbf{P}(s,t|\boldsymbol{z}_0)$ uniformly in $s,t\in[0,\tau]$, i.e., as $n\to\infty$,

$$\sup_{t\in[s,\tau]}\left\|\hat{\mathbf{P}}(s,t|\boldsymbol{z}_0)-\mathbf{P}(s,t|\boldsymbol{z}_0)\right\|_1\xrightarrow{P}0, \tag{2.44}$$

where $\|\mathbf{A}\|_1=\sup_i\sum_j|\mathrm{a}_{ij}|$ for a matrix $\mathbf{A}=(\mathrm{a}_{ij})$.

Proof. See the detailed proof for Theorem 2.2.3 in Andersen et al. (1991) and Andersen et al. (1993). ■

Theorem 2.2.4 (Section 3 in Andersen et al. (1991) and Page 514 in Andersen et al. (1993))**.** Under regularity conditions in Section 2.2.3, $n^{1/2}\left\{\hat{\mathbf{P}}(s,t|\boldsymbol{z}_0)-\mathbf{P}(s,t|\boldsymbol{z}_0)\right\}$ can be asymptotically rewritten as

$$\int_s^t\hat{\mathbf{P}}(s,u|\boldsymbol{z}_0)d\left\{\mathbf{W}_1(u)+\mathbf{W}_2(u)\right\}\mathbf{P}(u,t|\boldsymbol{z}_0), \tag{2.45}$$

where the $Q\times Q$ matrices $\mathbf{W}_1(t)$ and $\mathbf{W}_2(t)$ are asymptotically independent. They have (h,j) elements, $h\neq j$,

$$W_{1,hj}(t)=n^{1/2}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)^T\int_0^t\left\{\boldsymbol{z}_{0,hj}-\mathbf{e}_{hj}(\boldsymbol{\beta}_0,u)\right\}e^{\boldsymbol{\beta}_0^T\boldsymbol{z}_{0,hj}}\lambda_{0,hj}(u)du \tag{2.46}$$

and

$$W_{2,hj}(t) = \int_0^t \frac{J_h(u)e^{\boldsymbol{\beta}_0^T \boldsymbol{z}_{0,hj}} dM_{hj}(u)}{n^{1/2} S_{hj}^{(0)}(\boldsymbol{\beta}_0, u)}, \tag{2.47}$$

respectively, in which $J_h(u) = I\left\{\sum_{i=1}^n Y_{i,h}(u) > 0\right\}$ and $M_{hj}(u) = \sum_{i=1}^n N_{i,hj}(u) - \sum_{i=1}^n \int_0^u Y_{i,h}(v) e^{\boldsymbol{\beta}_0^T \boldsymbol{Z}_{i,hj}} \lambda_{0,hj}(v) dv$. Furthermore, we define $W_{\nu,hh}(t) = -\sum_{j \neq h} W_{\nu,hj}(t)$, for $h = 1, 2, \ldots, Q$ and $\nu = 1, 2$.

Corollary 2.2.4.1. Under regularity conditions in Theorem 2.2.4, the stochastic process $n^{1/2}\left\{\hat{\mathbf{P}}(s,t|\boldsymbol{z}_0) - \mathbf{P}(s,t|\boldsymbol{z}_0)\right\}$ converges weakly to a zero-mean Gaussian process whose variance function can be estimated by

$$\hat{\mathbf{V}}(s,t|\boldsymbol{z}_0) = \int_s^t \hat{\mathbf{P}}(u,t|\boldsymbol{z}_0)^T \oplus \hat{\mathbf{P}}(s,u|\boldsymbol{z}_0) d\left[\mathbf{W}_1(u) + \mathbf{W}_2(u)\right] \hat{\mathbf{P}}(u,t|\boldsymbol{z}_0) \oplus \hat{\mathbf{P}}(s,u|\boldsymbol{z}_0)^T, \tag{2.48}$$

where $\oplus$ denotes the Kronecker product and $[\mathbf{W}(u)]$ is the quadratic variation process of $\mathbf{W}(u)$.

Corollary 2.2.4.2. Under regularity conditions in Theorem 2.2.4, the variance of the process $\hat{\mathbf{P}}(s,t|\boldsymbol{z}_0)$ can be estimated by $\hat{\mathbf{var}}(\hat{\mathbf{P}}(s,t|\boldsymbol{z}_0))$, whose (hj,mr) element is given by $\hat{\mathrm{var}}_1(\hat{P}_{hj}(s,t|\boldsymbol{z}_0), \hat{P}_{mr}(s,t|\boldsymbol{z}_0)) + \hat{\mathrm{var}}_2(\hat{P}_{hj}(s,t|\boldsymbol{z}_0), \hat{P}_{mr}(s,t|\boldsymbol{z}_0))$, where

$$\begin{aligned} &\hat{\mathrm{var}}_1(\hat{P}_{hj}(s,t|\boldsymbol{z}_0), \hat{P}_{mr}(s,t|\boldsymbol{z}_0)) \\ &\quad = \left\{\int_s^t \sum_{g,l} \hat{P}_{hg}(s,u|\boldsymbol{z}_0) dD_{gl}(u) \hat{P}_{lj}(u,t)\right\} \\ &\qquad \times (n^{-1}\hat{\boldsymbol{\Sigma}}_3)\left\{\int_s^t \sum_{g,l} \hat{P}_{mg}(s,u|\boldsymbol{z}_0) dD_{gl}(u) \hat{P}_{lr}(u,t|\boldsymbol{z}_0)\right\}, \end{aligned} \tag{2.49}$$

and

$$\begin{aligned}
&\hat{\text{var}}_2(\hat{P}_{hj}(s,t|\boldsymbol{z}_0),\hat{P}_{mr}(s,t|\boldsymbol{z}_0)) \\
&\quad=\sum_{g\neq l}\int_s^t \hat{P}_{hg}(s,u|\boldsymbol{z}_0)\hat{P}_{mg}(s,u|\boldsymbol{z}_0)\left\{\hat{P}_{lj}(u,t)|\boldsymbol{z}_0-\hat{P}_{gj}(u,t|\boldsymbol{z}_0)\right\} \\
&\qquad\times\left\{\hat{P}_{lr}(u,t|\boldsymbol{z}_0)-\hat{P}_{gr}(u,t|\boldsymbol{z}_0)\right\}J_g(u)e^{2\cdot\hat{\boldsymbol{\beta}}^T\boldsymbol{z}_{0,gl}}\left\{nS_{gl}^{(0)}(\hat{\boldsymbol{\beta}},u)\right\}^{-2}dN_{gl}(u),
\end{aligned} \tag{2.50}$$

in which for $g\neq l$, $D_{gl}(t)=\int_0^t\left\{\boldsymbol{z}_{0,gl}-\mathbf{E}_{gl}(\hat{\boldsymbol{\beta}},u)\right\}J_g(u)e^{\hat{\boldsymbol{\beta}}^T\boldsymbol{z}_{0,gl}}\left\{nS_{gl}^{(0)}(\hat{\boldsymbol{\beta}},u)\right\}^{-1}dN_{gl}(u)$ while $D_{gg}(t)=-\sum_{l\neq g}D_{gl}(t)$, and $N_{gl}(u)=\sum_{i=1}^n N_{i,gl}(u)$.

Proof. The detailed proofs for Theorem 2.2.4, Corollary 2.2.4.1, and Corollary 2.2.4.2 can be found in Andersen et al. (1991), and Andersen et al. (1993). ■

2.3 Regularization Using Confidence Distribution

2.3.1 Regularized Likelihood Approach

When studying the dependence between response and covariates using regression analysis, often, many covariates are collected and to reduce possible modeling bias, a large parametric model is built. In such cases, variable selection is important in that it not only enhances the model interpretability with parsimonious representation, but also improves the prediction performance of the fitted model. Traditionally, the best-subset selection method is widely used to select significant predictors, but this procedure, according to Zou (2006), has two limitations: 1) it is computationally infeasible when the number of predictors is large; 2) it is extremely variable because of its inherent discreteness. The forward/backward/stepwise selection, which is used as a computational surrogate to best-subset selection, also suffers from lacking stability and in addition is often trapped into a local optimal solution rather than the global optimal solution. Furthermore, the statistical properties of selected predictors using these traditional methods are mostly unknown (Fan and Li, 2001).

In the past decades, a family of new techniques based on the regularized likelihood framework has been proposed to approach the problem of variable selection. Assume that $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_d)^T$ is a parameter of interest and $l_n(\boldsymbol{\beta})$ is a plausible loss function (e.g., least square function or log likelihood function). In the regularized likelihood estimation approach, it is of interest to consider the objective function

$$-n^{-1} l_n(\boldsymbol{\beta}) + \sum_{j=1}^{d} p_i(|\beta_j|) \tag{2.51}$$

where $p_i(.)$ is the penalty function indexed by the regularization parameter ρ_i that is possibly different for each β_j. By minimizing (2.51), the parameter estimation and variable selection can be simultaneously executed. In other words, those covariates whose regres-

sion coefficients are estimated as zero are automatically deleted.

Fan and Li (2001) advocated penalty functions that give estimators with three properties: (1) the resulting estimator is nearly unbiased, especially when the true coefficient $\boldsymbol{\beta}_j$ is large, to reduce model bias ("*unbiasedness*") ; (2) the resulting estimator automatically sets small estimated coefficients to zero to accomplish variable selection and reduce model complexity ("*sparsity*"); (3) the resulting estimator is continuous in the data to reduce instability in model prediction ("*continuity*").

It is known that the convex L_q penalty function with $q > 1$ does not satisfy the sparsity condition, whereas the convex L_1 penalty (also known as LASSO penalty (Tibshirani, 1996)) does not satisfy the unbiasedness condition, and the concave L_q penalty with $0 \leq q < 1$ does not satisfy the continuity condition. The SCAD penalty introduced by Fan and Li (2001) and the MCP penalty developed by Zhang et al. (2010) not only satisfy the above three conditions, but also enjoy oracle properties; namely, they perform as well as if the true underlying model were given in advance. In the language of Fan and Li (2001), denote the true model by $\mathcal{A} = \left\{j : \beta_j^* \neq 0\right\}$ and further assume that $|\mathcal{A}| = d_0 < d$, then the regularized estimator $\hat{\boldsymbol{\beta}}(\delta)$ obtained by using the oracle procedure δ satisfies the following oracle properties:

- identifying the right subset model: $\left\{j : \hat{\beta}_j(\delta) \neq 0\right\} = \mathcal{A}$;
- has the optimal estimation rate: $n^{1/2}\left(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\delta) - \boldsymbol{\beta}_{\mathcal{A}}^*\right) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^*)$, where $\boldsymbol{\Sigma}^*$ is the variance matrix knowing the true subset model.

The adaptive LASSO estimator proposed by Zou (2006) also has these nice properties given a good initial estimate is provided. In the adaptive LASSO penalty

$$p_i(|\beta_j|) = \rho_0 |\tilde{\beta}_j|^{-\phi}, \tag{2.52}$$

where ρ_0 and ϕ are tuning parameters, generally, $\tilde{\beta}_j$ is required to be $n^{1/2}$-consistent. For instance, $\tilde{\beta}_j$'s can be chosen as the un-regularized minimizer $\hat{\boldsymbol{\beta}}^{mle} = \operatorname{argmin}_{\boldsymbol{\beta}} \{-n^{-1} l_n(\boldsymbol{\beta})\}$.

Alternatively, if collinearity is a concern, the ridge estimator $\hat{\boldsymbol{\beta}}^{ridge}$ (van Wieringen, 2015) is suggested to serve as the initial estimate. The adaptive LASSO estimator (Zou, 2006) is computationally appealing as its penalty function is convex. The entire solution path to the adaptive LASSO optimization can be obtained as efficiently as a single un-regularized full model fitting — at the order of $O(nd^2)$. On the other hand, the noncovex penalty functions, such as SCAD or MCP, the global optimality is not guaranteed.

These regularized estimation approaches have been extensively studied and extended into univariate survival analysis using the Cox Proportional hazards model (Cox, 1972), such as the LASSO estimator (Tibshirani, 1997), the SCAD estimator (Fan and Li, 2002), the adaptive LASSO estimator (Zhang and Lu, 2007; Wang and Leng, 2007), and the MCP estimator (Zhang et al., 2010). However, only the SCAD estimator has been studied in multivariate survival analysis (Fan and Li, 2002; Cai et al., 2005). The desired properties, such as consistency and asymptotic normality of the SCAD estimators were established in both the proportional hazards frailty model and the marginal proportional hazards model. They further demonstrated that, with proper choices of tuning parameters, the SCAD estimators can correctly identify the true model, as if it were known in advance. Moreover, they provided sandwich formulas to obtain the standard errors of estimated regularized estimators. However, these approaches are not widely used in statistical analysis of biomedical research. A main reason for this could be the lack of software. In this dissertation, we propose to perform variable selection via a regularized confidence distribution approach, which is easy to be implemented using existing software packages.

2.3.2 Regularized Confidence Distribution Approach

Confidence distribution (CD) is a concept loosely referring to a distribution function that can represent confidence intervals of all levels for a parameter of interest (Xie et al., 2011). The concept of CD has a long history (e.g., Cox (1958), Fisher (1956), and Efron (1993)), but recent development has redefined the CD concept as a purely frequentist concept and

focused on providing inference tools for problems in modern applied statistics (Xie et al., 2011). Generally speaking, the CD approach is a statistical inference tool in providing a distribution estimate, instead of a point estimate, or an interval estimate, for a parameter of interest. For example, let $\mathbf{X} = \{X_i;\ i = 1, 2, \ldots, n\}$ denote an independent and identically distributed random sample drawn from a normal density with mean μ and variance 1, where μ is the parameter of interest. A point estimate and an interval estimate are given by $n^{-1}\sum_{i=1}^{n} X_i$ and $(n^{-1}\sum_{i=1}^{n} X_i - 1.96 \cdot n^{-1/2},\ n^{-1}\sum_{i=1}^{n} X_i + 1.96 \cdot n^{-1/2})$, respectively. Whereas the distribution estimate provided by the CD approach is $\mathcal{N}(n^{-1}\sum_{i=1}^{n} X_i,\ n^{-1})$, where $\mathcal{N}$ stands for normal distribution.

The CD approach can be used as a device to combine information from independent samples. Suppose $H_s(\theta) = H_s(\mathbf{X}_s, \theta)$, $s = 1, 2, \ldots, S$ are CD functions (i.e., sample-dependent continuous cumulative distribution functions) for the same parameter θ from S independent samples $\mathbf{X}_s$ and the sample size of $\mathbf{X}_s$ is n_s. Singh et al. (2005) proposed to combine CD functions using a coordinate-wise monotonic function from the S-dimensional cube $[0, 1]^S$ to the real line $\mathbb{R} = (-\infty, +\infty)$. Specifically, they suggested to combine S CD functions as

$$H^{(c)}(\theta) = G_c\left[g_c\left\{H_1(\theta), H_2(\theta), \ldots, H_S(\theta)\right\}\right], \tag{2.53}$$

where $g_c(u_1, u_2, \ldots, u_S)$ is a given continuous function on $[0, 1]^S \to \mathbb{R}$ which is monotonic (without loss of generality, say, increasing) in each coordinate, and G_c is completely determined by g_c, i.e., $G_c(t) = Pr\left\{g_c(U_1, U_2, \ldots, U_S) \le t\right\}$, where $U_1, U_2, \ldots, U_S$ are independent random variables following the uniform distribution $\mathcal{U}[0, 1]$. It can be shown that $H^{(c)}(\theta)$ is also a CD function for the parameter θ when the underlying true parameter values of the S individual CD functions $H_s(\theta)$ are the same (Xie et al., 2011).

The usefulness of the CD approach in combining information from multiple sources has been demonstrated in many practical situations with much success. For example, Xie et al. (2011) proposed robust meta-analysis approaches using the CD approach, with supporting asymptotic theories. Their proposed methodologies performed well even when data

are contaminated and have realistic sample sizes and number of studies. Liu et al. (2014) combined the p-value functions (i.e., distribution estimators of the unknown parameter) associated with the exact tests from multiple studies of discrete data. Their proposed exact approach was shown to be efficient and, generally, outperformed commonly used methods in discrete data, such as Mantel-Haenszel and Peto methods. Liu et al. (2015) proposed a meta-analysis method that can incorporates heterogeneous studies, which are excluded from conventional meta-analysis, by combining the confidence density functions derived from the summary statistics of individual studies. Their proposed meta analysis is shown to be asymptotically as efficient as the maximum likelihood approach using individual participant data from all studies.

Consider S independent studies with n_s participants in the s^{th} study, $s = 1, 2, \ldots, S$. Assume we are interested in making inference for a d-dimensional parameter vector $\boldsymbol{\theta}$, which is associated with the S studies. Under the general likelihood framework, in the s^{th} study ($s = 1, 2, \ldots, S$), denote the maximum likelihood estimator of $\boldsymbol{\theta}$ by $\hat{\boldsymbol{\theta}}_s$, namely, $\hat{\boldsymbol{\theta}}_s = \text{argmax}_{\boldsymbol{\theta}}\, \mathcal{L}_s(\boldsymbol{\theta})$; and express the observed information as $\mathcal{I}_s(\boldsymbol{\theta}) = -\partial^2 log\mathcal{L}_s(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T\partial\boldsymbol{\theta}$, where $\mathcal{L}_s(\boldsymbol{\theta})$ is the likelihood function. Then $\hat{\boldsymbol{\Sigma}}_s = \mathcal{I}_s^{-1}(\hat{\boldsymbol{\theta}}_s)$ is an estimate of the variance matrix for $\hat{\boldsymbol{\theta}}_s$. Under typical regularity conditions in likelihood inference, it follows that $\mathcal{I}_s/n_s \xrightarrow{P} \boldsymbol{I}_s$, and $n_s^{-1/2}\left\{\partial log\mathcal{L}_s(\boldsymbol{\theta})/\partial\boldsymbol{\theta}^T\right\} \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{I}_s)$, as $n_s \to \infty$. Here, $\boldsymbol{I}_s$ is the Fisher information. As discussed in Singh et al. (2007) and Liu et al. (2015), the density function of $\mathcal{N}(\hat{\boldsymbol{\theta}}_s, \hat{\boldsymbol{\Sigma}}_s)$ can serve as a confidence density (i.e., the density function of a confidence distribution) for the parameter $\boldsymbol{\theta}$. Denote the density function by $h_s(\boldsymbol{\theta}|\mathbf{S}_s)$, where $\mathbf{S}_s$ represents the sample in the s^{th} study. More specifically,

$$h_s(\boldsymbol{\theta}|\mathbf{S}_s) = \frac{1}{(2\pi)^{d/2}(\det(\hat{\boldsymbol{\Sigma}}_s))^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_s)^T\hat{\boldsymbol{\Sigma}}_s^{-1}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_s)\right\}, \tag{2.54}$$

for $s = 1, 2, \ldots, S$, where $\det(\mathbf{C})$ is the determinant of a matrix $\mathbf{C}$. The combined confidence density can be obtained by combining these confidence density functions $h_s(\boldsymbol{\theta}|\mathbf{S}_s)$

in the S studies as suggested in (2.53). Then the combined confidence density function can be used for regularization because this distribution estimator contains as much information as that in the full likelihood function constructed using all participants from S studies.

CHAPTER 3
PROPOSED DIVIDE-AND-COMBINE IN MULTIVARIATE SURVIVAL ANALYSIS

3.1 Introduction

The idea behind the divide-and-combine approach can be illustrated by a toy example of ordinary linear regression for univariate outcomes. Denote the regression parameter by $\boldsymbol{\alpha}$ and assume that the full data are divided into S subsets. Some algebraic calculations show that a weighted average of the subset ordinary least squares (OLS) estimators $\hat{\boldsymbol{\alpha}}_s$'s, with weight $\mathbf{X}_s^T\mathbf{X}_s$, is identical to the full-data estimator, i.e,

$$\hat{\boldsymbol{\alpha}}^{dc} = \left(\sum_{s=1}^{S} \mathbf{X}_s^T\mathbf{X}_s\right)^{-1} \sum_{s=1}^{S} \mathbf{X}_s^T\mathbf{X}_s\hat{\boldsymbol{\alpha}}_s = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \hat{\boldsymbol{\alpha}}^{full}, \tag{3.1}$$

where $\hat{\boldsymbol{\alpha}}^{dc}$ denotes the weighted-average estimator, $\mathbf{X}_s$, $\mathbf{y}_s$, and $\mathbf{X}$, $\mathbf{y}$ are the design matrices and the response vectors in subset s ($s = 1, 2, \ldots, S$) and the full data, respectively. Note that the weight $\mathbf{X}_s^T\mathbf{X}_s$ is proportional to $\boldsymbol{I}_s(\hat{\boldsymbol{\alpha}}_s)$ that is the inverse of $\hat{\text{var}}(\hat{\boldsymbol{\alpha}}_s)$, where $\boldsymbol{I}_s(.)$ denotes the observed information matrix of the log likelihood function. Thus, we can rewrite $\hat{\boldsymbol{\alpha}}^{dc}$ as

$$\hat{\boldsymbol{\alpha}}^{dc} = \left\{\sum_{s=1}^{S} \boldsymbol{I}_s(\hat{\boldsymbol{\alpha}}_s)\right\}^{-1} \sum_{s=1}^{S} \boldsymbol{I}_s(\hat{\boldsymbol{\alpha}}_s)\hat{\boldsymbol{\alpha}}_s. \tag{3.2}$$

This type of weighted-average estimator has been extensively studied in the meta analysis and other divide-and-combine literature (e.g., Lin and Zeng (2010) and Tang et al. (2016)). In this chapter, we extend this divide-and-combine and weighted average idea to the multivariate survival analysis. Section 3.2 setups models in multivariate survival analysis, including both marginal models and frailty models. The proposed divide-and-combine estimators are presented in Section 3.3.

3.2 Model Setups

Two classes of models in multivariate survival analysis are of interest in this dissertation, including the marginal proportional hazards model and the proportional hazards frailty model. In the marginal model approach, we assume that the marginal distributions of failure times of K different failure types follow proportional hazards models (Cox, 1972). The hazard function for the k^{th} type of failure is given by

$$\lambda_k(t|\boldsymbol{Z}_k) = \lambda_{0k}(t)e^{\boldsymbol{\beta}^T\boldsymbol{Z}_k(t)}, \tag{3.3}$$

where $\lambda_{0k}(.)$ is the unspecified baseline hazard function and $\boldsymbol{\beta}$ is a d-dimensional regression parameter whose associated covariate $\boldsymbol{Z}_k(.)$ is possibly time-varying. In the frailty model approach, the conditional distributions of failure times, given frailties, are formulated again by proportional hazards models (Cox, 1972). The hazard function for the k^{th} type of failure takes the form

$$\lambda_k(t|\boldsymbol{Z}_k, u) = u\lambda_0(t)e^{\boldsymbol{\beta}^T\boldsymbol{Z}_k}, \tag{3.4}$$

where $\lambda_0(.)$ is the unspecified baseline hazard function, u is the frailty with a density function $f_U(u|\theta)$ indexed by θ, and $\boldsymbol{\beta}$ represents a d-dimensional regression parameter. Of note, in our proportional hazards frailty model, $\boldsymbol{Z}_k$ is a time-fixed covariate. Both the marginal model in (3.3) and the frailty model in (3.4) can accommodate type-specific regression parameters by appropriately specifying $\boldsymbol{Z}_k$ for $k = 1, 2, \ldots, K$ (see more details in Sections 2.1.1 and 2.1.2).

By using the estimation and inference procedures described in Sections 2.1.1 and 2.1.2, we can estimate $\hat{\boldsymbol{\beta}}$ in the marginal model and $\hat{\boldsymbol{\gamma}} = (\hat{\theta}, \hat{\boldsymbol{\beta}}^T)^T$ in the frailty model, respectively. Our proposed divide-and-combine approach applies to both the marginal model and the frailty model. Thus, for the ease of notation, denote the unknown parameter by $\boldsymbol{\eta}$, where $\boldsymbol{\eta} = \boldsymbol{\beta}$ in the marginal proportional hazards model and $\boldsymbol{\eta} = \boldsymbol{\gamma} = (\theta, \boldsymbol{\beta}^T)^T$ in

the proportional hazards frailty model, and use $\hat{\boldsymbol{\eta}}^{full}$ to denote the full-data estimator from the (all-at-once) full data analysis. Specifically, $\hat{\boldsymbol{\eta}}^{full} = \hat{\boldsymbol{\beta}}$ for the marginal model and $\hat{\boldsymbol{\eta}}^{full} = \hat{\boldsymbol{\gamma}}$ for the frailty model. Throughout the remainder of this dissertation, we will use $\boldsymbol{\eta}$ to discuss our methodology without distinguishing marginal models or frailty models, unless necessary.

3.3 Divide-and-Combine Estimators for Regression Parameters

In multivariate survival analysis, we are interested in estimating regression parameters. However, when the sample size is extraordinarily large (when using the marginal model) or the estimation procedure is complex (when using the frailty model), it is difficult to finish the analysis within a reasonable time period. Thus to conquer these issues, we propose our divide-and-combine approach for estimating regression parameters as follows.

Divide Consider n independent clusters with cluster i potentially experiencing K distinct types of failures, $i = 1, 2, \ldots, n$. The observed full data can be denoted by $\mathbf{D}_i = \{X_{ik}, \delta_{ik}, \boldsymbol{Z}_{ik};\ k = 1, 2, \ldots, K\}$, where X_{ik} is the observed time, δ_{ik} is the censoring indicator, and $\boldsymbol{Z}_{ik}$ is a d-dimensional vector of covariates for $i = 1, 2, \ldots, n$. We randomly divide the full data by independent clusters into S subsets, $\mathbf{D}_{si} = \{X_{sik}, \delta_{sik}, \boldsymbol{Z}_{sik};\ k = 1, 2, \ldots, K\}$ for $s = 1, 2, \ldots, S$ and $i = 1, 2, \ldots, n_s$, where n_s is the number of independent clusters of the s^{th} subset with $n_s \gg d$. Note that $\sum_{s=1}^{S} n_s = n$, where n is the total sample size. Theoretically, the simple random splitting would yield homogeneous subsets and each subset would be a representative random sample of the full data. Thus it is reasonable to assume the same marginal proportional hazards model or the same proportional hazards frailty model holds in subsets as in the full data. In each subset s, we obtain the parameter estimator $\hat{\boldsymbol{\eta}}_s$ using the aforementioned marginal model approach or frailty model approach in Section 3.2 (see Sections 2.1.1 and 2.1.2 for more details). Throughout this dissertation, we assume that all n_s's diverge in the same order of $O(n/S)$ and $S = o(n^{1/2})$, following Zhang et al. (2013), Chen and Xie (2014), and Rosenblatt and Nadler (2016).

Additionally, we also consider stratified random splitting by the number of events per cluster to balance the distribution of the number of events across divided subsets. This method is implemented in the simulation studies (Chapter 6).

Combine We use a weighted average similar to (3.2) to combine the S estimators and define the divide-and-combine estimator $\hat{\boldsymbol{\eta}}^{dc}$ by

$$\hat{\boldsymbol{\eta}}^{dc} = \left\{ \sum_{s=1}^{S} \mathbf{W}_s(\hat{\boldsymbol{\eta}}_s) \right\}^{-1} \sum_{s=1}^{S} \mathbf{W}_s(\hat{\boldsymbol{\eta}}_s)\hat{\boldsymbol{\eta}}_s, \tag{3.5}$$

where $\mathbf{W}_s(.)$ is a weight function. Specifically, we consider $\mathbf{W}_s(.)$ with three choices: 1) the minus second derivative of the log likelihood, i.e., $\mathbf{W}_{1s}(\hat{\boldsymbol{\eta}}_s) = \mathcal{I}_s(\hat{\boldsymbol{\eta}}_s)$, motivated by the previous simple OLS regression example; 2) the inverse of the estimated variance of $\hat{\boldsymbol{\eta}}_s$, i.e., $\mathbf{W}_2(\hat{\boldsymbol{\eta}}_s) = \hat{\text{var}}_s^{-1}(\hat{\boldsymbol{\eta}}_s)$, motivated by the meta approach studied by Lin and Zeng (2010) and Tang et al. (2016); 3) the sample size, i.e., $\mathbf{W}_3(\hat{\boldsymbol{\eta}}_s) = n_s$, a simple and frequently used weight. Note that in multivariate survival analysis, when the intra-cluster association is non-trivial, $\mathcal{I}_s(\hat{\boldsymbol{\beta}}_s) \neq \hat{\text{var}}_s^{-1}(\hat{\boldsymbol{\beta}}_s)$ (see Sections 2.1.1 and 2.1.2).

Here, we state formally the following assumptions, which are the key to establishing the asymptotic properties for divide-and-combine estimators.

Homogeneity Assumptions Let $\boldsymbol{\eta}_0$ and $\boldsymbol{\eta}_{s0}$ ($s = 1, 2, \ldots, S$) be the true values of $\boldsymbol{\eta}$ in the full data and individual subsets. The homogeneity is two-fold:

(H1) Underlying parameters are the same across all subsets, i.e., $\boldsymbol{\eta}_{s0} = \boldsymbol{\eta}_0$;

(H2) Partitioned subsets are representatives of the full data, i.e., $\mathbf{A}_s(\boldsymbol{\eta}) = \mathbf{A}(\boldsymbol{\eta})$ and $\mathbf{B}_s(\boldsymbol{\eta}) = \mathbf{B}(\boldsymbol{\eta})$.

We assume the regularity conditions (M1) to (M6) and (F1) to (F10) (see Sections 2.1.1 and 2.1.2) hold in individual subsets for the marginal model approach and the frailty model approach, respectively. Then the aforementioned consistency and asymptotic normality properties in the marginal model or in the frailty model also hold for $\hat{\boldsymbol{\eta}}_s$, $s = 1, 2, \ldots, S$.

Under the homogeneity assumptions, it follows that $n^{1/2}(\hat{\boldsymbol{\eta}}^{full} - \boldsymbol{\eta}_0) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_1$ or $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_2$ when the marginal model approach or the frailty model approach is used, respectively. We establish the asymptotic properties of $\hat{\boldsymbol{\eta}}^{dc}$ in Theorem 3.3.1. Specifically, we show that the estimators $\hat{\boldsymbol{\eta}}^{dc}$ using the three weights are asymptotically equivalent to the estimator $\hat{\boldsymbol{\eta}}^{full}$ from the full data, in the sense that they all are consistent and converge to the same limiting distribution.

Theorem 3.3.1. Under regularity conditions (M1) to (M6) in Section 2.1.1, (F1) to (F10) in Section 2.1.2, and the homogeneity assumptions (H1) to (H2) in Section 3.3, the estimator $\hat{\boldsymbol{\eta}}^{dc}$ using weight $\mathbf{W}_{1s}(.)$, $\mathbf{W}_{2s}(.)$, or $\mathbf{W}_{3s}(.)$, satisfies the following as $n \to \infty$:

(1) (Consistency) $\hat{\boldsymbol{\eta}}^{dc} \xrightarrow{P} \boldsymbol{\eta}_0$;

(2) (Asymptotic Normality) $n^{1/2}(\hat{\boldsymbol{\eta}}^{dc} - \boldsymbol{\eta}_0) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma} = \{\mathbf{A}^{-1}(\boldsymbol{\eta}_0)\}\{\mathbf{B}(\boldsymbol{\eta}_0)\}\{\mathbf{A}^{-1}(\boldsymbol{\eta}_0)\}^T)$;

(3) The variance $\boldsymbol{\Sigma}$ corresponding to $\hat{\boldsymbol{\eta}}^{dc}$ can be consistently estimated by $\hat{\boldsymbol{\Sigma}}_{dc}$, where $\hat{\boldsymbol{\Sigma}}_{dc} = [\{\sum_{s=1}^{S}(n_s/n)\hat{\mathbf{A}}_s(\hat{\boldsymbol{\eta}}_s)\}^{-1}][\sum_{s=1}^{S}(n_s/n)\hat{\mathbf{B}}_s(\hat{\boldsymbol{\eta}}_s)][\{\sum_{s=1}^{S}(n_s/n)\hat{\mathbf{A}}_s(\hat{\boldsymbol{\eta}}_s)\}^{-1}]^T$ when $\mathbf{W}_s(.) = \mathbf{W}_{1s}(.)$; $\hat{\boldsymbol{\Sigma}}_{dc} = n\{\sum_{s=1}^{S} \hat{\text{var}}_s^{-1}(\hat{\boldsymbol{\eta}}_s)\}^{-1}$ when $\mathbf{W}_s(.) = \mathbf{W}_{2s}(.)$; and $\hat{\boldsymbol{\Sigma}}_{dc} = \sum_{s=1}^{S}(n_s^2/n)\hat{\text{var}}_s(\hat{\boldsymbol{\eta}}_s)$ when $\mathbf{W}_s(.) = \mathbf{W}_{3s}(.)$. Of note, $\hat{\text{var}}_s(\hat{\boldsymbol{\eta}}_s) = n_s^{-1}\{\hat{\mathbf{A}}_s^{-1}(\hat{\boldsymbol{\eta}}_s)\}\{\hat{\mathbf{B}}_s(\hat{\boldsymbol{\eta}}_s)\}\{\hat{\mathbf{A}}_s^{-1}(\hat{\boldsymbol{\eta}}_s)\}^T$, and $\hat{\mathbf{A}}_s(.)$ and $\hat{\mathbf{B}}_s(.)$ are defined in the same form as $\hat{\mathbf{A}}(.)$ and $\hat{\mathbf{B}}(.)$ in Sections 2.1.1 and 2.1.2, for each subset s, $s = 1, 2, \ldots, S$.

Proof. The homogeneity assumptions guarantee that, when the same weight function is used, $n_s^{-1}\mathbf{W}_s(.)$ converges to a constant in probability, and that $\hat{\boldsymbol{\eta}}_s$ converges to $\boldsymbol{\eta}_0$ in probability. It follows immediately that $\hat{\boldsymbol{\eta}}^{dc}$ is a consistent estimator. The asymptotic normality of $\hat{\boldsymbol{\eta}}^{dc}$ can be demonstrated using the Taylor series expansion. When deriving the asymptotic properties of $\hat{\boldsymbol{\eta}}^{dc}$, it is suggested that both the total sample size (n) and the number of subsets (S) can go to infinity, but S should be chosen such that its order is smaller than $n^{1/2}$, i.e., $S = o(n^{1/2})$. See Appendix A.6 for the detailed proof. ■

Under the special case of univariate failure time data, $\hat{\boldsymbol{\eta}}^{dc}$ with $\mathbf{W}_{1s}(.)$ and $\mathbf{W}_{2s}(.)$ are the same because $\mathcal{I}_s(\hat{\boldsymbol{\eta}}_s) = \hat{\text{var}}_s^{-1}(\hat{\boldsymbol{\eta}}_s)$, and it reduces to the inverse-variance meta

estimator considered in Lin and Zeng (2010). When the subsets are "equal splits" (i.e., $n_s = n/S$, for $s = 1, 2, \ldots, S$), $\hat{\boldsymbol{\eta}}^{dc}$ using $\mathbf{W}_{3s}(.)$ becomes the simple average of $\hat{\boldsymbol{\eta}}_s$'s, i.e., $\hat{\boldsymbol{\eta}}^{dc} = \sum_{s=1}^{S} \hat{\boldsymbol{\eta}}_s / S$.

When the sample size is large, theoretically $\hat{\boldsymbol{\eta}}^{dc}$ using all three weights are equivalent under the homogeneity assumptions (H1) to (H2). When the homogeneity assumption (H2) is violated such that $\mathbf{A}_s(\boldsymbol{\eta}_0) \neq \mathbf{A}(\boldsymbol{\eta}_0)$ or $\mathbf{B}_s(\boldsymbol{\eta}_0) \neq \mathbf{B}(\boldsymbol{\eta}_0)$ for some s ($1 \leq s \leq S$), we show in Appendix A.7 that the consistency and the asymptotic normality of $\hat{\boldsymbol{\eta}}^{dc}$ still hold. However, the asymptotic variance of $n^{1/2}(\hat{\boldsymbol{\eta}}^{dc} - \boldsymbol{\eta}_0)$ varies with weight and it satisfies that $\boldsymbol{\Sigma}_{\mathrm{w}2} \leq \boldsymbol{\Sigma}_{\mathrm{w}1}$, $\boldsymbol{\Sigma}_{\mathrm{w}2} \leq \boldsymbol{\Sigma}_{\mathrm{w}3}$, $\boldsymbol{\Sigma}_{\mathrm{w}1} = \boldsymbol{\Sigma}$, and $\boldsymbol{\Sigma}_{\mathrm{w}1} \neq \boldsymbol{\Sigma}_{\mathrm{w}3}$, where $\boldsymbol{\Sigma}_{\mathrm{w}i}$ represents the asymptotic variance using weight $\mathbf{W}_{is}(.)$ for $i = 1, 2, 3$. Note that $\boldsymbol{\Sigma}_{\mathrm{w}1} = \boldsymbol{\Sigma}$, and $\boldsymbol{\Sigma}$ is the asymptotic variance of $\hat{\boldsymbol{\eta}}^{full}$. Because the sandwich-type $\boldsymbol{\Sigma}$ may not be optimal (i.e., minimal) in terms of efficiency (in contrast to the inverse of Fisher's information as the asymptotic variance of the maximum likelihood estimator) (Kauermann and Carroll, 2001), it is possible to achieve that $\boldsymbol{\Sigma}_{\mathrm{w}2} \leq \boldsymbol{\Sigma}$. Also, in our simulation studies of finite samples, it is shown that, empirically, $\boldsymbol{\Sigma}_{\mathrm{w}2}$ is slightly smaller than $\boldsymbol{\Sigma}_{\mathrm{w}1}$, and as a result, the empirical coverage probability of the confidence interval is slightly under the nominal level. Finally, since $\mathbf{W}_{1s}(.)$ and $\mathbf{W}_{2s}(.)$ utilize more empirical information from the data in individual subsets (in the form of either $\mathcal{I}_s(\hat{\boldsymbol{\eta}}_s)$ or $\hat{\mathrm{var}}_s^{-1}(\hat{\boldsymbol{\eta}}_s)$), while $\mathbf{W}_{3s}(.)$ only utilizes the information of the subset sample size (a single number), one may expect that, empirically, $\hat{\boldsymbol{\eta}}^{dc}$ using $\mathbf{W}_{1s}(.)$ or $\mathbf{W}_{2s}(.)$ may outperform that using $\mathbf{W}_{3s}(.)$. Performance of $\hat{\boldsymbol{\eta}}^{dc}$ using these three weights is assessed in the simulation studies (Chapter 6).

CHAPTER 4
PROPOSED REGULARIZATION USING CONFIDENCE DISTRIBUTION

4.1 Introduction

In modern statistical analysis, as discussed in Section 2.3.1, variable selection can be achieved by regularized estimation whose objective function is typically constructed from the likelihood function using the original data in the sample(s). When it comes to the multivariate survival analysis, the intra-cluster association needs to be appropriately taken care of. As a result, the regularized estimation algorithm is complicated and thus the existing software packages cannot be directly applied. In this chapter, we propose a simple regularized estimation approach in multivariate survival analysis using the confidence distribution of parameters. The proposed confidence distribution approach uses just the asymptotic distribution of $\hat{\boldsymbol{\eta}}^{dc}$ obtained from divide-and-combine analysis in Chapter 3, and enables variable selection in multivariate survival analysis using existing software packages. Section 4.2 presents the proposed objective functions using the confidence distribution approach. The asymptotic properties of the regularized estimators are given in Section 4.3. Section 4.4 shows two useful asymptotic equivalences. The optimization of the proposed objective functions, including determination of tuning parameters, is discussed in Section 4.5.

4.2 Proposed Objective Functions

Our proposed regularized estimation is based on confidence distribution of parameters. Inference using confidence distribution has been discussed extensively (e.g., Efron (1993, 1998); Lehmann (1993); Singh et al. (2007); Xie and Singh (2013)). A confidence density is the density function representation of a confidence distribution. By Singh et al. (2007), we write the confidence density of the parameter $\boldsymbol{\eta}$ based on the asymptotic distribution of

$\hat{\boldsymbol{\eta}}^{dc}$ shown in Theorem 3.3.1,

$$h(\boldsymbol{\eta}) = \frac{1}{(2\pi)^{d/2}(\det(n^{-1}\hat{\boldsymbol{\Sigma}}_{dc}))^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\eta}-\hat{\boldsymbol{\eta}}^{dc})^T(n^{-1}\hat{\boldsymbol{\Sigma}}_{dc})^{-1}(\boldsymbol{\eta}-\hat{\boldsymbol{\eta}}^{dc})\right\}, \quad (4.1)$$

where $\det(\mathbf{C})$ is the determinant of a matrix $\mathbf{C}$. To facilitate statistical inference, we use adaptive LASSO (Zou, 2006). Other penalty functions such as SCAD (Fan and Li, 2001) or MCP (Zhang et al., 2010) can be easily adopted. After dropping constant terms, we propose to construct the objective function by adding penalty terms to the log confidence density in (4.1):

$$R(\boldsymbol{\eta}) = n(\boldsymbol{\eta}-\hat{\boldsymbol{\eta}}^{dc})^T\hat{\boldsymbol{\Sigma}}_{dc}^{-1}(\boldsymbol{\eta}-\hat{\boldsymbol{\eta}}^{dc}) + n\sum_{j=1}^{d}\rho_j|\beta_j|, \quad (4.2)$$

where $\rho_1, \rho_2, \ldots, \rho_d$ denote tuning parameters, and $|.|$ is the absolute value of a scalar. Note that we only apply penalty restrictions on the regression parameter $\boldsymbol{\beta}$, and not on the association parameter θ when the divide-and-combine estimator is obtained using the frailty model approach, i.e., $\hat{\boldsymbol{\eta}}^{dc} = \hat{\boldsymbol{\gamma}}^{dc}$. Interestingly, the objective function in (4.2) takes the same form as the objective function based on the least squares approximation of Wang and Leng (2007). Similar coincidence also happens in linear regression where the objective functions in maximum likelihood and least squares estimation are identical under the normality assumption.

