Spring 2012

Two-level lognormal frailty model and competing risks model with missing cause of failure

Xiongwen Tang
University of Iowa

TWO-