The regularized estimator of $\boldsymbol{\eta}$ is denoted by $\hat{\boldsymbol{\eta}}_{\rho}^{dc}$ obtained from minimizing (4.2) globally. With a proper choice of $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots, \rho_d)^T$ (see discussion in Section 4.5), some of the estimated coefficients will shrink to zero to achieve variable selection. Note that the objective function $R(\boldsymbol{\eta})$ in (4.2) has taken into account the intra-cluster association in multivariate survival analysis via $\hat{\boldsymbol{\Sigma}}_{dc}$. The following Theorems 4.3.1 and 4.3.2 show that the variance estimator of $\hat{\boldsymbol{\eta}}_{\rho}^{dc}$ does not require additional adjustment in the variance estimation, nor the joint estimation of regression coefficients and intra-cluster correlation in the regularized estimation (e.g., Cai et al. (2005) for multivariate unordered failure time

regression analysis; Rothman et al. (2010), Ibrahim et al. (2011), and Sofer et al. (2014) for multivariate regression analysis).

In our proposed method, the objective function $R(\boldsymbol{\eta})$ in (4.2) is only based on a pseudo sample of size d, instead of n, which leads to a substantial reduction in computational cost when $n \gg d$. Compared to some previous divide-and-combine approaches that perform variable selections in each subset before combination (e.g., Chen and Xie (2014); Tang et al. (2016)), our proposed method avoids the possibly inconsistent variable selection from different subsets before the final combination step. Furthermore, our proposed regularized estimation method is solely based on the confidence distribution of regression parameters, which lends itself to a unified approach for variable selection in a large family of regression models as long as consistent estimators and their well-established asymptotic distributions are tractable.

4.3 Regularized Estimators in Multivariate Survival Analysis

Marginal Models

In the marginal model, $\boldsymbol{\eta} = \boldsymbol{\beta}$. The proposed objective function in (4.2) can be rewritten as

$$R(\boldsymbol{\beta}) = n(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{dc})^T \hat{\boldsymbol{\Sigma}}_{dc}^{-1} (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{dc}) + n \sum_{j=1}^{d} \rho_j |\beta_j|, \tag{4.3}$$

then the regularized estimator $\hat{\boldsymbol{\eta}}_{\rho}^{dc}$ becomes $\hat{\boldsymbol{\beta}}_{\rho}^{dc}$ in the marginal model. Next we establish the asymptotic properties of $\hat{\boldsymbol{\beta}}_{\rho}^{dc}$ in Theorem 4.3.1. Without loss of generality, we assume that only the first d_0 predictors are informative, i.e.,

$$\boldsymbol{\beta}_0 = \begin{pmatrix} \boldsymbol{\beta}_{0a} \\ \boldsymbol{\beta}_{0b} \end{pmatrix}, \quad \boldsymbol{\beta}_{0a} = \begin{pmatrix} \beta_{01} \\ \vdots \\ \beta_{0d_0} \end{pmatrix} \neq \mathbf{0}, \quad \text{and} \quad \boldsymbol{\beta}_{0b} = \begin{pmatrix} \beta_{0d_0+1} \\ \vdots \\ \beta_{0d} \end{pmatrix} = \mathbf{0}. \tag{4.4}$$

Similarly, we write

$$\hat{\boldsymbol{\beta}}_{\rho}^{dc} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_{\rho a}^{dc} \\ \hat{\boldsymbol{\beta}}_{\rho b}^{dc} \end{pmatrix}. \tag{4.5}$$

Theorem 4.3.1. Under regularity conditions (M1) to (M6) in Section 2.1.1 and homogeneity assumptions (H1) to (H2) in Section 3.3, given $a_n = \max\{\rho_j, j \leq d_0\}$ and $b_n = \min\{\rho_j, j > d_0\}$, the regularized estimator $\hat{\boldsymbol{\beta}}_{\rho}^{dc}$ satisfies the following as $n \to \infty$:

(1) (Estimation Consistency) If $n^{1/2}a_n \xrightarrow{P} 0$, $\hat{\boldsymbol{\beta}}_{\rho}^{dc} \xrightarrow{P} \boldsymbol{\beta}_0$;

(2) (Selection Consistency) If $n^{1/2}a_n \xrightarrow{P} 0$ and $n^{1/2}b_n \xrightarrow{P} \infty$, $\Pr(\hat{\boldsymbol{\beta}}_{\rho b}^{dc} = \mathbf{0}) \to 1$;

(3) (Oracle Property) If $n^{1/2}a_n \xrightarrow{P} 0$ and $n^{1/2}b_n \xrightarrow{P} \infty$, $n^{1/2}(\hat{\boldsymbol{\beta}}_{\rho a}^{dc} - \boldsymbol{\beta}_{0a}) \xrightarrow{D} \mathcal{N}(\mathbf{0}, ([\boldsymbol{\Sigma}^{-1}]_{d_0 d_0})^{-1})$, in which $[\boldsymbol{\Sigma}^{-1}]_{d_0 d_0}$ is the leading $d_0 \times d_0$ submatrix of $\boldsymbol{\Sigma}^{-1}$;

(4) $[\boldsymbol{\Sigma}^{-1}]_{d_0 d_0}$ can be consistently estimated by $[\hat{\boldsymbol{\Sigma}}_{dc}^{-1}]_{d_0 d_0}$, where $\hat{\boldsymbol{\Sigma}}_{dc}$ is defined in Theorem 3.3.1.

Proof. Following Fan and Li (2001), projecting the ball $\left\{\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u} : ||\mathbf{u}||_2 \leq C\right\}$ to the objective function $R(\boldsymbol{\beta})$, a $n^{1/2}$-consistent local minimizer is implied by the fact that the projection is locally bounded with probability tending to one. The estimation consistency follows. The selection consistency is demonstrated by contradiction. The asymptotic normality and the oracle property is shown by decomposing the objective function into zero components and nonzero components. See Appendix A.8 for the detailed proof. ■

In Theorem 4.3.1, the estimation consistency ensures the consistency of the estimators of the nonzero coefficients; the selection consistency guarantees the zero coefficients must be estimated as zero, with probability tending to one. Together, they both imply that the proposed confidence distribution based regularized estimator can identify the true model consistently. The oracle estimator is defined as an estimator that knows in advance which coefficients are zero and which coefficients are not, and maximum likelihood is applied using only the nonzero covariates. The oracle property in Theorem 4.3.1 ensures that our

proposed regularized estimator performs as well as the oracle estimator (i.e., oralce maximum likelihood estimator), and that $\hat{\boldsymbol{\beta}}_{\rho a}^{dc}$ has the same asymptotic distribution as that of the oracle estimator.

Frailty Models

In the frailty model, $\boldsymbol{\eta} = (\theta, \boldsymbol{\beta}^T)^T$. Decompose the regularized estimator $\hat{\boldsymbol{\eta}}_{\rho}^{dc}$ obtained from minimizing (4.2), as

$$\hat{\boldsymbol{\eta}}_{\rho}^{dc} = \begin{pmatrix} \hat{\boldsymbol{\eta}}_{\rho a}^{dc} \\ \hat{\boldsymbol{\beta}}_{\rho b}^{dc} \end{pmatrix}, \quad \text{and} \quad \hat{\boldsymbol{\eta}}_{\rho a}^{dc} = \begin{pmatrix} \hat{\theta}^{dc} \\ \hat{\boldsymbol{\beta}}_{\rho a}^{dc} \end{pmatrix}, \tag{4.6}$$

and let the true values be

$$\boldsymbol{\eta}_0 = \begin{pmatrix} \boldsymbol{\eta}_{0a} \\ \boldsymbol{\beta}_{0b} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\eta}_{0a} \\ \mathbf{0} \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\eta}_{0a} = \begin{pmatrix} \theta_0 \\ \boldsymbol{\beta}_{0a} \end{pmatrix} \neq \mathbf{0}. \tag{4.7}$$

Theorem 4.3.2. Under regularity conditions (F1) to (F10) in Section 2.1.2 and homogeneity assumptions (H1) to (H2) in Section 3.3, given $a_n = \max\{\rho_j, j \leq d_0\}$ and $b_n = \min\{\rho_j, j > d_0\}$, the regularized estimator $\hat{\boldsymbol{\eta}}_{\rho}^{dc}$ satisfies the following as $n \to \infty$:

(1) (Estimation Consistency) If $n^{1/2} a_n \xrightarrow{P} 0$, $\hat{\boldsymbol{\eta}}_{\rho}^{dc} \xrightarrow{P} \boldsymbol{\eta}_0$;

(2) (Selection Consistency) If $n^{1/2} a_n \xrightarrow{P} 0$ and $n^{1/2} b_n \xrightarrow{P} \infty$, $\Pr(\hat{\boldsymbol{\beta}}_{\rho b}^{dc} = \mathbf{0}) \to 1$;

(3) (Oracle Property) If $n^{1/2} a_n \xrightarrow{P} 0$ and $n^{1/2} b_n \xrightarrow{P} \infty$, $n^{1/2}(\hat{\boldsymbol{\eta}}_{\rho a}^{dc} - \boldsymbol{\eta}_{0a}) \xrightarrow{D} \mathcal{N}(\mathbf{0}, ([\boldsymbol{\Sigma}^{-1}]_{(1+d_0)(1+d_0)})^{-1})$, in which $[\boldsymbol{\Sigma}^{-1}]_{(1+d_0)(1+d_0)}$ is the leading $(1+d_0) \times (1+d_0)$ submatrix of $\boldsymbol{\Sigma}^{-1}$;

(4) $[\boldsymbol{\Sigma}^{-1}]_{(1+d_0)(1+d_0)}$ can be consistently estimated by $[\hat{\boldsymbol{\Sigma}}_{dc}^{-1}]_{(1+d_0)(1+d_0)}$, where $\hat{\boldsymbol{\Sigma}}_{dc}$ is defined in Theorem 3.3.1.

Proof. The proof is similar to that of Theorem 4.3.1. See Appendix A.9 for the detailed proof. ■

Theorem 4.3.2 ensures that the true model can be selected consistently and that $\hat{\boldsymbol{\eta}}_{\rho a}^{dc}$ has the same asymptotic distribution as that of the oracle estimator.

4.4 Asymptotic Equivalence

The regularized estimation we discussed above in Section 4.3 is achieved by using the asymptotic distribution of the divide-and-combine estimator $\hat{\boldsymbol{\eta}}^{dc}$. However, the regularized estimation can certainly be applied to the confidence distribution of $\boldsymbol{\eta}$ based on the asymptotic distribution of the full-data estimator $\hat{\boldsymbol{\eta}}^{full}$. Define the regularized full-data estimator $\hat{\boldsymbol{\eta}}_{\rho}^{full}$ as

$$\hat{\boldsymbol{\eta}}_{\rho}^{full} = \underset{\boldsymbol{\eta}}{\operatorname{argmin}} \left\{ n(\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}^{full})^T \hat{\boldsymbol{\Sigma}}_{full}^{-1} (\boldsymbol{\eta} - \hat{\boldsymbol{\eta}}^{full}) + n \sum_{j=1}^{d} \rho_j^{full} |\beta_j| \right\}, \tag{4.8}$$

where $\hat{\boldsymbol{\Sigma}}_{full}$ is the estimator for variance $\boldsymbol{\Sigma}$ in the full data analysis. Since $\hat{\boldsymbol{\eta}}^{full}$ is asymptotically equivalent to $\hat{\boldsymbol{\eta}}^{dc}$ (Theorem 3.3.1), we can establish estimation consistency, selection consistency, and oracle properties for $\hat{\boldsymbol{\eta}}_{\rho}^{full}$ as in Theorems 4.3.1 and 4.3.2. Then we establish the following Theorem 4.4.1.

Theorem 4.4.1. Under conditions in Theorems 4.3.1 and 4.3.2, the regularized full-data estimator $\hat{\boldsymbol{\eta}}_{\rho}^{full}$ is asymptotically equivalent to the regularized divide-and-combine estimator $\hat{\boldsymbol{\eta}}_{\rho}^{dc}$ in the sense that they have the same properties of estimation consistency, selection consistency, and oracle properties.

Proof. The proof for Theorem 4.4.1 is straightforward and thus is omitted here. ■

We now consider a special case in the marginal proportional hazards model where only one failure could be potentially observed, i.e., $\hat{\boldsymbol{\beta}}_{\rho}^{dc}$ is obtained by using divide-and-combine statistics from the univariate Cox proportional hazards model (Cox, 1972). Define $\hat{\boldsymbol{\beta}}_{z}$ as the adaptive LASSO estimator from the full-data analysis under the Cox proportional hazards

model by Zhang and Lu (2007), i.e.,

$$\hat{\boldsymbol{\beta}}_z = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, R_z(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ -log\mathcal{PL}(\boldsymbol{\beta}) + n \sum_{j=1}^{d} \rho_j^z |\beta_j| \right\}, \tag{4.9}$$

where ρ_j^z denotes the tuning parameter for $j = 1, 2, \ldots, d$, and $log\mathcal{PL}(\boldsymbol{\beta})$ is the log partial likelihood in the Cox proportional hazards model (Cox, 1972). In the following Theorem 4.4.2, we show that the difference between two objective functions $R_z(\boldsymbol{\beta})$ and $R(\boldsymbol{\beta})/2$ (see definition of $R(\boldsymbol{\beta})$ in (4.3)) is asymptotically ignorable. Therefore, the minimizers $\hat{\boldsymbol{\beta}}_z = \operatorname{argmin}_{\boldsymbol{\beta}} R_z(\boldsymbol{\beta})$ and $\hat{\boldsymbol{\beta}}_\rho^{dc} = \operatorname{argmin}_{\boldsymbol{\beta}} R(\boldsymbol{\beta})$ are asymptotically equivalent.

Theorem 4.4.2. Under the univariate Cox proportional hazards model, and the regularity conditions (M1) to (M6) in Section 2.1.1 and homogeneity assumptions (H1) to (H2) in Section 3.3, $R_z(\boldsymbol{\beta}) = R(\boldsymbol{\beta})/2 + o_p(1)$, provided that $\rho_j^z = \rho_j/2$, for $j = 1, 2, \ldots, d$.

Proof. The asymptotic equivalence can be demonstrated by the Taylor series expansion and by acknowledging that $\hat{\boldsymbol{\Sigma}}_{full}^{-1} = n^{-1}\mathcal{I}(\hat{\boldsymbol{\beta}}^{full})$ in the univariate case. See Appendix A.10 for the detailed proof. ■

As indicated in the proof of Theorem 4.4.2, $\hat{\boldsymbol{\Sigma}}_{full}^{-1}$ generally does not equal $n^{-1}\mathcal{I}(\hat{\boldsymbol{\beta}}^{full})$ when the intra-cluster association in multivariate survival analysis is non-trivial. Therefore, the asymptotic equivalence between the partial likelihood function and the confidence density function established in Theorem 4.4.2 may not be directly extended to multivariate failure time data. It follows immediately that the regularized estimators obtained by minimizing both functions, generally, are not asymptotically equivalent. Nevertheless, under a more general framework, for example, the generalized methods of moments (GMM) (Hansen, 1982), it is possible to obtain a similar asymptotic equivalence between the objective function using GMM and the confidence density function for multivariate survival analysis, which is a potential direction for future research.

4.5 Optimization in Regularized Estimation

The objective function in (4.2) is strictly convex and therefore can be solved by many standard optimizers, such as R packages, `glmnet` (Friedman et al., 2010), `CVXR` (Fu et al., 2017), and `lars` (Efron et al., 2004). However, it is not straightforward to directly apply these optimizers in our case because of the annoying portion $\hat{\mathbf{\Sigma}}_{dc}^{-1}$ inserted within the quadratic term in the objective function. To facilitate the optimization, similar to Zhang and Lu (2007), we rewrite the objective function in (4.2) as

$$R(\boldsymbol{\beta}) = (\Gamma\boldsymbol{\beta} - \Psi)^T (\Gamma\boldsymbol{\beta} - \Psi) + n \sum_{j=1}^{d} \rho_j |\beta_j|, \tag{4.10}$$

where $(n^{-1}\hat{\mathbf{\Sigma}}_{dc})^{-1} = \Gamma^T\Gamma$ and $\Psi = \Gamma\hat{\boldsymbol{\beta}}^{dc}$, in which Γ can be easily obtained using the singular value decomposition. Now the objective function in (4.10) becomes a typical problem in the convex optimization, and those optimizers can be used without any difficulties.

The tuning parameter $\boldsymbol{\rho}$ for the adaptive LASSO (Zou, 2006) penalty function can be chosen by an exhaustive search in a d-dimensional Euclidean space, which is, however, computationally challenging and practically infeasible. A simple solution suggested by Zou (2006) is to replace each tuning parameter ρ_j by

$$\rho_j = \rho_0 |\hat{\beta}_j^{dc}|^{-\phi}, \text{ for } j = 1, 2, \ldots, d, \tag{4.11}$$

where ϕ is some prespecified positive number, for example $\phi = 1$ for simplicity. It can be verified that the tuning parameter ρ_j satisfies all the technical requirements for adaptive LASSO (Zou, 2006) as long as $n^{1/2}\rho_0 \xrightarrow{P} 0$ and $n^{(1+\phi)/2}\rho_0 \xrightarrow{P} \infty$.

Tuning parameters usually can be chosen using some model selection criteria, such as cross validation (CV), Akaike information criterion (AIC), or Bayesian information criterion (BIC). Wang and Leng (2007) recommended BIC instead of the commonly used CV since the latter approach tends to generate overfitted models when a finite dimensional

model truly exists. Specifically, we consider minimizing the following to obtain the optimal $\boldsymbol{\rho}$

$$BIC_\rho = n(\hat{\boldsymbol{\beta}}_\rho^{dc} - \hat{\boldsymbol{\beta}}^{dc})^T \hat{\boldsymbol{\Sigma}}_{dc}^{-1} (\hat{\boldsymbol{\beta}}_\rho^{dc} - \hat{\boldsymbol{\beta}}^{dc}) + (\log n) df_\rho, \tag{4.12}$$

where df_ρ is the number of nonzero coefficients in $\hat{\boldsymbol{\beta}}_\rho^{dc}$, a simple estimate for the degrees of freedom (Zou et al., 2007).

CHAPTER 5
PROPOSED DIVIDE-AND-COMBINE IN MULTISTATE SURVIVAL ANALYSIS

5.1 Introduction

Heart failure and atrial fibrillation have emerged as new cardiovascular epidemics over the past decades (Braunwald, 1997). Although the association between heart failure and atrial fibrillation has been appreciated since a century ago (Mackenzie, 1914), the causal relationship between these two conditions has not yet been fully determined (Anter et al., 2009). The prognostic significance of atrial fibrillation in patients with heart failure remains controversial because some argues that atrial fibrillation is a marker rather than an independent risk factor of some adverse cardiovascular outcomes (see Anter et al. (2009) for a detailed review of heart failure and atrial fibrillation). To characterize the disease courses of cardiovascular disease patients while taking into account the complicated relationship between heart failure and atrial fibrillation, in this chapter, we propose a five-state Markov stochastic model to study the dynamic process of cardiovascular diseases (CVDs).

In cardiovascular disease studies, the progress of a disease can be defined through several dynamic states, such as multiple hospitalizations due to different non-fatal cardiovascular diseases, and death or loss of follow-up. The proposed five-state model, as illustrated in Figure 5.1, contains four hospitalization states due to non-fatal cardiovascular diseases and one death state. Four hospitalization states include being hospitalized once due to heart failure (HF-1), being hospitalized twice and both due to heart failure (HF-2), being hospitalized twice and one due to heart failure the other due to atrial fibrillation (AF), and being hospitalized three times and two due to heart failure the other one due to atrial fibrillation (HF-2+AF). As described in Section 2.2.1, a patient can potentially visit any of the four hospitalization states depending on their starting state. During the entire follow-up period,

a patient may die or may be censored at any time because of the loss of follow-up.

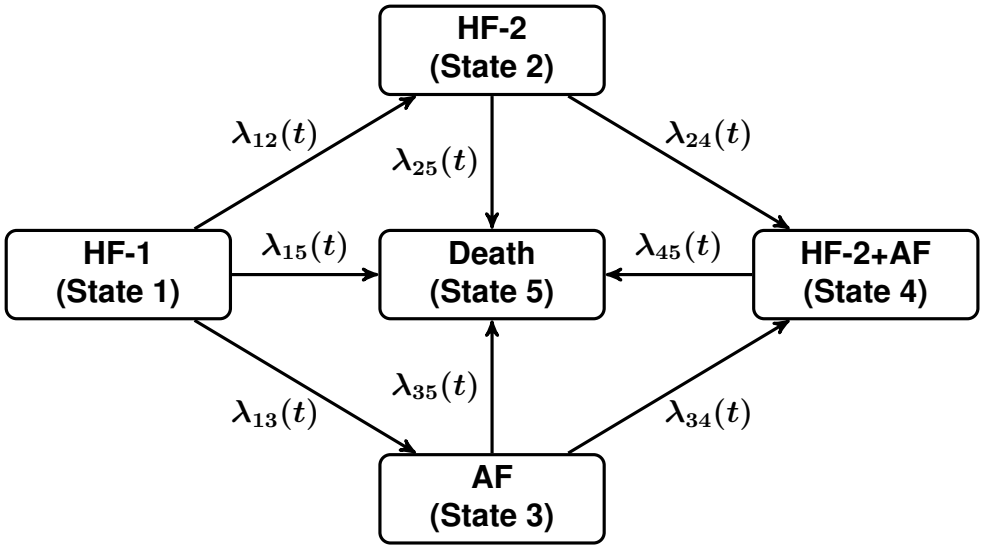

Figure 5.1: The five-state model for cardiovascular disease patients.

In the rest of this chapter, Section 5.2 presents the Andersen-type Cox Markov model in multistate survival analysis. The proposed divide-and-combine estimators for cumulative hazards are presented in Section 5.3. Section 5.4 shows the asymptotic equivalence of predicted transition probabilities using either divide-and-combine statistics or full-data statistics.

5.2 Model Setups

Recall that the random process $X(t)$ records the state being occupied at time t by a certain subject, and $X(t)$ takes values from the state space $\mathcal{Q} = \{1, 2, \ldots, 5\}$. Then the hazard for transition $h \to j$ is defined as

$$\lambda_{hj}(t) = \lim_{\Delta t \to 0} \frac{Pr(X(t + \Delta t) = j | X(t) = h)}{\Delta t} \tag{5.1}$$

for $h, j = 1, 2, \ldots, 5$ and $h \neq j$. Correspondingly, the matrix of transition hazards of the proposed five-state model is given by

$$\boldsymbol{\lambda}(t) = \begin{pmatrix} \lambda_{11}(t) & \lambda_{12}(t) & \lambda_{13}(t) & 0 & \lambda_{15}(t) \\ 0 & \lambda_{22}(t) & 0 & \lambda_{24}(t) & \lambda_{25}(t) \\ 0 & 0 & \lambda_{33}(t) & \lambda_{34}(t) & \lambda_{35}(t) \\ 0 & 0 & 0 & \lambda_{44}(t) & \lambda_{45}(t) \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{5.2}$$

where $\lambda_{hh}(t) = -\sum_{j \neq h} \lambda_{hj}(t)$ for $h, j = 1, 2, \ldots, 5$. Note that $h \to j$, hereinafter, refers to the direct transition from state h to state j, unless otherwise specified.

To adjust for different risk profiles in different subjects, we relate the transition hazards to covariates through a Cox proportional hazards regression model (Cox, 1972). The hazard function for transition $h \to j$ is

$$\lambda_{hj}(t|\boldsymbol{Z}_{hj}) = \lambda_{0,hj}(t) e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{hj}}, \tag{5.3}$$

in which $\lambda_{0,hj}(t)$, $h, j = 1, 2, \ldots, 5$ and $h \neq j$, is the unspecified baseline hazard function for transition $h \to j$, $\boldsymbol{\beta}$ is a d-dimensional regression parameter, and $\boldsymbol{Z}_{hj}$ is a time-independent covariate vector (see Section 2.2.3 for a discussion of possible time-dependent covariates). In fact, by using proper transition-specific covariates, model (5.3) can accommodate transition-specific regression coefficients, as elaborated in the simulation studies and the real data example (Chapter 6).

By using the estimation approach described in Section 2.2.3, the cumulative hazard function in transition $h \to j$ ($h, j = 1, 2, \ldots, 5$ and $h \neq j$) for a future subject with a covariate $\boldsymbol{z}_{0,hj}$, can be estimated by

$$\hat{\Lambda}_{hj}(t|\boldsymbol{z}_{0,hj}) = \sum_{i=1}^{n} \int_0^t \frac{e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{z}_{0,hj}} dN_{i,hj}(s)}{n S_{hj}^{(0)}(\hat{\boldsymbol{\beta}}, s)}, \tag{5.4}$$

where $\hat{\boldsymbol{\beta}}$ is the estimated regression coefficients, $S_{hj}^{(0)}(\boldsymbol{\beta}, s) = n^{-1}\sum_{i=1}^{n} Y_{i,h}(s)e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{i,hj}}$, and $N_{i,hj}(s)$ and $Y_{i,h}(s)$ are the counting process and the at risk process for subject i in a sample of size n, respectively. See detailed definitions in Section 2.2.3. The corresponding matrix for estimated cumulative hazards is given by

$$\hat{\boldsymbol{\Lambda}}(t|\boldsymbol{z}_0) = \begin{pmatrix} \hat{\Lambda}_{11}(t|\boldsymbol{z}_0) & \hat{\Lambda}_{12}(t|\boldsymbol{z}_0) & \hat{\Lambda}_{13}(t|\boldsymbol{z}_0) & 0 & \hat{\Lambda}_{15}(t|\boldsymbol{z}_0) \\ 0 & \hat{\Lambda}_{22}(t|\boldsymbol{z}_0) & 0 & \hat{\Lambda}_{24}(t|\boldsymbol{z}_0) & \hat{\Lambda}_{25}(t|\boldsymbol{z}_0) \\ 0 & 0 & \hat{\Lambda}_{33}(t|\boldsymbol{z}_0) & \hat{\Lambda}_{34}(t|\boldsymbol{z}_0) & \hat{\Lambda}_{35}(t|\boldsymbol{z}_0) \\ 0 & 0 & 0 & \hat{\Lambda}_{44}(t|\boldsymbol{z}_0) & \hat{\Lambda}_{45}(t|\boldsymbol{z}_0) \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{5.5}$$

in which $\hat{\Lambda}_{hh}(t|\boldsymbol{z}_0) = -\sum_{j\neq h}\hat{\Lambda}_{hj}(t|\boldsymbol{z}_0)$ for $h, j = 1, 2, \ldots, 5$. Note that $\boldsymbol{z}_0$ is the basic covariate that can be easily extended into the transition-specific structure $\boldsymbol{z}_{0,hj}$ for $h, j = 1, 2, \ldots, 5$ and $h \neq j$.

5.3 Divide-and-Combine Estimators for Cumulative Hazards

Instead of estimating regression coefficients in multivariate survival analysis, we are more interested in estimating cumulative hazards and predicting transition probabilities in multistate survival analysis. Therefore, we propose a divide-and-combine estimator for the cumulative hazard in multistate survival analysis by using the Andersen-type Cox Markov model (see more details in Section 2.2). The proposed divide-and-combine estimator in multistate survival analysis is motivated by and similar to the $\hat{\boldsymbol{\eta}}^{dc}$ in multivariate survival analysis. We introduce our divide-and-combine approach for estimating cumulative hazards as follows.

Divide Consider n independent subjects and use the Markov random process $X_i(t)$ to record the state being occupied by subject i at time t. $X_i(t)$ takes values from a state space $\mathcal{Q} = \{1, 2, \ldots, 5\}$. Denote the observed full data by $\mathbf{D}_i(t) = \{X_i(t), \boldsymbol{Z}_i; t \in [0, \tau]\}$ for $i = 1, 2, \ldots, n$, where τ is the end of the follow-up period and $\boldsymbol{Z}_i$ is a d-dimensional vector

of basic covariates that can be extended into transition-specific structure $\boldsymbol{Z}_{i,hj}$ for transition $h \to j$, $h, j = 1, 2, \ldots, 5$ and $h \neq j$. We divide the full data by independent subjects into S subsets, $\mathbf{D}_{si}(t) = \{X_{si}(t), \boldsymbol{Z}_{si}; t \in [0, \tau]\}$ for $s = 1, 2, \ldots, S$ and $i = 1, 2, \ldots, n_s$, where n_s is the number of independent subjects of the s^{th} subset with $n_s \gg d$. Note that $n = \sum_{s=1}^{S} n_s$. Because different subjects may stop at different "final states" (i.e., the states being occupied at the end of the follow-up period), the simple random splitting method may result in biased distribution of disease processes in divided subsets. Therefore, we implement a stratified random splitting method, stratified on the "final states", to make the "final states" evenly distributed across S subsets. We demonstrate this splitting method in the simulation studies and the real data example (Chapter 6). As in multivariate survival analysis, throughout this dissertation, we assume that all n_s's diverge in the same order of $O(n/S)$ and $S = o(n^{1/2})$, following Zhang et al. (2013), Chen and Xie (2014), and Rosenblatt and Nadler (2016).

Similar to the heuristic justification in multivariate survival analysis, in multistate survival analysis, the divided subsets generated by stratified random splitting would be homogeneous and each of them would be a representative random sample of the full data. Thus we can reasonably assume the same Andersen-type Cox Markov model in subsets as in the full data. In each subset s, we can estimate the transition-specific baseline cumulative hazards and thus obtain the cumulative hazard estimators of individual transitions for a subject with a covariate $\boldsymbol{z}_{0,hj}$ using (5.4) (see Section 2.2.3 for more details), i.e., $\hat{\Lambda}_{s,hj}(t|\boldsymbol{z}_{0,hj})$ for $h, j = 1, 2, \ldots, 5$ and $h \neq j$.

Combine We combine the S estimators obtained from subsets and obtain the divide-and-combine estimator by the following procedure:

$$\hat{\Lambda}_{hj}^{dc}(t|\boldsymbol{z}_{0,hj}) = n^{-1} \sum_{s=1}^{S} n_s \hat{\Lambda}_{s,hj}(t|\boldsymbol{z}_{0,hj}), \tag{5.6}$$

for $h, j = 1, 2, \ldots, 5$ and $h \neq j$.

Before studying the asymptotic properties of $\hat{\Lambda}_{hj}^{dc}(t|\boldsymbol{z}_{0,hj})$, we state formally the following assumptions, which is the key to establishing the following Theorems 5.3.1 and 5.3.2.

Homogeneity Assumptions Let $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_{s0}$ $(s = 1, 2, \ldots, S)$ be the true values of $\boldsymbol{\beta}$ in the full data and individual subsets. Accordingly, $\mathrm{s}_{s,hj}^{(r)}(\boldsymbol{\beta}, t)$ and $\Lambda_{s,hj}(t|\boldsymbol{z}_{0,hj})$ denote the counterparts of $\mathrm{s}_{hj}^{(r)}(\boldsymbol{\beta}, t)$ and $\Lambda_{hj}(t|\boldsymbol{z}_{0,hj})$ in the s^{th} subset for $s = 1, 2, \ldots, S$. See definitions of $\mathrm{s}_{hj}^{(r)}(\boldsymbol{\beta}, t)$ and $\Lambda_{hj}(t|\boldsymbol{z}_{0,hj})$ in Section 2.2.3. The homogeneity is two-fold:

(C1) Underlying parameters are the same across subsets, i.e., $\boldsymbol{\beta}_{s0} = \boldsymbol{\beta}_0$, and $\Lambda_{s,hj}(t|\boldsymbol{z}_{0,hj}) = \Lambda_{hj}(t|\boldsymbol{z}_{0,hj})$;

(C2) Limiting processes $\mathrm{s}_{s,hj}^{(r)}(\boldsymbol{\beta}, t)$ are identical across all subsets, i.e., $\mathrm{s}_{s,hj}^{(r)}(\boldsymbol{\beta}, t) = \mathrm{s}_{hj}^{(r)}(\boldsymbol{\beta}, t)$.

Assume that regularity conditions in Section 2.2.3 hold in each of the divided subsets, as shown in Theorems 2.2.1 and 2.2.2, the estimators $\hat{\Lambda}_{s,hj}(t|\boldsymbol{z}_{0,hj})$ $(s = 1, 2. \ldots, S)$ are uniformly consistent and asymptotically normal. Under homogeneity assumptions (C1) to (C3), it follows that $\hat{\Lambda}_{hj}^{full}$ is also uniformly consistent and $n^{1/2}\left\{\hat{\Lambda}_{hj}^{full}(t|\boldsymbol{z}_{0,hj}) - \Lambda_{hj}(t|\boldsymbol{z}_{0,hj})\right\}$ is asymptotically normally distributed with mean zero and a variance of $b^2(t|\boldsymbol{z}_{0,hj}) + \boldsymbol{a}^T(t|\boldsymbol{z}_{0,hj})\boldsymbol{\Sigma}_3\boldsymbol{a}(t|\boldsymbol{z}_{0,hj})$, where $\hat{\Lambda}_{hj}^{full}(t|\boldsymbol{z}_{0,hj})$ is the full-data estimator using the entire data set, $b^2(t|\boldsymbol{z}_{0,hj})$, $\boldsymbol{a}(t|\boldsymbol{z}_{0,hj})$, and $\boldsymbol{\Sigma}_3$ are defined in the form as in Theorem 2.2.2. We establish the asymptotic properties of $\hat{\Lambda}_{hj}^{dc}(t|\boldsymbol{z}_{0,hj})$ in Theorems 5.3.1 and 5.3.2, which imply that the proposed divide-and-combine estimator $\hat{\Lambda}_{hj}^{dc}(t|\boldsymbol{z}_{0,hj})$ is asymptotically equivalent to the full-data estimator $\hat{\Lambda}_{hj}^{full}(t|\boldsymbol{z}_{0,hj})$, in the sense that they converge weakly to the same Gaussian process.

Theorem 5.3.1. Under regularity conditions in Section 2.2.3 and homogeneity assumptions (C1) to (C3) in Section 5.3, $\hat{\Lambda}_{hj}^{dc}(t|\boldsymbol{z}_{0,hj})$ converges in probability to $\Lambda_{hj}(t|\boldsymbol{z}_{0,hj})$ uniformly in $t \in [0, \tau]$, i.e., as $n \to \infty$,

$$\sup_{t\in[0,\tau]} \left|\hat{\Lambda}_{hj}^{dc}(t|\boldsymbol{z}_{0,hj}) - \Lambda_{hj}(t|\boldsymbol{z}_{0,hj})\right| \xrightarrow{P} 0. \tag{5.7}$$

Theorem 5.3.2. Under regularity conditions in Section 2.2.3 and homogeneity assumptions (C1) to (C3) in Section 5.3, the random process $n^{1/2}\left\{\hat{\Lambda}_{hj}^{dc}(t|\boldsymbol{z}_{0,hj}) - \Lambda_{hj}(t|\boldsymbol{z}_{0,hj})\right\} \xrightarrow{D} \mathcal{GP}(0, \Omega_{hj}(t))$, where $\mathcal{GP}$ denotes Gaussian process, and the variance function is given by

$$\Omega_{hj}(t) = b^2(t|\boldsymbol{z}_{0,hj}) + \boldsymbol{a}^T(t|\boldsymbol{z}_{0,hj})\boldsymbol{\Sigma}_3\boldsymbol{a}(t|\boldsymbol{z}_{0,hj}), \tag{5.8}$$

in which $b^2(t|\boldsymbol{z}_{0,hj})$, $\boldsymbol{a}(t|\boldsymbol{z}_{0,hj})$, and $\boldsymbol{\Sigma}_3$ are defined in the same form as in Theorem 2.2.2.

Corollary 5.3.2.1. Under conditions in Theorem 5.3.2, the variance function of the random process $n^{1/2}\left\{\hat{\Lambda}_{hj}^{dc}(t|\boldsymbol{z}_{0,hj}) - \Lambda_{hj}(t|\boldsymbol{z}_{0,hj})\right\}$ can be uniformly consistently estimated by

$$\hat{\Omega}_{hj}^{dc}(t) = n^{-1}\sum_{s=1}^{S} n_s\left\{\hat{b}_s^2(t|\boldsymbol{z}_{0,hj}) + \hat{\boldsymbol{a}}_s^T(t|\boldsymbol{z}_{0,hj})\hat{\boldsymbol{\Sigma}}_{s3}\hat{\boldsymbol{a}}_s(t|\boldsymbol{z}_{0,hj})\right\}, \tag{5.9}$$

where

$$\begin{aligned}
\hat{b}_s^2(t|\boldsymbol{z}_{0,hj}) &= \sum_{i=1}^{n_s}\int_0^t \frac{e^{2\cdot\hat{\boldsymbol{\beta}}_s^T\boldsymbol{z}_{0,hj}}dN_{si,hj}(u)}{n_s\left\{S_{s,hj}^{(0)}(\hat{\boldsymbol{\beta}}_s,u)\right\}^2},\\
\hat{\boldsymbol{a}}_s(t|\boldsymbol{z}_{0,hj}) &= \sum_{i=1}^{n_s}\int_0^t \frac{\left\{\boldsymbol{z}_{0,hj} - \mathbf{E}_{s,hj}(\hat{\boldsymbol{\beta}}_s,u)\right\}e^{\hat{\boldsymbol{\beta}}_s^T\boldsymbol{z}_{0,hj}}dN_{si,hj}(u)}{n_s S_{s,hj}^{(0)}(\hat{\boldsymbol{\beta}}_s,u)},\\
\hat{\boldsymbol{\Sigma}}_{s3}^{-1} &= n_s^{-1}\sum_{h\neq j}\sum_{i=1}^{n_s}\int_0^\tau \mathbf{V}_{s,hj}(\hat{\boldsymbol{\beta}}_s,t)dN_{si,hj}(t),
\end{aligned} \tag{5.10}$$

in which $\hat{\boldsymbol{\beta}}_s$ is the estimated regression coefficients in the s^{th} subset, and $N_{si,hj}(t)$, $S_{s,hj}^{(0)}(\boldsymbol{\beta},t)$, $\mathbf{E}_{s,hj}(\boldsymbol{\beta},t)$, and $\mathbf{V}_{s,hj}(\boldsymbol{\beta},t)$ are the counterparts in the s^{th} subset of $N_{i,hj}(t)$, $S_{hj}^{(0)}(\boldsymbol{\beta},t)$, $\mathbf{E}_{hj}(\boldsymbol{\beta},t)$, and $\mathbf{V}_{hj}(\boldsymbol{\beta},t)$ (see definitions in Section 2.2.3).

Proof. Along the lines of the proof for $\hat{\boldsymbol{\eta}}^{dc}$ in multivariate survival analysis, the asymptotic properties of $\hat{\Lambda}_{hj}^{dc}(t|\boldsymbol{z}_{0,hj})$ in Theorems 5.3.1 and 5.3.2, and Corollary 5.3.2.1 can be easily established. See the detailed proof in Appendix A.11. ■

Theorems 5.3.1 and 5.3.2 imply the asymptotic equivalence between $\hat{\Lambda}_{hj}^{dc}(t|\boldsymbol{z}_{0,hj})$ and

$\hat{\Lambda}_{hj}^{full}(t|\boldsymbol{z}_{0,hj})$. It follows immediately that the corresponding matrices $\hat{\boldsymbol{\Lambda}}^{dc}(t|\boldsymbol{z}_0)$ and $\hat{\boldsymbol{\Lambda}}^{full}(t|\boldsymbol{z}_0)$ are also asymptotically equivalent because $\hat{\boldsymbol{\Lambda}}(t,|\boldsymbol{z}_0)$ is an element-wise aggregation of $\hat{\Lambda}_{hj}(t|\boldsymbol{z}_0)$, $h,j = 1,2,\ldots,5$. $\hat{\boldsymbol{\Lambda}}^{dc}(t|\boldsymbol{z}_0)$ and $\hat{\boldsymbol{\Lambda}}^{full}(t|\boldsymbol{z}_0)$ are defined in the same structure as (5.5) with their elements are replaced by $\hat{\Lambda}_{hj}^{dc}(t|\boldsymbol{z}_{0,hj})$ and $\hat{\Lambda}_{hj}^{full}(t|\boldsymbol{z}_{0,hj})$ for $h,j = 1,2,\ldots,5$ and $h \neq j$, respectively.

5.4 Prediction of Transition Probabilities Using Estimated Cumulative Hazards

In the proposed five-state model, the transition probability matrix for a subject with a basic covariate $\boldsymbol{z}_0$ in the time interval $(u,t]$[1] is given by

$$\mathbf{P}(u,t|\boldsymbol{z}_0) = \begin{pmatrix} P_{11}(u,t|\boldsymbol{z}_0) & P_{12}(u,t|\boldsymbol{z}_0) & P_{13}(u,t|\boldsymbol{z}_0) & 0 & P_{15}(u,t|\boldsymbol{z}_0) \\ 0 & P_{22}(u,t|\boldsymbol{z}_0) & 0 & P_{24}(u,t|\boldsymbol{z}_0) & P_{25}(u,t|\boldsymbol{z}_0) \\ 0 & 0 & P_{33}(u,t|\boldsymbol{z}_0) & P_{34}(u,t|\boldsymbol{z}_0) & P_{35}(u,t|\boldsymbol{z}_0) \\ 0 & 0 & 0 & P_{44}(u,t|\boldsymbol{z}_0) & P_{45}(u,t|\boldsymbol{z}_0) \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}, \quad (5.11)$$

where $P_{hj}(u,t|\boldsymbol{z}_0)$ $(h,j = 1,2,\ldots,5)$ denotes the transition probability from state h to state j, including both direct and indirect transitions. Of note, $\boldsymbol{z}_0$ is a basic covariate that can be extended into the transition-specific structure $\boldsymbol{z}_{0,hj}$ ($h,j = 1,2,\ldots,5$ and $h \neq j$).

The transition probability matrix $\mathbf{P}(u,t|\boldsymbol{z}_0)$ can be estimated by the Aalen-Johansen estimator (Andersen et al., 1991)

$$\hat{\mathbf{P}}(u,t|\boldsymbol{z}_0) = \prod_{v\in(u,t]} \left\{\mathbf{I} + d\hat{\boldsymbol{\Lambda}}(v|\boldsymbol{z}_0)\right\}, \quad (5.12)$$

where $\prod$ is the sign of product integral (See Section 2.2.2 for more details), $\mathbf{I}$ is the identity matrix, and $\hat{\boldsymbol{\Lambda}}(v|\boldsymbol{z}_0)$ is the estimated cumulative hazard matrix. According to Corollaries 2.2.4.1 and 2.2.4.2, the variance estimator for the estimated transition probability matrix

[1]Interval $(u,t]$ instead of the previous $(s,t]$, is used here to avoid confusion with s, the notation of the divided subset.

is $\hat{\mathbf{var}}(\hat{\mathbf{P}}(u,t|\boldsymbol{z}_0))$, whose (hj,mr) element is given by $\hat{\text{var}}_1(\hat{P}_{hj}(u,t|\boldsymbol{z}_0),\hat{P}_{mr}(u,t|\boldsymbol{z}_0)) + \hat{\text{var}}_2(\hat{P}_{hj}(u,t|\boldsymbol{z}_0),\hat{P}_{mr}(u,t|\boldsymbol{z}_0))$, which is defined in the same form as in Corollary 2.2.4.2.

Using $\hat{\mathbf{\Lambda}}^{dc}(t|\boldsymbol{z}_0)$ obtained from the divide-and-combine analysis, we can predict the cumulative hazard matrix by $\hat{\mathbf{P}}^{dc}(u,t|\boldsymbol{z}_0)$ and obtain its variance estimator $\hat{\mathbf{var}}(\hat{\mathbf{P}}^{dc}(u,t|\boldsymbol{z}_0))$. Similarly, by using the full-data analysis, we obtain $\hat{\mathbf{P}}^{full}(u,t|\boldsymbol{z}_0)$ and $\hat{\mathbf{var}}(\hat{\mathbf{P}}^{full}(u,t|\boldsymbol{z}_0))$.

Theorem 5.4.1. Under conditions in Theorem 5.3.1, $\hat{\mathbf{P}}^{dc}(u,t|\boldsymbol{z}_0)$ is asymptotically equivalent to $\hat{\mathbf{P}}^{full}(u,t|\boldsymbol{z}_0)$, in the sense that they both are uniformly consistent and converge weakly to the same Gaussian process.

Proof. Because the necessary ingredients for the calculation of $\hat{\mathbf{P}}(u,t|\boldsymbol{z}_0)$ and $\hat{\mathbf{var}}(\hat{\mathbf{P}}(u,t|\boldsymbol{z}_0))$ are just the estimates of the cumulative transition hazards and of their variance matrices, irrespective of how these estimates were obtained. Since $\hat{\mathbf{\Lambda}}^{dc}(t|\boldsymbol{z}_0)$ is asymptotically equivalent to $\hat{\mathbf{\Lambda}}^{full}(t|\boldsymbol{z}_0)$ as implied by Theorems 5.3.1 and 5.3.2, following the same arguments in the proofs of Theorems 2.2.3 and 2.2.4, and Corollaries 2.2.4.1 and 2.2.4.2, we can prove Theorem 5.4.1. Similar to the divide-and-combine approach in multivariate survival analysis, in multistate survival analysis, both the total sample size (n) and the number of subsets (S) can also go to infinity, but S should be chosen such that its order is smaller than $n^{1/2}$, i.e., $S = o(n^{1/2})$. ■

Theorem 5.4.1 implies that when the sample size n is large enough, by using the divide-and-combine analysis, we can predict the transition probabilities as accurately as those by using the full-data analysis. However, the proposed divide-and-combine is only performed in the estimation of transition hazards but not the estimation of transition probabilities. As a result, the savings in computational cost may not be as significant as those in multivariate survival analysis. The performance of $\hat{\mathbf{P}}^{dc}(u,t|\boldsymbol{z}_0)$, including the asymptotic equivalence with $\hat{\mathbf{P}}^{full}(u,t|\boldsymbol{z}_0)$ and the reduction in computation time, is evaluated in the simulation studies (Chapter 6).

CHAPTER 6
SIMULATION STUDIES AND DATA ANALYSES

6.1 Introduction

In this chapter, we present simulation studies to demonstrate the performances of the proposed divide-and-combine estimators in multivariate survival analysis and multistate survival analysis, respectively, and use real data examples to illustrate both methodologies. The proposed regularized estimators in multivariate survival analysis are also assessed using simulation studies and real data analyses. In particular, the comparisons of the proposed divide-and-combine estimators with the full-data estimators in both multivariate survival analysis and multistate survival analysis are provided in Sections 6.2.1 and 6.2.2, and Section 6.2.3, respectively. The variable selection ability of the proposed regularized estimators is summarized in Sections 6.2.1 and 6.2.2, too.

6.2 Simulation Studies

6.2.1 Marginal Models in Multivariate Survival Analysis

A simulation study of $n = 100,000$ independent clusters was conducted to evaluate the asymptotic equivalence between the divide-and-combine estimator $\hat{\boldsymbol{\beta}}^{dc}$ and the full-data estimator $\hat{\boldsymbol{\beta}}^{full}$. The statistical properties of the regularized estimators $\hat{\boldsymbol{\beta}}_{\rho}^{dc}$ and $\hat{\boldsymbol{\beta}}_{\rho}^{full}$, including estimation consistency, selection consistency, and oracle properties were also assessed.

We simulated multivariate unordered failure time data from the Clayton-Oakes distribution (Clayton and Cuzick, 1985; Oakes, 1989) with a marginal Weibull distribution for $K = 3$ types of failures. We parameterized the marginal baseline hazard function by $\lambda_{0k}(t) = h_{0k}\lambda_0(t) = h_{ok}\xi h(ht)^{\xi-1}$, where $\lambda_0(t) = \xi h(ht)^{\xi-1}$ is denoted by $t \sim$

Weibull(shape = ξ, scale = h) and set $\xi = h = 2$. We chose $h_{01} = e^0$, $h_{02} = e^{0.01}$, and $h_{03} = e^{0.05}$ to differentiate three types of failures. Therefore, according to (2.6), the joint survival function of failure times T_{i1}, T_{i2}, and T_{i3} in the i^{th} cluster ($i = 1, 2, \ldots, n$) is given by

$$Pr(T_{i1} > t_{i1}, T_{i2} > t_{i2}, T_{i3} > t_{i3} | \boldsymbol{Z}_{i1}, \boldsymbol{Z}_{i2}, \boldsymbol{Z}_{i3}) = \left[\sum_{k=1}^{3} \exp\left\{ -(1-\nu) h_{ok} (h t_{ik})^{\xi} e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{ik}} \right\} - 2 \right]^{1/(1-\nu)}, \tag{6.1}$$

where we chose $\nu = 13/7$, corresponding to Kendall's tau $\kappa = 0.3$, for a moderate intra-cluster correlation. Specifically, we generated K-variate correlated unordered failure times for n independent clusters by the following steps.

(S1) Generate α_i from Gamma(shape = $\frac{1}{\nu - 1}$, rate = 1);

(S2) Generate K i.i.d. random variables U_{ik} from Uniform(0, 1) for the conditional survival functions $S_{ik}(t_{ik}|\alpha_i)$, $k = 1, 2, \ldots, K$;

(S3) Solve for $S_{ik}(t_{ik}|\alpha_i)$ in terms of $S_{ik}(t_{ik})$ and α_i. Here, $S_{ik}(t_{ik}) = \exp\left\{ -h_{0k} (h t_{ik})^{\xi} e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{ik}} \right\}$, after some derivation, one gets

$$T_{ik} = \left\{ \frac{\frac{1}{1-\nu} \cdot log(1 - \frac{logU_{ik}}{\alpha_i})}{-e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{ik}} \cdot h_{0k} h^{\xi}} \right\}^{1/\xi} \tag{6.2}$$

for the k^{th} type of failure in the i^{th} cluster, $k = 1, 2, \ldots, K$, $i = 1, 2, \ldots, n$. (See Chen (1998) for the detailed derivation.)

To allow for varying covariate effects on different failure types, in the i^{th} cluster, data were simulated according to type-specific regression parameters: $\boldsymbol{\beta}_1^* = (\mathbf{0.8}_4, \mathbf{0.6}_4, \mathbf{0}_{92})^T$, $\boldsymbol{\beta}_2^* = (\mathbf{0.8}_4, \mathbf{0.4}_4, \mathbf{0}_{92})^T$, and $\boldsymbol{\beta}_3^* = (\mathbf{0.8}_4, \mathbf{0.2}_4, \mathbf{0}_{92})^T$, corresponding to covariate $\boldsymbol{C}_{ik} = (C_{ik1}, \ldots, C_{ik100})^T$ ($k = 1, 2, 3$) generated from a multivariate normal distribution with standard normal marginals and an equal correlation of 0.2. Each of $\boldsymbol{\beta}_k^*$, $k = 1, 2, 3$, contains

8 nonzeros and 92 zeros. We considered the marginal proportional hazards model (3.3) with a design matrix $\boldsymbol{Z}_i = (\boldsymbol{Z}_{i1}, \boldsymbol{Z}_{i2}, \boldsymbol{Z}_{i3})^T$, where

$$\begin{pmatrix} \boldsymbol{Z}_{i1}^T \\ \boldsymbol{Z}_{i2}^T \\ \boldsymbol{Z}_{i3}^T \end{pmatrix} = \begin{pmatrix} C_{i11}, \cdots, C_{i14}, C_{i15}, C_{i16}, C_{i17}, C_{i18}, 0, 0, 0, 0, 0, 0, 0, 0, C_{i19}, \cdots, C_{i1100} \\ C_{i21}, \cdots, C_{i24}, 0, 0, 0, 0, C_{i25}, C_{i26}, C_{i27}, C_{i28}, 0, 0, 0, 0, C_{i29}, \cdots, C_{i2100} \\ C_{i31}, \cdots, C_{i34}, 0, 0, 0, 0, 0, 0, 0, 0, C_{i35}, C_{i36}, C_{i37}, C_{i38}, C_{i39}, \cdots, C_{i3100} \end{pmatrix},$$

and a common regression parameter vector $\boldsymbol{\beta} = (\mathbf{0.8}_4, \mathbf{0.6}_4, \mathbf{0.4}_4, \mathbf{0.2}_4, \mathbf{0}_{92})^T$. Note that $\boldsymbol{\beta}^T \boldsymbol{Z}_{i1} = \boldsymbol{\beta}_1^{*T} \boldsymbol{C}_{i1}$, $\boldsymbol{\beta}^T \boldsymbol{Z}_{i2} = \boldsymbol{\beta}_2^{*T} \boldsymbol{C}_{i2}$, and $\boldsymbol{\beta}^T \boldsymbol{Z}_{i3} = \boldsymbol{\beta}_3^{*T} \boldsymbol{C}_{i3}$. This illustrates that using a common $\boldsymbol{\beta}$ notation for the multivariate (K-variate) marginal proportional hazards models does not preclude type-specific regression parameters.

To approximate the event rate (13%) in the motivating example (i.e., MIDAS data) and further study the low event rate issue, fixed administrative censoring times of 0.031 and 0.056 were used to yield event rates of 5% and 10%, respectively. In the divide-and-combine analysis, three sets of partition ratios were used: $n_s/n = (\mathbf{2}/\mathbf{8}_3, \mathbf{1}/\mathbf{8}_2)$, $n_s/n = (\mathbf{2}/\mathbf{15}_5, \mathbf{1}/\mathbf{15}_5)$, and $n_s/n = (\mathbf{2}/\mathbf{30}_{10}, \mathbf{1}/\mathbf{30}_{10})$, corresponding to $S = 5$, 10, and 20 subsets, respectively. Note that our simulation studies were designed following the recommendation of Vittinghoff and McCulloch (2007), such that the number of events per covariate is at least 5 to 10 in each subset. In each configuration, we ran simulations 500 times. All simulations were carried out on a Linux cluster via parallel computing (S cores) with one subset allocated to one core. The average computation time was calculated based on 50 simulations performed on Intel® Xeon ® E5-2680 v4 @2.40GHz. All statistical analyses regarding fitting marginal proportional hazards models were performed using R package, `survival` (Therneau, 2020), and regularized estimation was conducted using R package, `glmnet` (Friedman et al., 2010).

Simulation Results

We first assessed the performance of the divide-and-combine estimator $\hat{\boldsymbol{\beta}}^{dc}$, compared to the full-data estimator $\hat{\boldsymbol{\beta}}^{full}$ in average computation time (Time), mean of biasedness (Bias), empirical standard error (ESE), mean asymptotic standard error (ASE) using the theoretical formula and the associated standard error, and empirical coverage probability (CovP) of the 95% Wald-type confidence interval (Table 6.1). As expected, the computation time of $\hat{\boldsymbol{\beta}}^{dc}$ is shorter than that for $\hat{\boldsymbol{\beta}}^{full}$. More subsets save more time. In terms of biasedness, ESE, ASE, and CovP, the performance of $\hat{\boldsymbol{\beta}}^{dc}$ using all three weights is generally close to that of $\hat{\boldsymbol{\beta}}^{full}$. The bias of $\hat{\boldsymbol{\beta}}^{dc}$ is generally small; ESE and ASE are close to each other; and CovP is close to the nominal 95% level. However, the estimator $\hat{\boldsymbol{\beta}}^{dc}$ using $\mathbf{W}_{3s}(\hat{\boldsymbol{\beta}}_s) = n_s$ appears to have large bias and a poor coverage probability when the magnitude of $\boldsymbol{\beta}$ is large (say, 0.6, 0.8) and the number of subsets (S) is big. We conjecture that $\mathbf{W}_{1s}(.)$ and $\mathbf{W}_{2s}(.)$ outperformed $\mathbf{W}_{3s}(.)$ might be because $\mathbf{W}_{1s}(.)$ and $\mathbf{W}_{2s}(.)$ utilize more empirical information from the data in individual subsets than $\mathbf{W}_{3s}(.)$ does as we discussed in Section 3.3. Under both event rates of 5% and 10%, the asymptotic equivalence between $\hat{\boldsymbol{\beta}}^{dc}$ and $\hat{\boldsymbol{\beta}}^{full}$ as discussed above can be observed.

We next investigated the performances of the regularized estimators $\hat{\boldsymbol{\beta}}_{\rho}^{dc}$ and $\hat{\boldsymbol{\beta}}_{\rho}^{full}$. In addition to computation time, biasedness, ESE, ASE, and CovP, we also evaluated the performance of variable selection (selection consistency) using sensitivity (Sn), the percentage of true-nonzero parameters being selected, and specificity (Sp), the percentage of true-zero parameters being selected (Table 6.2). When we calculated computation time, we only calculated the time of the regularization process using the confidence distribution approach, including the determination of tuning parameters. Because the dimension of the "data" is greatly reduced from the original data of $n = 100,000$ to the dimension of the multivariate normal distribution associated with $\hat{\boldsymbol{\beta}}^{dc}$ and $\hat{\boldsymbol{\beta}}^{full}$, the computation time of this step is small ($<$ 1 second) and about the same across all regularized estimators $\hat{\boldsymbol{\beta}}_{\rho}^{dc}$ and $\hat{\boldsymbol{\beta}}_{\rho}^{full}$. In terms of biasedness, ESE, ASE, and CovP, the performance of $\hat{\boldsymbol{\beta}}_{\rho}^{dc}$ using all three weights

is generally close to that of $\hat{\boldsymbol{\beta}}_{\rho}^{full}$. The bias of these estimators is generally small, ESE and ASE are close to each other, and CovP is close to the nominal 95% level, with the exception of $\hat{\boldsymbol{\beta}}_{\rho}^{dc}$ using $\mathbf{W}_{3s}(\hat{\boldsymbol{\beta}}_s) = n_s$. This is because the bias of the previous estimator $\hat{\boldsymbol{\beta}}^{dc}$ using $\mathbf{W}_{3s}(\hat{\boldsymbol{\beta}}_s) = n_s$ is large when the magnitude of $\boldsymbol{\beta}$ is big, and the number of subsets is large. For the performance of variable selection, the sensitivity and specificity are almost 100% for all estimators. The performance of $\hat{\boldsymbol{\beta}}_{\rho}^{dc}$ is similar to that of $\hat{\boldsymbol{\beta}}_{\rho}^{full}$, irrespective of event rates (5% or 10%).

We also conducted simulation studies under the same setting as above using a stratified random splitting approach to make the number of events evenly distributed across S subsets. Specifically, we stratified the full data into 4 strata by the number of events per cluster (0, 1, 2, and 3), then randomly split the data in each stratum into S sets according to the pre-specified partition ratios (n_s/n), and formed each subset by combining one set in each stratum. Results are summarized in Tables 6.3 and 6.4. The performance of our proposed method using the stratified random splitting is similar to that using the simple random splitting approach.

Table 6.1: Performances of $\hat{\boldsymbol{\beta}}^{dc}$ and $\hat{\boldsymbol{\beta}}^{full}$ for estimating $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ in marginal models with simple random splitting.

$\boldsymbol{\beta}_0$		5 Subsets			10 Subsets			20 Subsets			Full Data
		$\mathbf{W}_{1s}(.)$	$\hat{\boldsymbol{\beta}}^{dc}$ $\mathbf{W}_{2s}(.)$	$\mathbf{W}_{3s}(.)$	$\mathbf{W}_{1s}(.)$	$\hat{\boldsymbol{\beta}}^{dc}$ $\mathbf{W}_{2s}(.)$	$\mathbf{W}_{3s}(.)$	$\mathbf{W}_{1s}(.)$	$\hat{\boldsymbol{\beta}}^{dc}$ $\mathbf{W}_{2s}(.)$	$\mathbf{W}_{3s}(.)$	$\hat{\boldsymbol{\beta}}^{full}$
	Event Rate = 0.05										
	Time[1] (min)	1.3	1.3	1.3	0.5	0.5	0.5	0.2	0.2	0.2	12.7
0.8	Bias (ESE) ($\times 10^{-3}$)	1.0 (9.5)	1.2 (9.6)	6.1 (9.6)	0.3 (9.6)	0.5 (9.9)	12.7 (10.1)	-2.9 (9.3)	-3.1 (10.0)	25.7 (10.2)	1.3 (9.5)
	ASE (SE) ($\times 10^{-3}$)	9.5 (0.1)	9.3 (0.1)	9.6 (0.1)	9.5 (0.1)	9.1 (0.1)	9.7 (0.1)	9.4 (0.1)	8.8 (0.1)	10.0 (0.1)	9.5 (0.1)
	CovP (%)	95.3	94.1	90.1	93.9	92.3	74.1	94.2	89.3	27.9	95.3
0.6	Bias (ESE) ($\times 10^{-3}$)	0.8 (13.8)	1.0 (14.1)	5.4 (14.1)	-0.1 (13.4)	0.0 (14.0)	10.6 (14.0)	-2.1 (13.7)	-2.4 (14.6)	22.3 (14.7)	1.1 (13.8)
	ASE (SE) ($\times 10^{-3}$)	13.6 (0.2)	13.4 (0.2)	13.7 (0.2)	13.5 (0.2)	13.1 (0.2)	13.9 (0.2)	13.4 (0.2)	12.5 (0.2)	14.3 (0.2)	13.7 (0.2)
	CovP (%)	95.2	94.2	93.0	95.4	93.6	88.1	94.0	90.5	65.3	95.1
0.4	Bias (ESE) ($\times 10^{-3}$)	0.2 (15.5)	0.3 (15.9)	2.9 (15.8)	-0.2 (16.2)	0.1 (16.8)	6.0 (16.8)	-1.8 (15.7)	-1.6 (16.8)	12.5 (17.0)	0.4 (15.5)
	ASE (SE) ($\times 10^{-3}$)	15.6 (0.2)	15.4 (0.2)	15.8 (0.2)	15.6 (0.2)	15.1 (0.2)	16.0 (0.2)	15.4 (0.2)	14.4 (0.2)	16.4 (0.2)	15.7 (0.2)
	CovP (%)	95.5	94.5	94.5	94.3	92.8	92.3	93.5	90.6	87.7	95.6
0.2	Bias (ESE) ($\times 10^{-3}$)	0.8 (18.5)	0.9 (18.6)	1.7 (18.7)	-0.3 (18.2)	-0.2 (18.8)	1.7 (18.7)	-0.8 (18.8)	-0.7 (19.9)	4.5 (20.0)	0.9 (18.5)
	ASE (SE) ($\times 10^{-3}$)	18.3 (0.3)	18.0 (0.3)	18.5 (0.3)	18.2 (0.3)	17.6 (0.3)	18.7 (0.3)	18.1 (0.3)	16.9 (0.3)	19.1 (0.3)	18.3 (0.3)
	CovP (%)	94.5	94.1	94.5	94.7	93.2	94.8	93.5	91.1	93.5	94.5
0.0	Bias (ESE) ($\times 10^{-3}$)	0.0 (9.0)	0.0 (9.2)	0.0 (9.2)	0.0 (9.1)	0.0 (9.4)	0.0 (9.4)	0.0 (9.1)	0.0 (9.7)	0.0 (9.8)	0.0 (9.0)
	ASE (SE) ($\times 10^{-3}$)	9.0 (0.1)	8.9 (0.1)	9.1 (0.1)	9.0 (0.1)	8.7 (0.1)	9.2 (0.1)	8.9 (0.1)	8.3 (0.1)	9.4 (0.1)	9.0 (0.1)
	CovP (%)	94.8	93.9	94.6	94.8	93.0	94.4	94.4	90.6	94.0	94.9
	Event Rate = 0.10										
	Time[1] (min)	2.1	2.2	2.1	0.8	0.8	0.8	0.3	0.3	0.3	24.0
0.8	Bias (ESE) ($\times 10^{-3}$)	0.7 (6.9)	0.7 (6.9)	3.8 (7.0)	0.5 (6.6)	0.7 (6.7)	7.7 (6.8)	-0.6 (6.7)	-0.3 (7.0)	15.4 (7.0)	0.8 (6.9)
	ASE (SE) ($\times 10^{-3}$)	6.8 (0.0)	6.7 (0.0)	6.8 (0.0)	6.7 (0.0)	6.6 (0.0)	6.8 (0.0)	6.7 (0.0)	6.4 (0.0)	6.9 (0.0)	6.8 (0.1)
	CovP (%)	94.5	94.2	90.6	95.0	94.2	79.5	95.2	92.6	40.1	94.5
0.6	Bias (ESE) ($\times 10^{-3}$)	0.4 (10.3)	0.5 (10.5)	3.1 (10.4)	0.5 (10.2)	0.7 (10.5)	6.4 (10.5)	-0.2 (10.4)	-0.2 (10.8)	13.3 (10.9)	0.5 (10.3)
	ASE (SE) ($\times 10^{-3}$)	10.1 (0.1)	9.9 (0.1)	10.1 (0.1)	10.0 (0.1)	9.8 (0.1)	10.2 (0.1)	10.0 (0.1)	9.5 (0.1)	10.3 (0.1)	10.1 (0.1)
	CovP (%)	94.6	94.2	93.4	94.8	93.6	89.8	94.3	91.8	72.4	94.7
0.4	Bias (ESE) ($\times 10^{-3}$)	-0.1 (11.2)	0.0 (11.3)	1.5 (11.3)	0.0 (11.1)	0.1 (11.4)	3.6 (11.4)	-0.4 (11.0)	-0.2 (11.3)	7.8 (11.5)	-0.1 (11.2)
	ASE (SE) ($\times 10^{-3}$)	10.9 (0.1)	10.8 (0.1)	11.0 (0.1)	10.9 (0.1)	10.7 (0.1)	11.1 (0.1)	10.9 (0.1)	10.4 (0.1)	11.2 (0.1)	11.0 (0.1)
	CovP (%)	94.1	94.0	94.2	94.0	93.2	93.5	95.0	92.8	88.8	94.2
0.2	Bias (ESE) ($\times 10^{-3}$)	-0.1 (12.1)	0.0 (12.2)	0.6 (12.2)	0.1 (12.3)	0.1 (12.5)	1.6 (12.5)	-0.3 (12.1)	0.0 (12.5)	3.0 (12.6)	-0.1 (12.1)
	ASE (SE) ($\times 10^{-3}$)	12.1 (0.1)	12.0 (0.1)	12.1 (0.1)	12.1 (0.1)	11.8 (0.1)	12.2 (0.1)	12.0 (0.1)	11.5 (0.1)	12.4 (0.1)	12.1 (0.1)
	CovP (%)	95.3	94.3	95.2	94.4	93.7	94.0	94.5	93.1	93.6	95.2
0.0	Bias (ESE) ($\times 10^{-3}$)	0.0 (6.3)	0.0 (6.4)	0.0 (6.4)	0.0 (6.4)	0.0 (6.5)	0.0 (6.5)	0.0 (6.4)	0.0 (6.6)	0.0 (6.7)	0.0 (6.3)
	ASE (SE) ($\times 10^{-3}$)	6.3 (0.0)	6.3 (0.0)	6.4 (0.0)	6.3 (0.0)	6.2 (0.0)	6.4 (0.0)	6.3 (0.0)	6.0 (0.0)	6.5 (0.0)	6.3 (0.0)
	CovP (%)	94.8	94.4	94.7	94.7	93.6	94.5	94.6	92.5	94.2	94.9

[1] Average computation time in minutes.
Abbreviations: ASE: asymptotic standard error using the theoretical formula; ESE: empirical standard error; SE: standard error of ASE; CovP: empirical coverage probability of 95% confidence interval.
Note: $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$; Intra-cluster association is Kendall tau $\kappa = 0.3$; Sample size is $100,000$; Event rates of top panel and bottom panel are 5% and 10%, respectively.

Table 6.2: Performances of $\hat{\boldsymbol{\beta}}_{\rho}^{dc}$ and $\hat{\boldsymbol{\beta}}_{\rho}^{full}$ for estimating $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ in marginal models with simple random splitting.

$\boldsymbol{\beta}_0$		5 Subsets			10 Subsets			20 Subsets			Full Data
		$\mathbf{W}_{1s}(.)$	$\hat{\boldsymbol{\beta}}_{\rho}^{dc}$ $\mathbf{W}_{2s}(.)$	$\mathbf{W}_{3s}(.)$	$\mathbf{W}_{1s}(.)$	$\hat{\boldsymbol{\beta}}_{\rho}^{dc}$ $\mathbf{W}_{2s}(.)$	$\mathbf{W}_{3s}(.)$	$\mathbf{W}_{1s}(.)$	$\hat{\boldsymbol{\beta}}_{\rho}^{dc}$ $\mathbf{W}_{2s}(.)$	$\mathbf{W}_{3s}(.)$	$\hat{\boldsymbol{\beta}}_{\rho}^{full}$
	Event Rate = 0.05										
	Time[1] (min)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.8	Bias (ESE) ($\times 10^{-3}$)	1.0 (9.5)	1.1 (9.6)	6.2 (9.7)	-0.1 (9.5)	0.1 (9.8)	12.1 (9.9)	-3.6 (9.7)	-3.8 (10.3)	25.1 (10.5)	1.3 (9.5)
	ASE (SE) ($\times 10^{-3}$)	9.5 (0.1)	9.3 (0.1)	9.6 (0.1)	9.4 (0.1)	9.1 (0.1)	9.7 (0.1)	9.4 (0.1)	8.8 (0.1)	10.0 (0.1)	9.5 (0.1)
	CovP (%)	94.5	93.5	90.2	94.6	93.2	74.8	92.5	88.5	29.2	94.3
	Sn (%)	100	100	100	100	100	100	100	100	100	100
0.6	Bias (ESE) ($\times 10^{-3}$)	0.2 (13.6)	0.3 (13.9)	4.8 (13.9)	-0.2 (13.7)	0.1 (14.1)	10.5 (14.3)	-2.6 (13.7)	-2.9 (14.7)	21.8 (14.9)	0.5 (13.6)
	ASE (SE) ($\times 10^{-3}$)	13.6 (0.2)	13.4 (0.2)	13.7 (0.2)	13.5 (0.2)	13.1 (0.2)	13.9 (0.2)	13.4 (0.2)	12.5 (0.2)	14.3 (0.2)	13.7 (0.2)
	CovP (%)	95.0	94.0	92.6	95.0	92.7	87.3	94.1	89.5	66.2	95.0
	Sn (%)	100	100	100	100	100	100	100	100	100	100
0.4	Bias (ESE) ($\times 10^{-3}$)	0.2 (15.3)	0.3 (15.6)	2.8 (15.5)	0.1 (15.9)	0.4 (16.4)	6.4 (16.4)	-2.1 (15.2)	-2.0 (16.5)	12.3 (16.4)	0.4 (15.3)
	ASE (SE) ($\times 10^{-3}$)	15.6 (0.2)	15.4 (0.2)	15.8 (0.2)	15.6 (0.2)	15.0 (0.2)	16.0 (0.2)	15.4 (0.2)	14.4 (0.2)	16.4 (0.2)	15.7 (0.2)
	CovP (%)	95.4	94.5	94.9	94.7	92.5	92.8	95.5	91.2	88.7	95.4
	Sn (%)	100	100	100	100	100	100	100	100	100	100
0.2	Bias (ESE) ($\times 10^{-3}$)	0.2 (18.9)	0.3 (19.3)	0.9 (19.3)	-0.1 (18.1)	0.0 (18.7)	1.9 (18.7)	-0.3 (18.2)	-0.4 (19.7)	5.0 (19.4)	0.2 (18.9)
	ASE (SE) ($\times 10^{-3}$)	18.3 (0.3)	18.0 (0.3)	18.5 (0.3)	18.2 (0.3)	17.6 (0.3)	18.7 (0.3)	18.1 (0.3)	16.9 (0.3)	19.1 (0.3)	18.4 (0.3)
	CovP (%)	94.4	93.0	94.2	95.0	93.1	94.3	95.2	90.6	93.8	94.3
	Sn (%)	100	100	100	100	100	100	100	100	100	100
0.0	Bias (ESE) ($\times 10^{-3}$)	0.0 (0.5)	0.0 (0.6)	0.0 (0.5)	0.0 (0.4)	0.0 (0.6)	0.0 (0.4)	0.0 (0.4)	0.0 (0.9)	0.0 (0.5)	0.0 (0.1)
	Sp (%)	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.8	99.9	99.9
	Event Rate = 0.10										
	Time[1] (min)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.8	Bias (ESE) ($\times 10^{-3}$)	0.5 (6.7)	0.6 (6.8)	3.6 (6.8)	0.2 (6.7)	0.4 (6.8)	7.3 (6.8)	-0.6 (6.9)	-0.3 (7.1)	15.5 (7.2)	0.7 (6.7)
	ASE (SE) ($\times 10^{-3}$)	6.8 (0.0)	6.7 (0.0)	6.8 (0.0)	6.7 (0.0)	6.6 (0.0)	6.8 (0.0)	6.7 (0.0)	6.4 (0.0)	6.9 (0.0)	6.8 (0.0)
	CovP (%)	95.4	94.8	91.7	95.0	94.0	80.8	94.2	92.2	39.2	95.4
	Sn (%)	100	100	100	100	100	100	100	100	100	100
0.6	Bias (ESE) ($\times 10^{-3}$)	0.6 (10.2)	0.7 (10.3)	3.2 (10.4)	0.3 (10.0)	0.4 (10.2)	6.2 (10.3)	-0.4 (10.0)	-0.3 (10.3)	13.0 (10.5)	0.7 (10.2)
	ASE (SE) ($\times 10^{-3}$)	10.1 (0.1)	9.9 (0.1)	10.1 (0.1)	10.0 (0.1)	9.8 (0.1)	10.2 (0.1)	10.0 (0.1)	9.5 (0.1)	10.3 (0.1)	10.1 (0.1)
	CovP (%)	95.0	94.6	93.0	95.1	94.0	90.3	95.2	93.2	74.8	95.2
	Sn (%)	100	100	100	100	100	100	100	100	100	100
0.4	Bias (ESE) ($\times 10^{-3}$)	0.5 (10.9)	0.6 (11.0)	2.1 (11.0)	0.2 (10.9)	0.4 (11.2)	3.9 (11.2)	-0.5 (10.6)	-0.2 (11.0)	7.7 (11.1)	0.5 (10.9)
	ASE (SE) ($\times 10^{-3}$)	10.9 (0.1)	10.8 (0.1)	11.0 (0.1)	10.9 (0.1)	10.7 (0.1)	11.1 (0.1)	10.9 (0.1)	10.4 (0.1)	11.2 (0.1)	11.0 (0.1)
	CovP (%)	95.2	94.7	94.7	95.0	94.2	93.8	94.9	93.6	90.3	95.2
	Sn (%)	100	100	100	100	100	100	100	100	100	100
0.2	Bias (ESE) ($\times 10^{-3}$)	-0.2 (12.0)	-0.1 (12.2)	0.4 (12.1)	-0.3 (12.6)	-0.2 (12.8)	1.2 (12.9)	-0.1 (12.0)	0.1 (12.5)	3.3 (12.6)	-0.1 (12.0)
	ASE (SE) ($\times 10^{-3}$)	12.1 (0.1)	12.0 (0.1)	12.1 (0.1)	12.1 (0.1)	11.8 (0.1)	12.2 (0.1)	12.0 (0.1)	11.5 (0.1)	12.4 (0.1)	12.1 (0.2)
	CovP (%)	94.9	94.5	94.7	93.5	92.9	93.1	94.8	93.0	93.8	94.8
	Sn (%)	100	100	100	100	100	100	100	100	100	100
0.0	Bias (ESE) ($\times 10^{-3}$)	0.0 (0.2)	0.0 (0.3)	0.0 (0.2)	0.0 (0.2)	0.0 (0.3)	0.0 (0.2)	0.0 (0.2)	0.0 (0.4)	0.0 (0.2)	0.0 (0.2)
	Sp (%)	100	100	100	100	99.9	100	100	99.9	100	100

[1] Average computation time in minutes.
Abbreviations: ASE: asymptotic standard error using the theoretical formula; ESE: empirical standard error; SE: standard error of ASE; CovP: empirical coverage probability of 95% confidence interval; Sn: sensitivity; Sp: specificity.
Note: $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$; Intra-cluster association is Kendall tau $\kappa = 0.3$; Sample size is $100,000$; Event rates of top panel and bottom panel are 5% and 10%, respectively.

Table 6.3: Performances of $\hat{\boldsymbol{\beta}}^{dc}$ and $\hat{\boldsymbol{\beta}}^{full}$ for estimating $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ in marginal models with stratified random splitting.

$\boldsymbol{\beta}_0$		5 Subsets			10 Subsets			20 Subsets			Full Data
			$\hat{\boldsymbol{\beta}}^{dc}$			$\hat{\boldsymbol{\beta}}^{dc}$			$\hat{\boldsymbol{\beta}}^{dc}$		$\hat{\boldsymbol{\beta}}^{full}$
		$\mathbf{W}_{1s}(.)$	$\mathbf{W}_{2s}(.)$	$\mathbf{W}_{3s}(.)$	$\mathbf{W}_{1s}(.)$	$\mathbf{W}_{2s}(.)$	$\mathbf{W}_{3s}(.)$	$\mathbf{W}_{1s}(.)$	$\mathbf{W}_{2s}(.)$	$\mathbf{W}_{3s}(.)$	
	Event Rate = 0.05										
	Time[1] (min)	1.3	1.3	1.3	0.4	0.4	0.4	0.2	0.2	0.2	12.5
0.8	Bias (ESE) ($\times 10^{-3}$)	1.5 (9.4)	1.6 (9.6)	6.7 (9.6)	0.0 (9.4)	0.2 (9.7)	12.3 (9.8)	-3.0 (9.3)	-3.2 (9.9)	25.7 (10.2)	1.9 (9.4)
	ASE (SE) ($\times 10^{-3}$)	9.5 (0.1)	9.3 (0.1)	9.6 (0.1)	9.5 (0.1)	9.1 (0.1)	9.7 (0.1)	9.4 (0.1)	8.8 (0.1)	10.0 (0.1)	9.5 (0.1)
	CovP (%)	95.0	94.2	90.6	94.4	93.5	74.9	94.0	90.0	27.2	95.2
0.6	Bias (ESE) ($\times 10^{-3}$)	0.8 (14.2)	0.8 (14.5)	5.3 (14.4)	-0.2 (13.6)	-0.2 (14.2)	10.4 (14.2)	-2.3 (13.7)	-2.5 (15.0)	22.1 (14.8)	1.0 (14.2)
	ASE (SE) ($\times 10^{-3}$)	13.6 (0.2)	13.4 (0.2)	13.7 (0.2)	13.5 (0.2)	13.1 (0.2)	13.9 (0.2)	13.4 (0.2)	12.5 (0.2)	14.3 (0.2)	13.7 (0.2)
	CovP (%)	94.2	93.2	92.1	95.0	93.0	88.5	93.9	89.3	66.5	94.3
0.4	Bias (ESE) ($\times 10^{-3}$)	0.4 (15.7)	0.4 (16.0)	3.0 (16.0)	-0.3 (16.2)	-0.1 (16.9)	5.8 (16.7)	-1.9 (15.8)	-1.8 (16.9)	12.5 (17.0)	0.5 (15.8)
	ASE (SE) ($\times 10^{-3}$)	15.6 (0.2)	15.4 (0.2)	15.8 (0.2)	15.6 (0.2)	15.1 (0.2)	16.0 (0.2)	15.4 (0.2)	14.4 (0.2)	16.4 (0.2)	15.7 (0.2)
	CovP (%)	95.2	94.1	94.6	94.0	92.0	91.3	93.8	90.3	88.9	95.2
0.2	Bias (ESE) ($\times 10^{-3}$)	0.0 (18.3)	0.1 (18.7)	0.9 (18.6)	0.2 (18.3)	0.2 (18.8)	2.3 (18.8)	-0.8 (18.8)	-0.8 (20.2)	4.3 (20.0)	0.1 (18.3)
	ASE (SE) ($\times 10^{-3}$)	18.3 (0.3)	18.0 (0.3)	18.4 (0.3)	18.2 (0.3)	17.6 (0.3)	18.7 (0.3)	18.1 (0.3)	16.9 (0.3)	19.1 (0.3)	18.3 (0.3)
	CovP (%)	94.7	94.0	94.7	95.0	93.7	94.8	93.7	89.8	93.0	95.0
0.0	Bias (ESE) ($\times 10^{-3}$)	0.0 (9.1)	0.0 (9.2)	0.0 (9.2)	0.0 (9.1)	0.0 (9.4)	0.0 (9.4)	0.0 (9.1)	0.0 (9.7)	0.0 (9.9)	0.0 (9.1)
	ASE (SE) ($\times 10^{-3}$)	9.0 (0.1)	8.9 (0.1)	9.1 (0.1)	9.0 (0.1)	8.7 (0.1)	9.2 (0.1)	8.9 (0.1)	8.3 (0.1)	9.4 (0.1)	9.0 (0.1)
	CovP (%)	94.9	94.1	94.6	94.8	93.1	94.5	94.4	90.5	94.0	94.9
	Event Rate = 0.10										
	Time[1] (min)	2.3	2.3	2.3	0.7	0.7	0.7	0.3	0.3	0.3	24.7
0.8	Bias (ESE) ($\times 10^{-3}$)	0.9 (7.0)	1.0 (7.0)	4.0 (7.0)	0.5 (6.6)	0.7 (6.8)	7.7 (6.7)	-0.6 (6.8)	-0.3 (7.1)	15.4 (7.1)	1.0 (7.0)
	ASE (SE) ($\times 10^{-3}$)	6.8 (0.0)	6.7 (0.0)	6.8 (0.0)	6.7 (0.0)	6.6 (0.0)	6.8 (0.0)	6.7 (0.0)	6.4 (0.0)	6.9 (0.0)	6.8 (0.1)
	CovP (%)	93.8	93.5	90.1	95.3	94.2	81.2	95.3	92.6	39.9	93.8
0.6	Bias (ESE) ($\times 10^{-3}$)	0.5 (10.1)	0.6 (10.3)	3.1 (10.3)	0.4 (10.1)	0.6 (10.3)	6.5 (10.4)	-0.2 (10.4)	-0.1 (10.8)	13.2 (11.0)	0.6 (10.1)
	ASE (SE) ($\times 10^{-3}$)	10.1 (0.1)	9.9 (0.1)	10.1 (0.1)	10.0 (0.1)	9.8 (0.1)	10.2 (0.1)	10.0 (0.1)	9.5 (0.1)	10.3 (0.1)	10.1 (0.1)
	CovP (%)	95.2	94.2	93.1	94.8	93.3	89.7	94.0	92.0	72.8	95.1
0.4	Bias (ESE) ($\times 10^{-3}$)	0.0 (11.0)	0.0 (11.1)	1.5 (11.1)	0.3 (10.8)	0.4 (11.0)	4.0 (11.1)	-0.2 (11.0)	0.0 (11.5)	8.0 (11.5)	0.0 (11.0)
	ASE (SE) ($\times 10^{-3}$)	10.9 (0.1)	10.8 (0.1)	11.0 (0.1)	10.9 (0.1)	10.7 (0.1)	11.1 (0.1)	10.9 (0.1)	10.4 (0.1)	11.2 (0.1)	11.0 (0.1)
	CovP (%)	94.5	94.1	94.6	95.5	94.9	93.3	94.9	92.6	88.8	94.5
0.2	Bias (ESE) ($\times 10^{-3}$)	-0.1 (12.1)	-0.1 (12.2)	0.5 (12.3)	0.0 (12.1)	0.1 (12.4)	1.4 (12.4)	-0.3 (12.1)	-0.1 (12.7)	3.1 (12.7)	-0.1 (12.2)
	ASE (SE) ($\times 10^{-3}$)	12.1 (0.1)	12.0 (0.1)	12.1 (0.1)	12.1 (0.1)	11.8 (0.1)	12.2 (0.1)	12.0 (0.1)	11.5 (0.1)	12.4 (0.1)	12.1 (0.1)
	CovP (%)	94.8	94.6	94.6	95.0	94.0	94.8	94.7	92.5	93.3	94.8
0.0	Bias (ESE) ($\times 10^{-3}$)	0.0 (6.4)	0.0 (6.4)	0.0 (6.4)	0.0 (6.4)	0.0 (6.5)	0.0 (6.5)	0.0 (6.4)	0.0 (6.6)	0.0 (6.7)	0.0 (6.4)
	ASE (SE) ($\times 10^{-3}$)	6.3 (0.0)	6.3 (0.0)	6.4 (0.0)	6.3 (0.0)	6.2 (0.0)	6.4 (0.0)	6.3 (0.0)	6.0 (0.0)	6.5 (0.0)	6.3 (0.0)
	CovP (%)	94.8	94.5	94.7	94.8	93.7	94.5	94.5	92.5	94.1	94.9

[1] Average computation time in minutes.
Abbreviations: ASE: asymptotic standard error using the theoretical formula; ESE: empirical standard error; SE: standard error of ASE; CovP: empirical coverage probability of 95% confidence interval.
Note: $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$; Intra-cluster association is Kendall tau $\kappa = 0.3$; Sample size is $100,000$; Event rates of top panel and bottom panel are 5% and 10%, respectively.

Table 6.4: Performances of $\hat{\boldsymbol{\beta}}_{\rho}^{dc}$ and $\hat{\boldsymbol{\beta}}_{\rho}^{full}$ for estimating $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ in marginal models with stratified random splitting.

$\boldsymbol{\beta}_0$		5 Subsets			10 Subsets			20 Subsets			Full Data
		$\hat{\boldsymbol{\beta}}^{dc}$			$\hat{\boldsymbol{\beta}}^{dc}$			$\hat{\boldsymbol{\beta}}^{dc}$			$\hat{\boldsymbol{\beta}}^{full}$
		$\mathbf{W}_{1s}(.)$	$\mathbf{W}_{2s}(.)$	$\mathbf{W}_{3s}(.)$	$\mathbf{W}_{1s}(.)$	$\mathbf{W}_{2s}(.)$	$\mathbf{W}_{3s}(.)$	$\mathbf{W}_{1s}(.)$	$\mathbf{W}_{2s}(.)$	$\mathbf{W}_{3s}(.)$	
	Event Rate = 0.05										
	Time[q] (min)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.8	Bias (ESE) ($\times 10^{-3}$)	1.2 (9.7)	1.3 (9.9)	6.4 (9.9)	-0.1 (9.7)	0.0 (10.1)	12.3 (10.1)	-3.6 (9.7)	-3.7 (10.2)	25.0 (10.6)	1.5 (9.7)
	ASE (SE) ($\times 10^{-3}$)	9.5 (0.1)	9.3 (0.1)	9.6 (0.1)	9.4 (0.1)	9.1 (0.1)	9.7 (0.1)	9.4 (0.1)	8.8 (0.1)	10.0 (0.1)	9.5 (0.1)
	CovP (%)	93.8	93.3	89.7	93.8	92.0	74.4	92.3	88.1	30.1	93.8
	Sn (%)	100	100	100	100	100	100	100	100	100	100
0.6	Bias (ESE) ($\times 10^{-3}$)	0.9 (13.9)	1.0 (14.2)	5.4 (14.1)	0.0 (13.8)	0.0 (14.2)	10.7 (14.4)	-2.6 (13.7)	-2.9 (14.6)	21.7 (15.0)	1.2 (14.0)
	ASE (SE) ($\times 10^{-3}$)	13.6 (0.2)	13.4 (0.2)	13.8 (0.2)	13.5 (0.2)	13.1 (0.2)	13.9 (0.2)	13.4 (0.2)	12.5 (0.2)	14.3 (0.2)	13.7 (0.2)
	CovP (%)	94.8	94.0	92.3	94.6	93.0	87.5	94.0	89.9	66.9	94.9
	Sn (%)	100	100	100	100	100	100	100	100	100	100
0.4	Bias (ESE) ($\times 10^{-3}$)	0.3 (15.7)	0.5 (15.8)	2.9 (15.9)	0.0 (15.7)	0.2 (16.2)	6.1 (16.3)	-2.0 (15.3)	-1.8 (16.5)	12.2 (16.7)	0.5 (15.6)
	ASE (SE) ($\times 10^{-3}$)	15.6 (0.2)	15.4 (0.2)	15.8 (0.2)	15.6 (0.2)	15.0 (0.2)	16.0 (0.2)	15.4 (0.2)	14.4 (0.2)	16.4 (0.2)	15.7 (0.2)
	CovP (%)	94.5	94.0	94.4	94.5	93.5	92.8	95.6	91.0	87.7	94.8
	Sn (%)	100	100	100	100	100	100	100	100	100	100
0.2	Bias (ESE) ($\times 10^{-3}$)	0.6 (18.8)	0.6 (19.1)	1.3 (19.0)	-0.1 (18.0)	0.0 (18.8)	1.9 (18.6)	-0.3 (18.3)	0.1 (19.7)	4.9 (19.5)	0.6 (18.8)
	ASE (SE) ($\times 10^{-3}$)	18.3 (0.3)	18.0 (0.3)	18.5 (0.3)	18.2 (0.3)	17.6 (0.3)	18.7 (0.3)	18.1 (0.3)	16.9 (0.3)	19.1 (0.3)	18.3 (0.3)
	CovP (%)	94.2	93.0	94.0	95.2	93.1	95.2	95.0	91.2	93.8	94.2
	Sn (%)	100	100	100	100	100	100	100	100	100	100
0.0	Bias (ESE) ($\times 10^{-3}$)	0.0 (0.5)	0.0 (0.5)	0.0 (0.6)	0.0 (0.4)	0.0 (0.4)	0.0 (0.4)	0.0 (0.4)	0.0 (0.9)	0.0 (0.5)	0.0 (0.5)
	Sp (%)	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
	Event Rate = 0.10										
	Time[1] (min)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.8	Bias (ESE) ($\times 10^{-3}$)	0.7 (6.6)	0.8 (6.7)	3.8 (6.7)	0.1 (6.8)	0.3 (6.9)	7.4 (6.9)	-0.6 (6.8)	-0.2 (6.9)	15.6 (7.1)	0.8 (6.6)
	ASE (SE) ($\times 10^{-3}$)	6.8 (0.0)	6.7 (0.0)	6.8 (0.0)	6.7 (0.0)	6.6 (0.0)	6.8 (0.0)	6.7 (0.0)	6.4 (0.0)	6.9 (0.0)	6.8 (0.1)
	CovP (%)	95.1	94.6	91.5	94.8	94.0	80.5	94.4	93.4	38.1	95.2
	Sn (%)	100	100	100	100	100	100	100	100	100	100
0.6	Bias (ESE) ($\times 10^{-3}$)	0.9 (10.1)	1.1 (10.2)	3.6 (10.2)	0.5 (10.1)	0.7 (10.3)	6.6 (10.4)	-0.4 (10.0)	-0.3 (10.4)	12.9 (10.5)	1.0 (10.1)
	ASE (SE) ($\times 10^{-3}$)	10.1 (0.1)	9.9 (0.1)	10.1 (0.1)	10.0 (0.1)	9.8 (0.1)	10.2 (0.1)	10.0 (0.1)	9.5 (0.1)	10.3 (0.1)	10.1 (0.1)
	CovP (%)	95.0	94.8	93.5	94.7	93.3	89.9	94.9	92.5	75.8	95.1
	Sn (%)	100	100	100	100	100	100	100	100	100	100
0.4	Bias (ESE) ($\times 10^{-3}$)	0.4 (10.9)	0.5 (11.0)	2.0 (10.9)	0.6 (10.8)	0.8 (10.9)	4.2 (11.0)	-0.7 (10.8)	-0.4 (11.3)	7.5 (11.3)	0.5 (10.9)
	ASE (SE) ($\times 10^{-3}$)	10.9 (0.1)	10.8 (0.1)	11.0 (0.1)	10.9 (0.1)	10.7 (0.1)	11.1 (0.1)	10.9 (0.1)	10.4 (0.1)	11.2 (0.1)	11.0 (0.1)
	CovP (%)	95.7	95.4	95.4	95.0	94.3	93.0	94.8	92.8	89.3	95.6
	Sn (%)	100	100	100	100	100	100	100	100	100	100
0.2	Bias (ESE) ($\times 10^{-3}$)	-0.3 (12.0)	-0.2 (12.1)	0.4 (12.1)	-0.4 (12.5)	-0.3 (12.8)	1.1 (12.8)	-0.1 (12.1)	0.2 (12.5)	3.2 (12.5)	-0.2 (12.0)
	ASE (SE) ($\times 10^{-3}$)	12.1 (0.1)	12.0 (0.1)	12.1 (0.1)	12.1 (0.1)	11.8 (0.1)	12.2 (0.1)	12.0 (0.1)	11.5 (0.1)	12.4 (0.1)	12.1 (0.1)
	CovP (%)	94.8	94.4	94.8	94.0	92.9	93.9	95.2	92.7	94.0	95.0
	Sn (%)	100	100	100	100	100	100	100	100	100	100
0.0	Bias (ESE) ($\times 10^{-3}$)	0.0 (0.2)	0.0 (0.2)	0.0 (0.2)	0.0 (0.2)	0.0 (0.3)	0.0 (0.3)	0.0 (0.2)	0.0 (0.3)	0.0 (0.3)	0.0 (0.1)
	Sp (%)	100	100	100	100	99.9	99.9	100	100	99.9	100

[1] Average computation time in minutes.
Abbreviations: ASE: asymptotic standard error using the theoretical formula; ESE: empirical standard error; SE: standard error of ASE; CovP: empirical coverage probability of 95% confidence interval; Sn: sensitivity; Sp: specificity.
Note: $\boldsymbol{\beta}_0$ is the true value of $\boldsymbol{\beta}$; Intra-cluster association is Kendall tau $\kappa = 0.3$; Sample size is $100,000$; Event rates of top panel and bottom panel are 5% and 10%, respectively.

6.2.2 Frailty Models in Multivariate Survival Analysis

In frailty models, the estimation and inference procedures are excessively complex, thus a simulation study with reduced numbers of independent clusters ($n = 400$ and 1000) was conducted to assess the asymptotic equivalence between the divide-and-combine estimator $\hat{\boldsymbol{\gamma}}^{dc}$ and the full-data estimator $\hat{\boldsymbol{\gamma}}^{full}$. The statistical properties of the regularized estimators $\hat{\boldsymbol{\gamma}}_{\rho}^{dc}$ and $\hat{\boldsymbol{\gamma}}_{\rho}^{full}$ were also evaluated, such as estimation consistency, selection consistency, and oracle properties. Note that $\boldsymbol{\gamma}$ contains both the dependence parameter and the regression parameters, i.e., $\boldsymbol{\gamma} = (\theta, \boldsymbol{\beta}^T)^T$.

The multivariate unordered failure time data were generated from the gamma frailty model (Hougaard, 2000) with a conditional Weibull distribution for $K = 3$ types of failures. Conditional on the frailty u, the baseline hazard function was parameterized by $u\lambda_0(t) = u\xi h(ht)^{\xi-1}$ where $\lambda_0(t) = \xi h(ht)^{\xi-1}$ is denoted by $t \sim$ Weibull(shape = ξ, scale = h) and set $\xi = h = 2$. The frailty u was assumed to follow a gamma distribution with mean of one and variance of θ, whose density function, as discussed in (2.15), is given by

$$f_U(u|\theta) = \frac{u^{1/\theta-1}e^{-u/\theta}}{\theta^{1/\theta}\Gamma(1/\theta)}. \tag{6.3}$$

Therefore, according to (2.16), the joint survival function of failure times T_{i1}, T_{i2}, and T_{i3} in the i^{th} cluster ($i = 1, 2, \ldots, n$) is given by

$$\begin{aligned} &Pr(T_{i1} > t_{i1}, T_{i2} > t_{i2}, T_{i3} > t_{i3} | \boldsymbol{Z}_{i1}, \boldsymbol{Z}_{i2}, \boldsymbol{Z}_{i3}) \\ &\qquad = \left[\sum_{k=1}^{3} \left\{ 1 + \theta \left(ht_{ik}\right)^{\xi} e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{ik}} \right\} - 2 \right]^{-1/\theta}, \end{aligned} \tag{6.4}$$

where $\theta = 6/7$ is chosen, corresponding to Kendall's tau $\kappa = 0.3$ for a same moderate intra-cluster association as in the simulation studies for marginal models. Specifically, we generated K-variate correlated unorderd failure times for n independent clusters by the following steps.

(S1) Generate u_i from Gamma(shape = $\frac{1}{\theta}$, rate = $\frac{1}{\theta}$);

(S2) Generate K i.i.d. random variables U_{ik} from Uniform(0, 1) for the conditional survival function $S_{ik}(t_{ik}|u_i)$, $k = 1, 2, \ldots, K$;

(S3) Recognizing that $S_{ik}(t_{ik}|u_i)$ is the function defined in (2.17), following the principle in the Monte Carlo method, we have $U_{ik} = \exp\left\{-u_i\Lambda_0(T_{ik})e^{\boldsymbol{\beta}^T\boldsymbol{Z}_{ik}}\right\}$. Given that $\Lambda_0(T_{ik}) = (hT_{ik})^{\xi}$, one gets

$$T_{ik} = \left\{\frac{logU_{ik}}{-e^{\boldsymbol{\beta}^T\boldsymbol{Z}_{ik}} \cdot u_i h^{\xi}}\right\}^{1/\xi} \tag{6.5}$$

for the k^{th} type of failure in the i^{th} cluster, $k = 1, 2, \ldots, K$, $i = 1, 2, \ldots, n$.

We considered the proportional hazards frailty model (3.4) with a shared regression parameter $\boldsymbol{\beta} = (\mathbf{0.8}_2, \mathbf{0.4}_2, \mathbf{0}_6)^T$ consisting of four nonzeros and six zeros. In the i^{th} cluster, the covariate $\boldsymbol{Z}_{ik}, k = 1, 2, 3$, were generated from a multivariate normal distribution with standard normal marginals and an equal correlation of 0.2. A fixed administrative censoring time of 0.48 was used to yield an event rate of 50%. Because relatively small sample sizes were considered in the simulation study for frailty models, a simpler partition algorithm was used: $n_s/n = (\mathbf{2}/\mathbf{5}_2, \mathbf{1}/\mathbf{5}_1)$ and $n_s/n = (\mathbf{2}/\mathbf{8}_3, \mathbf{1}/\mathbf{8}_2)$, corresponding to $S = 3$ and 5 subsets, respectively. Similar to the simulation study for the marginal models, we ran simulations 500 times in each configuration. All simulations were carried out on a Linux cluster via parallel computing (S cores) with one subset allocated to one core. The average computation time was calculated based on 50 simulations performed on Intel® Xeon ® E5-2680 v4 @2.40GHz. All statistical analyses regarding fitting proportional hazards frailty models were performed using R package, `frailtySurv` (Monaco et al., 2018), and regularized estimation was conducted using R package, `glmnet` (Friedman et al., 2010).

Simulation Results

The performance of the divide-and-combine estimator $\hat{\boldsymbol{\gamma}}^{dc}$ was assessed, compared to the full-data estimator $\hat{\boldsymbol{\gamma}}^{full}$ in regard to average computation time (Time), mean of biasedness (Bias), empirical standard error (ESE), mean asymptotic standard error (ASE) using the theoretical formula and its associated standard error, and empirical coverage probability (CovP) of the 95% Wald-type confidence interval, with results summarized in Table 6.5. When $n = 1000$, fitting a proportional hazards frailty model in the full data could not be finished within a reasonable time period—it took more than 420 minutes for each simulation run and thus the results of $\hat{\boldsymbol{\gamma}}^{full}$ for $n = 1000$ were not reported. The computational barriers in the full-data analysis for $n = 1000$ were easily overcome by using the divide-and-combine analysis. The estimation was completed without loosing any efficiency in 24 minutes when the full data were split into $S = 3$ subsets. When any of the three weights ($\mathbf{W}_{1s}(.)$, $\mathbf{W}_{2s}(.)$, and $\mathbf{W}_{3s}(.)$) is used for $\hat{\boldsymbol{\gamma}}^{dc}$, the bias is generally small, ESE and ASE are close to each other, and CovP is close to the nominal 95% level. The computation time was further shortened to 6 minutes while keeping similar estimation efficiency for $\boldsymbol{\beta}$ when the full data were split into $S = 5$ subsets. However, the CovP for θ when $\mathbf{W}_{1s}(.)$ or $\mathbf{W}_{2s}(.)$ is used, is less than the nominal level. The similar trends were also observed when $n = 400$, but under the reduced sample size the estimation algorithm in divided subsets was less stable due to the small sample size. As a result, the performance of $\hat{\boldsymbol{\gamma}}^{dc}$ using $\mathbf{W}_{1s}(.)$ or $\mathbf{W}_{2s}(.)$ generally, was not as good as expected, especially when estimating the dependence parameter θ.

The performances of the regularized estimators $\hat{\boldsymbol{\gamma}}_{\rho}^{dc}$ and $\hat{\boldsymbol{\gamma}}_{\rho}^{full}$ were also investigated (Table 6.6). Because regularization was only applied on $\boldsymbol{\beta}$ but not θ, thus only $\hat{\boldsymbol{\beta}}_{\rho}^{dc}$ and $\hat{\boldsymbol{\beta}}_{\rho}^{full}$ were summarized in Table 6.6. As in the simulation studies for marginal models, computation time, biasedness, ESE, ASE, CovP, as well as sensitivity and specificity were assessed in $\hat{\boldsymbol{\beta}}_{\rho}^{dc}$ and $\hat{\boldsymbol{\beta}}_{\rho}^{full}$ to evaluate their performances. The computation time was only measured in the regularization step. Because the dimensionality of the "data" is signif-

icantly reduced from n to d, the average computation time of this step is small (i.e., < 1 second) across different scenarios. When $n = 1000$, $\hat{\boldsymbol{\beta}}_{\rho}^{dc}$ using $\mathbf{W}_{1s}(.)$ and $\mathbf{W}_{3s}(.)$ achieved acceptable statistical performance as reflected by the near nominal-level coverage probability and the almost 100% sensitivity and specificity. When $\hat{\boldsymbol{\beta}}_{\rho}^{dc}$ using $\mathbf{W}_{2s}(.)$ was used in the divide-and-combine analysis, there was some amount of efficiency lost, especially when the magnitude of $\boldsymbol{\beta}$ is large and the number of subsets is big. When the sample size was reduced to $n = 400$, the similar patterns were observed but the performance was relatively worse than that in $n = 1000$.

We also conducted simulation studies under the same setting as above using a stratified random splitting approach to make the number of events evenly distributed across S subsets. Specifically, we stratified the full data into 4 strata by the number of events per cluster (0, 1, 2, and 3), then randomly split the data in each stratum into S sets according to the pre-specified partition ratios (n_s/n), and formed each subset by combining one set in each stratum. Results are summarized in Tables 6.7 and 6.8. The performance of our proposed method using the stratified random splitting is similar to that using the simple random splitting approach.

Table 6.5: Performances of $\hat{\boldsymbol{\gamma}}^{dc}$ and $\hat{\boldsymbol{\gamma}}^{full}$ for estimating $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0 = (\theta_0, \boldsymbol{\beta}_0^T)^T$ in frailty models with simple random splitting.

γ_0			3 Subsets			5 Subsets			Full Data
			$\hat{\gamma}^{dc}$			$\hat{\gamma}^{dc}$			$\hat{\gamma}^{full}$
			$\mathbf{W}_{1s}(.)$	$\mathbf{W}_{2s}(.)$	$\mathbf{W}_{3s}(.)$	$\mathbf{W}_{1s}(.)$	$\mathbf{W}_{2s}(.)$	$\mathbf{W}_{3s}(.)$	
		n = 400							
		Time[1] (min)	1.7	1.7	1.7	0.7	0.7	0.7	20.7
θ_0	6/7	Bias (ESE) ($\times 10^{-3}$)	-88.8 (118.6)	-112.4 (124.2)	-2.0 (123.8)	-184.3 (119.6)	-227.8 (132.5)	-7.0 (127.8)	-0.7 (122.3)
		ASE (SE) ($\times 10^{-3}$)	138.3 (14.7)	133.9 (14.5)	149.1 (15.8)	132.6 (14.9)	123.5 (15.0)	155.6 (18.0)	143.2 (14.9)
		CovP (%)	89.5	85.3	97.3	69.8	54.0	96.6	97.3
$\boldsymbol{\beta}_0$	0.8	Bias (ESE) ($\times 10^{-3}$)	-15.7 (58.4)	-22.9 (61.8)	14.1 (63.3)	-36.1 (57.5)	-51.0 (62.9)	25.5 (66.8)	6.5 (61.0)
		ASE (SE) ($\times 10^{-3}$)	63.7 (4.6)	60.7 (4.4)	66.7 (5.2)	63.2 (4.6)	57.2 (4.1)	69.9 (6.5)	64.5 (4.8)
		CovP (%)	95.2	91.5	96.4	92.3	82.9	95.4	96.1
	0.4	Bias (ESE) ($\times 10^{-3}$)	-7.5 (55.7)	-11.8 (57.3)	7.1 (58.7)	-16.7 (53.5)	-23.8 (58.2)	13.7 (59.4)	2.9 (56.9)
		ASE (SE) ($\times 10^{-3}$)	56.5 (3.8)	53.9 (3.6)	58.8 (4.1)	55.9 (3.8)	50.9 (3.5)	60.7 (4.6)	57.1 (3.8)
		CovP (%)	95.4	92.6	95.2	94.3	88.1	95.2	95.1
	0.0	Bias (ESE) ($\times 10^{-3}$)	0.6 (53.6)	0.7 (56.0)	0.8 (56.5)	0.0 (52.1)	-0.1 (57.0)	0.1 (58.2)	0.7 (54.7)
		ASE (SE) ($\times 10^{-3}$)	53.9 (3.3)	51.5 (3.2)	55.8 (3.6)	53.5 (3.3)	48.7 (3.1)	57.5 (3.8)	54.4 (3.4)
		CovP (%)	95.2	92.9	94.8	95.6	90.2	94.7	95.2
		n = 1000							
		Time[1] (min)	23.7	23.5	23.3	5.6	5.6	5.6	—[2]
θ_0	6/7	Bias (ESE) ($\times 10^{-3}$)	-33.7 (76.2)	-45.3 (75.6)	3.2 (79.0)	-69.4 (73.4)	-89.2 (74.7)	3.8 (77.3)	—
		ASE (SE) ($\times 10^{-3}$)	88.9 (5.9)	87.9 (5.9)	91.6 (6.3)	87.7 (5.7)	85.5 (5.7)	93.1 (6.2)	—
		CovP (%)	94.8	94.3	98.2	90.1	83.5	98.3	—
$\boldsymbol{\beta}_0$	0.8	Bias (ESE) ($\times 10^{-3}$)	-4.6 (38.6)	-8.0 (38.9)	7.5 (39.7)	-11.8 (39.1)	-17.9 (40.1)	12.5 (41.6)	—
		ASE (SE) ($\times 10^{-3}$)	40.3 (1.9)	39.5 (1.8)	41.0 (2.0)	40.1 (1.9)	38.6 (1.8)	41.6 (2.1)	—
		CovP (%)	94.6	93.9	94.7	93.5	91.5	94.0	—
	0.4	Bias (ESE) ($\times 10^{-3}$)	-1.9 (34.1)	-3.5 (34.4)	4.3 (34.9)	-5.9 (34.0)	-9.1 (35.3)	6.2 (35.7)	—
		ASE (SE) ($\times 10^{-3}$)	35.8 (1.5)	35.1 (1.5)	36.3 (1.5)	35.6 (1.4)	34.3 (1.4)	36.7 (1.5)	—
		CovP (%)	95.8	95.2	96.4	95.8	93.5	95.4	—
	0.0	Bias (ESE) ($\times 10^{-3}$)	-0.4 (34.0)	-0.5 (34.4)	-0.5 (34.7)	-0.5 (34.1)	-0.4 (34.9)	-0.5 (35.6)	—
		ASE (SE) ($\times 10^{-3}$)	34.2 (1.4)	33.6 (1.4)	34.6 (1.4)	34.1 (1.4)	32.9 (1.3)	35.0 (1.4)	—
		CovP (%)	95.1	94.5	95.0	94.7	93.6	94.2	—

[1] Average computation time in minutes.

[2] Computation in full data is infeasible due to unreasonably long time (i.e., > 420 minutes).

Abbreviations: ASE: asymptotic standard error using the theoretical formula; ESE: empirical standard error; CovP: empirical coverage probability of 95% confidence interval.

Notes: $\boldsymbol{\gamma}_0$ is the true value of $\boldsymbol{\gamma}$; Intra-cluster association is Kendall tau $\kappa = 0.3$; Event rate is 50%; Sample sizes of top panel and bottom panel are 400 and 1000, respectively.

Table 6.6: Performances of $\hat{\boldsymbol{\gamma}}_\rho^{dc}$ and $\hat{\boldsymbol{\gamma}}_\rho^{full}$ for estimating $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0 = (\theta_0, \boldsymbol{\beta}_0^T)^T$ in frailty models with simple random splitting.

$\boldsymbol{\beta}_0$[1]		3 Subsets			5 Subsets			Full Data
		$\hat{\boldsymbol{\gamma}}_\rho^{dc}$			$\hat{\boldsymbol{\gamma}}_\rho^{dc}$			$\hat{\boldsymbol{\gamma}}_\rho^{full}$
		$\mathbf{W}_{1s}(.)$	$\mathbf{W}_{2s}(.)$	$\mathbf{W}_{3s}(.)$	$\mathbf{W}_{1s}(.)$	$\mathbf{W}_{2s}(.)$	$\mathbf{W}_{3s}(.)$	
	n = 400							
	Time[2] (min)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.8	Bias (ESE) ($\times 10^{-3}$)	-37.8 (56.9)	-44.5 (60.8)	-11.1 (61.2)	-57.8 (56.8)	-71.3 (62.6)	-2.9 (64.4)	-15.9 (59.2)
	ASE (SE) ($\times 10^{-3}$)	63.7 (4.6)	60.7 (4.4)	66.7 (5.2)	63.2 (4.6)	57.2 (4.1)	69.9 (6.5)	64.5 (4.8)
	CovP (%)	92.0	86.9	96.7	86.0	73.7	97.1	95.5
	Sn (%)	100	100	100	100	100	100	100
0.4	Bias (ESE) ($\times 10^{-3}$)	-30.1 (57.6)	-34.1 (59.9)	-17.5 (60.5)	-38.9 (55.4)	-45.2 (61.5)	-12.5 (60.3)	-19.8 (58.4)
	ASE (SE) ($\times 10^{-3}$)	56.5 (3.8)	53.9 (3.6)	58.8 (4.1)	55.9 (3.8)	50.9 (3.5)	60.7 (4.6)	57.1 (3.8)
	CovP (%)	91.6	87.2	94.1	89.2	79.8	94.4	92.9
	Sn (%)	100	100	100	100	100	100	100
0.0	Bias (ESE) ($\times 10^{-3}$)	0.4 (15.5)	0.7 (18.1)	0.4 (17.0)	0.5 (14.5)	1.1 (21.9)	0.8 (18.1)	0.5 (15.6)
	Sp (%)	97.8	96.8	97.6	98.0	95.1	97.4	97.8
	n = 1000							
	Time[2] (min)	0.0	0.0	0.0	0.0	0.0	0.0	—[3]
0.8	Bias (ESE) ($\times 10^{-3}$)	-15.3 (37.7)	-18.4 (38.3)	-3.8 (38.9)	-22.8 (38.0)	-28.3 (39.1)	0.4 (40.4)	—
	ASE (SE) ($\times 10^{-3}$)	40.3 (1.9)	39.5 (1.8)	41.0 (2.0)	40.1 (1.9)	38.6 (1.8)	41.6 (2.1)	—
	CovP (%)	93.6	91.6	95.0	92.2	89.2	95.6	—
	Sn (%)	100	100	100	100	100	100	—
0.4	Bias (ESE) ($\times 10^{-3}$)	-12.3 (34.3)	-13.8 (34.8)	-6.6 (35.1)	-16.4 (34.2)	-19.3 (35.6)	-5.0 (35.8)	—
	ASE (SE) ($\times 10^{-3}$)	35.8 (1.5)	35.1 (1.5)	36.3 (1.5)	35.6 (1.4)	34.3 (1.4)	36.7 (1.5)	—
	CovP (%)	94.7	94.5	95.1	94.6	91.2	95.6	—
	Sn (%)	100	100	100	100	100	100	—
0.0	Bias (ESE) ($\times 10^{-3}$)	0.2 (6.6)	0.1 (8.0)	0.1 (7.0)	0.2 (7.1)	0.3 (9.0)	0.2 (8.2)	—
	Sp (%)	99.2	98.9	99.1	99.1	98.6	98.9	—

[1] Regularization only applies to $\boldsymbol{\beta}$ but not θ, thus only $\boldsymbol{\beta}$ is summarized in this table.

[2] Average computation time in minutes.

[3] Regularized estimation in full data is not available due to the absence of $\hat{\boldsymbol{\gamma}}^{full}$.

Abbreviations: ASE: asymptotic standard error using the theoretical formula; ESE: empirical standard error; CovP: empirical coverage probability of 95% confidence interval; Sn: sensitivity; Sp: specificity.

Notes: $\boldsymbol{\gamma}_0$ is the true value of $\boldsymbol{\gamma}$; Intra-cluster association is Kendall tau $\kappa = 0.3$; Event rate is 50%; Sample sizes of top panel and bottom panel are 400 and 1000, respectively.

Table 6.7: Performances of $\hat{\boldsymbol{\gamma}}^{dc}$ and $\hat{\boldsymbol{\gamma}}^{full}$ for estimating $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0 = (\theta_0, \boldsymbol{\beta}_0^T)^T$ in frailty models with stratified random splitting.

$\boldsymbol{\gamma}_0$			3 Subsets			5 Subsets			Full Data
			$\hat{\boldsymbol{\gamma}}^{dc}$			$\hat{\boldsymbol{\gamma}}^{dc}$			$\hat{\boldsymbol{\gamma}}^{full}$
			$\mathbf{W}_{1s}(.)$	$\mathbf{W}_{2s}(.)$	$\mathbf{W}_{3s}(.)$	$\mathbf{W}_{1s}(.)$	$\mathbf{W}_{2s}(.)$	$\mathbf{W}_{3s}(.)$	
		n = 400							
		Time[1] (min)	1.7	1.7	1.7	0.7	0.6	0.6	20.8
θ_0	6/7	Bias (ESE) ($\times 10^{-3}$)	-99.5 (121.3)	-122.8 (126.1)	-9.1 (122.9)	-198.4 (120.2)	-243.6 (131.4)	-16.7 (130.1)	-7.5 (122.1)
		ASE (SE) ($\times 10^{-3}$)	136.5 (14.9)	132.2 (14.9)	147.5 (15.6)	130.2 (14.6)	120.9 (14.6)	154.6 (20.3)	142.0 (15.1)
		CovP (%)	88.6	82.2	96.3	64.5	50.6	96.1	95.7
$\boldsymbol{\beta}_0$	0.8	Bias (ESE) ($\times 10^{-3}$)	-19.0 (60.7)	-27.0 (62.4)	11.2 (64.5)	-39.3 (58.6)	-53.4 (63.6)	23.9 (68.9)	2.2 (62.4)
		ASE (SE) ($\times 10^{-3}$)	63.6 (4.6)	60.6 (4.4)	66.5 (5.4)	63.0 (4.6)	56.9 (4.1)	69.7 (6.7)	64.3 (4.6)
		CovP (%)	95.1	92.1	95.7	92.0	80.6	95.5	95.8
	0.4	Bias (ESE) ($\times 10^{-3}$)	-7.4 (57.5)	-11.3 (59.5)	7.5 (61.5)	-17.5 (55.5)	-23.9 (59.7)	13.3 (62.6)	2.2 (57.9)
		ASE (SE) ($\times 10^{-3}$)	56.4 (3.7)	53.9 (3.6)	58.6 (4.1)	55.9 (3.7)	51.0 (3.5)	60.6 (4.8)	56.9 (3.8)
		CovP (%)	93.7	92.2	94.1	93.7	87.7	93.6	94.5
	0.0	Bias (ESE) ($\times 10^{-3}$)	-0.2 (54.8)	-0.2 (56.4)	0.0 (57.7)	-0.7 (53.4)	-0.6 (59.3)	-1.3 (59.6)	-0.4 (55.5)
		ASE (SE) ($\times 10^{-3}$)	54.0 (3.3)	51.6 (3.2)	55.8 (3.5)	53.4 (3.3)	48.7 (3.1)	57.5 (3.9)	54.5 (3.3)
		CovP (%)	94.7	93.0	94.3	95.2	89.3	94.3	94.8
		n = 1000							
		Time[1] (min)	24.3	24.0	23.8	5.7	5.6	5.6	—[2]
θ_0	6/7	Bias (ESE) ($\times 10^{-3}$)	-42.3 (78.4)	-53.1 (79.1)	-4.1 (80.4)	-74.7 (78.6)	-95.2 (79.9)	0.2 (81.0)	—
		ASE (SE) ($\times 10^{-3}$)	88.0 (6.1)	86.9 (6.1)	90.8 (6.3)	87.1 (6.0)	84.9 (6.0)	92.5 (6.4)	—
		CovP (%)	91.6	89.8	95.3	86.2	79.4	96.0	—
$\boldsymbol{\beta}_0$	0.8	Bias (ESE) ($\times 10^{-3}$)	-4.1 (40.6)	-7.4 (40.9)	7.6 (41.7)	-11.4 (40.1)	-17.5 (41.6)	12.4 (42.0)	—
		ASE (SE) ($\times 10^{-3}$)	40.2 (1.9)	39.4 (1.9)	40.9 (2.0)	40.1 (1.9)	38.5 (1.8)	41.5 (2.1)	—
		CovP (%)	94.7	93.1	93.7	93.4	89.9	93.7	—
	0.4	Bias (ESE) ($\times 10^{-3}$)	-2.7 (36.6)	-4.3 (37.0)	3.2 (37.4)	-7.5 (35.9)	-10.8 (36.2)	4.7 (37.5)	—
		ASE (SE) ($\times 10^{-3}$)	35.7 (1.5)	35.1 (1.5)	36.3 (1.5)	35.6 (1.5)	34.4 (1.4)	36.7 (1.6)	—
		CovP (%)	94.0	93.7	94.0	93.9	92.0	94.0	—
	0.0	Bias (ESE) ($\times 10^{-3}$)	-1.0 (34.6)	-0.8 (35.1)	-1.1 (35.4)	-1.0 (34.0)	-1.0 (35.1)	-1.2 (35.6)	—
		ASE (SE) ($\times 10^{-3}$)	34.1 (1.3)	33.5 (1.3)	34.6 (1.4)	34.0 (1.3)	32.8 (1.3)	34.9 (1.4)	—
		CovP (%)	95.1	94.1	94.8	95.2	93.1	94.5	—

[1] Average computation time in minutes.

[2] Computation in full data is infeasible due to unreasonably long time (i.e., > 420 minutes).

Abbreviations: ASE: asymptotic standard error using the theoretical formula; ESE: empirical standard error; CovP: empirical coverage probability of 95% confidence interval.

Notes: $\boldsymbol{\gamma}_0$ is the true value of $\boldsymbol{\gamma}$; Intra-cluster association is Kendall tau $\kappa = 0.3$; Event rate is 50%; Sample sizes of top panel and bottom panel are 400 and 1000, respectively.

Table 6.8: Performances of $\hat{\boldsymbol{\gamma}}_\rho^{dc}$ and $\hat{\boldsymbol{\gamma}}_\rho^{full}$ for estimating $\boldsymbol{\gamma} = \boldsymbol{\gamma}_0 = (\theta_0, \boldsymbol{\beta}_0^T)^T$ in frailty models with stratified random splitting.

$\boldsymbol{\beta}_0$[1]		3 Subsets			5 Subsets			Full Data
		$\hat{\boldsymbol{\gamma}}_\rho^{dc}$			$\hat{\boldsymbol{\gamma}}_\rho^{dc}$			$\hat{\boldsymbol{\gamma}}_\rho^{full}$
		$\mathbf{W}_{1s}(.)$	$\mathbf{W}_{2s}(.)$	$\mathbf{W}_{3s}(.)$	$\mathbf{W}_{1s}(.)$	$\mathbf{W}_{2s}(.)$	$\mathbf{W}_{3s}(.)$	
	n = 400							
	Time[2] (min)	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.8	Bias (ESE) ($\times 10^{-3}$)	-41.8 (59.4)	-48.6 (61.4)	-14.6 (62.4)	-61.6 (57.3)	-74.4 (62.8)	-6.4 (65.1)	-21.1 (61.4)
	ASE (SE) ($\times 10^{-3}$)	63.6 (4.6)	60.6 (4.4)	66.5 (5.4)	63.0 (4.6)	56.9 (4.1)	69.7 (6.7)	64.3 (4.6)
	CovP (%)	90.9	86.0	95.3	84.1	70.8	97.2	94.9
	Sn (%)	100	100	100	100	100	100	100
0.4	Bias (ESE) ($\times 10^{-3}$)	-31.0 (57.4)	-34.0 (60.5)	-18.1 (61.1)	-40.6 (55.5)	-46.7 (61.9)	-15.4 (61.9)	-21.7 (58.0)
	ASE (SE) ($\times 10^{-3}$)	56.4 (3.7)	53.9 (3.6)	58.6 (4.1)	55.9 (3.7)	51.0 (3.5)	60.6 (4.8)	56.9 (3.8)
	CovP (%)	90.1	86.7	92.2	88.8	80.6	93.8	92.3
	Sn (%)	100	100	100	100	100	100	100
0.0	Bias (ESE) ($\times 10^{-3}$)	0.5 (16.2)	0.7 (20.1)	0.5 (17.3)	0.5 (16.1)	1.1 (24.3)	0.6 (17.7)	0.5 (16.7)
	Sp (%)	97.7	96.3	97.5	97.7	94.1	97.6	97.7
	n = 1000							
	Time[2] (min)	0.0	0.0	0.0	0.0	0.0	0.0	—[3]
0.8	Bias (ESE) ($\times 10^{-3}$)	-15.0 (39.9)	-18.1 (40.2)	-3.8 (40.9)	-22.0 (39.3)	-27.9 (41.0)	0.6 (40.9)	—
	ASE (SE) ($\times 10^{-3}$)	40.2 (1.9)	39.4 (1.9)	40.9 (2.0)	40.1 (1.9)	38.5 (1.8)	41.5 (2.1)	—
	CovP (%)	93.0	91.0	94.8	91.2	86.5	95.1	—
	Sn (%)	100	100	100	100	100	100	—
0.4	Bias (ESE) ($\times 10^{-3}$)	-13.7 (36.4)	-15.3 (37.0)	-8.3 (37.3)	-18.4 (35.8)	-21.8 (36.1)	-7.3 (37.5)	—
	ASE (SE) ($\times 10^{-3}$)	35.7 (1.5)	35.1 (1.5)	36.3 (1.5)	35.6 (1.5)	34.4 (1.4)	36.7 (1.6)	—
	CovP (%)	92.4	90.8	93.0	91.3	88.9	93.1	—
	Sn (%)	100	100	100	100	100	100	—
0.0	Bias (ESE) ($\times 10^{-3}$)	0.0 (3.7)	0.0 (4.3)	0.0 (4.7)	-0.1 (3.5)	0.0 (6.3)	-0.1 (4.6)	—
	Sp (%)	99.6	99.4	99.4	99.6	99.1	99.5	—

[1] Regularization only applies to $\boldsymbol{\beta}$ but not θ, thus only $\boldsymbol{\beta}$ is summarized in this table.

[2] Average computation time in minutes.

[3] Regularized estimation in full data is not available due to the absence of $\hat{\boldsymbol{\gamma}}^{full}$.

Abbreviations: ASE: asymptotic standard error using the theoretical formula; ESE: empirical standard error; CovP: empirical coverage probability of 95% confidence interval; Sn: sensitivity; Sp: specificity.

Notes: $\boldsymbol{\gamma}_0$ is the true value of $\boldsymbol{\gamma}$; Intra-cluster association is Kendall tau $\kappa = 0.3$; Event rate is 50%; Sample sizes of top panel and bottom panel are 400 and 1000, respectively.

6.2.3 Multistate Survival Analysis

We performed a simulation study of $n = 100,000$ independent subjects to evaluate the asymptotic equivalence of the divide-and-combine estimator $\hat{\mathbf{P}}^{dc}(u, t|\boldsymbol{z}_0)$ and the full-data estimator $\hat{\mathbf{P}}^{full}(u, t|\boldsymbol{z}_0)$.

We simulated Markov multistate processes for our proposed five-state model, as illustrated in Figure 5.1, by simulating multivariate ordered failure times. Hence, for each of the n independent subjects, we simulated failure times T_{12} for transition $1 \to 2$, T_{13} for transition $1 \to 3$, T_{15} for transition $1 \to 5$, T_{15}^2 for transition $1 \to 2 \to 5$, T_{15}^3 for transition $1 \to 3 \to 5$, T_{14}^2 for transition $1 \to 2 \to 4$, T_{14}^3 for transition $1 \to 3 \to 4$, T_{15}^{24} for transition $1 \to 2 \to 4 \to 5$, and T_{15}^{34} for transitions $1 \to 3 \to 4 \to 5$. Note that in Markov models, which are of interest in this dissertation, all times are measured from the same time origin, i.e., the start of the follow-up period. It is also worth explaining that, for example, T_{12} is the failure time of a subject from state 1 to state 2, whereas T_{15}^2 is the failure time of a subject from state 1 to state 5 through state 2. We describe the procedures to simulate multivariate ordered failure times below. For the ease of notation, we assume the hazard function for transition $j \to k$ follows a Weibull distribution and takes the form $\lambda_{jk}(t) = a_{jk}t^{b_{jk}}$ ($j, k = 1, 2, \ldots, 5$ and $j \neq k$). This hazard function can be easily incorporated with covariates and modified to the conventional form by letting $a_{jk} = \xi_{jk}h_{jk}^{\xi_{jk}}e^{\boldsymbol{\beta}^T\boldsymbol{Z}_{jk}}$ and $b_{jk} = \kappa_{jk} - 1$ ($j, k = 1, 2, \ldots, 5$ and $j \neq k$).

Consider transition $1 \to 2$ with hazard $\lambda_{12}(t)$. The survival function is expressed by $S_{12}(t) = e^{-\Lambda_{12}(t)}$, where

$$\Lambda_{12}(t) = \int_0^t \lambda_{12}(u)du = \int_0^t a_{12}u^{b_{12}}du = \frac{a_{12}}{b_{12}+1}t^{b_{12}+1}. \tag{6.6}$$

It is known that $S_{12}(T_{12}) = U_{12} \sim$ Uniform(0, 1), then by using (6.6), the failure time for

transition $1 \to 2$ is given by

$$T_{12} = \left(-\frac{b_{12}+1}{a_{12}} log U_{12} \right)^{1/(b_{12}+1)}. \tag{6.7}$$

Similarly, we have

$$T_{13} = \left(-\frac{b_{13}+1}{a_{13}} log U_{13} \right)^{1/(b_{13}+1)}, \tag{6.8}$$

and

$$T_{15} = \left(-\frac{b_{15}+1}{a_{15}} log U_{15} \right)^{1/(b_{15}+1)}. \tag{6.9}$$

Consider another transition $1 \to 2 \to 5$ with hazard $\lambda_{25}(t)$ for transition $2 \to 5$. The survival function for the subject having visited state 2 at time T_{12} and being at risk of state 5 is expressed by $S_{25}(t|T_{12}) = e^{-\{\Lambda_{25}(t)-\Lambda_{25}(T_{12})\}}$, where

$$\Lambda_{25}(t) - \Lambda_{25}(T_{12}) = \int_{T_{12}}^{t} a_{25} u^{b_{25}} du = \frac{a_{25}}{b_{25}+1} t^{b_{25}+1} - \frac{a_{25}}{b_{25}+1} (T_{12})^{b_{25}+1}. \tag{6.10}$$

Given that $S_{25}(T_{15}^2|T_{12}) = U_{25} \sim$ Uniform(0, 1), by (6.10), the failure time for transition $1 \to 2 \to 5$ is given by

$$T_{15}^2 = \left(-\frac{b_{25}+1}{a_{25}} log U_{25} + (T_{12})^{b_{25}+1} \right)^{1/(b_{25}+1)}. \tag{6.11}$$

Similarly, we have

$$T_{15}^3 = \left(-\frac{b_{35}+1}{a_{35}} log U_{35} + (T_{13})^{b_{35}+1} \right)^{1/(b_{35}+1)}, \tag{6.12}$$

$$T_{14}^2 = \left(-\frac{b_{24}+1}{a_{24}} log U_{24} + (T_{12})^{b_{24}+1} \right)^{1/(b_{24}+1)}, \tag{6.13}$$

and

$$T_{14}^3 = \left(-\frac{b_{34}+1}{a_{34}} log U_{34} + (T_{13})^{b_{34}+1} \right)^{1/(b_{34}+1)}. \tag{6.14}$$

Next we consider the most complicated transition $1 \rightarrow 2 \rightarrow 4 \rightarrow 5$ with hazard $\lambda_{45}(t)$ for transition $4 \rightarrow 5$. The survival function for the subject having visited state 2 at time T_{12} and state 4 at time T_{14}^2 and being at risk of state 5 is expressed by $S_{45}(t|T_{14}^2) = e^{-\left\{\Lambda_{45}(t)-\Lambda_{45}(T_{14}^2)\right\}}$, where

$$\Lambda_{45}(t) - \Lambda_{45}(T_{14}^2) = \int_{T_{14}^2}^{t} a_{45}u^{b_{45}}du = \frac{a_{45}}{b_{45}+1}t^{b_{45}+1} - \frac{a_{45}}{b_{45}+1}(T_{14}^2)^{b_{45}+1}. \tag{6.15}$$

Given that $S_{45}(T_{15}^{24}|T_{14}^2) = U_{45} \sim$ Uniform(0, 1), by (6.15), the failure time for transition $1 \rightarrow 2 \rightarrow 4 \rightarrow 5$ is given by

$$T_{15}^{24} = \left(-\frac{b_{45}+1}{a_{45}}logU_{45} + (T_{14}^2)^{b_{45}+1}\right)^{1/(b_{45}+1)}. \tag{6.16}$$

Similarly, we have

$$T_{15}^{34} = \left(-\frac{b_{45}+1}{a_{45}}logU_{45} + (T_{14}^3)^{b_{45}+1}\right)^{1/(b_{45}+1)}. \tag{6.17}$$

Of note, U_{45} appears in both T_{15}^{24} and T_{15}^{34} but this does not impact the independent generation of failure times in the same subject because these two "potential" failure times T_{15}^{24} and T_{15}^{34} cannot happen at the same time.

We adopted a censoring mechanism in which subjects have the same probability to be censored at any time. We also chose a threshold τ that is the end of the follow-up period to stop the observation. Let $U_{cen} \sim$ Uniform(0, α) where $\alpha > \tau$, then the censoring time $C = \min(U_{cen}, \tau)$.

In multistate models, observed times T_q and censoring indicators δ_q ($q = 1, 2, \ldots, Q$) for each of $Q = 5$ states are not straightforward as in the conventional survival analysis. Thus we explicitly define the notation in the following. State 1 is the start of the entire multistate processes and thus $T_1 = 0$. For state 2, $T_2 = \min(T_{12}, T_{13}, T_{15}, C)$ and $\delta_2 = I(T_2 = T_{12})$. For state 3, $T_3 = \min(T_{12}, T_{13}, T_{15}, C)$ and $\delta_3 = I(T_3 = T_{13})$. For state 4, if

$\delta_2 = 1$ and $\delta_3 = 0$, then $T_4 = \min(T_{14}^2, T_{15}^2, C)$ and $\delta_4 = I(T_4 = T_{14}^2)$; if $\delta_2 = 0$ and $\delta_3 = 1$, then $T_4 = \min(T_{14}^3, T_{15}^3, C)$ and $\delta_4 = I(T_4 = T_{14}^3)$; if $\delta_2 = 0$ and $\delta_3 = 0$, then $T_4 = 0$ and $\delta_4 = 0$. For state 5, if $\delta_2 = 0$ and $\delta_3 = 0$, $T_5 = \min(T_{15}, C)$ and $\delta_5 = I(T_5 = T_{15})$; if $\delta_2 = 1$ and $\delta_3 = 0$ and $\delta_4 = 0$, $T_5 = \min(T_{15}^2, C)$ and $\delta_5 = I(T_5 = T_{15}^2)$; if $\delta_2 = 1$ and $\delta_3 = 0$ and $\delta_4 = 1$, $T_5 = \min(T_{15}^{24}, C)$ and $\delta_5 = I(T_5 = T_{15}^{24})$; if $\delta_2 = 0$ and $\delta_3 = 1$ and $\delta_4 = 0$, $T_5 = \min(T_{15}^3, C)$ and $\delta_5 = I(T_5 = T_{15}^3)$; if $\delta_2 = 0$ and $\delta_3 = 1$ and $\delta_4 = 1$, $T_5 = \min(T_{15}^{34}, C)$ and $\delta_5 = I(T_5 = T_{15}^{34})$.

In our simulation study, multivariate ordered failure times for n independent subjects were generated by the following steps.

(S1) Generate $U_{i,12}$, $U_{i,13}$, $U_{i,15}$, $U_{i,24}$, $U_{i,25}$, $U_{i,34}$, $U_{i,35}$, and $U_{i,45}$ from Uniform(0, 1);

(S2) Generate $U_{i,cen}$ from Uniform(0, $\alpha = 2.3$) and obtain the censoring times $C_i = \min(U_{i,cen}, \tau = 1.5)$ — parameters are chosen such that 25% of the subjects are left in state 1 at time τ, i.e., $P_{11}(0, \tau) = 0.25$;

(S3) Based on the procedures described above, solve for the observed times T_{i1}, T_{i2}, T_{i3}, T_{i4}, T_{i5}, and censoring indicators δ_{i1}, δ_{i2}, δ_{i3}, δ_{i4}, δ_{i5}, for the i^{th} subject, $i = 1, 2, \ldots, n$.

In the Andersen-type Cox Markov model (5.3) considered in this simulation study, we set $h_{12} = 0.38$, $h_{13} = 0.08$, $h_{15} = 0.32$, $h_{24} = 0.06$, $h_{25} = 0.31$, $h_{34} = 0.30$, $h_{35} = 0.29$, $h_{45} = 0.14$, and $\xi_{12} = \xi_{13} = \xi_{15} = \xi_{24} = \xi_{25} = \xi_{34} = \xi_{35} = \xi_{45} = 1$, to differentiate baseline hazards of different transitions. To allow for varying covariate effects for different transitions, in the i^{th} subject, for transitions $h \to j$, $h, j = 1, 2, \ldots, 5$ and $h \neq j$, data were simulated according to transition-specific regression parameters: $\boldsymbol{\beta}_{12}^* = (\mathbf{0.9}_4, \mathbf{0.8}_4, \mathbf{0}_{92})^T$, $\boldsymbol{\beta}_{13}^* = (\mathbf{0.9}_4, \mathbf{0.7}_4, \mathbf{0}_{92})^T$, $\boldsymbol{\beta}_{15}^* = (\mathbf{0.9}_4, \mathbf{0.6}_4, \mathbf{0}_{92})^T$, $\boldsymbol{\beta}_{24}^* = (\mathbf{0.9}_4, \mathbf{0.5}_4, \mathbf{0}_{92})^T$, $\boldsymbol{\beta}_{25}^* = (\mathbf{0.9}_4, \mathbf{0.4}_4, \mathbf{0}_{92})^T$, $\boldsymbol{\beta}_{34}^* = (\mathbf{0.9}_4, \mathbf{0.3}_4, \mathbf{0}_{92})^T$, $\boldsymbol{\beta}_{35}^* = (\mathbf{0.9}_4, \mathbf{0.2}_4, \mathbf{0}_{92})^T$, and $\boldsymbol{\beta}_{45}^* = (\mathbf{0.9}_4, \mathbf{0.1}_4, \mathbf{0}_{92})^T$, corresponding to covariate $\mathbf{C}_{i,hj} = (C_{i,hj1}, C_{i,hj2}, \ldots, C_{i,hj100})^T$ ($h, j = 1, 2, \ldots, 5$ and $h \neq j$) generated from a multivariate normal distribution with standard normal marginals

and an equal correlation of 0.2. Each of $\boldsymbol{\beta}_{hj}^{*}, h, j = 1, 2, \ldots, 5$ and $h \neq j$, contains 8 nonzeros and 92 zeros. As a result, model (5.3) used in this dissertation can accommodate type-specific regression parameters by using a common regression parameter vector $\boldsymbol{\beta} = (\mathbf{0.9}_4, \mathbf{0.8}_4, \mathbf{0.7}_4, \mathbf{0.6}_4, \mathbf{0.5}_4, \mathbf{0.4}_4, \mathbf{0.3}_4, \mathbf{0.2}_4, \mathbf{0.1}_4, \mathbf{0}_{92})^T$ and a design matrix $\boldsymbol{Z}_i = (\boldsymbol{Z}_{i,12}, \boldsymbol{Z}_{i,13}, \boldsymbol{Z}_{i,15}, \boldsymbol{Z}_{i,24}, \boldsymbol{Z}_{i,25}, \boldsymbol{Z}_{i,34}, \boldsymbol{Z}_{i,35}, \boldsymbol{Z}_{i,45})^T$, where

$$\begin{pmatrix} \boldsymbol{Z}_{i,12}^T \\ \boldsymbol{Z}_{i,13}^T \\ \boldsymbol{Z}_{i,15}^T \\ \boldsymbol{Z}_{i,24}^T \\ \boldsymbol{Z}_{i,25}^T \\ \boldsymbol{Z}_{i,34}^T \\ \boldsymbol{Z}_{i,35}^T \\ \boldsymbol{Z}_{i,45}^T \end{pmatrix} = \begin{pmatrix} C_{i,121}, \cdots, C_{i,124}, \cdots\cdots\cdots\cdots\cdots C_{i,129}, \cdots, C_{i,12100} \\ C_{i,131}, \cdots, C_{i,134}, \cdots\cdots\cdots\cdots\cdots C_{i,139}, \cdots, C_{i,13100} \\ C_{i,151}, \cdots, C_{i,154}, \cdots\cdots\cdots\cdots\cdots C_{i,159}, \cdots, C_{i,15100} \\ C_{i,241}, \cdots, C_{i,244}, \cdots\cdots\cdots\cdots\cdots C_{i,249}, \cdots, C_{i,24100} \\ C_{i,251}, \cdots, C_{i,254}, \cdots\cdots\cdots\cdots\cdots C_{i,259}, \cdots, C_{i,25100} \\ C_{i,341}, \cdots, C_{i,344}, \cdots\cdots\cdots\cdots\cdots C_{i,349}, \cdots, C_{i,34100} \\ C_{i,351}, \cdots, C_{i,354}, \cdots\cdots\cdots\cdots\cdots C_{i,359}, \cdots, C_{i,35100} \\ C_{i,451}, \cdots, C_{i,454}, \underbrace{\cdots\cdots\cdots\cdots\cdots}_{\text{Transition-Specific Portion}} C_{i,459}, \cdots, C_{i,45100} \end{pmatrix}.$$

The transition-specific portion in the design matrix is given by

$$\begin{pmatrix} C_{i,125}, C_{i,126}, C_{i,127}, C_{i,128}, \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4 \\ \mathbf{0}_4, C_{i,135}, C_{i,136}, C_{i,137}, C_{i,138}, \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4 \\ \mathbf{0}_4, \mathbf{0}_4, C_{i,155}, C_{i,156}, C_{i,157}, C_{i,158}, \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4 \\ \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4, C_{i,245}, C_{i,246}, C_{i,247}, C_{i,248}, \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4 \\ \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4, C_{i,255}, C_{i,256}, C_{i,257}, C_{i,258}, \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4 \\ \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4, C_{i,345}, C_{i,346}, C_{i,347}, C_{i,348}, \mathbf{0}_4, \mathbf{0}_4 \\ \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4, C_{i,355}, C_{i,356}, C_{i,357}, C_{i,358}, \mathbf{0}_4 \\ \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4, \mathbf{0}_4, C_{i,455}, C_{i,456}, C_{i,457}, C_{i,458} \end{pmatrix}.$$

Note that $\boldsymbol{\beta}^T \boldsymbol{Z}_{i,12} = \boldsymbol{\beta}_{12}^{*T} \boldsymbol{C}_{i,12}$, $\boldsymbol{\beta}^T \boldsymbol{Z}_{i,13} = \boldsymbol{\beta}_{13}^{*T} \boldsymbol{C}_{i,13}$, $\boldsymbol{\beta}^T \boldsymbol{Z}_{i,15} = \boldsymbol{\beta}_{15}^{*T} \boldsymbol{C}_{i,15}$, $\boldsymbol{\beta}^T \boldsymbol{Z}_{i,24} = \boldsymbol{\beta}_{24}^{*T} \boldsymbol{C}_{i,24}$, $\boldsymbol{\beta}^T \boldsymbol{Z}_{i,25} = \boldsymbol{\beta}_{25}^{*T} \boldsymbol{C}_{i,25}$, $\boldsymbol{\beta}^T \boldsymbol{Z}_{i,34} = \boldsymbol{\beta}_{34}^{*T} \boldsymbol{C}_{i,34}$, $\boldsymbol{\beta}^T \boldsymbol{Z}_{i,35} = \boldsymbol{\beta}_{35}^{*T} \boldsymbol{C}_{i,35}$, and $\boldsymbol{\beta}^T \boldsymbol{Z}_{i,45} = \boldsymbol{\beta}_{45}^{*T} \boldsymbol{C}_{i,45}$. This illustrates that using a common $\boldsymbol{\beta}$ notation for the transition-specific proportional hazards models does not preclude transition-specific regression parameters.

In the divide-and-combine analysis, to balance the "final states" of subjects at the end

of follow-up period, we used a stratified random splitting method to partition the full data. Specifically, we stratified the full data into 5 strata by the "final states" that subjects occupied when $t = \tau$ (1, 2, 3, 4, and 5), then randomly split the data in each stratum into S sets according to the pre-specified partition ratios (n_s/n), and formed each subset by combining one set in each stratum. In this simulation study, three sets of partition ratios were used: $n_s/n = (\mathbf{2/8}_3, \mathbf{1/8}_2)$, $n_s/n = (\mathbf{2/15}_5, \mathbf{1/15}_5)$, and $n_s/n = (\mathbf{2/30}_{10}, \mathbf{1/30}_{10})$, corresponding to $S = 5$, 10, and 20 subsets, respectively. In each configuration, we ran simulations 500 times. All simulations were carried out on a Linux cluster (CPU specification: Intel® Xeon ® E5-2680 v4 @2.40GHz) via parallel computing (S cores) with one subset allocated to one core. The average computation time was calculated based on all 500 simulation runs. All statistical analyses regarding fitting transition-specific proportional hazards models were performed using R package, `survival` (Therneau, 2020), and estimation of cumulative hazards and prediction of transition probabilities were conducted using R package, `mstate` (de Wreede et al., 2011).

Simulation Results

The performance of the divide-and-combine estimator $\hat{\mathbf{P}}^{dc}(u, t|\boldsymbol{z}_0)$ was assessed, compared to the full-data estimator $\hat{\mathbf{P}}^{full}(u, t|\boldsymbol{z}_0)$ in terms of average computation time (Time), mean of biasedness (Bias), empirical standard error (ESE), mean asymptotic standard error (ASE) using the theoretical formula and the associated standard error, and empirical coverage probability (CovP) of the 95% Wald-type confidence interval. Note that $\boldsymbol{z}_0$ is the covariate vector of a future subject whose all covariates are at the mean levels (continuous variables). In this dissertation, we are particularly interested in predicting probabilities of advancing to worse conditions at a future time t for an average subject given that they have only been hospitalized once due to heart failure at time u, i.e., $P_{11}(u, t|\boldsymbol{z}_0)$, $P_{12}(u, t|\boldsymbol{z}_0)$, $P_{13}(u, t|\boldsymbol{z}_0)$, $P_{14}(u, t|\boldsymbol{z}_0)$, and $P_{15}(u, t|\boldsymbol{z}_0)$. Comparisons between $\hat{\mathbf{P}}^{dc}(u, t|\boldsymbol{z}_0)$ and $\hat{\mathbf{P}}^{full}(u, t|\boldsymbol{z}_0)$ for predicting these probabilities are summarized in

Table 6.9. The true transition probabilities $\mathbf{P}_0(u, t|\boldsymbol{z}_0)$ can be calculated based on the Kolmogorov forward equation in (2.39). See the detailed derivations for these transition probabilities of interest in Appendix A.12. The computation times of $\hat{\mathbf{P}}^{dc}(u, t|\boldsymbol{z}_0)$ were 57 minutes, 58 minutes, and 55 minutes when the full data were divided into $S = 5$, 10, and 20 subsets in the divide-and-combine analysis, whereas the computing of $\hat{\mathbf{P}}^{full}(u, t|\boldsymbol{z}_0)$ took a longer time of 95 minutes. The computational savings in multistate survival analysis using the divide-and-combine approach were not as significant as we observed in multivariate survival analysis (multivariate: divide-and-combine analysis vs. full-data analysis is $\sim 1 : 10$; multistate: divide-and-combine analysis vs. full-data analysis is $\sim 6 : 10$; when $S = 5$). As discussed in Section 5.4, the less computational savings may be caused by the fact that the divide-and-combine in multistate survival analysis was only used for estimating the cumulative hazards, which is only an intermediate step for predicting the transition probabilities.

In terms of biasedness, ESE, ASE, and CovP, the performance of $\hat{\mathbf{P}}^{dc}(u, t|\boldsymbol{z}_0)$ for different combinations of u and t is generally close to that of $\hat{\mathbf{P}}^{full}(u, t|\boldsymbol{z}_0)$. The bias of $\hat{\mathbf{P}}^{dc}(u, t|\boldsymbol{z}_0)$ is generally small; ESE and ASE are close to each other; and CovP is close to the nominal 95% level. However, when the number of subsets (S) is large or the prediction is made at an earlier time u, $\hat{\mathbf{P}}^{dc}(u, t|\boldsymbol{z}_0)$ has a relatively large bias and thus a poor coverage probability. We conjecture that the unsatisfactory performance when S is large might be because the homogeneity assumptions which are the key to establishing the large sample property of $\hat{\mathbf{P}}^{dc}(u, t|\boldsymbol{z}_0)$ may be violated to some extent in practice as the number of subsets increases. The issue with the prediction time u needs more research in the future.

Table 6.9: Performances of $\hat{\mathbf{P}}^{dc}(u,t|\boldsymbol{z}_0)$ and $\hat{\mathbf{P}}^{full}(u,t|\boldsymbol{z}_0)$ for predicting $\mathbf{P}(u,t|\boldsymbol{z}_0) = \mathbf{P}_0(u,t|\boldsymbol{z}_0)$ in multistate models with stratified random splitting.

$\mathbf{P}_0$[1]			5 Subsets			10 Subsets			20 Subsets			Full Data		
			$\hat{\mathbf{P}}^{dc}$			$\hat{\mathbf{P}}^{dc}$			$\hat{\mathbf{P}}^{dc}$			$\hat{\mathbf{P}}^{full}$		
	u	t	Bias (ESE) ($\times 10^{-3}$)	ASE (SE) ($\times 10^{-3}$)	CovP (%)	Bias (ESE) ($\times 10^{-3}$)	ASE (SE) ($\times 10^{-3}$)	CovP (%)	Bias (ESE) ($\times 10^{-3}$)	ASE (SE) ($\times 10^{-3}$)	CovP (%)	Bias (ESE) ($\times 10^{-3}$)	ASE (SE) ($\times 10^{-3}$)	CovP (%)
P_{11}	0	$\tau/2$	0.4 (3.3)	2.8 (0.0)	89.2	-2.5 (3.3)	2.9 (0.0)	81.8	-5.5 (3.3)	2.9 (0.0)	52.4	0.0 (3.3)	2.8 (0.0)	89.6
	0	τ	-0.3 (4.3)	3.8 (0.0)	91.0	-2.1 (4.2)	3.8 (0.0)	87.0	-4.2 (4.4)	3.8 (0.0)	76.0	-0.2 (4.3)	3.7 (0.0)	90.8
	$\tau/2$	$3\tau/4$	-0.5 (5.0)	4.7 (0.1)	93.4	-0.7 (4.6)	4.7 (0.1)	95.4	-1.5 (4.9)	4.8 (0.1)	94.0	-0.2 (4.9)	4.7 (0.1)	93.4
	$\tau/2$	τ	-0.9 (6.6)	6.2 (0.1)	93.4	-1.3 (6.1)	6.2 (0.1)	95.0	-2.0 (6.5)	6.2 (0.1)	93.2	-0.5 (6.6)	6.1 (0.1)	92.8
P_{12}	0	$\tau/2$	0.0 (2.5)	2.0 (0.0)	88.2	0.3 (2.6)	2.0 (0.0)	88.0	0.7 (2.5)	2.0 (0.0)	88.2	0.0 (2.5)	2.0 (0.0)	88.2
	0	τ	0.3 (4.0)	3.2 (0.0)	88.4	-1.0 (3.9)	3.2 (0.0)	88.0	-1.7 (3.8)	3.2 (0.0)	86.0	0.1 (4.0)	3.2 (0.0)	88.8
	$\tau/2$	$3\tau/4$	0.4 (3.7)	3.5 (0.1)	93.6	0.0 (3.7)	3.5 (0.1)	94.4	0.0 (3.7)	3.6 (0.1)	93.8	0.2 (3.7)	3.5 (0.1)	93.4
	$\tau/2$	τ	0.5 (5.3)	4.8 (0.1)	91.6	-0.2 (5.0)	4.8 (0.1)	94.2	-0.3 (4.9)	4.8 (0.1)	94.0	0.3 (5.3)	4.8 (0.1)	91.6
P_{13}	0	$\tau/2$	-0.2 (0.7)	0.6 (0.0)	89.0	0.1 (0.7)	0.6 (0.0)	90.6	0.4 (0.7)	0.6 (0.0)	88.0	-0.1 (0.7)	0.6 (0.0)	89.6
	0	τ	-0.1 (1.0)	0.9 (0.0)	91.2	-0.2 (1.0)	0.9 (0.0)	91.0	-0.2 (1.0)	0.9 (0.0)	90.0	0.0 (1.0)	0.9 (0.0)	91.8
	$\tau/2$	$3\tau/4$	-0.1 (0.9)	1.0 (0.0)	95.8	0.1 (0.9)	1.0 (0.0)	96.8	0.4 (1.0)	1.0 (0.0)	93.8	0.0 (0.9)	1.0 (0.0)	96.0
	$\tau/2$	τ	-0.1 (1.3)	1.3 (0.0)	95.4	0.1 (1.3)	1.3 (0.0)	95.8	0.3 (1.4)	1.3 (0.0)	92.6	0.0 (1.3)	1.3 (0.0)	95.8
P_{14}	0	$\tau/2$	0.0 (0.2)	0.2 (0.0)	91.4	0.3 (0.2)	0.2 (0.0)	75.4	0.6 (0.3)	0.2 (0.0)	27.6	0.0 (0.2)	0.2 (0.0)	92.0
	0	τ	-0.1 (0.7)	0.6 (0.0)	92.2	0.5 (0.7)	0.6 (0.0)	85.4	1.2 (0.8)	0.6 (0.0)	55.4	0.0 (0.7)	0.6 (0.0)	92.2
	$\tau/2$	$3\tau/4$	0.0 (0.1)	0.1 (0.0)	98.4	0.1 (0.1)	0.1 (0.0)	94.8	0.1 (0.1)	0.2 (0.0)	87.0	0.0 (0.1)	0.1 (0.0)	98.2
	$\tau/2$	τ	0.0 (0.4)	0.4 (0.0)	94.2	0.2 (0.4)	0.4 (0.0)	94.0	0.4 (0.4)	0.4 (0.0)	86.4	0.0 (0.4)	0.4 (0.0)	94.4
P_{15}	0	$\tau/2$	-0.2 (2.3)	2.0 (0.0)	91.4	1.8 (2.5)	2.0 (0.0)	78.0	3.8 (2.4)	2.0 (0.0)	54.6	0.0 (2.3)	2.0 (0.0)	91.8
	0	τ	0.2 (3.8)	3.2 (0.0)	90.6	2.7 (4.1)	3.3 (0.0)	80.8	4.9 (3.9)	3.3 (0.0)	66.8	0.2 (3.8)	3.2 (0.0)	90.0
	$\tau/2$	$3\tau/4$	0.1 (3.3)	3.2 (0.1)	94.0	0.6 (3.3)	3.2 (0.1)	93.8	1.0 (3.3)	3.2 (0.1)	92.8	0.0 (3.2)	3.1 (0.1)	94.8
	$\tau/2$	τ	0.5 (4.7)	4.5 (0.1)	93.4	1.2 (4.8)	4.5 (0.1)	92.6	1.6 (4.7)	4.5 (0.1)	93.4	0.2 (4.7)	4.5 (0.1)	93.6

[1] Transition probabilities starting from state 1 are of primary interest and thus only these probabilities are summarized in this table.

Abbreviations: ASE: asymptotic standard error using the theoretical formula; ESE: empirical standard error; SE: standard error of ASE; CovP: empirical coverage probability of 95% confidence interval.

Notes: $\mathbf{P}_0(u,t|\boldsymbol{z}_0)$ is the true value of $\mathbf{P}(u,t|\boldsymbol{z}_0)$; Transition probabilities are predicted at time u for time t, in which $\tau = 1.5$ is the end of the follow-up period; Sample size is $100,000$; The rate of staying in state 1 at the end of the follow-up period is 25%.

6.3 MIDAS Data Analyses

6.3.1 Marginal Models in Multivariate Survival Analysis

MIDAS data was used to illustrate the proposed divide-and-combine approach in the marginal proportional hazards model. We analyzed data on 2,117,763 patients who had at least one hospitalization record due to cardiovascular diseases during 1995 to 2014 in the MIDAS database. Times from index cardiovascular hospitalization to the subsequent hospitalizations due to myocardial infarction (MI), heart failure (HF), and stroke were calculated and censored by death or the end of study (December 31, 2014), whichever occurred first. The median follow-up was 11.3 years; a total of 427,288 (20%) patients were re-admitted to hospitals due to one of MI, HF, or stroke, whereas 154,670 (7%) patients and 22,413 (1%) patients were re-admitted to hospitals due to occurrences of two or all of MI, HF or stroke, after the index hospitalizations. Covariates of interest include patient clinical characteristics at the index admissions, including age, gender, race, length of stay, comorbidity conditions, medical procedures received, and health insurance payer. Patient ZIP code level socioeconomic information, including general health status, health care coverage, education attainment, poverty status, median household income, blood cholesterol screening history, high cholesterol diagnosis, high blood pressure diagnosis, angina diagnosis, stroke diagnosis, obesity diagnosis, is also considered. In addition, some hospital characteristics of index admissions, i.e., teaching status, location, and size are included. Moreover, we modeled type-specific and gender-specific regression coefficients for selected covariates (i.e., age, length of stay, general health status, health care coverage, education attainment, poverty status, median household income, blood cholesterol screening history, high cholesterol diagnosis, high blood pressure diagnosis, angina diagnosis, obesity diagnosis, hospital size) by using interaction terms of these covariates with failure types and genders. In order to test the performance of variable selection of our proposed method, we added ten unrelated and randomly generated noise variables into the data set. The final data set ended up with

6,353,289 rows of records with 121 covariates for data analysis.

After several unsuccessful attempts to analyze the entire data set all at once in a single computer (CPU specification: Intel® Core™ i7-4790 @3.60GHz), we used our proposed divide-and-combine approach (with simple random splitting) and randomly partitioned full data set into $S = 20$ subsets with a partition ratio of $n_s/n = (\mathbf{2}/\mathbf{30}_{10}, \mathbf{1}/\mathbf{30}_{10})$. The data analysis was finished in 120 minutes on a Linux cluster via parallel computing (20 cores) with one subset allocated to one core (CPU specification: Intel® Xeon® E5-2680 v4 @2.40GHz). We show the estimated regularized regression coefficients ($\hat{\boldsymbol{\beta}}_{\rho}^{dc}$ using $\mathbf{W}_{1s}(\hat{\boldsymbol{\beta}}_s) = \mathcal{I}_s(\hat{\boldsymbol{\beta}}_s)$) and associated 95% confidence intervals (CIs) in Figure 6.1. To explore whether or not one may gain advantages by using a large data set over a smaller subset (a common practical compromise between large data sets and limited computing capability), we applied the same marginal regression analysis by randomly selecting a subset of 140,766 patients, with $\hat{\boldsymbol{\beta}}_{\rho}^{full}$ and its 95% CIs obtained in this subset analysis summarized in Figure 6.1, too. Among the 121 covariates, 27 covariates identified as significant risk factors by the divide-and-combine analysis are not selected or determined as insignificant by the subset analysis. For example, medical procedure cardiac ablation shows a "significant" protective effect on lowering the occurrence of hospitalization due to MI, HF, or stroke in the divide-and-combine analysis ($\hat{\beta}_{\rho}^{dc}$= -0.103, 95% CI: -0.200, -0.006), and its estimate $\hat{\beta}_{\rho}^{full}$ is -0.161 (95% CI: -0.546, 0.224) in the subset analysis. In addition, 10 random noise variables are estimated as zero by both divide-and-combine and subset analyses, indicating the selection consistency of our proposed regularization approach. Therefore, bigger data seem to provide more reliable estimates and enable better selection of plausible covariates than the smaller data set, especially for those with weak to moderate effects. Nevertheless, one should note that the present MIDAS analysis is only an illustration of our proposed method. These estimates may not be appropriate for clinical guidance.

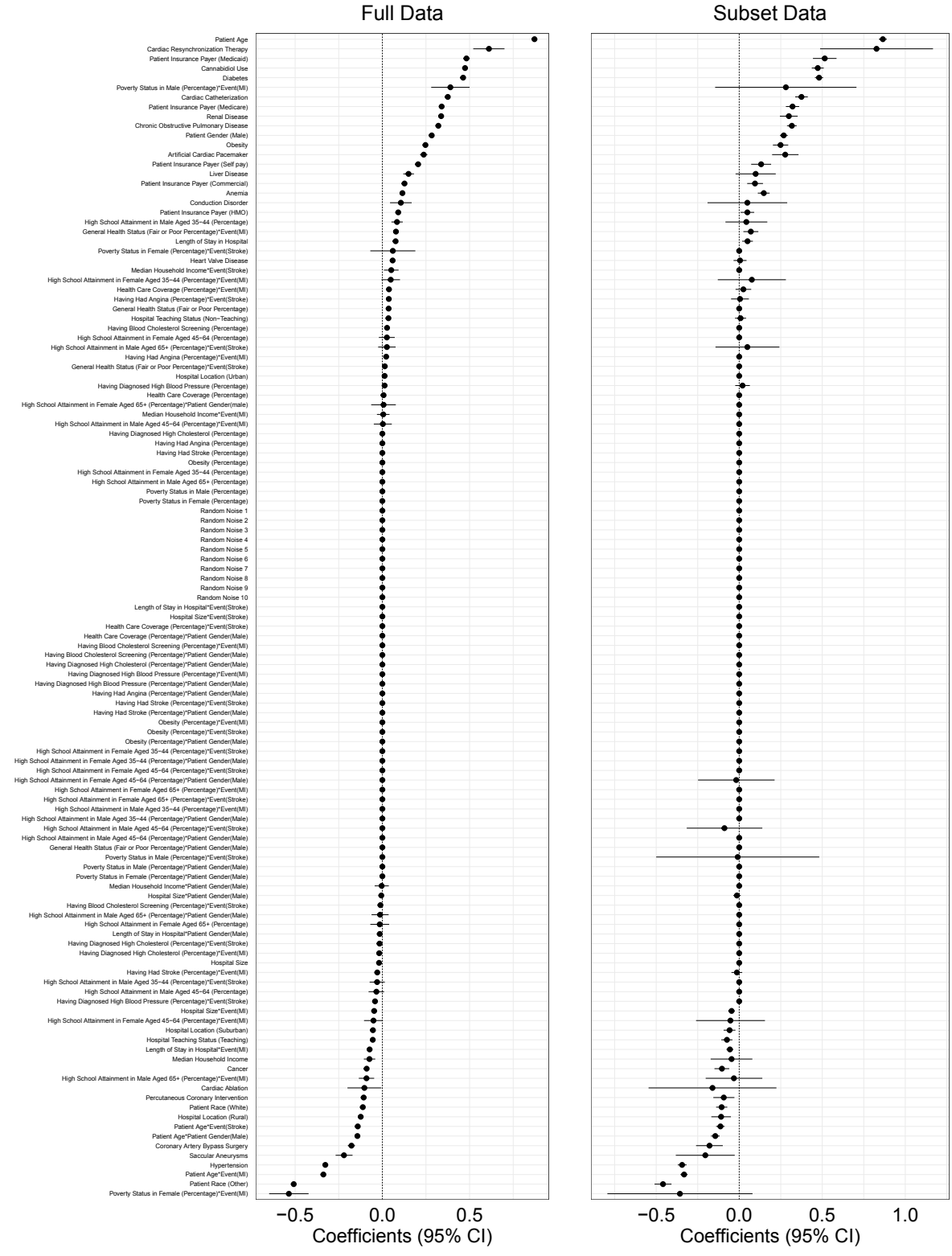

Figure 6.1: Estimated regularized regression coefficients in full data ($\hat{\boldsymbol{\beta}}_{\rho}^{dc}$) and random subset data ($\hat{\boldsymbol{\beta}}_{\rho}^{full}$) using marginal models.

6.3.2 Frailty Models in Multivariate Survival Analysis

Due to the complexity of the estimation algorithms in frailty models, we employed a random subset of MIDAS data used in the data analysis for marginal models to illustrate the proposed divide-and-combine approach in the proportional hazards frailty model. We analyzed data on 10,343 patients who had at least one hospitalization record due to cardiovascular diseases during 1995 to 2014 in the MIDAS database. Times from index cardiovascular hospitalization to the subsequent hospitalizations due to myocardial infarction (MI), heart failure (HF), and stroke were calculated and censored by death or the end of study (December 31, 2014), whichever occurred first. The median follow-up was 11.5 years; a total of 2,203 (21%) patients were re-admitted to hospitals due to one of MI, HF, or stroke, whereas 719 (7%) patients and 115 (1%) patients were re-admitted to hospitals due to occurrences of two or all of MI, HF or stroke, after the index hospitalizations. Covariates of interest include patient clinical characteristics at the index admissions, such as age and length of stay. Patient ZIP code level socioeconomic information, including health care coverage and median household income, is also considered. In addition, the hospital size of index admissions is included. Moreover, we modeled type-specific regression coefficients for all five covariates by specifying a type-specific design matrix. The performance of variable selection of the proposed regularized estimation approach was also tested by adding five unrelated and randomly generated noise variables into the data set. The final data set ended up with 31,029 rows of records with 20 covariates for data analysis.

After several unsuccessful attempts to analyze the full data all at once in a single computer (CPU specification: Intel® Core™ i7-4790 @3.60GHz), we used our proposed divide-and-combine approach (with simple random splitting) and randomly partitioned full data into $S = 20$ subsets with a partition ratio of $n_s/n = (\mathbf{2}/\mathbf{30}_{10}, \mathbf{1}/\mathbf{30}_{10})$. The divide-and-combine analysis was finished in 265 minutes on a Linux cluster via parallel computing (20 cores) with one subset allocated to one core (CPU specification: Intel® Xeon® E5-2680 v4 @2.40GHz). The estimated regularized regression coefficients ($\hat{\boldsymbol{\gamma}}_\rho^{dc}$ using

$\mathbf{W}_{1s}(\hat{\boldsymbol{\gamma}}_s) = \mathcal{I}_s(\hat{\boldsymbol{\gamma}}_s)$) and associated asymptotic standard error are shown in Table 6.10. Note that regularization was only applied on $\boldsymbol{\beta}$ but not θ, thus only $\hat{\boldsymbol{\beta}}_{\rho}^{dc}$ was summarized in Table 6.10. As expected, all five noise variables are estimated as zero, which indicates the selection consistency of our proposed regularization method. The effects of hospital size and percentage of health care coverage for three types of failures are estimated as zero. Median household income shows a protective effect on lowering the occurrence of hospitalization due to MI, HF, or stroke; on the contrary, staying longer in the hospital and being older increase the risk of hospitalization.

Table 6.10: Estimated regularized regression coefficients ($\hat{\gamma}_\rho^{dc}$) using divide-and-combine analysis in frailty models.

β_0[1]	$\hat{\gamma}_\rho^{dc}$ Estimate (ASE[2])
Noise Variable 1	0.00 (—)
Noise Variable 2	0.00 (—)
Noise Variable 3	0.00 (—)
Noise Variable 4	0.00 (—)
Noise Variable 5	0.00 (—)
Length of Stay*Event(MI)	0.09 (0.032)
Length of Stay*Event(HF)	0.26 (0.030)
Length of Stay*Event(Stroke)	0.17 (0.029)
Age*Event(MI)	0.56 (0.033)
Age*Event(HF)	1.51 (0.044)
Age*Event(Stroke)	0.65 (0.032)
Hospital Size*Event(MI)	0.00 (—)
Hospital Size*Event(HF)	0.00 (—)
Hospital Size*Event(Stroke)	0.00 (—)
Percentage of Health Care Coverage*Event(MI)	0.00 (—)
Percentage of Health Care Coverage*Event(HF)	0.00 (—)
Percentage of Health Care Coverage*Event(Stroke)	0.00 (—)
Median Household Income*Event(MI)	0.00 (—)
Median Household Income*Event(HF)	-0.06 (0.044)
Median Household Income*Event(Stroke)	-0.02 (0.034)

[1] Regularization only applies to β but not θ, thus only β is summarized in this table.

[2] ASE: asymptotic standard error using the theoretical formula.
Note: β_0 is the true value of β.

6.3.3 Multistate Survival Analysis

MIDAS data was utilized to demonstrate the proposed divide-and-combine approach in multistate survival analysis. As discussed in Section 5.1, the proposed five state Markov model (illustrated in Figure 5.1) can be used to study the disease processes in cardiovascular disease patients while taking into account the complicated relationship between heart failure and atrial fibrillation. We analyzed data on an incident cohort of 203,334 patients who had at least one hospitalization record due to heart failure from 2000 to 2017 and had no hospitalization due to heart failure within a five-year period before the incident heart failure hospitalization. Times from incident hospitalization due to heart failure (HF-1, i.e., state 1) to the subsequent second hospitalization due to heart failure (HF-2, i.e., state 2) or atrial fibrillation (AF, i.e., state 3), and the third hospitalization due to atrial fibrillation or heart failure (HF-2+AF, i.e., state 4) were calculated. Death (state 5) could happen to patients in any of the four hospitalization states, and thus the death times were calculated accordingly. All the patients were censored by the end of the study (December 31, 2017). The median follow-up was 11.7 years; after the incident hospitalization due to heart failure, a total of 48,718 (24%) patients and 6,546 (3%) patients were re-admitted to hospitals once due to heart failure (state 2) and atrial fibrillation (state 3), respectively; whereas 6,788 (3%) patients were re-admitted twice due to one heart failure and one atrial fibrillation (state 4), after the incident hospitalization. Among all the patients, 41, 162 (21%) patients stayed in state 1 without experiencing any admission after the incident hospitalization; while 100,120 (49%) patients were dead. Covariates of interest include patient clinical characteristics at the incident hospitalization, including age, gender, race, length of stay, comorbidity conditions, medical procedures received, and health insurance payer. Patient ZIP code level socioeconomic information, including general health status, health care coverage, education attainment, poverty status, median household income, blood cholesterol screening history, high cholesterol diagnosis, high blood pressure diagnosis, angina diagnosis, stroke diagnosis, obesity diagnosis, is also considered. In addition, some hos-

pital characteristics of incident hospitalization, i.e., teaching status, location, and size are included. A transition-specific design matrix was specified to include transition-specific effects of all covariates. The final data set ended up with 808,356 rows of records with 416 covariates for data analysis.

The full data were analyzed using the proposed divide-and-combine approach (with stratified random splitting) and were randomly partitioned into $S = 5$ subsets with a partition ratio of $n_s/n = (\mathbf{2}/\mathbf{8}_3, \mathbf{1}/\mathbf{8}_2)$. The divide-and-combine analysis was finished in 18 minutes on a Linux cluster via parallel computing (5 cores) with one subset allocated to one core (CPU specification: Intel® Xeon® E5-2680 v4 @2.40GHz). As a comparator, the full-data analysis was also performed in the full data, which was finished in 42 minutes on the same Linux cluster. We compared the estimated transition probabilities, $\hat{\mathbf{P}}^{dc}(u, t|\boldsymbol{z}_0)$ and $\hat{\mathbf{P}}^{full}(u, t|\boldsymbol{z}_0)$ obtained from using the divide-and-combine analysis and full-data analysis, respectively. The probabilities were predicted at time u for a future subject at time t, whose covariates ($\boldsymbol{z}_0$) are at the mean levels (continuous variables) or the reference levels (binary variables). We summarized the estimated transition probabilities from state 1 to states 2, 3, 4, 5, and their associated asymptotic standard errors (ASE) in Table 6.11. The results obtained from both divide-and-combine and full-data analyses are close enough, which shows the advantage by using the divide-and-combine analysis. It not only saves computational costs, but also achieves estimates as accurate as the full-data analysis.

Table 6.11: Estimated transition probabilities using divide-and-combine analysis ($\hat{\mathbf{P}}^{dc}(u,t|\boldsymbol{z}_0)$) and full-data analysis ($\hat{\mathbf{P}}^{full}(u,t|\boldsymbol{z}_0)$) in multistate models.

$\mathbf{P}_0$[1]			$\hat{\mathbf{P}}^{dc}$	$\hat{\mathbf{P}}^{full}$
	u	t	Estimate (ASE[2])	Estimate (ASE)
P_{11}	0	$\tau/2$	0.28 (0.006)	0.28 (0.006)
	0	τ	0.23 (0.005)	0.23 (0.005)
	$\tau/2$	$3\tau/4$	0.87 (0.008)	0.87 (0.008)
	$\tau/2$	τ	0.83 (0.010)	0.83 (0.010)
P_{12}	0	$\tau/2$	0.29 (0.007)	0.29 (0.007)
	0	τ	0.29 (0.007)	0.29 (0.007)
	$\tau/2$	$3\tau/4$	0.05 (0.005)	0.05 (0.005)
	$\tau/2$	τ	0.06 (0.006)	0.06 (0.006)
P_{13}	0	$\tau/2$	0.04 (0.003)	0.04 (0.003)
	0	τ	0.04 (0.003)	0.04 (0.003)
	$\tau/2$	$3\tau/4$	0.02 (0.003)	0.02 (0.003)
	$\tau/2$	τ	0.02 (0.005)	0.03 (0.005)
P_{14}	0	$\tau/2$	0.03 (0.003)	0.03 (0.003)
	0	τ	0.04 (0.004)	0.04 (0.004)
	$\tau/2$	$3\tau/4$	0.00 (0.000)	0.00 (0.000)
	$\tau/2$	τ	0.00 (0.000)	0.00 (0.000)
P_{15}	0	$\tau/2$	0.36 (0.007)	0.36 (0.007)
	0	τ	0.40 (0.007)	0.40 (0.007)
	$\tau/2$	$3\tau/4$	0.06 (0.006)	0.06 (0.006)
	$\tau/2$	τ	0.08 (0.007)	0.09 (0.007)

[1] Transition probabilities starting from state 1 are of primary interest and thus only these probabilities are summarized in this table. They are predicted at time u for time t, in which $\tau = 6574$ days is the end of the follow-up period.

[2] ASE: asymptotic standard error using the theoretical formula.

Note: $\mathbf{P}_0(u,t|\boldsymbol{z}_0)$ is the true value of $\mathbf{P}(u,t|\boldsymbol{z}_0)$.

CHAPTER 7
DISCUSSION AND FUTURE WORK

7.1 Discussion

In this dissertation, we propose divide-and-combine approaches for multivariate survival analysis and multistate survival analysis to analyze large-scale multivariate ordered and unordered failure time data. Specifically, a divide-and-combine estimator $\hat{\boldsymbol{\eta}}^{dc}$ for estimating regression parameters in multivariate survival analysis, and a divide-and-combine estimator $\hat{\boldsymbol{\Lambda}}^{dc}(t|\boldsymbol{z}_0)$ for estimating cumulative hazards in multistate survival analysis, are proposed. It is shown that the proposed divide-and-combine estimators $\hat{\boldsymbol{\eta}}^{dc}$ and $\hat{\boldsymbol{\Lambda}}^{dc}(t|\boldsymbol{z}_0)$ are statistically efficient in the sense that they are consistent and asymptotically equivalent to the full-data estimators $\hat{\boldsymbol{\eta}}^{full}$ and $\hat{\boldsymbol{\Lambda}}^{full}(t|\boldsymbol{z}_0)$. We also propose a confidence distribution approach to perform regularized estimation in multivariate survival analysis and construct the objective function based on the asymptotic distribution of $\hat{\boldsymbol{\eta}}^{dc}$. Because the objective function using the confidence distribution has taken into account the intra-cluster association in the multivariate unordered failure time data, we do not need additional variance adjustment in the regularized estimation. In contrast to the typical regularized estimation whose objective function is constructed using original data in the sample(s), our confidence distribution based regularization substantially reduces the dimensionality of data from n to d, which escalates the computational efficiency. Moreover, since our regularized estimation is performed after the combination step, we avoid the possibly inconsistent variable selection if the regularized estimation is performed on the individual subsets. We show that the proposed regularized estimator in the multivariate survival analysis $\hat{\boldsymbol{\eta}}_{\rho}^{dc}$ possesses estimation consistency, selection consistency, and oracle properties. We also demonstrate that the proposed divide-and-combine approaches tremendously reduce computation time in an

extraordinarily large data set. Generally, more subsets can save more time in computation.

Besides the statistical efficiency and the computational efficiency mentioned above, the proposed divide-and-combine approaches are also communication-efficient as it only requires one-round communication: individual parameter estimates in subsets are broadcast once to a "central" core to form the final estimates. Unlike other divide-and-combine algorithms with recursive communication design (e.g., Shamir et al. (2014); Wang et al. (2019)), our proposed approaches are simple and easy to be implemented using existing software packages.

Our simulation studies in multivariate survival analysis show that the empirical performance of $\hat{\boldsymbol{\eta}}^{dc}$ and $\hat{\boldsymbol{\eta}}_{\rho}^{dc}$ using $\mathbf{W}_{1s}(.)$ and $\mathbf{W}_{2s}(.)$ is better than that using $\mathbf{W}_{3s}(.)$ when the magnitude of regression coefficients is large and the number of subsets is big. This might be because $\mathbf{W}_{1s}(.)$ and $\mathbf{W}_{2s}(.)$ utilize more empirical information from the data in individual subsets (in the form of either $\mathcal{I}_s(\hat{\boldsymbol{\eta}}_s)$ or $\hat{\text{var}}_s^{-1}(\hat{\boldsymbol{\eta}}_s)$), while $\mathbf{W}_{3s}(.)$ only utilizes the information of the subset sample size (a single number). Therefore, in practice, we suggest to use $\mathbf{W}_{1s}(.)$ or $\mathbf{W}_{2s}(.)$ when applying the divide-and-combine approach in multivariate survival analysis.

The simulation studies in multistate survival analysis indicate that the empirical performance of $\hat{\mathbf{P}}^{dc}(u, t|\boldsymbol{z}_0)$ is better when the number of subsets (S) is smaller and the prediction is made at a later time. We conjecture that the unsatisfactory performance when S is large may result from violation of the homogeneity assumptions in practice.

Theoretically, in both multivariate survival analysis and multistate survival analysis, the number of subsets (S) is allowed to go to infinity with the order of $S = o(n^{1/2})$. In practice, we suggest to choose S such that the number of events per covariate is at least 5 to 10, following the recommendation of Vittinghoff and McCulloch (2007). The simulation studies show that our proposed approaches work well when the number of events per covariate is at least 5 to 10 in each divided subset, in spite of the low event rate.

When there are rare events and/or rare exposure(s), the simple random splitting may

result in no events or no exposure(s) in some subsets, and the divide-and-combine approach may not work. Thus we also implement the stratified random splitting in the divide step, which can be used to ensure events or exposures to evenly distributed across subsets and avoid too few events in some subset, for instance, to meet the condition of "5 to 10 events per covariate in each subset". In situations where the event is rare and the condition can hardly be met, there are other possible solutions. For example, we can perhaps adopt the Firth penalty term in our divide-and-combine analysis which has shown promising results in the univariate failure time data when there are unbalanced covariates (including rare covariates), large parameter effects, and heavy censoring (Heinze and Schemper, 2001; Nagashima and Sato, 2017). Another approach is to adopt the principle of case-control design (e.g., case-cohort design) to handle rare events and achieve computational efficiency, instead of the divide-and-combine approaches. Some references on the case-cohort design for multivariate unordered failure time data are Lu and Shih (2006), Kang and Cai (2009), Kim et al. (2018), among others.

7.2 Future Work

As suggested in the simulation studies for multistate survival analysis, the savings in the computational cost in the proposed divide-and-combine estimators for combining cumulative hazards are not as significant as those in multivariate survival analysis. It is under research to propose the divide-and-combine estimators for combining transition probabilities, which may save more computation time. Another topic for future research in multistate survival analysis is to study the impact of the prediction time for the performance of the proposed divide-and-combine approaches.

Regarding the rare events/rare exposure issues in both multivariate survival analysis and multistate survival analysis, the firth penalty as discussed above, is also a research topic in the future because it can be incorporated into the divide-and-combine framework.

BIBLIOGRAPHY

Aalen, O. O. and Johansen, S. (1978). An empirical transition matrix for non-homogeneous markov chains based on censored observations. *Scandinavian Journal of Statistics*, pages 141–150.

Aho, A. V. and Hopcroft, J. E. (1974). *The design and analysis of computer algorithms*. Pearson Education India.

Andersen, P. K. and Gill, R. D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10(4):1100–1120.

Andersen, P. K., Gill, R. D., Borgan, Ø., and Keiding, N. (1993). *Statistical models based on counting processes*. Springer.

Andersen, P. K., Hansen, L. S., and Keiding, N. (1991). Non-and semi-parametric estimation of transition probabilities from censored observation of a non-homogeneous markov process. *Scandinavian Journal of Statistics*, 18(2):153–167.

Andersen, P. K. and Keiding, N. (2002). Multi-state models for event history analysis. *Statistical methods in medical research*, 11(2):91–115.

Anter, E., Jessup, M., and Callans, D. J. (2009). Atrial fibrillation and heart failure: treatment considerations for a dual epidemic. *Circulation*, 119(18):2516–2525.

Battey, H., Fan, J., Liu, H., Lu, J., and Zhu, Z. (2018). Distributed testing and estimation under sparse high dimensional models. *Annals of statistics*, 46(3):1352.

Braunwald, E. (1997). Cardiovascular medicine at the turn of the millennium: triumphs, concerns, and opportunities. *New England Journal of Medicine*, 337(19):1360–1369.

Breslow, N. E. (1972). Discussion of the paper by d. r. cox. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):216–217.

Byar, D. P. (1980). *The Veterans Administration Study of Chemoprophylaxis for Recurrent Stage I Bladder Tumours: Comparisons of Placebo, Pyridoxine and Topical Thiotepa*, pages 363–370. Springer US, Boston, MA.

Cai, J., Fan, J., Li, R., and Zhou, H. (2005). Variable selection for multivariate failure time data. *Biometrika*, 92(2):303–316.

Chen, M.-C. (1998). *On modeling and inference for multivariate failure time data*. PhD

thesis. Department of Biostatistics, John Hopkins University.

Chen, X. and Xie, M. (2014). A split-and-conquer approach for analysis of extraordinarily large data. *Statistica Sinica*, 24(4):1655–1684.

Clayton, D. and Cuzick, J. (1985). Multivariate generalizations of the proportional hazards model. *Journal of the Royal Statistical Society. Series A (General)*, 148(2):82–117.

Cortese, G. and Andersen, P. K. (2010). Competing risks and time-dependent covariates. *Biometrical Journal*, 52(1):138–158.

Cox, D. (1958). Some problems connected with statistical inference. *The Annals of Mathematical Statistics*, 29(2):357–372.

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

Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

Cox, D. R. (2013). Discussion. *International Statistical Review*, 81(1):40–41.

de Wreede, L., Fiocco, M., and Putter, H. (2011). mstate: An r package for the analysis of competing risks and multi-state models. *Journal of Statistical Software, Articles*, 38(7):1–30.

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

Diabetic Retinopathy Study Research Group (1981). *Diabetic retinopathy study. Report Number 6. Design, methods, and baseline results. Report Number 7. A modification of the Airlie House classification of diabetic retinopathy.*, volume 21.

D'Agostino, R. B., Vasan, R. S., Pencina, M. J., Wolf, P. A., Cobain, M., Massaro, J. M., and Kannel, W. B. (2008). General cardiovascular risk profile for use in primary care. *Circulation*, 117(6):743–753.

Efron, B. (1993). Bayes and likelihood calculations from confidence intervals. *Biometrika*, 80(1):3–26.

Efron, B. (1998). R. a. fisher in the 21st century. *Statistical Science*, 13(2):95–114.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, 32(2):407–499.

Fan, J., Han, F., and Liu, H. (2014). Challenges of big data analysis. *National science review*, 1(2):293–314.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Fan, J. and Li, R. (2002). Variable selection for cox's proportional hazards model and frailty model. *Annals of Statistics*, pages 74–99.

Fisher, R. A. (1956). *Statistical methods and scientific inference (3rd ed., 1973)*. Hafner Publishing Co.

Fleming, T. R. and Harrington, D. P. (2011). *Counting processes and survival analysis*, volume 169. John Wiley & Sons.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Fu, A., Narasimhan, B., and Boyd, S. (2017). Cvxr: An r package for disciplined convex optimization. *arXiv preprint arXiv:1711.07582*.

Gorfine, M., Zucker, D. M., and Hsu, L. (2006). Prospective survival analysis with a general semiparametric shared frailty model: A pseudo full likelihood approach. *Biometrika*, 93(3):735–741.

Ha, I. D., Lee, Y., and Song, J. (2001). Hierarchical likelihood approach for frailty models. *Biometrika*, 88(1):233–233.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054.

Heinze, G. and Schemper, M. (2001). A solution to the problem of monotone likelihood in cox regression. *Biometrics*, 57(1):114–119.

Hippisley-Cox, J., Coupland, C., and Brindle, P. (2017). Development and validation of qrisk3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *bmj*, 357:j2099.

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

Hougaard, P. (2000). *Analysis of multivariate survival data*. Springer Science & Business Media.

Huster, W. J., Brookmeyer, R., and Self, S. G. (1989). Modelling paired survival data with covariates. *Biometrics*, pages 145–156.

Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics*, 67(2):495–503.

Jordan, M. I., Lee, J. D., and Yang, Y. (2019). Communication-efficient distributed statistical inference. *Journal of the American Statistical Association*, 114(526):668–681.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, volume 360. John Wiley & Sons.

Kang, S. and Cai, J. (2009). Marginal hazards model for case-cohort studies with multiple disease outcomes. *Biometrika*, 96(4):887–901.

Kauermann, G. and Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, 96(456):1387–1396.

Kim, S., Zeng, D., and Cai, J. (2018). Analysis of multiple survival events in generalized case-cohort designs. *Biometrics*, 74(4):1250–1260.

Klein, J. P. (1992). Semiparametric estimation of random effects using the cox model based on the em algorithm. *Biometrics*, 48(3):795–806.

Kleiner, A., Talwalkar, A., Sarkar, P., and Jordan, M. I. (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):795–816.

Kostis, W. J., Demissie, K., Marcella, S. W., Shao, Y.-H., Wilson, A. C., and Moreyra, A. E. (2007). Weekend versus weekday admission and mortality from myocardial infarction. *New England Journal of Medicine*, 356(11):1099–1109.

Laney, D. (2001). 3d data management: Controlling data volume, velocity and variety. *META group research note*, 6(70):1.

Lee, E. W., Wei, L. J., Amato, D. A., and Leurgans, S. (1992). *Cox-Type Regression Analysis for Large Numbers of Small Groups of Correlated Failure Time Observations*, pages 237–247. Springer Netherlands, Dordrecht.

Lee, J. D., Sun, Y., Liu, Q., and Taylor, J. E. (2015). Communication-efficient sparse regression: a one-shot approach. *arXiv preprint arXiv:1503.04337*.

Lehmann, E. L. (1993). The fisher, neyman-pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, 88(424):1242–1249.

Lin, D. Y. (1994). Cox regression analysis of multivariate failure time data: The marginal approach. *Statistics in Medicine*, 13(21):2233–2247.

Lin, D. Y. and Zeng, D. (2010). On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika*, 97(2):321–332.

Lin, N. and Xi, R. (2011). Aggregated estimating equation estimation. *Statistics and Its Interface*, 4(1):73–83.

Liu, D., Liu, R. Y., and Xie, M. (2015). Multivariate meta-analysis of heterogeneous studies using only summary statistics: Efficiency and robustness. *Journal of the American Statistical Association*, 110(509):326–340.

Liu, D., Liu, R. Y., and Xie, M.-g. (2014). Exact meta-analysis approach for discrete data and its application to $2\times$ 2 tables with rare events. *Journal of the American Statistical Association*, 109(508):1450–1465.

Lu, S.-E. and Shih, J. H. (2006). Case-cohort designs and analysis for clustered failure time data. *Biometrics*, 62(4):1138–1148.

Mackenzie, J. (1914). Diseases of the heart, 3rd edn. london: Frowde. *Hodder and Stoughton*, 101:102–103.

Mahé, C. and Chevret, S. (1999). Estimating regression parameters and degree of dependence for multivariate failure time data. *Biometrics*, 55(4):1078–1084.

McCullagh, P. (1989). *Generalized linear models*. Routledge.

Monaco, J. V., Gorfine, M., and Hsu, L. (2018). General semiparametric shared frailty model: Estimation and simulation with frailtySurv. *Journal of Statistical Software*, 86(4):1–42.

Nagashima, K. and Sato, Y. (2017). Information criteria for firth's penalized partial likelihood approach in cox regression models. *Statistics in medicine*, 36(21):3422–3436.

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

Parner, E. (1998). Asymptotic theory for the correlated gamma-frailty model. *Ann. Statist.*, 26(1):183–214.

Piepoli, M. F., Hoes, A. W., Agewall, S., Albus, C., Brotons, C., Catapano, A. L., Cooney, M.-T., Corra, U., Cosyns, B., Deaton, C., et al. (2016). 2016 european guidelines on cardiovascular disease prevention in clinical practice: The sixth joint task force of the european society of cardiology and other societies on cardiovascular disease prevention in clinical practice (constituted by representatives of 10 societies and by invited experts) developed with the special contribution of the european association for cardiovascular prevention & rehabilitation (eacpr). *European heart journal*, 37(29):2315–2381.

Pulver, A. E., Liang, K.-Y., and Vogler, G. P. (1991). Estimating effects of proband characteristics on familial risk: Ii. the association between age at onset and familial risk in the maryland schizophrenia sample. *Genetic Epidemiology*, 8(5):339–350.

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

Rondeau, V., Filleul, L., and Joly, P. (2006). Nested frailty models using maximum penalized likelihood estimation. *Statistics in Medicine*, 25(23):4036–4052.

Rosenblatt, J. D. and Nadler, B. (2016). On the optimality of averaging in distributed statistical learning. *Information and Inference: A Journal of the IMA*, 5(4):379–404.

Rothman, A. J., Levina, E., and Zhu, J. (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics*, 19(4):947–962.

Schifano, E. D., Wu, J., Wang, C., Yan, J., and Chen, M.-H. (2016). Online updating of statistical inference in the big data setting. *Technometrics*, 58(3):393–403.

Shamir, O., Srebro, N., and Zhang, T. (2014). Communication-efficient distributed optimization using an approximate newton-type method. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1000–1008, Beijing, China. PMLR.

Singh, K., Xie, M., and Strawderman, W. E. (2005). Combining information from independent sources through confidence distributions. *The Annals of Statistics*, 33(1):159–183.

Singh, K., Xie, M., and Strawderman, W. E. (2007). *Confidence distribution (CD) – distribution estimator of a parameter*, volume Volume 54 of *Lecture Notes–Monograph Series*, pages 132–150. Institute of Mathematical Statistics, Beachwood, Ohio, USA.

Sofer, T., Dicker, L., and Lin, X. (2014). Variable selection for high dimensional multivariate outcomes. *Statistica Sinica*, 24(4):1633.

Spiekerman, C. F. and Lin, D. Y. (1998). Marginal regression models for multivariate failure time data. *Journal of the American Statistical Association*, 93(443):1164–1175.

Swerdel, J. N., Rhoads, G. G., Cheng, J. Q., Cosgrove, N. M., Moreyra, A. E., Kostis, J. B., Kostis, W. J., Group, M. I. D. A. S. M. . S., Group, M. I. D. A. S. M. . S., Cabrera, J., et al. (2016). Ischemic stroke rate increases in young adults: Evidence for a generational effect? *Journal of the American Heart Association*, 5(12):e004245.

Tang, L., Zhou, L., and Song, P. X. (2016). Method of divide-and-combine in regularised generalised linear models for big data. *arXiv preprint arXiv:1611.06208*.

Therneau, T. M. (2020). *A Package for Survival Analysis in R*. R package version 3.2-3.

Therneau, T. M., Grambsch, P. M., and Pankratz, V. S. (2003). Penalized survival models and frailty. *Journal of Computational and Graphical Statistics*, 12(1):156–175.

Therneau, T. M. and Lumley, T. (2015). Package 'survival'. *R Top Doc*, 128:112.

Thomas, H., Diamond, J., Vieco, A., Chaudhuri, S., Shinnar, E., Cromer, S., Perel, P., Mensah, G. A., Narula, J., Johnson, C. O., et al. (2018). Global atlas of cardiovascular disease. *Global heart*, 13(3).

Tian, L., Wang, R., Cai, T., and Wei, L.-J. (2011). The highest confidence density region and its usage for joint inferences about constrained parameters. *Biometrics*, 67(2):604–610.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in medicine*, 16(4):385–395.

Van de Geer, S., Bühlmann, P., Ritov, Y., Dezeure, R., et al. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202.

van Wieringen, W. N. (2015). Lecture notes on ridge regression. *arXiv preprint arXiv:1509.09169*.

Vittinghoff, E. and McCulloch, C. E. (2007). Relaxing the rule of ten events per variable in logistic and cox regression. *American journal of epidemiology*, 165(6):710–718.

Wang, C., Chen, M.-H., Schifano, E., Wu, J., and Yan, J. (2016). Statistical methods and computing for big data. *Statistics and its interface*, 9(4):399.

Wang, H. and Leng, C. (2007). Unified lasso estimation by least squares approximation. *Journal of the American Statistical Association*, 102(479):1039–1048.

Wang, Y., Hong, C., Palmer, N., Di, Q., Schwartz, J., Kohane, I., and Cai, T. (2019). A fast divide-and-conquer sparse Cox regression. *Biostatistics*. kxz036.

Wei, L. J., Lin, D. Y., and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84(408):1065–1073.

Wellings, J., Kostis, J. B., Sargsyan, D., Cabrera, J., Kostis, W. J., System, M. I. D. A., and

Group, S. (2018). Risk factors and trends in incidence of heart failure following acute myocardial infarction. *The American journal of cardiology*, 122(1):1–5.

Xie, M. and Singh, K. (2013). Confidence distribution, the frequentist distribution estimator of a parameter: A review. *International Statistical Review*, 81(1):3–39.

Xie, M., Singh, K., and Strawderman, W. E. (2011). Confidence distributions and a unifying framework for meta-analysis. *Journal of the American Statistical Association*, 106(493):320–333.

Zeng, D., Lin, D. Y., and Lin, X. (2008). Semiparametric transformation models with random effects for clustered failure time data. *Statistica Sinica*, 18(1):355–377.

Zhang, C.-H. et al. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2):894–942.

Zhang, H. H. and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, 94(3):691–703.

Zhang, Y., Duchi, J. C., and Wainwright, M. J. (2013). Communication-efficient algorithms for statistical optimization. *The Journal of Machine Learning Research*, 14(1):3321–3363.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *Ann. Statist.*, 35(5):2173–2192.

Zucker, D. M., Gorfine, M., and Hsu, L. (2008). Pseudo-full likelihood estimation for prospective survival analysis with a general semiparametric shared frailty model: Asymptotic theory. *Journal of Statistical Planning and Inference*, 138(7):1998 – 2016.

APPENDIX

A.1 Derivation of Pseudo-Partial Likelihood from Full Likelihood

Inserting the model (2.2) into the full likelihood (2.1), we obtain

$$\mathcal{L}(\boldsymbol{\beta}, \lambda_{0k}(.)) = \prod_{i=1}^{n}\prod_{k=1}^{K}\left\{\lambda_{0k}(X_{ik})e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{ik}(X_{ik})}\right\}^{\delta_{ik}} \times \exp\left\{-\int_0^{X_{ik}} \lambda_{0k}(t)e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{ik}(t)}dt\right\}. \tag{A.1.1}$$

If we discretize $\lambda_{0k}(.)$ to mass points at L_k uncensored failure times for the k^{th} type of failure, $0 < t_{k[1]} < t_{k[2]} < \cdots < t_{k[L_k]}$, the logarithm of the full likelihood in (A.1.1) becomes

$$log\mathcal{L}(\boldsymbol{\beta}, \lambda_{0k}(.)) = \sum_{i=1}^{n}\sum_{k=1}^{K}\left[\delta_{ik}\left\{log\lambda_{0k}(X_{ik}) + \boldsymbol{\beta}^T \boldsymbol{Z}_{ik}(X_{ik})\right\} - \sum_{l=1}^{L_k} I(X_{ik} \geq t_{k[l]})\lambda_{0k}(t_{k[l]})e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{ik}(t_{k[l]})}\right]. \tag{A.1.2}$$

For given $\boldsymbol{\beta}$, we can maximize $log\mathcal{L}(\boldsymbol{\beta}, \lambda_{0k}(.))$ in (A.1.2) over $\lambda_{0k}(.)$ by setting

$$\frac{\partial}{\partial\lambda_{0k}(t_{k[l]})}log\mathcal{L}(\boldsymbol{\beta}, \lambda_{0k}(.)) = \frac{d_{k[l]}}{\lambda_{0k}(t_{k[l]})} - \sum_{i=1}^{n} I(X_{ik} \geq t_{k[l]})e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{ik}(t_{k[l]})} = 0, \tag{A.1.3}$$

where $d_{k[l]}$ is the number of failures at the failure time $t_{k[l]}$. The resulting maximizer is given by

$$\hat{\lambda}_{0k}(t_{k[l]}) = \frac{d_{k[l]}}{\sum_{i=1}^{n} I(X_{ik} \geq t_{k[l]})e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{ik}(t_{k[l]})}} \tag{A.1.4}$$

Plugging $\hat{\lambda}_{0k}(t_{k[l]})$ back into (A.1.2) and assuming no ties (i.e., $d_{k[l]} = 1$), we have

$$
\begin{aligned}
& log\mathcal{L}(\boldsymbol{\beta}, \hat{\lambda}_{0k}(.)) \\
& \quad = \sum_{k=1}^{K} \sum_{l=1}^{L_k} \left\{ log \frac{1}{\sum_{j=1}^{n} I(X_{jk} \geq t_{k[l]}) e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{jk}(t_{k[l]})}} \right\} \\
& \quad \quad + \sum_{i=1}^{n} \sum_{k=1}^{K} \left[\delta_{ik} \left\{ \boldsymbol{\beta}^T \boldsymbol{Z}_{ik}(X_{ik}) \right\} \right] \\
& \quad \quad - \sum_{k=1}^{K} \sum_{l=1}^{L_k} \frac{\sum_{i=1}^{n} I(X_{ik} \geq t_{k[l]}) e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{ik}(t_{k[l]})}}{\sum_{j=1}^{n} I(X_{jk} \geq t_{k[l]}) e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{jk}(t_{k[l]})}} \\
& \quad = \sum_{i=1}^{n} \sum_{k=1}^{K} \left[\delta_{ik} \left\{ log \frac{e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{ik}(X_{ik})}}{\sum_{j=1}^{n} I(X_{jk} \geq X_{ik}) e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{jk}(X_{ik})}} \right\} \right] - K \cdot L_k \\
& \quad = \sum_{i=1}^{n} \sum_{k=1}^{K} \left[\delta_{ik} \left\{ log \frac{e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{ik}(X_{ik})}}{n S_k^{(0)}(\boldsymbol{\beta}, X_{ik})} \right\} \right] - K \cdot L_k \\
& \quad = log\mathcal{PL}(\boldsymbol{\beta}) + O_p(1),
\end{aligned}
\tag{A.1.5}
$$

that is the pseudo-partial log-likelihood used in the marginal proportional hazards model.

A.2 Matrices for Asymptotic Variance Estimators in Frailty Models

(1) $\hat{\mathbf{V}}(\hat{\boldsymbol{\gamma}}) = n^{-1}\sum_{i=1}^{n}\xi_i\xi_i^T$, in which ξ_i is a $(1+d)$-variate vector with the 1^{st} element given by

$$\xi_{i1} = \frac{\int_0^\infty u_i^{A_i(\tau)}\hat{H}_i f_U'(u_i|\hat{\theta})du_i}{\int_0^\infty u_i^{A_i(\tau)}\hat{H}_i f_U(u_i|\hat{\theta})du_i}, \tag{A.2.1}$$

and the r^{th} element, $r = 2, 3, \ldots, (1+d)$ given by

$$\begin{aligned}\xi_{ir} =& \sum_{k=1}^{K}\delta_{ik}Z_{ikr} \\ &- \frac{\left\{\sum_{k=1}^{K}\hat{\Lambda}_0(\tau)e^{\hat{\boldsymbol{\beta}}^T\boldsymbol{Z}_{ik}}Z_{ikr}\right\}\int_0^\infty u_i^{A_i(\tau)+1}\hat{H}_i f_U(u_i|\hat{\theta})du_i}{\int_0^\infty u_i^{A_i(\tau)}\hat{H}_i f_U(u_i|\hat{\theta})du_i},\end{aligned} \tag{A.2.2}$$

where $\hat{H}_i = \exp\left\{-u_i\sum_{k=1}^{K}\hat{\Lambda}_0(\tau)e^{\hat{\boldsymbol{\beta}}^T\boldsymbol{Z}_{ik}}\right\}$.

(2) $\hat{\mathbf{G}}(\hat{\boldsymbol{\gamma}})$ is a matrix with rl^{th} element, $r, l = 1, 2, \ldots, (1+d)$ given by

$$\begin{aligned}&\hat{G}_{rl}(\hat{\boldsymbol{\gamma}}) \\ &= n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\delta_{ik}\left[\pi_r(X_{ik},\hat{\boldsymbol{\gamma}})\pi_l(X_{ik},\hat{\boldsymbol{\gamma}})\left\{R(X_{ik},\hat{\boldsymbol{\gamma}})\right\}^{-2}\right. \\ &\left.\left(\prod_{t\le X_{ik}-}\left[1+\sum_{j=1}^{n}\sum_{m=1}^{K}\left\{\delta_{jm}P(t)+Q_{jm}(t,X_{ik}-)\right\}I(X_{jm}\le t)\right]\right)^2\right],\end{aligned} \tag{A.2.3}$$

where

$$\begin{aligned}&\pi_r(s,\hat{\boldsymbol{\gamma}}) \\ &= n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}I(s<X_{ik}\le\tau)\left[\left(\prod_{t\le X_{ik}}\left[1+\right.\right.\right. \\ &\left.\left.\left.\sum_{j=1}^{n}\sum_{m=1}^{K}\left\{\delta_{jm}P(t)+Q_{jm}(t,X_{ik})\right\}I(X_{jm}\le t)\right]\right)^{-1}T_{ikr}(\hat{\boldsymbol{\gamma}},X_{ik})\right],\end{aligned} \tag{A.2.4}$$

$$T_{ik1}(\hat{\boldsymbol{\gamma}}, X_{ik}) = e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{Z}_{ik}} \left\{ \frac{\phi_{i2}(\hat{\boldsymbol{\gamma}}, \hat{\Lambda}_0(\tau), \tau)\phi_{i1}^{(\theta)}(\hat{\boldsymbol{\gamma}}, \hat{\Lambda}_0(\tau), \tau)}{\phi_{i1}^2(\hat{\boldsymbol{\gamma}}, \hat{\Lambda}_0(\tau), \tau)} - \frac{\phi_{i2}^{(\theta)}(\hat{\boldsymbol{\gamma}}, \hat{\Lambda}_0(\tau), \tau)}{\phi_{i1}(\hat{\boldsymbol{\gamma}}, \hat{\Lambda}_0(\tau), \tau)} \right\}, \tag{A.2.5}$$

$$\begin{aligned} T_{ikr}(\hat{\boldsymbol{\gamma}}, X_{ik}) = -\Bigg\{ & \frac{\phi_{i2}(\hat{\boldsymbol{\gamma}}, \hat{\Lambda}_0(\tau), \tau)}{\phi_{i1}(\hat{\boldsymbol{\gamma}}, \hat{\Lambda}_0(\tau), \tau)} e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{Z}_{ik}} Z_{ikr} \\ & - \frac{\phi_{i2}(\hat{\boldsymbol{\gamma}}, \hat{\Lambda}_0(\tau), \tau)}{\phi_{i1}(\hat{\boldsymbol{\gamma}}, \hat{\Lambda}_0(\tau), \tau)} e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{Z}_{ik}} \sum_{m=1}^{K} \hat{\Lambda}_0(X_{ij}) e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{Z}_{im}} Z_{imr} \\ & + \frac{\phi_{i2}^2(\hat{\boldsymbol{\gamma}}, \hat{\Lambda}_0(\tau), \tau)}{\phi_{i1}^2(\hat{\boldsymbol{\gamma}}, \hat{\Lambda}_0(\tau), \tau)} e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{Z}_{ik}} \sum_{m=1}^{K} \hat{\Lambda}_0(X_{ij}) e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{Z}_{im}} Z_{imr} \Bigg\} \end{aligned} \tag{A.2.6}$$

for $r = 2, 3, \ldots, (1 + d)$,

$$\begin{aligned} P(t) = n^{-2} \left\{ R(t, \hat{\boldsymbol{\gamma}}) \right\}^{-2} & \sum_{i=1}^{n} \sum_{k=1}^{K} I(X_{ik} > t) e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{Z}_{ik}} \nu_{1i}(t) \\ & \left\{ \sum_{m=1}^{K} I(X_{im} \geq t) e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{Z}_{im}} \right\}, \end{aligned} \tag{A.2.7}$$

$$\begin{aligned} Q_{ik}(s, t) = n^{-2} e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{Z}_{ik}} \sum_{j=1}^{n} \sum_{m=1}^{K} \delta_{jm} I(s < X_{jm} \leq t) \Bigg[& \left\{ R(X_{jm}, \hat{\boldsymbol{\gamma}}) \right\}^{-2} \\ & \nu_{1i}(X_{jm}) \left\{ \sum_{u=1}^{K} I(X_{iu} \geq X_{jm}) e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{Z}_{iu}} \right\} \Bigg], \end{aligned} \tag{A.2.8}$$

$$R(t, \hat{\boldsymbol{\gamma}}) = n^{-1} \sum_{i=1}^{n} \Psi(w_i | t, \hat{\boldsymbol{\gamma}}) \sum_{k=1}^{K} I(X_{ik} \geq t) e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{Z}_{ik}}, \tag{A.2.9}$$

$$\Psi(w_i | t, \hat{\boldsymbol{\gamma}}) = \frac{\phi_{i2}(\hat{\boldsymbol{\gamma}}, \hat{\Lambda}_0(t), t)}{\phi_{i1}(\hat{\boldsymbol{\gamma}}, \hat{\Lambda}_0(t), t)}, \tag{A.2.10}$$

$$\nu_{1i}(t) = \frac{\phi_{i3}(\hat{\boldsymbol{\gamma}}, \hat{\Lambda}_0(t), t)}{\phi_{i1}(\hat{\boldsymbol{\gamma}}, \hat{\Lambda}_0(t), t)} - \left\{ \frac{\phi_{i2}(\hat{\boldsymbol{\gamma}}, \hat{\Lambda}_0(t), t)}{\phi_{i1}(\hat{\boldsymbol{\gamma}}, \hat{\Lambda}_0(t), t)} \right\}^2, \tag{A.2.11}$$

$$\begin{aligned} \phi_{im}(\hat{\boldsymbol{\gamma}}, \hat{\Lambda}_0(t), t) = \int_0^\infty & u_i^{A_i(t)+(m-1)} \\ & \exp\left\{ -u_i \sum_{k=1}^{K} \hat{\Lambda}_0(t) e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{Z}_{ik}} \right\} f_U(u_i|\hat{\theta}) du_i, \end{aligned} \tag{A.2.12}$$

$$\begin{aligned} \phi_{im}^{(\theta)}(\hat{\boldsymbol{\gamma}}, \hat{\Lambda}_0(t), t) = \int_0^\infty & u_i^{A_i(t)+(m-1)} \\ & \exp\left\{ -u_i \sum_{k=1}^{K} \hat{\Lambda}_0(t) e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{Z}_{ik}} \right\} f'_U(u_i|\hat{\theta}) du_i, \end{aligned} \tag{A.2.13}$$

and

$$\hat{\Lambda}_0(t) = \sum_{i=1}^{n} \sum_{k=1}^{K} \frac{\delta_{ik} I(X_{ik} \leq t)}{\sum_{j=1}^{n} \sum_{m=1}^{k} \Psi(w_i|X_{ik}-, \hat{\boldsymbol{\gamma}}) I(X_{jm} \geq X_{ik}) e^{\hat{\boldsymbol{\beta}}^T \boldsymbol{Z}_{jm}}}. \tag{A.2.14}$$

(3) $\hat{\mathbf{C}}(\hat{\boldsymbol{\gamma}})$ is a matrix with rl^{th} element, $r, l = 1, 2, \ldots, (1+d)$ given by

$$\hat{C}_{rl}(\hat{\boldsymbol{\gamma}}) = n^{-1} \sum_{i=1}^{n} (\xi_{ir}\mu_{il} + \xi_{il}\mu_{ir}), \tag{A.2.15}$$

where

$$
\begin{aligned}
&\mu_{ir} = \\
&\sum_{k=1}^{K}\Bigg(\delta_{ik}\prod_{t\le X_{ik}-}\left[1+\sum_{j=1}^{n}\sum_{m=1}^{K}\left\{\delta_{jm}P(t)+Q_{jm}(t,X_{ik}-)\right\}I(X_{jm}\le t)\right] \\
&\qquad\times\left\{R(X_{X_{ik}},\hat{\boldsymbol{\gamma}})\right\}^{-1}\pi_r(X_{ik},\hat{\boldsymbol{\gamma}}) \\
&\qquad-\sum_{j=1}^{n}\sum_{m=1}^{K}\delta_{jm}\frac{e^{\hat{\boldsymbol{\beta}}^T\boldsymbol{z}_{ik}}I(X_{ik}\ge X_{jm})\Psi(w_i|\hat{\boldsymbol{\gamma}},X_{jm}-)}{\sum_{u=1}^{n}\sum_{v=i}^{K}\Psi(w_i|\hat{\boldsymbol{\gamma}},X_{jm}-)I(X_{uv}\ge X_{jm})e^{\hat{\boldsymbol{\beta}}^T\boldsymbol{z}_{uv}}} \\
&\qquad\times\prod_{t\le X_{jm}-}\left[1+\sum_{u=1}^{n}\sum_{v=1}^{K}\left\{\delta_{uv}P(t)+Q_{uv}(t,X_{jm}-)\right\}I(X_{uv}\le t)\right] \\
&\qquad\times\left\{R(X_{jm},\hat{\boldsymbol{\gamma}})\right\}^{-1}\pi_r(X_{jm},\hat{\boldsymbol{\gamma}})\Bigg).
\end{aligned}
\tag{A.2.16}
$$

A.3 Derivation of Transition Probability Matrix

In Markov multistate models, using the total probability theorem, the transition probability from state h to state j in the time interval $(s, t]$ can be calculated as

$$P_{i,hj}(s,t) = \sum_{q=1}^{Q} P_{i,hq}(s,u) P_{i,qj}(u,t), \tag{A.3.1}$$

where Q is the total number of states that subject i can potentially visit. (A.3.1) is also called the Chapman-Kolmogorov equation.

By using (A.3.1), we have

$$\begin{aligned} \mathbf{P}_i(s,t+\Delta t) - \mathbf{P}_i(s,t) &= \mathbf{P}_i(s,t)\mathbf{P}_i(t,t+\Delta t) - \mathbf{P}_i(s,t) \\ &= \mathbf{P}_i(s,t)\left\{\mathbf{P}_i(t,t+\Delta t) - \mathbf{I}\right\} \\ &\approx \mathbf{P}_i(s,t)\boldsymbol{\lambda}_i(t)\Delta t, \end{aligned} \tag{A.3.2}$$

where $\boldsymbol{\lambda}_i(t) = \lim_{\Delta t \to 0} \Delta t^{-1} \left\{\mathbf{P}_i(t,t+\Delta t) - \mathbf{I}\right\}$. Hence we have the Kolmogorov forward equation

$$\frac{\partial}{\partial t}\mathbf{P}_i(s,t) = \mathbf{P}_i(s,t)\boldsymbol{\lambda}_i(t). \tag{A.3.3}$$

By integrating both sides, (A.3.3) can be equivalently expressed as

$$\mathbf{P}_i(s,t) = \mathbf{I} + \int_s^t \mathbf{P}_i(s,u) d\boldsymbol{\Lambda}_i(u). \tag{A.3.4}$$

Note that $\boldsymbol{\Lambda}_i(t)$ is the element-wise integral of $\boldsymbol{\lambda}_i(t)$.

To solve $\mathbf{P}_i(s,t)$ from (A.3.4), we may partition the time interval $(s,t]$ into G sufficiently small sub-intervals $(s = t_0, t_1], (t_1, t_2], \ldots, (t_{G-1}, t = t_G]$, and by using (A.3.2), write the transition probability $\mathbf{P}_i(s,t)$ as

$$\mathbf{P}_i(s,t) = \mathbf{P}_i(t_0,t_1)\mathbf{P}_i(t_1,t_2)\cdots\mathbf{P}_i(t_{G-1},t_G). \tag{A.3.5}$$

Plugging (A.3.4) into (A.3.5), we obtain

$$\begin{aligned}\mathbf{P}_i(s,t) &= \prod_{g=1}^{G}\left[\mathbf{I} + \{\mathbf{\Lambda}_i(t_g) - \mathbf{\Lambda}_i(t_{g-1})\}\right] \\ &= \prod_{u\in(s,t]}\left\{\mathbf{I} + d\mathbf{\Lambda}_i(u)\right\},\end{aligned} \tag{A.3.6}$$

where $\prod$ is the sign of product integral, which has the same relation to a product as the well-known integral has to a sum. It can be understood using the simple example below.

In univariate survival analysis, if we divide the time interval $(0,t]$ into G sub-intervals $(0=t_0,t_1],(t_1,t_2],\ldots,(t_{G-1},t=t_G]$, the survival function can be expressed by a product of conditional survival functions

$$\begin{aligned}S(t) &= Pr(T>t) \\ &= Pr(T>t_1)Pr(T>t_2|T>t_1)\cdots Pr(T>t|T>t_{G-1}) \\ &= \prod_{g=1}^{G} S(t_g|t_{g-1}).\end{aligned} \tag{A.3.7}$$

Given $S(t)=e^{-\Lambda(t)}$, we can derive that $dS(t)=-S(t)d\Lambda(t)$, which can be approximated by $S(t_g)-S(t_{g-1})\approx -S(t_{g-1})\{\Lambda(t_g)-\Lambda(t_{g-1})\}$. Dividing both sides by $S(t_{g-1})$, we have

$$S(t_g|t_{g-1}) \approx 1-\{\Lambda(t_g)-\Lambda(t_{g-1})\}. \tag{A.3.8}$$

Now plugging (A.3.8) in (A.3.7), we get

$$S(t) = \prod_{g=1}^{G}\left[1-\{\Lambda(t_g)-\Lambda(t_{g-1})\}\right]. \tag{A.3.9}$$

If we let the lengths of G sub-intervals go to zero uniformly, by definition the survival

function in (A.3.9) can be expressed as the product integral

$$S(t) = \prod_{u \in (0,t]} \{1 - d\Lambda(u)\}. \tag{A.3.10}$$

Using the approximation of $\exp(-\lambda(u)du) \approx 1 - \lambda(u)du$, we can see the equality between survival functions expressed in the product integral and expressed in the conventional way below

$$S(t) = \prod_{u \in (0,t]} \{1 - d\Lambda(u)\} = \prod_{u \in (0,t]} \{1 - \lambda(u)du\} = \exp(-\int_0^t \lambda(u)du) = S(t). \tag{A.3.11}$$

A.4 Construction of Likelihood in Markov Multistate Models

Assume each subject can potentially visit each state once and denote the failure time of transition $h \to j$ for subject i by T_{hj}^i for $i = 1, 2, \ldots, n$, and $h, j = 1, 2, \ldots, Q$, and $h \neq j$. Using the similar idea in Cox partial likelihood (Cox, 1975), the likelihood function in the Markov multistate models can be constructed as

$$\mathcal{L} = \prod_{i=1}^{n} \prod_{h \neq j} Pr\left\{X_i(T_{hj}^i) = j, X_i(0) = h\right\} \tag{A.4.1}$$

We can rewrite $Pr\left\{X_i(T_{hj}^i) = j, X_i(0) = h\right\}$ as

$$Pr\left\{X_i(T_{hj}^i) = j | X_i(T_{hj}^i-) = h\right\} Pr\left\{X_i(T_{hj}^i-) = h\right\} Pr\left\{X_i(0) = h\right\}, \tag{A.4.2}$$

where $Pr\left\{X_i(T_{hj}^i) = j | X_i(T_{hj}^i-) = h\right\}$ and $Pr\left\{X_i(T_{hj}^i-) = h\right\}$ are transition-specific hazard and survival functions for transition $h \to j$, respectively.

Then the likelihood in (A.4.1) becomes

$$\mathcal{L} = \prod_{i=1}^{n} \pi_{\{X_i(0)\}} \prod_{h \neq j} \prod_{N_{i,hj}(\tau) \neq 0} \lambda_{i,hj}(T_{hj}^i) \exp\left\{-\int_0^{T_{hj}^i} \lambda_{i,hj}(t) Y_{i,h}(t) dt\right\}, \tag{A.4.3}$$

where $\pi_{\{X_i(0)\}}$ is the initial distribution of subject i, which is oftentimes fixed and can be omitted during optimization. The core part of the likelihood in (A.4.3) takes the same form as the full likelihood in multivariate survival analysis as shown in (2.1). It follows immediately that the likelihood in (A.4.3) can be derived into the the partial likelihood as demonstrated in Appendix A.1. Then the estimation approach used in Section 2.2.3 is justified.

A.5 Derivation of the Expression of $P_{12}(s,t|\boldsymbol{z}_0)$

By manipulating the differential equation

$$\frac{\partial}{\partial t}P_{12}(s,t|\boldsymbol{z}_0) = \lambda_{12}(t|\boldsymbol{z}_0)P_{11}(s,t|\boldsymbol{z}_0) - \lambda_{23}(t|\boldsymbol{z}_0)P_{12}(s,t|\boldsymbol{z}_0), \tag{A.5.1}$$

we have

$$\begin{aligned} &\frac{\partial}{\partial t}P_{12}(s,t|\boldsymbol{z}_0)\exp\left\{\int_s^t \lambda_{23}(u|\boldsymbol{z}_0)du\right\} \\ &+ \lambda_{23}(t|\boldsymbol{z}_0)P_{12}(s,t|\boldsymbol{z}_0)\exp\left\{\int_s^t \lambda_{23}(u|\boldsymbol{z}_0)du\right\} \\ &= \lambda_{12}(t|\boldsymbol{z}_0)P_{11}(s,t|\boldsymbol{z}_0)\exp\left\{\int_s^t \lambda_{23}(u|\boldsymbol{z}_0)du\right\}. \end{aligned} \tag{A.5.2}$$

Recognizing the left-hand side as a partial derivative, we obtain

$$\begin{aligned} &\frac{\partial}{\partial t}\left[P_{12}(s,t|\boldsymbol{z}_0)\exp\left\{\int_s^t \lambda_{23}(u|\boldsymbol{z}_0)du\right\}\right] \\ &\qquad = \lambda_{12}(t|\boldsymbol{z}_0)P_{11}(s,t|\boldsymbol{z}_0)\exp\left\{\int_s^t \lambda_{23}(u|\boldsymbol{z}_0)du\right\}, \end{aligned} \tag{A.5.3}$$

and hence

$$\begin{aligned} P_{12}(s,t|\boldsymbol{z}_0) &= \exp\left\{-\int_s^t \lambda_{23}(u|\boldsymbol{z}_0)du\right\} \\ &\quad \times \int_s^t \lambda_{12}(u|\boldsymbol{z}_0)P_{11}(s,u|\boldsymbol{z}_0)\exp\left\{\int_s^u \lambda_{23}(v|\boldsymbol{z}_0)dv\right\}du \\ &= \int_s^t P_{11}(s,u|\boldsymbol{z}_0)\lambda_{12}(u|\boldsymbol{z}_0)\exp\left\{-\int_u^t \lambda_{23}(v|\boldsymbol{z}_0)dv\right\}du \\ &= \int_s^t P_{11}(s,u|\boldsymbol{z}_0)\lambda_{12}(u|\boldsymbol{z}_0)P_{22}(u,t|\boldsymbol{z}_0)du. \end{aligned} \tag{A.5.4}$$

A.6 Proof of Theorem 3.3.1: Asymptotic Properties of $\hat{\boldsymbol{\eta}}^{dc}$

Under regularity conditions (M1) to (M6) and (F1) to (F10) plus the homogeneity assumptions (H1) to (H2), and asymptotic properties for $\hat{\boldsymbol{\beta}}_s$ (Theorem 2.1.1) and $\hat{\boldsymbol{\gamma}}_s$ (Theorem 2.1.2), we have $\hat{\boldsymbol{\eta}}_s \xrightarrow{P} \boldsymbol{\eta}_0$, $n_s^{-1}\mathcal{I}_s(\hat{\boldsymbol{\eta}}_s) = \hat{\mathbf{A}}_s(\hat{\boldsymbol{\eta}}_s) \xrightarrow{P} \mathbf{A}(\boldsymbol{\eta}_0)$, $\hat{\mathbf{B}}_s(\hat{\boldsymbol{\eta}}_s) \xrightarrow{P} \mathbf{B}(\boldsymbol{\eta}_0)$, and $n_s^{-1/2}\mathbf{U}_s(\boldsymbol{\eta}_0) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{B}(\boldsymbol{\eta}_0))$ as $n_s \to \infty$. These asymptotic properties are explicitly shown in Spiekerman and Lin (1998) and Gorfine et al. (2006). We introduce the following Lemma A.6.1 before proving Theorem 3.3.1.

Lemma A.6.1. Under conditions in Theorem 3.3.1, we have that $n_s^{-1}\mathbf{W}_s(\hat{\boldsymbol{\eta}}_s) \xrightarrow{P} \mathbf{w}(\boldsymbol{\eta}_0)$ as $n_s \to \infty$, where

$$\mathbf{w}(\boldsymbol{\eta}_0) = \begin{cases} \mathbf{w}_1(\boldsymbol{\eta}_0) = \mathbf{A}(\boldsymbol{\eta}_0), & \text{if } \mathbf{W}_s(.) = \mathbf{W}_{1s}(.); \\ \mathbf{w}_2(\boldsymbol{\eta}_0) = \{\mathbf{A}(\boldsymbol{\eta}_0)\}^T\{\mathbf{B}^{-1}(\boldsymbol{\eta}_0)\}\{\mathbf{A}(\boldsymbol{\eta}_0)\}, & \text{if } \mathbf{W}_s(.) = \mathbf{W}_{2s}(.); \\ \mathbf{w}_3(\boldsymbol{\eta}_0) = 1, & \text{if } \mathbf{W}_s(.) = \mathbf{W}_{3s}(.). \end{cases} \tag{A.6.1}$$

Proof. Given that $\hat{\boldsymbol{\eta}}_s \xrightarrow{P} \boldsymbol{\eta}_0$, $n_s^{-1}\mathcal{I}_s(\hat{\boldsymbol{\eta}}_s) = \hat{\mathbf{A}}_s(\hat{\boldsymbol{\eta}}_s) \xrightarrow{P} \mathbf{A}(\boldsymbol{\eta}_0)$, $\hat{\mathbf{B}}_s(\hat{\boldsymbol{\eta}}_s) \xrightarrow{P} \mathbf{B}(\boldsymbol{\eta}_0)$ as $n_s \to \infty$, and $\hat{\text{var}}_s^{-1}(\hat{\boldsymbol{\eta}}_s) = n_s\{\hat{\mathbf{A}}_s(\hat{\boldsymbol{\eta}}_s)\}^T\{\hat{\mathbf{B}}_s^{-1}(\hat{\boldsymbol{\eta}}_s)\}\{\hat{\mathbf{A}}_s(\hat{\boldsymbol{\eta}}_s)\}$, we can verify the following

$$\begin{aligned} &n_s^{-1}\mathbf{W}_{1s}(\hat{\boldsymbol{\eta}}_s) = n_s^{-1}\mathcal{I}_s(\hat{\boldsymbol{\eta}}_s) \xrightarrow{P} \mathbf{A}(\boldsymbol{\eta}_0) = \mathbf{w}_1(\boldsymbol{\eta}_0), \\ &n_s^{-1}\mathbf{W}_{2s}(\hat{\boldsymbol{\eta}}_s) = n_s^{-1}\hat{\text{var}}_s^{-1}(\hat{\boldsymbol{\eta}}_s) \xrightarrow{P} \{\mathbf{A}(\boldsymbol{\eta}_0)\}^T\{\mathbf{B}^{-1}(\boldsymbol{\eta}_0)\}\{\mathbf{A}(\boldsymbol{\eta}_0)\} = \mathbf{w}_2(\boldsymbol{\eta}_0), \\ &n_s^{-1}\mathbf{W}_{3s}(\hat{\boldsymbol{\eta}}_s) = 1 = \mathbf{w}_3(\boldsymbol{\eta}_0), \end{aligned} \tag{A.6.2}$$

as $n_s \to \infty$. This completes the proof of Lemma A.6.1. ■

We now show Theorem 3.3.1. In light of Lemma A.6.1 and given that $\hat{\boldsymbol{\eta}}_s \xrightarrow{P} \boldsymbol{\eta}_0$, it can

be seen that

$$\begin{aligned}\hat{\boldsymbol{\eta}}^{dc} &= \left\{\sum_{s=1}^{S}(n_s/n)n_s^{-1}\mathbf{W}_s(\hat{\boldsymbol{\eta}})\right\}^{-1}\sum_{s=1}^{S}(n_s/n)n_s^{-1}\mathbf{W}_s(\hat{\boldsymbol{\eta}})\hat{\boldsymbol{\eta}}_s \\ &\xrightarrow{P} \left\{\sum_{s=1}^{S}(n_s/n)\mathbf{w}(\boldsymbol{\eta}_0)\right\}^{-1}\sum_{s=1}^{S}(n_s/n)\mathbf{w}(\boldsymbol{\eta}_0)\boldsymbol{\eta}_0 \\ &= \boldsymbol{\eta}_0,\end{aligned} \tag{A.6.3}$$

as $n \to \infty$. Before showing the asymptotic normality, we apply the Taylor series expansion and obtain

$$\begin{aligned}n_s^{-1}\mathbf{U}_s(\hat{\boldsymbol{\eta}}_s) =& n_s^{-1}\mathbf{U}_s(\boldsymbol{\eta}_0) - n_s^{-1}\mathcal{I}_s(\boldsymbol{\eta}_0)(\hat{\boldsymbol{\eta}}_s - \boldsymbol{\eta}_0) \\ &+ \sum_{j=1}^{d}(2n_s)^{-1}(\boldsymbol{\eta}^* - \boldsymbol{\eta}_0)^T \left.\frac{\partial^2\mathbf{U}_s(\boldsymbol{\eta})}{\partial\eta_j\partial\boldsymbol{\eta}}\right|_{\boldsymbol{\eta}=\boldsymbol{\eta}^*}(\boldsymbol{\eta}^* - \boldsymbol{\eta}_0),\end{aligned} \tag{A.6.4}$$

where $\boldsymbol{\eta}^*$ lies between $\boldsymbol{\eta}_0$ and $\hat{\boldsymbol{\eta}}_s$. Notice that $\mathbf{U}_s(\hat{\boldsymbol{\eta}}_s) = \mathbf{0}$, $n_s^{-1}\mathcal{I}_s(\boldsymbol{\eta}_0) = \mathbf{A}(\boldsymbol{\eta}_0) + O_p(n_s^{-1/2})$, $n_s^{-1}\partial^2\mathbf{U}_s(\boldsymbol{\eta})/\partial\eta_j\partial\boldsymbol{\eta}|_{\boldsymbol{\eta}=\boldsymbol{\eta}^*} = O_p(1)$, and $\hat{\boldsymbol{\eta}}_s = \boldsymbol{\eta}_0 + O_p(n_s^{-1/2})$, then we have

$$n_s^{1/2}(\hat{\boldsymbol{\eta}}_s - \boldsymbol{\eta}_0) = \mathbf{A}^{-1}(\boldsymbol{\eta}_0)n_s^{-1/2}\mathbf{U}_s(\boldsymbol{\eta}_0) + O_p(n_s^{-1/2}). \tag{A.6.5}$$

To show the asymptotic normality, by using (A.6.5), we write

$$\begin{aligned}&n^{1/2}(\hat{\boldsymbol{\eta}}^{dc} - \boldsymbol{\eta}_0) \\ &= n^{1/2}\left\{\sum_{s=1}^{S}\mathbf{W}_s(\hat{\boldsymbol{\eta}}_s)\right\}^{-1}\left\{\sum_{s=1}^{S}\mathbf{W}_s(\hat{\boldsymbol{\eta}}_s)(\hat{\boldsymbol{\eta}}_s - \boldsymbol{\eta}_0)\right\} \\ &= \left\{\sum_{s=1}^{S}(n_s/n)n_s^{-1}\mathbf{W}_s(\hat{\boldsymbol{\eta}}_s)\right\}^{-1}\left[\sum_{s=1}^{S}(n_s/n)^{1/2}n_s^{-1}\mathbf{W}_s(\hat{\boldsymbol{\eta}}_s)\left\{n_s^{1/2}(\hat{\boldsymbol{\eta}}_s - \boldsymbol{\eta}_0)\right\}\right] \\ &= \left\{\sum_{s=1}^{S}(n_s/n)n_s^{-1}\mathbf{W}_s(\hat{\boldsymbol{\eta}}_s)\right\}^{-1}\left[\sum_{s=1}^{S}(n_s/n)^{1/2}n_s^{-1}\mathbf{W}_s(\hat{\boldsymbol{\eta}}_s)\mathbf{A}^{-1}(\boldsymbol{\eta}_0)n_s^{-1/2}\mathbf{U}_s(\boldsymbol{\eta}_0) + O_p(n^{-1/2}S)\right].\end{aligned} \tag{A.6.6}$$

Because $S = o(n^{1/2})$, $O_p(n^{-1/2}S)$ is $o_p(1)$, and as $n \to \infty$, $n_s^{-1}\mathbf{W}_s(\hat{\boldsymbol{\eta}}_s) \xrightarrow{P} \mathbf{w}(\boldsymbol{\eta}_0)$

(Lemma A.6.1) and $n_s^{-1/2}\mathbf{U}_s(\boldsymbol{\eta}_0) \xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{B}(\boldsymbol{\eta}_0))$, (A.6.6) converges in distribution to a multivariate normal distribution with mean $\mathbf{0}$ and variance $\boldsymbol{\Sigma} = \{\mathbf{A}^{-1}(\boldsymbol{\eta}_0)\}\{\mathbf{B}(\boldsymbol{\eta}_0)\}\{\mathbf{A}^{-1}(\boldsymbol{\eta}_0)\}^T$.

A.7 Proof for Varied Variances When Homogeneity Assumption (H2) Is Violated

Under regularity conditions (M1) to (M6) and (F1) to (F10) and given that the homogeneity assumption (H2) is violated, we have $\hat{\boldsymbol{\eta}}_s \xrightarrow{P} \boldsymbol{\eta}_0$, $n_s^{-1}\mathcal{I}_s(\hat{\boldsymbol{\eta}}_s) = \hat{\mathbf{A}}_s(\hat{\boldsymbol{\eta}}_s) \xrightarrow{P} \mathbf{A}_s(\boldsymbol{\eta}_0)$ and $\hat{\mathbf{B}}_s(\hat{\boldsymbol{\eta}}_s) \xrightarrow{P} \mathbf{B}_s(\boldsymbol{\eta}_0)$ as $n_s \to \infty$. Similar to Lemma A.6.1, we have $n_s^{-1}\mathbf{W}_{1s}(\hat{\boldsymbol{\eta}}_s) \xrightarrow{P} \mathbf{A}_s(\boldsymbol{\eta}_0)$, $n_s^{-1}\mathbf{W}_{2s}(\hat{\boldsymbol{\eta}}_s) \xrightarrow{P} \{\mathbf{A}_s^{-1}(\boldsymbol{\eta}_0)\}^T\{\mathbf{B}_s(\boldsymbol{\eta}_0)\}\{\mathbf{A}_s^{-1}(\boldsymbol{\eta}_0)\}$, and $n_s^{-1}\mathbf{W}_{3s}(\hat{\boldsymbol{\eta}}_s) = 1$ as $n_s \to \infty$. Given that $\hat{\boldsymbol{\eta}}^{dc} = \{\sum_{s=1}^{S}(n_s/n)n_s^{-1}\mathbf{W}_s(\hat{\boldsymbol{\eta}})\}^{-1}\sum_{s=1}^{S}(n_s/n)n_s^{-1}\mathbf{W}_s(\hat{\boldsymbol{\eta}})\hat{\boldsymbol{\eta}}_s$, it is shown that the divide-and-combine estimator satisfies $\hat{\boldsymbol{\eta}}^{dc} \xrightarrow{P} \boldsymbol{\eta}_0$ conceding $\hat{\boldsymbol{\eta}}_s \xrightarrow{P} \boldsymbol{\eta}_0$ as $n \to \infty$. Denote $c_s = n_s/n$. With similar arguments in the proof for Theorem 3.3.1, we can see that $n^{1/2}(\hat{\boldsymbol{\eta}}^{dc} - \boldsymbol{\eta}_0)$ converges in distribution to a multivariate normal distribution with mean $\mathbf{0}$, and variance $\boldsymbol{\Sigma}_{\text{w1}}$, $\boldsymbol{\Sigma}_{\text{w2}}$, or $\boldsymbol{\Sigma}_{\text{w3}}$, when $\mathbf{W}_s(.) = \mathbf{W}_{1s}(.)$, $\mathbf{W}_s(.) = \mathbf{W}_{2s}(.)$, or $\mathbf{W}_s(.) = \mathbf{W}_{3s}(.)$, respectively, where

$$
\begin{aligned}
\boldsymbol{\Sigma}_{\text{w1}} &= \lim_{n\to\infty}\left[\left\{\sum_{s=1}^{S} c_s\mathbf{A}_s(\boldsymbol{\beta}_0)\right\}^{-1}\right]\left[\left\{\sum_{s=1}^{S} c_s\mathbf{B}_s(\boldsymbol{\beta}_0)\right\}\right]\left[\left\{\sum_{s=1}^{S} c_s\mathbf{A}_s(\boldsymbol{\beta}_0)\right\}^{-1}\right]^T, \\
\boldsymbol{\Sigma}_{\text{w2}} &= \lim_{n\to\infty}\left\{\sum_{s=1}^{S} c_s\mathbf{A}_s^T(\boldsymbol{\beta}_0)\mathbf{B}_s^{-1}(\boldsymbol{\beta}_0)\mathbf{A}_s(\boldsymbol{\beta}_0)\right\}^{-1}, \\
\boldsymbol{\Sigma}_{\text{w3}} &= \lim_{n\to\infty}\left[\sum_{s=1}^{S} c_s\left\{\mathbf{A}_s^{-1}(\boldsymbol{\beta}_0)\right\}\left\{\mathbf{B}_s(\boldsymbol{\beta}_0)\right\}\left\{\mathbf{A}_s^{-1}(\boldsymbol{\beta}_0)\right\}^T\right].
\end{aligned}
\tag{A.7.1}
$$

Note that $\sum_{s=1}^{S} c_s\mathbf{A}_s(\boldsymbol{\beta}_0) = \mathbf{A}(\boldsymbol{\beta}_0)$ and $\sum_{s=1}^{S} c_s\mathbf{B}_s(\boldsymbol{\beta}_0) = \mathbf{B}(\boldsymbol{\beta}_0)$, we can conclude $\boldsymbol{\Sigma}_{\text{w1}} = \boldsymbol{\Sigma}$. It is shown below that $\boldsymbol{\Sigma}_{\text{w2}} \leq \boldsymbol{\Sigma}_{\text{w1}}$, $\boldsymbol{\Sigma}_{\text{w2}} \leq \boldsymbol{\Sigma}_{\text{w3}}$, and the relation between $\boldsymbol{\Sigma}_{\text{w3}}$ and $\boldsymbol{\Sigma}_{\text{w1}}$ is indeterminate.

We now introduce the following Lemma A.7.1 before showing the inequality of variances $\boldsymbol{\Sigma}_{\text{w2}} \leq \boldsymbol{\Sigma}_{\text{w1}}$.

Lemma A.7.1. For any $d \times d$ square matrices $\mathbf{A}_s$ and $\mathbf{B}_s$ ($s = 1, 2, \ldots, S$) with $\mathbf{B}_s$'s are

positive definite, we have

$$\left(\sum_{s=1}^{S}\mathbf{A}_s\right)\left(\sum_{s=1}^{S}\mathbf{B}_s\right)^{-1}\left(\sum_{s=1}^{S}\mathbf{A}_s\right)^{T} \leq \sum_{s=1}^{S}\mathbf{A}_s\mathbf{B}_s^{-1}\mathbf{A}_s^{T}. \tag{A.7.2}$$

The equality holds if and only if $\mathbf{A}_m\mathbf{B}_m^{-1} = \mathbf{A}_n\mathbf{B}_n^{-1}$ for any m and n ($1 \leq m, n \leq S$).

Proof. We first prove the inequality under a special scenario when $S = 2$, then Lemma A.7.1 follows immediately on the principle of mathematical induction. When $S = 2$, the inequality becomes

$$(\mathbf{A}_1 + \mathbf{A}_2)(\mathbf{B}_1 + \mathbf{B}_2)^{-1}(\mathbf{A}_1 + \mathbf{A}_2)^{T} \leq \mathbf{A}_1\mathbf{B}_1^{-1}\mathbf{A}_1^{T} + \mathbf{A}_2\mathbf{B}_2^{-1}\mathbf{A}_2^{T}, \tag{A.7.3}$$

the equality holds if and only if $\mathbf{A}_1\mathbf{B}_1^{-1} = \mathbf{A}_2\mathbf{B}_2^{-1}$.

Since $\mathbf{B}_s > 0$, we can find a nonsingular matrix $\mathbf{P}$ such that $\mathbf{B}_1 = \mathbf{P}diag\{e_1, e_2, \ldots, e_d\}\mathbf{P}^T$ and $\mathbf{B}_2 = \mathbf{P}diag\{f_1, f_2, \ldots, f_d\}\mathbf{P}^T$, where $e_i > 0$ and $f_i > 0$ ($i = 1, 2, \ldots, d$). By redefining $\mathbf{A}_s$ as $\mathbf{A}_s(\mathbf{P}^T)^{-1}$ ($s = 1, 2$), it suffices to prove (A.7.3) when $\mathbf{B}_1 = diag\{e_1, e_2, \ldots, e_d\}$ and $\mathbf{B}_2 = diag\{f_1, f_2, \ldots, f_d\}$. Let $\mathbf{a}_{si}$ be the i^{th} column of $\mathbf{A}_s$ ($s = 1, 2$ and $i = 1, 2, \ldots, d$), then (A.7.3) becomes

$$\sum_{i=1}^{d}(e_i + f_i)^{-1}(\mathbf{a}_{1i} + \mathbf{a}_{2i})(\mathbf{a}_{1i} + \mathbf{a}_{2i})^{T} \leq \sum_{i=1}^{d}(e_i^{-1}\mathbf{a}_{1i}\mathbf{a}_{1i}^{T} + f_i^{-1}\mathbf{a}_{2i}\mathbf{a}_{2i}^{T}). \tag{A.7.4}$$

The inequality (A.7.4) holds as long as we can show that $(e_i+f_i)^{-1}(\mathbf{a}_{1i}+\mathbf{a}_{2i})(\mathbf{a}_{1i}+\mathbf{a}_{2i})^T \leq e_i^{-1}\mathbf{a}_{1i}\mathbf{a}_{1i}^T + f_i^{-1}\mathbf{a}_{2i}\mathbf{a}_{2i}^T$ or equivalently $e_if_i(\mathbf{a}_{1i}\mathbf{a}_{2i}^T + \mathbf{a}_{2i}\mathbf{a}_{1i}^T) \leq e_i^2\mathbf{a}_{2i}\mathbf{a}_{2i}^T + f_i^2\mathbf{a}_{1i}\mathbf{a}_{1i}^T$ for all i ($i = 1, 2, \ldots, d$). By the union-intersection principle, the desired inequality holds if, for any vector $\boldsymbol{\eta}$,

$$\boldsymbol{\eta}^T\left\{e_if_i(\mathbf{a}_{1i}\mathbf{a}_{2i}^T + \mathbf{a}_{2i}\mathbf{a}_{1i}^T)\right\}\boldsymbol{\eta} \leq \boldsymbol{\eta}^T(e_i^2\mathbf{a}_{2i}\mathbf{a}_{2i}^T + f_i^2\mathbf{a}_{1i}\mathbf{a}_{1i}^T)\boldsymbol{\eta}. \tag{A.7.5}$$

After rearrangement, (A.7.5) can be equivalently written as $2e_if_i(\boldsymbol{\eta}^T\mathbf{a}_{1i})(\boldsymbol{\eta}^T\mathbf{a}_{2i}) \leq$

$(e_i\boldsymbol{\eta}^T\mathbf{a}_{2i})^2+(f_i\boldsymbol{\eta}^T\mathbf{a}_{1i})^2$, which holds from the Cauchy-Schwartz inequality. The foregoing inequality becomes equality if and only if $e_i^{-1}\mathbf{a}_{1i}=f_i^{-1}\mathbf{a}_{2i}$. Thus the equality in (A.7.3) holds if and only if $\mathbf{A}_1\mathbf{B}_1^{-1}=\mathbf{A}_2\mathbf{B}_2^{-1}$. This completes the proof of Lemma A.7.1. ■

Now we show the inequality of variances. Redefine $\mathbf{A}_s$ as $c_s\mathbf{A}_s^T$ and $\mathbf{B}_s$ as $c_s\mathbf{B}_s$, by Lemma A.7.1, we have $(\sum_{s=1}^S c_s\mathbf{A}_s)^T(\sum_{s=1}^S c_s\mathbf{B}_s)^{-1}(\sum_{s=1}^S c_s\mathbf{A}_s)\le\sum_{s=1}^S c_s\mathbf{A}_s^T\mathbf{B}_s^{-1}\mathbf{A}_s$. It follows that $(\sum_{s=1}^S c_s\mathbf{A}_s^T\mathbf{B}_s^{-1}\mathbf{A}_s)^{-1}\le\{(\sum_{s=1}^S c_s\mathbf{A}_s)^T(\sum_{s=1}^S c_s\mathbf{B}_s)^{-1}(\sum_{s=1}^S c_s\mathbf{A}_s)\}^{-1}$, because of $\mathbf{B}_s>0$, which implies $\boldsymbol{\Sigma}_{\mathrm{w}2}\le\boldsymbol{\Sigma}_{\mathrm{w}1}$.

As regards the inequality $\boldsymbol{\Sigma}_{\mathrm{w}2}\le\boldsymbol{\Sigma}_{\mathrm{w}3}$, it suffices to reach the conclusion by showing the following

$$\left(\sum_{s=1}^S c_s\mathbf{A}_s^T\mathbf{B}_s^{-1}\mathbf{A}_s\right)^{-1}\le\sum_{s=1}^S c_s\left(\mathbf{A}_s^T\mathbf{B}_s^{-1}\mathbf{A}_s\right)^{-1}. \tag{A.7.6}$$

After rearrangement, (A.7.6) can be equivalently written as

$$\sum_{s=1}^S c_s^2\mathbf{I}+\sum_{\substack{s,t=1\\s<t}}^S c_sc_t\left\{\left(\mathbf{A}_s^T\mathbf{B}_s^{-1}\mathbf{A}_s\right)\left(\mathbf{A}_t^T\mathbf{B}_t^{-1}\mathbf{A}_t\right)^{-1}+\left(\mathbf{A}_t^T\mathbf{B}_t^{-1}\mathbf{A}_t\right)\left(\mathbf{A}_s^T\mathbf{B}_s^{-1}\mathbf{A}_s\right)^{-1}\right\}\ge\mathbf{I}. \tag{A.7.7}$$

Because $\left(\mathbf{A}_s^T\mathbf{B}_s^{-1}\mathbf{A}_s\right)\left(\mathbf{A}_t^T\mathbf{B}_t^{-1}\mathbf{A}_t\right)^{-1}+\left(\mathbf{A}_t^T\mathbf{B}_t^{-1}\mathbf{A}_t\right)\left(\mathbf{A}_s^T\mathbf{B}_s^{-1}\mathbf{A}_s\right)^{-1}\ge2\mathbf{I}$ for $s,t=1,2,\ldots,S$, in (A.7.7), the left-hand side $\ge(\sum_{s=1}^S c_s^2+\sum_{\substack{s,t=1\\s<t}}^S 2c_sc_t)\mathbf{I}=(\sum_{s=1}^S c_s)^2\mathbf{I}=$ the right-hand side. This completes the proof.

The indeterminate relation between $\boldsymbol{\Sigma}_{\mathrm{w}1}$ and $\boldsymbol{\Sigma}_{\mathrm{w}3}$ is demonstrated by the following example: when $S=2$, $d=2$, and the corresponding matrices are

$$\mathbf{A}_1=\begin{pmatrix}13, & 4\\ 4, & 17\end{pmatrix},\ \mathbf{A}_2=\begin{pmatrix}11, & 4\\ 4, & 9\end{pmatrix},\ \mathbf{B}_1=\begin{pmatrix}14, & 5\\ 5, & 16\end{pmatrix},\ \mathbf{B}_2=\begin{pmatrix}14, & -5\\ -5, & 16\end{pmatrix}, \tag{A.7.8}$$

it can be shown that, in this case, $(\boldsymbol{\Sigma}_{\mathrm{w}1}-\boldsymbol{\Sigma}_{\mathrm{w}3})$ is indefinite (with eigenvalues of 0.00015 and -0.00527). In other words, the relative order between $\boldsymbol{\Sigma}_{\mathrm{w}1}$ and $\boldsymbol{\Sigma}_{\mathrm{w}3}$ is indeterminate.

A.8 Proof of Theorem 4.3.1: Asymptotic Properties of $\hat{\boldsymbol{\beta}}_{\rho}^{dc}$

In the marginal model, $\boldsymbol{\eta} = \boldsymbol{\beta}$. We shall use $\boldsymbol{\beta}$ throughout this proof. Since the objective function $R(\boldsymbol{\beta})$ in (4.3) is a strictly convex function for $\boldsymbol{\beta}$, a local consistent minimizer is the global consistent minimizer. Thus the estimation consistency follows immediately as long as we can show the existence of a local consistent minimizer. Following Fan and Li (2001) and letting $\mathbf{u} = (u_1, u_2, \ldots, u_d)^T$, the existence of a local consistent minimizer is implied by the fact that for any given $\epsilon > 0$, there exists a large constant C such that

$$\lim_{n\to\infty} P\left\{ \inf_{\mathbf{u}\in\mathbb{R}^d:||\mathbf{u}||_2=C} R(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) > R(\boldsymbol{\beta}_0) \right\} > 1-\epsilon, \tag{A.8.1}$$

where $||\mathbf{a}||_2 = (\mathbf{a}^T\mathbf{a})^{1/2}$ for a column vector $\mathbf{a}$.

By the definition of $R(\boldsymbol{\beta})$ in (4.3), some algebraic manipulations show that

$$\begin{aligned}
&R(\boldsymbol{\beta}_0 + n^{-1/2}\mathbf{u}) - R(\boldsymbol{\beta}_0) \\
&= \mathbf{u}^T\hat{\boldsymbol{\Sigma}}_{dc}^{-1}\mathbf{u} + 2\mathbf{u}^T\hat{\boldsymbol{\Sigma}}_{dc}^{-1}\left\{n^{1/2}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}^{dc})\right\} + n\sum_{j=1}^{d}\rho_j(|\beta_{0j} + n^{-1/2}u_j| - |\beta_{0j}|) \\
&\geq \mathbf{u}^T\hat{\boldsymbol{\Sigma}}_{dc}^{-1}\mathbf{u} + 2\mathbf{u}^T\hat{\boldsymbol{\Sigma}}_{dc}^{-1}\left\{n^{1/2}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}^{dc})\right\} + n\sum_{j=1}^{d_0}\rho_j(|\beta_{0j} + n^{-1/2}u_j| - |\beta_{0j}|) \\
&\geq \mathbf{u}^T\hat{\boldsymbol{\Sigma}}_{dc}^{-1}\mathbf{u} + 2\mathbf{u}^T\hat{\boldsymbol{\Sigma}}_{dc}^{-1}\left\{n^{1/2}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}^{dc})\right\} - n\sum_{j=1}^{d_0}\rho_j|n^{-1/2}u_j| \\
&\geq \mathbf{u}^T\hat{\boldsymbol{\Sigma}}_{dc}^{-1}\mathbf{u} + 2\mathbf{u}^T\hat{\boldsymbol{\Sigma}}_{dc}^{-1}\left\{n^{1/2}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}^{dc})\right\} - d_0(n^{1/2}a_n)||\mathbf{u}||_2,
\end{aligned} \tag{A.8.2}$$

followed by $\boldsymbol{\beta}_{0b} = \mathbf{0}$, the triangle inequality, and $a_n = \max\{\rho_j, j \leq d_0\}$, respectively. According to the condition $n^{1/2}a_n \xrightarrow{P} 0$, the third term in (A.8.2) is $o_p(1)$. Based on Theorem 3.3.1, $\hat{\boldsymbol{\beta}}^{dc}$ and $\hat{\boldsymbol{\Sigma}}_{dc}$ are consistent, thus the second term in (A.8.2) is bounded by $2C||\hat{\boldsymbol{\Sigma}}_{dc}^{-1}n^{1/2}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}^{dc})||_2$, which is linear in terms of C with a coefficient $2||\hat{\boldsymbol{\Sigma}}_{dc}^{-1}n^{1/2}(\boldsymbol{\beta}_0 - \hat{\boldsymbol{\beta}}^{dc})||_2 = O_p(1)$. As the variance $\boldsymbol{\Sigma}$ and its estimate $\hat{\boldsymbol{\Sigma}}_{dc}$ are positive semidefinite, the first term in (A.8.2) is larger than $\mu_{min}(\hat{\boldsymbol{\Sigma}}_{dc}^{-1})C^2 \xrightarrow{P} \mu_{min}(\boldsymbol{\Sigma}^{-1})C^2$, where $\mu_{min}(.)$ refers to the

minimal eigenvalue. It follows that, with probability tending to one, the first term in (A.8.2) is larger than $\mu_{min}(\mathbf{\Sigma}^{-1})C^2$ which is quadratic in terms of C. By choosing a sufficiently large C, the first term will dominate the other two terms. Hence, by choosing a sufficiently large C, (A.8.1) holds and the proof of estimation consistency is completed.

The selection consistency can be shown by contradiction. We want to show that $Pr(\hat{\beta}_{\rho_j}^{dc} = 0) \to 1$ for any $d_0 < j \leq d$. Suppose $\hat{\beta}_{\rho_j}^{dc} \neq 0$ for some $d_0 < j \leq d$, then by definition

$$n^{-1/2} \left. \frac{\partial R(\boldsymbol{\beta})}{\partial \beta_j} \right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_{\rho}^{dc}} = 2\hat{\mathbf{\Sigma}}_{dc(j)}^{-1} n^{1/2} (\hat{\boldsymbol{\beta}}_{\rho}^{dc} - \hat{\boldsymbol{\beta}}^{dc}) + n^{1/2} \rho_j sgn(\hat{\boldsymbol{\beta}}_{\rho_j}^{dc}) = 0, \tag{A.8.3}$$

where $\hat{\mathbf{\Sigma}}_{dc(j)}^{-1}$ represents the j^{th} row of $\hat{\mathbf{\Sigma}}_{dc}^{-1}$ and $sgn(.)$ is the sign function. It can be shown that the first term on the right hand side of (A.8.3) is $O_p(1)$. Based on the condition $n^{1/2} b_n \xrightarrow{P} \infty$, we have $n^{1/2}\rho_j \geq n^{1/2} b_n \xrightarrow{P} \infty$. Then to satisfy (A.8.3), with probability tending to one, $\hat{\beta}_{\rho_j}^{dc} = 0$, which however, contradicts the assumed condition that $\hat{\beta}_{\rho_j}^{dc} \neq 0$. As a result, with probability tending to one, $\hat{\beta}_{\rho_j}^{dc} = 0$ for any $d_0 < j \leq d$. This completes the proof of selection consistency.

Before proving the oracle property, for notational ease, we write $\hat{\mathbf{\Sigma}}_{dc} = \hat{\mathbf{\Sigma}}$, suppressing the divide-and-combine notation. We also decompose $\mathbf{\Sigma}$ and $\mathbf{\Sigma}^{-1}$ into block matrices:

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{aa} & \mathbf{\Sigma}_{ab} \\ \mathbf{\Sigma}_{ba} & \mathbf{\Sigma}_{bb} \end{pmatrix}, \quad \mathbf{\Sigma}^{-1} = \mathbf{\Omega} = \begin{pmatrix} \mathbf{\Omega}_{aa} & \mathbf{\Omega}_{ab} \\ \mathbf{\Omega}_{ba} & \mathbf{\Omega}_{bb} \end{pmatrix}, \tag{A.8.4}$$

where $\boldsymbol{M}_{aa}$ is the leading $a \times a$ submatrix of $\boldsymbol{M}$. Decomposing (4.3), we have

$$\begin{aligned} R(\boldsymbol{\beta}) =& n \left\{ \begin{pmatrix} \boldsymbol{\beta}_a \\ \boldsymbol{\beta}_b \end{pmatrix} - \begin{pmatrix} \hat{\boldsymbol{\beta}}_a^{dc} \\ \hat{\boldsymbol{\beta}}_b^{dc} \end{pmatrix} \right\}^T \begin{pmatrix} \hat{\mathbf{\Omega}}_{aa} & \hat{\mathbf{\Omega}}_{ab} \\ \hat{\mathbf{\Omega}}_{ba} & \hat{\mathbf{\Omega}}_{bb} \end{pmatrix} \left\{ \begin{pmatrix} \boldsymbol{\beta}_a \\ \boldsymbol{\beta}_b \end{pmatrix} - \begin{pmatrix} \hat{\boldsymbol{\beta}}_a^{dc} \\ \hat{\boldsymbol{\beta}}_b^{dc} \end{pmatrix} \right\} \\ &+ n \sum_{j=1}^{d_0} \rho_j |\beta_j| + n \sum_{j=d_0+1}^{d} \rho_j |\beta_j|. \end{aligned} \tag{A.8.5}$$

Taking partial derivative of $R(\boldsymbol{\beta})$ and evaluating at the global minimizers, by definition, we have

$$\frac{\partial R(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_a^T}\bigg|_{\boldsymbol{\beta}=\begin{pmatrix}\hat{\boldsymbol{\beta}}_{\rho a}^{dc}\\ \mathbf{0}\end{pmatrix}} = 2n\hat{\boldsymbol{\Omega}}_{aa}(\hat{\boldsymbol{\beta}}_{\rho a}^{dc} - \hat{\boldsymbol{\beta}}_a^{dc}) + 2n\hat{\boldsymbol{\Omega}}_{ab}(\mathbf{0} - \hat{\boldsymbol{\beta}}_b^{dc}) + nD(\hat{\boldsymbol{\beta}}_{\rho a}^{dc}) = 0, \quad \text{(A.8.6)}$$

where $D(\hat{\boldsymbol{\beta}}_{\rho a}^{dc}) = (\rho_1 sgn(\hat{\beta}_{\rho_1}^{dc}), \rho_2 sgn(\hat{\beta}_{\rho_2}^{dc}), \ldots, \rho_{d_0} sgn(\hat{\beta}_{\rho_{d_0}}^{dc}))^T$. Reorganize (A.8.6), we have $\hat{\boldsymbol{\beta}}_{\rho a}^{dc} = \hat{\boldsymbol{\beta}}_a^{dc} + (\hat{\boldsymbol{\Omega}}_{aa})^{-1}\hat{\boldsymbol{\Omega}}_{ab}\hat{\boldsymbol{\beta}}_b^{dc} - 1/2(\hat{\boldsymbol{\Omega}}_{aa})^{-1}D(\hat{\boldsymbol{\beta}}_{\rho a}^{dc})$, which leads to

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{\rho a}^{dc} - \boldsymbol{\beta}_{0a}) = n^{1/2}(\hat{\boldsymbol{\beta}}_a^{dc} - \boldsymbol{\beta}_{0a}) + (\hat{\boldsymbol{\Omega}}_{aa})^{-1}\hat{\boldsymbol{\Omega}}_{ab}(n^{1/2}\hat{\boldsymbol{\beta}}_b^{dc}) - 1/2(\hat{\boldsymbol{\Omega}}_{aa})^{-1}n^{1/2}D(\hat{\boldsymbol{\beta}}_{\rho a}^{dc}). \quad \text{(A.8.7)}$$

According to the condition $n^{1/2}a_n \xrightarrow{P} 0$, we have $n^{1/2}\rho_j \leq n^{1/2}a_n \xrightarrow{P} 0$. Thus the third term in (A.8.7) is $o_p(1)$. Then, we can rewrite (A.8.7) as

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{\rho a}^{dc} - \boldsymbol{\beta}_{0a}) = \left\{1, (\hat{\boldsymbol{\Omega}}_{aa})^{-1}\hat{\boldsymbol{\Omega}}_{ab}\right\} \cdot n^{1/2}\begin{pmatrix}\hat{\boldsymbol{\beta}}_a^{dc} - \boldsymbol{\beta}_{0a}\\ \hat{\boldsymbol{\beta}}_b^{dc} - \mathbf{0}\end{pmatrix} + o_p(1). \quad \text{(A.8.8)}$$

Given that

$$n^{1/2}\begin{pmatrix}\hat{\boldsymbol{\beta}}_a^{dc} - \boldsymbol{\beta}_{0a}\\ \hat{\boldsymbol{\beta}}_b^{dc} - \mathbf{0}\end{pmatrix} \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \begin{pmatrix}\boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab}\\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb}\end{pmatrix}\right), \quad \text{(A.8.9)}$$

and that $\hat{\boldsymbol{\Omega}}_{aa} \xrightarrow{P} \boldsymbol{\Omega}_{aa}$, $\hat{\boldsymbol{\Omega}}_{ab} \xrightarrow{P} \boldsymbol{\Omega}_{ab}$, (A.8.8) can be derived into

$$n^{1/2}(\hat{\boldsymbol{\beta}}_{\rho a}^{dc} - \boldsymbol{\beta}_{0a}) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \left\{1, (\boldsymbol{\Omega}_{aa})^{-1}\boldsymbol{\Omega}_{ab}\right\}\begin{pmatrix}\boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab}\\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb}\end{pmatrix}\left\{1, (\boldsymbol{\Omega}_{aa})^{-1}\boldsymbol{\Omega}_{ab}\right\}^T\right). \quad \text{(A.8.10)}$$

Providing the fact that

$$\boldsymbol{\Omega} = \begin{pmatrix}\boldsymbol{\Omega}_{aa} & \boldsymbol{\Omega}_{ab}\\ \boldsymbol{\Omega}_{ba} & \boldsymbol{\Omega}_{bb}\end{pmatrix} = \begin{pmatrix}\boldsymbol{N} & -\boldsymbol{N}\boldsymbol{\Sigma}_{ab}(\boldsymbol{\Sigma}_{bb})^{-1}\\ -(\boldsymbol{\Sigma}_{bb})^{-1}\boldsymbol{\Sigma}_{ba}\boldsymbol{N} & (\boldsymbol{\Sigma}_{bb})^{-1} + (\boldsymbol{\Sigma}_{bb})^{-1}\boldsymbol{\Sigma}_{ba}\boldsymbol{N}\boldsymbol{\Sigma}_{ab}(\boldsymbol{\Sigma}_{bb})^{-1}\end{pmatrix}, \quad \text{(A.8.11)}$$

where $\boldsymbol{N} = (\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}(\boldsymbol{\Sigma}_{bb})^{-1}\boldsymbol{\Sigma}_{ba})^{-1}$, the proof of the oracle property is completed by verifying that

$$
\begin{aligned}
&\left\{1, (\boldsymbol{\Omega}_{aa})^{-1}\boldsymbol{\Omega}_{ab}\right\} \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix} \left\{1, (\boldsymbol{\Omega}_{aa})^{-1}\boldsymbol{\Omega}_{ab}\right\}^T \\
&= \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\left(\boldsymbol{\Sigma}_{bb}\right)^{-1}\boldsymbol{\Sigma}_{ba} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{ab}\left(\boldsymbol{\Sigma}_{bb}\right)^{-1} + \boldsymbol{\Sigma}_{ab}\left(\boldsymbol{\Sigma}_{bb}\right)^{-1}\boldsymbol{\Sigma}_{bb}\boldsymbol{\Sigma}_{ab}\left(\boldsymbol{\Sigma}_{bb}\right)^{-1} \qquad \text{(A.8.12)} \\
&= \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\left(\boldsymbol{\Sigma}_{bb}\right)^{-1}\boldsymbol{\Sigma}_{ba} \\
&= \left(\left[\boldsymbol{\Sigma}^{-1}\right]_{aa}\right)^{-1}.
\end{aligned}
$$

A.9 Proof of Theorem 4.3.2: Asymptotic Properties of $\hat{\boldsymbol{\eta}}_{\rho}^{dc}$

Theorem 4.3.2 for the frailty model can be established along the same lines in the proof of Theorem 4.3.1 for the marginal model. Note that in the frailty model, $\boldsymbol{\eta} = (\theta, \boldsymbol{\beta}^T)^T$. Following the same arguments, the estimation consistency is implied by the fact that for any given $\epsilon > 0$, there exists a large constant C such that

$$\lim_{n\to\infty} P\left\{ \inf_{\mathbf{u}\in\mathbb{R}^{1+d}:||\mathbf{u}||_2=C} R(\boldsymbol{\eta}_0 + n^{-1/2}\mathbf{u}) > R(\boldsymbol{\eta}_0) \right\} > 1-\epsilon, \tag{A.9.1}$$

where $||\mathbf{a}||_2 = (\mathbf{a}^T\mathbf{a})^{1/2}$ for a column vector $\mathbf{a}$, and $\mathbf{u} = (u_1, u_2, \ldots, u_{(1+d)})^T$.

Similarly, based on $R(\boldsymbol{\eta})$ in (4.2), we can derive that

$$\begin{aligned}
&R(\boldsymbol{\eta}_0 + n^{-1/2}\mathbf{u}) - R(\boldsymbol{\eta}_0) \\
&= \mathbf{u}^T\hat{\boldsymbol{\Sigma}}_{dc}^{-1}\mathbf{u} + 2\mathbf{u}^T\hat{\boldsymbol{\Sigma}}_{dc}^{-1}\left\{n^{1/2}(\boldsymbol{\eta}_0 - \hat{\boldsymbol{\eta}}^{dc})\right\} + n\sum_{j=1}^{d}\rho_j(|\beta_{0j} + n^{-1/2}u_j| - |\beta_{0j}|) \\
&\geq \mathbf{u}^T\hat{\boldsymbol{\Sigma}}_{dc}^{-1}\mathbf{u} + 2\mathbf{u}^T\hat{\boldsymbol{\Sigma}}_{dc}^{-1}\left\{n^{1/2}(\boldsymbol{\eta}_0 - \hat{\boldsymbol{\eta}}^{dc})\right\} + n\sum_{j=1}^{d_0}\rho_j(|\beta_{0j} + n^{-1/2}u_j| - |\beta_{0j}|) \\
&\geq \mathbf{u}^T\hat{\boldsymbol{\Sigma}}_{dc}^{-1}\mathbf{u} + 2\mathbf{u}^T\hat{\boldsymbol{\Sigma}}_{dc}^{-1}\left\{n^{1/2}(\boldsymbol{\eta}_0 - \hat{\boldsymbol{\eta}}^{dc})\right\} - n\sum_{j=1}^{d_0}\rho_j|n^{-1/2}u_j| \\
&\geq \mathbf{u}^T\hat{\boldsymbol{\Sigma}}_{dc}^{-1}\mathbf{u} + 2\mathbf{u}^T\hat{\boldsymbol{\Sigma}}_{dc}^{-1}\left\{n^{1/2}(\boldsymbol{\eta}_0 - \hat{\boldsymbol{\eta}}^{dc})\right\} - d_0(n^{1/2}a_n)||\mathbf{u}||_2,
\end{aligned} \tag{A.9.2}$$

followed by $\boldsymbol{\beta}_{0b} = \mathbf{0}$, the triangle inequality, and $a_n = \max\{\rho_j, j \leq d_0\}$, respectively. Due to the same arguments in the proof of Theorem 4.3.1, it can be shown that, with probability tending to one, the first term in (A.9.2) is larger than a quadratic term of C, and that by choosing a sufficiently large C the first term will dominate the other two terms. This completes the proof of estimation consistency.

The selection consistency can be shown by contradiction. We want to show that $Pr(\hat{\beta}_{\rho_j}^{dc} =$

$0) \to 1$ for any $d_0 < j \leq d$. Suppose $\hat{\beta}_{\rho_j}^{dc} \neq 0$ for some $d_0 < j \leq d$, then by definition

$$n^{-1/2} \left. \frac{\partial R(\boldsymbol{\eta})}{\partial \beta_j} \right|_{\boldsymbol{\eta}=\hat{\boldsymbol{\eta}}_{\rho}^{dc}} = 2\hat{\boldsymbol{\Sigma}}_{dc(1+j)}^{-1} n^{1/2} (\hat{\boldsymbol{\eta}}_{\rho}^{dc} - \hat{\boldsymbol{\eta}}^{dc}) + n^{1/2} \rho_j sgn(\hat{\boldsymbol{\beta}}_{\rho_j}^{dc}) = 0, \tag{A.9.3}$$

where $\hat{\boldsymbol{\Sigma}}_{dc(1+j)}^{-1}$ represents the $(1+j)^{th}$ row of $\hat{\boldsymbol{\Sigma}}_{dc}^{-1}$ and $sgn(.)$ is the sign function. Following the same rationale in the proof of Theorem 4.3.1, we can show that, with probability tending to one, $\hat{\beta}_{\rho_j}^{dc} = 0$, for any $d_0 < j \leq d$. This completes the proof of selection consistency.

As in the proof of Theorem 4.3.1, we write $\hat{\boldsymbol{\Sigma}}_{dc} = \hat{\boldsymbol{\Sigma}}$ and decompose $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$ into block matrices:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}, \quad \boldsymbol{\Sigma}^{-1} = \boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{aa} & \boldsymbol{\Omega}_{ab} \\ \boldsymbol{\Omega}_{ba} & \boldsymbol{\Omega}_{bb} \end{pmatrix}, \tag{A.9.4}$$

where $\boldsymbol{M}_{aa}$ is the leading $a \times a$ submatrix of $\boldsymbol{M}$. Decomposing (4.2), we have

$$\begin{aligned} R(\boldsymbol{\eta}) =& n \left\{ \begin{pmatrix} \boldsymbol{\eta}_a \\ \boldsymbol{\beta}_b \end{pmatrix} - \begin{pmatrix} \hat{\boldsymbol{\eta}}_a^{dc} \\ \hat{\boldsymbol{\beta}}_b^{dc} \end{pmatrix} \right\}^T \begin{pmatrix} \hat{\boldsymbol{\Omega}}_{aa} & \hat{\boldsymbol{\Omega}}_{ab} \\ \hat{\boldsymbol{\Omega}}_{ba} & \hat{\boldsymbol{\Omega}}_{bb} \end{pmatrix} \left\{ \begin{pmatrix} \boldsymbol{\eta}_a \\ \boldsymbol{\beta}_b \end{pmatrix} - \begin{pmatrix} \hat{\boldsymbol{\eta}}_a^{dc} \\ \hat{\boldsymbol{\beta}}_b^{dc} \end{pmatrix} \right\} \\ &+ n \sum_{j=1}^{d_0} \rho_j |\beta_j| + n \sum_{j=d_0+1}^{d} \rho_j |\beta_j|. \end{aligned} \tag{A.9.5}$$

Taking partial derivative of $R(\boldsymbol{\eta})$ and evaluating at the global minimizers, by definition, we have

$$\left. \frac{\partial R(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}_a^T} \right|_{\boldsymbol{\eta}=\begin{pmatrix} \hat{\boldsymbol{\eta}}_{\rho a}^{dc} \\ \mathbf{0} \end{pmatrix}} = 2n\hat{\boldsymbol{\Omega}}_{aa}(\hat{\boldsymbol{\eta}}_{\rho a}^{dc} - \hat{\boldsymbol{\eta}}_a^{dc}) + 2n\hat{\boldsymbol{\Omega}}_{ab}(\mathbf{0} - \hat{\boldsymbol{\beta}}_b^{dc}) + nD(\hat{\boldsymbol{\beta}}_{\rho a}^{dc}) = 0, \tag{A.9.6}$$

where $D(\hat{\boldsymbol{\beta}}_{\rho a}^{dc}) = (\rho_1 sgn(\hat{\beta}_{\rho_1}^{dc}), \rho_2 sgn(\hat{\beta}_{\rho_2}^{dc}), \ldots, \rho_{d_0} sgn(\hat{\beta}_{\rho_{d_0}}^{dc}))^T$. Reorganize (A.9.6), we

have $\hat{\boldsymbol{\eta}}_{\rho a}^{dc} = \hat{\boldsymbol{\eta}}_{a}^{dc} + (\hat{\boldsymbol{\Omega}}_{aa})^{-1}\hat{\boldsymbol{\Omega}}_{ab}\hat{\boldsymbol{\beta}}_{b}^{dc} - 1/2(\hat{\boldsymbol{\Omega}}_{aa})^{-1}D(\hat{\boldsymbol{\beta}}_{\rho a}^{dc})$, which leads to

$$n^{1/2}(\hat{\boldsymbol{\eta}}_{\rho a}^{dc} - \boldsymbol{\eta}_{0a}) = n^{1/2}(\hat{\boldsymbol{\eta}}_{a}^{dc} - \boldsymbol{\eta}_{0a}) + (\hat{\boldsymbol{\Omega}}_{aa})^{-1}\hat{\boldsymbol{\Omega}}_{ab}(n^{1/2}\hat{\boldsymbol{\beta}}_{b}^{dc}) - 1/2(\hat{\boldsymbol{\Omega}}_{aa})^{-1}n^{1/2}D(\hat{\boldsymbol{\beta}}_{\rho a}^{dc}). \tag{A.9.7}$$

The condition $n^{1/2}a_n \xrightarrow{P} 0$ guarantees the third term in (A.9.7) is $o_p(1)$. Thus, (A.9.7) can be rewritten as

$$n^{1/2}(\hat{\boldsymbol{\eta}}_{\rho a}^{dc} - \boldsymbol{\eta}_{0a}) = \left\{1, (\hat{\boldsymbol{\Omega}}_{aa})^{-1}\hat{\boldsymbol{\Omega}}_{ab}\right\} \cdot n^{1/2} \begin{pmatrix} \hat{\boldsymbol{\eta}}_{a}^{dc} - \boldsymbol{\eta}_{0a} \\ \hat{\boldsymbol{\beta}}_{b}^{dc} - \mathbf{0} \end{pmatrix} + o_p(1). \tag{A.9.8}$$

Given that

$$n^{1/2} \begin{pmatrix} \hat{\boldsymbol{\eta}}_{a}^{dc} - \boldsymbol{\eta}_{0a} \\ \hat{\boldsymbol{\beta}}_{b}^{dc} - \mathbf{0} \end{pmatrix} \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}\right), \tag{A.9.9}$$

and that $\hat{\boldsymbol{\Omega}}_{aa} \xrightarrow{P} \boldsymbol{\Omega}_{aa}$, $\hat{\boldsymbol{\Omega}}_{ab} \xrightarrow{P} \boldsymbol{\Omega}_{ab}$, (A.9.8) can be derived into

$$n^{1/2}(\hat{\boldsymbol{\eta}}_{\rho a}^{dc} - \boldsymbol{\eta}_{0a}) \xrightarrow{D} \mathcal{N}\left(\mathbf{0}, \left\{1, (\boldsymbol{\Omega}_{aa})^{-1}\boldsymbol{\Omega}_{ab}\right\} \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix} \left\{1, (\boldsymbol{\Omega}_{aa})^{-1}\boldsymbol{\Omega}_{ab}\right\}^T\right). \tag{A.9.10}$$

As in the proof of Theorem 4.3.1, providing the fact that

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{aa} & \boldsymbol{\Omega}_{ab} \\ \boldsymbol{\Omega}_{ba} & \boldsymbol{\Omega}_{bb} \end{pmatrix} = \begin{pmatrix} \boldsymbol{N} & -\boldsymbol{N}\boldsymbol{\Sigma}_{ab}(\boldsymbol{\Sigma}_{bb})^{-1} \\ -(\boldsymbol{\Sigma}_{bb})^{-1}\boldsymbol{\Sigma}_{ba}\boldsymbol{N} & (\boldsymbol{\Sigma}_{bb})^{-1} + (\boldsymbol{\Sigma}_{bb})^{-1}\boldsymbol{\Sigma}_{ba}\boldsymbol{N}\boldsymbol{\Sigma}_{ab}(\boldsymbol{\Sigma}_{bb})^{-1} \end{pmatrix}, \tag{A.9.11}$$

where $\boldsymbol{N} = (\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}(\boldsymbol{\Sigma}_{bb})^{-1}\boldsymbol{\Sigma}_{ba})^{-1}$, the proof of the oracle property is completed by verifying that

$$\left\{1, (\boldsymbol{\Omega}_{aa})^{-1}\boldsymbol{\Omega}_{ab}\right\} \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix} \left\{1, (\boldsymbol{\Omega}_{aa})^{-1}\boldsymbol{\Omega}_{ab}\right\}^T = \left(\left[\boldsymbol{\Sigma}^{-1}\right]_{aa}\right)^{-1}. \tag{A.9.12}$$

A.10 Proof of Theorem 4.4.2: Asymptotic Equivalence between $R_z(\boldsymbol{\beta})$ and $R(\boldsymbol{\beta})/2$

We apply the Taylor series expansion on $R_z(\boldsymbol{\beta})$ and obtain that

$$\begin{aligned}
R_z(\boldsymbol{\beta}) =& - log\mathcal{PL}(\boldsymbol{\beta}) + n\sum_{j=1}^{d}\rho_j^z|\beta_j| \\
=& - log\mathcal{PL}(\hat{\boldsymbol{\beta}}^{full}) - (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^{full})^T\left\{\frac{\partial}{\partial\boldsymbol{\beta}^T}log\mathcal{PL}(\hat{\boldsymbol{\beta}}^{full})\right\} \\
& - 1/2(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^{full})^T\left\{\frac{\partial^2}{\partial\boldsymbol{\beta}^T\partial\boldsymbol{\beta}}log\mathcal{PL}(\hat{\boldsymbol{\beta}}^{full})\right\}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^{full}) \\
& + n\sum_{j=1}^{d}\rho_j^z|\beta_j| + o_p(1). \qquad \text{(A.10.1)}
\end{aligned}$$

Notice that $\partial log\mathcal{PL}(\hat{\boldsymbol{\beta}}^{full})/\partial\boldsymbol{\beta}^T = 0$, $-\partial^2 log\mathcal{PL}(\hat{\boldsymbol{\beta}}^{full})/\partial\boldsymbol{\beta}^T\partial\boldsymbol{\beta} = \mathcal{I}(\hat{\boldsymbol{\beta}}^{full})$, and that $log\mathcal{PL}(\hat{\boldsymbol{\beta}}^{full})$ is a constant. Of note, in the univariate survival analysis ($K=1$), $\mathcal{I}(\hat{\boldsymbol{\beta}}^{full}) = n\hat{\boldsymbol{\Sigma}}_{full}^{-1}$, but in the multivariate survival analysis ($K>1$), $\mathcal{I}(\hat{\boldsymbol{\beta}}^{full})$ generally does not equal $n\hat{\boldsymbol{\Sigma}}_{full}^{-1}$ when the intra-cluster association is non-trivial. We can rewrite (A.10.1) as

$$\begin{aligned}
R_z(\boldsymbol{\beta}) =& -log\mathcal{PL}(\boldsymbol{\beta}) + n\sum_{j=1}^{d}\rho_j^z|\beta_j| \\
=& \; n/2(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^{full})^T\hat{\boldsymbol{\Sigma}}_{full}^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^{full}) + n\sum_{j=1}^{d}\rho_j^z|\beta_j| + o_p(1) \\
=& \; n/2(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^{dc})^T\hat{\boldsymbol{\Sigma}}_{dc}^{-1}(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^{dc}) + n/2\sum_{j=1}^{d}\rho_j|\beta_j| + o_p(1) \\
=& \; Q(\boldsymbol{\beta})/2 + o_p(1). \qquad \text{(A.10.2)}
\end{aligned}$$

The third equality is due to the asymptotic equivalence between $\hat{\boldsymbol{\beta}}^{full}$ and $\hat{\boldsymbol{\beta}}^{dc}$ (Theorem 3.3.1) and providing that $\rho_j^z = \rho_j/2$ for $j = 1, 2, \ldots, d$. This completes the proof of Theorem 4.4.2.

A.11 Proofs of Theorems 5.3.1 and 5.3.2: Asymptotic Properties of $\hat{\Lambda}_{hj}^{dc}(t|\boldsymbol{z}_{0,hj})$

under regularity conditions in Section 2.2.3, Theorem 2.2.1 ensures that in $t \in [0, \tau]$, as $n \to \infty$,

$$\sup_{t\in[0,\tau]} \left|\hat{\Lambda}_{hj}(t|\boldsymbol{z}_{0,hj}) - \Lambda_{hj}(t|\boldsymbol{z}_{0,hj})\right| \xrightarrow{P} 0. \tag{A.11.1}$$

By the triangle inequality, we have

$$\left|\hat{\Lambda}_{hj}^{dc}(t|\boldsymbol{z}_{0,hj}) - \Lambda_{hj}(t|\boldsymbol{z}_{0,hj})\right| < n^{-1}\sum_{s=1}^{S} n_s \left|\hat{\Lambda}_{s,hj}(t|\boldsymbol{z}_{0,hj}) - \Lambda_{hj}(t|\boldsymbol{z}_{0,hj})\right|, \tag{A.11.2}$$

thus it follows that

$$\sup_{t\in[0,\tau]} \left|\hat{\Lambda}_{hj}^{dc}(t|\boldsymbol{z}_{0,hj}) - \Lambda_{hj}(t|\boldsymbol{z}_{0,hj})\right| < n^{-1}\sum_{s=1}^{S} n_s \sup_{t\in[0,\tau]} \left|\hat{\Lambda}_{s,hj}(t|\boldsymbol{z}_{0,hj}) - \Lambda_{hj}(t|\boldsymbol{z}_{0,hj})\right| \xrightarrow{P} 0. \tag{A.11.3}$$

This completes the proof for uniform consistency in Theorem 5.3.1.

Before showing the asymptotic normality of $\hat{\Lambda}_{hj}^{dc}(t|\boldsymbol{z}_{0,hj})$, following the proof of Theorem 2.2.2, we recognize that $n^{1/2}\left\{\hat{\Lambda}_{s,hj}(t|\boldsymbol{z}_{0,hj}) - \Lambda_{hj}(t|\boldsymbol{z}_{0,hj})\right\}$ can be written into two asymptotically independent terms, i.e.,

$$\begin{aligned}
&n_s^{1/2}\left\{\hat{\Lambda}_{s,hj}(t|\boldsymbol{z}_{0,hj}) - \Lambda_{hj}(t|\boldsymbol{z}_{0,hj})\right\} \\
&\quad = n_s^{1/2}\sum_{i=1}^{n_s}\int_0^t \frac{e^{\boldsymbol{\beta}_0^T\boldsymbol{z}_{0,hj}} dM_{si,hj}(u)}{n_s S_{s,hj}^{(0)}(\boldsymbol{\beta}_0, u)} \\
&\qquad + \left[\sum_{i=1}^{n_s}\int_0^t \frac{\{\boldsymbol{z}_{0,hj} - \mathbf{E}_{s,hj}(\boldsymbol{\beta}^*, u)\}^T e^{\boldsymbol{\beta}^{*T}\boldsymbol{z}_{0,hj}} dN_{si,hj}(u)}{n_s S_{s,hj}^{(0)}(\boldsymbol{\beta}^*, u)}\right] n_s^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),
\end{aligned} \tag{A.11.4}$$

where $\boldsymbol{\beta}^*$ lies between $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}_s$, and $M_{si,hj}(t) = N_{si,hj}(t) - \int_0^t Y_{si,h}(u)e^{\boldsymbol{\beta}_0^T\boldsymbol{Z}_{si,hj}}\lambda_{0,hj}(u)du$. Note that $N_{si,hj}(t)$, $S_{s,hj}^{(0)}(\boldsymbol{\beta}, t)$, and $\mathbf{E}_{s,hj}(\boldsymbol{\beta}, t)$ are the counterparts in the s^{th} subset of $N_{i,hj}(t)$, $S_{hj}^{(0)}(\boldsymbol{\beta}, t)$, and $\mathbf{E}_{hj}(\boldsymbol{\beta}, t)$ (see definitions in Section 2.2.3).

By using (A.6.5) and (A.11.4), we can write

$$
\begin{aligned}
& n^{1/2}\left\{\hat{\Lambda}_{hj}^{dc}(t|\boldsymbol{z}_{0,hj})-\Lambda_{hj}(t|\boldsymbol{z}_{0,hj})\right\} \\
&\quad=\sum_{s=1}^{S}(n_s/n)^{1/2}n_s^{1/2}\left\{\hat{\Lambda}_{s,hj}(t|\boldsymbol{z}_{0,hj})-\Lambda_{hj}(t|\boldsymbol{z}_{0,hj})\right\} \\
&\quad=\sum_{s=1}^{S}(n_s/n)^{1/2}\left\{n_s^{1/2}\sum_{i=1}^{n_s}\int_0^t\frac{e^{\boldsymbol{\beta}_0^T\boldsymbol{z}_{0,hj}}dM_{si,hj}(u)}{n_sS_{s,hj}^{(0)}(\boldsymbol{\beta}_0,u)}\right\} \\
&\qquad+\sum_{s=1}^{S}(n_s/n)^{1/2}\left(\left[\sum_{i=1}^{n_s}\int_0^t\frac{\left\{\boldsymbol{z}_{0,hj}-\mathbf{E}_{s,hj}(\boldsymbol{\beta}^*,u)\right\}^T e^{\boldsymbol{\beta}^{*T}\boldsymbol{z}_{0,hj}}dN_{si,hj}(u)}{n_sS_{s,hj}^{(0)}(\boldsymbol{\beta}^*,u)}\right]\right. \\
&\qquad\qquad\left.\times\boldsymbol{\Sigma}_3n_s^{-1/2}\mathbf{U}_s(\boldsymbol{\beta}_0)\right) \\
&\qquad+O_p(n^{-1/2}S).
\end{aligned}
\tag{A.11.5}
$$

Because $S=o(n^{1/2})$, $O_p(n^{-1/2}S)=o_p(1)$, and as $n\to\infty$, $n_s^{-1/2}\mathbf{U}_s(\boldsymbol{\beta}_0)\xrightarrow{D}\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}_3^{-1})$. Note that here $\boldsymbol{\Sigma}_3$ is assumed to be identical across subsets, which is generally true under homogeneity assumptions (C1) to (C3) in Section 5.3. Then the asymptotic normality in Theorem 5.3.2 follows.

A.12 True Transition Probabilities in Proposed Five-State Model

The transition-specific hazard function for a subject with covariate $\boldsymbol{z}_0$ is

$$\lambda_{jk}(t|\boldsymbol{z}_{0,jk}) = \xi_{jk} h_{jk}^{\xi_{jk}} t^{\xi_{jk}-1} e^{\boldsymbol{\beta}^T \boldsymbol{z}_{0,jk}}, \tag{A.12.1}$$

for $j, k = 1, 2, \ldots, 5$ and $j \neq k$. Note that $\boldsymbol{z}_0$ is the basic covariate that can be extended into the transition-specific structure $\boldsymbol{z}_{0,jk}$. For notational ease, we write the transition-specific hazard function as $\lambda_{jk}(t) = a_{jk} t^{b_{jk}}$. Of note, this hazard function can be easily converted into (A.12.1) by letting $a_{jk} = \xi_{jk} h_{jk}^{\xi_{jk}} e^{\boldsymbol{\beta}^T \boldsymbol{z}_{0,jk}}$ and $b_{jk} = \xi_{jk} - 1$. Moreover, when deriving the true transition probabilities $\mathbf{P}(s, t|\boldsymbol{z}_0)$ for the proposed five-state model considered in the simulation studies, we suppress $\boldsymbol{z}_0$, and use $\lambda_{jk}(t)$ and $\mathbf{P}(s, t)$ in the following.

Since transition probabilities starting from state 1 are evaluated in the simulation studies, we only provide their true probabilities here. Note that most of the true transition probabilities have no analytic solutions and thus should be evaluated numerically.

Given the Kolmogorov forward equation

$$\mathbf{P}(s, t) = \mathbf{I} + \int_s^t \mathbf{P}(s, u) d\boldsymbol{\Lambda}(u) \tag{A.12.2}$$

and motivated by the results of $P_{11}(s, t)$ in (2.41) and $P_{12}(s, t)$ in (2.43), we derive

$$\begin{aligned} P_{11}(s, t) &= \exp\left[-\int_s^t \{\lambda_{12}(u) + \lambda_{13}(u) + \lambda_{15}(u)\} \, du \right] \\ &= \exp\left(-\frac{a_{12}}{b_{12}+1} t^{b_{12}+1} - \frac{a_{13}}{b_{13}+1} t^{b_{13}+1} - \frac{a_{15}}{b_{15}+1} t^{b_{15}+1} \right. \\ &\qquad \left. + \frac{a_{12}}{b_{12}+1} s^{b_{12}+1} + \frac{a_{13}}{b_{13}+1} s^{b_{13}+1} + \frac{a_{15}}{b_{15}+1} s^{b_{15}+1} \right), \end{aligned} \tag{A.12.3}$$

$$
\begin{aligned}
P_{12}(s,t) &= \int_s^t P_{11}(s,u)\lambda_{12}(u)P_{22}(u,t)du \\
&= \int_s^t \exp\left(-\frac{a_{12}}{b_{12}+1}u^{b_{12}+1} - \frac{a_{13}}{b_{13}+1}u^{b_{13}+1} - \frac{a_{15}}{b_{15}+1}u^{b_{15}+1}\right. \\
&\qquad \left. + \frac{a_{12}}{b_{12}+1}s^{b_{12}+1} + \frac{a_{13}}{b_{13}+1}s^{b_{13}+1} + \frac{a_{15}}{b_{15}+1}s^{b_{15}+1}\right) \\
&\qquad \times a_{12}u^{b_{12}} \\
&\qquad \times \exp\left(-\frac{a_{24}}{b_{24}+1}t^{b_{24}+1} - \frac{a_{25}}{b_{25}+1}t^{b_{25}+1}\right. \\
&\qquad\qquad \left. + \frac{a_{24}}{b_{24}+1}u^{b_{24}+1} + \frac{a_{25}}{b_{25}+1}u^{b_{25}+1}\right) du,
\end{aligned}
\tag{A.12.4}
$$

$$
\begin{aligned}
P_{13}(s,t) &= \int_s^t P_{11}(s,u)\lambda_{13}(u)P_{33}(u,t)du \\
&= \int_s^t \exp\left(-\frac{a_{12}}{b_{12}+1}u^{b_{12}+1} - \frac{a_{13}}{b_{13}+1}u^{b_{13}+1} - \frac{a_{15}}{b_{15}+1}u^{b_{15}+1}\right. \\
&\qquad \left. + \frac{a_{12}}{b_{12}+1}s^{b_{12}+1} + \frac{a_{13}}{b_{13}+1}s^{b_{13}+1} + \frac{a_{15}}{b_{15}+1}s^{b_{15}+1}\right) \\
&\qquad \times a_{13}u^{b_{13}} \\
&\qquad \times \exp\left(-\frac{a_{34}}{b_{34}+1}t^{b_{34}+1} - \frac{a_{35}}{b_{35}+1}t^{b_{35}+1}\right. \\
&\qquad\qquad \left. + \frac{a_{34}}{b_{34}+1}u^{b_{34}+1} + \frac{a_{35}}{b_{35}+1}u^{b_{35}+1}\right) du,
\end{aligned}
\tag{A.12.5}
$$

$$
\begin{aligned}
P_{24}(s,t) &= \int_s^t P_{22}(s,u)\lambda_{24}(u)P_{44}(u,t)du \\
&= \int_s^t \exp\left(-\frac{a_{24}}{b_{24}+1}u^{b_{24}+1} - \frac{a_{25}}{b_{25}+1}u^{b_{25}+1}\right. \\
&\qquad \left. + \frac{a_{24}}{b_{24}+1}s^{b_{24}+1} + \frac{a_{25}}{b_{25}+1}s^{b_{25}+1}\right) \\
&\qquad \times a_{24}u^{b_{24}} \\
&\qquad \times \exp\left(-\frac{a_{45}}{b_{45}+1}t^{b_{45}+1} + \frac{a_{45}}{b_{45}+1}u^{b_{45}+1}\right) du,
\end{aligned}
\tag{A.12.6}
$$

$$
\begin{aligned}
P_{34}(s,t) &= \int_s^t P_{33}(s,u)\lambda_{34}(u)P_{44}(u,t)du \\
&= \int_s^t \exp\left(-\frac{a_{34}}{b_{34}+1}u^{b_{34}+1} - \frac{a_{35}}{b_{35}+1}u^{b_{35}+1}\right. \\
&\qquad \left. + \frac{a_{34}}{b_{34}+1}s^{b_{34}+1} + \frac{a_{35}}{b_{35}+1}s^{b_{35}+1}\right) \\
&\qquad \times a_{34}u^{b_{34}} \\
&\qquad \times \exp\left(-\frac{a_{45}}{b_{45}+1}t^{b_{45}+1} + \frac{a_{45}}{b_{45}+1}u^{b_{45}+1}\right)du,
\end{aligned}
\tag{A.12.7}
$$

$$
\begin{aligned}
P_{45}(s,t) &= \int_s^t P_{44}(s,u)\lambda_{45}(u)P_{55}(u,t)du \\
&= \int_s^t \exp\left(-\frac{a_{45}}{b_{45}+1}u^{b_{45}+1} + \frac{a_{45}}{b_{45}+1}s^{b_{45}+1}\right) \times a_{45}u^{b_{45}}du.
\end{aligned}
\tag{A.12.8}
$$

Then by using $P_{45}(s,t)$, we have

$$
\begin{aligned}
&P_{25}(s,t) \\
&= \int_s^t P_{22}(s,u)\left\{\lambda_{25}(u)P_{55}(u,t) + \lambda_{24}(u)P_{45}(u,t)\right\}du \\
&= \int_s^t \exp\left(-\frac{a_{24}}{b_{24}+1}u^{b_{24}+1} - \frac{a_{25}}{b_{25}+1}u^{b_{25}+1}\right. \\
&\qquad \left. + \frac{a_{24}}{b_{24}+1}s^{b_{24}+1} + \frac{a_{25}}{b_{25}+1}s^{b_{25}+1}\right) \\
&\qquad \times \left[a_{25}u^{b_{25}} + a_{24}u^{b_{24}} \times \left\{\int_u^t \exp\left(-\frac{a_{45}}{b_{45}+1}v^{b_{45}+1} + \frac{a_{45}}{b_{45}+1}u^{b_{45}+1}\right)\right.\right. \\
&\qquad \left.\left. \times a_{45}v^{b_{45}}dv\right\}\right]du,
\end{aligned}
\tag{A.12.9}
$$

$$
\begin{aligned}
&P_{35}(s,t) \\
&\quad = \int_s^t P_{33}(s,u)\left\{\lambda_{35}(u)P_{55}(u,t)+\lambda_{34}(u)P_{45}(u,t)\right\}du \\
&\quad = \int_s^t \exp\left(-\frac{a_{34}}{b_{34}+1}u^{b_{34}+1}-\frac{a_{35}}{b_{35}+1}u^{b_{35}+1}\right. \\
&\qquad\qquad \left.+\frac{a_{34}}{b_{34}+1}s^{b_{34}+1}+\frac{a_{35}}{b_{35}+1}s^{b_{35}+1}\right) \\
&\qquad \times\left[a_{35}u^{b_{35}}+a_{34}u^{b_{34}}\times\left\{\int_u^t \exp\left(-\frac{a_{45}}{b_{45}+1}v^{b_{45}+1}+\frac{a_{45}}{b_{45}+1}u^{b_{45}+1}\right)\right.\right. \\
&\qquad\qquad \left.\left.\times a_{45}v^{b_{45}}dv\right\}\right]du.
\end{aligned}
\tag{A.12.10}
$$

Next by using $P_{24}(s,t)$ and $P_{34}(s,t)$, we derive

$$
\begin{aligned}
&P_{14}(s,t) \\
&\quad = \int_s^t P_{11}(s,u)\left\{\lambda_{12}(u)P_{24}(u,t)+\lambda_{13}(u)P_{34}(u,t)\right\}du \\
&\quad = \int_s^t \exp\left(-\frac{a_{12}}{b_{12}+1}u^{b_{12}+1}-\frac{a_{13}}{b_{13}+1}u^{b_{13}+1}-\frac{a_{15}}{b_{15}+1}u^{b_{15}+1}\right. \\
&\qquad\qquad \left.+\frac{a_{12}}{b_{12}+1}s^{b_{12}+1}+\frac{a_{13}}{b_{13}+1}s^{b_{13}+1}+\frac{a_{15}}{b_{15}+1}s^{b_{15}+1}\right) \\
&\qquad \times\left[a_{12}u^{b_{12}}\times\exp\left(\frac{a_{24}}{b_{24}+1}u^{b_{24}+1}+\frac{a_{25}}{b_{25}+1}u^{b_{25}+1}-\frac{a_{45}}{b_{45}+1}t^{b_{45}+1}\right)\right. \\
&\qquad\quad \times\left\{\int_u^t \exp\left(-\frac{a_{24}}{b_{24}+1}v^{b_{24}+1}-\frac{a_{25}}{b_{25}+1}v^{b_{25}+1}\right)\times a_{24}v^{b_{24}}\right. \\
&\qquad\qquad \left.\times\exp\left(\frac{a_{45}}{b_{45}+1}v^{b_{45}+1}\right)dv\right\} \\
&\qquad\quad + a_{13}u^{b_{13}}\times\exp\left(\frac{a_{34}}{b_{34}+1}u^{b_{34}+1}+\frac{a_{35}}{b_{35}+1}u^{b_{35}+1}-\frac{a_{45}}{b_{45}+1}t^{b_{45}+1}\right) \\
&\qquad\quad \times\left\{\int_u^t \exp\left(-\frac{a_{34}}{b_{34}+1}v^{b_{34}+1}-\frac{a_{35}}{b_{35}+1}v^{b_{35}+1}\right)\times a_{34}v^{b_{34}}\right. \\
&\qquad\qquad \left.\left.\times\exp\left(\frac{a_{45}}{b_{45}+1}v^{b_{45}+1}\right)dv\right\}\right]du.
\end{aligned}
\tag{A.12.11}
$$

Finally by using $P_{25}(s,t)$ and $P_{35}(s,t)$, we derive

$$
\begin{aligned}
&P_{15}(s,t)\\
&\quad=\int_s^t P_{11}(s,u)\left\{\lambda_{15}(u)P_{55}(u,t)+\lambda_{12}(u)P_{25}(u,t)+\lambda_{13}(u)P_{35}(u,t)\right\}du\\
&\quad=\int_s^t \exp\left(-\frac{a_{12}}{b_{12}+1}u^{b_{12}+1}-\frac{a_{13}}{b_{13}+1}u^{b_{13}+1}-\frac{a_{15}}{b_{15}+1}u^{b_{15}+1}\right.\\
&\qquad\qquad\left.+\frac{a_{12}}{b_{12}+1}s^{b_{12}+1}+\frac{a_{13}}{b_{13}+1}s^{b_{13}+1}+\frac{a_{15}}{b_{15}+1}s^{b_{15}+1}\right)\\
&\qquad\times\left[a_{15}u^{b_{15}}+a_{12}u^{b_{12}}\times\exp\left(\frac{a_{24}}{b_{24}+1}u^{b_{24}+1}+\frac{a_{25}}{b_{25}+1}u^{b_{25}+1}\right)\right.\\
&\qquad\quad\times\left\{\int_u^t\exp\left(-\frac{a_{24}}{b_{24}+1}v^{b_{24}+1}-\frac{a_{25}}{b_{25}+1}v^{b_{25}+1}\right)\right.\\
&\qquad\qquad\times\left[a_{25}v^{b_{25}}+a_{24}v^{b_{24}}\times\exp\left(\frac{a_{45}}{b_{45}+1}v^{b_{45}+1}\right)\right.\\
&\qquad\qquad\quad\left.\left.\left.\times\left\{\int_v^t\exp\left(-\frac{a_{45}}{b_{45}+1}w^{b_{45}+1}\right)\times a_{45}w^{b_{45}}dw\right\}\right]dv\right\}\\
&\qquad+a_{13}u^{b_{13}}\times\exp\left(\frac{a_{34}}{b_{34}+1}u^{b_{34}+1}+\frac{a_{35}}{b_{35}+1}u^{b_{35}+1}\right)\\
&\qquad\times\left\{\int_u^t\exp\left(-\frac{a_{34}}{b_{34}+1}v^{b_{34}+1}-\frac{a_{35}}{b_{35}+1}v^{b_{35}+1}\right)\right.\\
&\qquad\qquad\times\left[a_{35}v^{b_{35}}+a_{34}v^{b_{34}}\times\exp\left(\frac{a_{45}}{b_{45}+1}v^{b_{45}+1}\right)\right.\\
&\qquad\qquad\quad\left.\left.\left.\left.\times\left\{\int_v^t\exp\left(-\frac{a_{45}}{b_{45}+1}w^{b_{45}+1}\right)\times a_{45}w^{b_{45}}dw\right\}\right]dv\right\}\right]du.
\end{aligned}
\tag{A.12.12}
$$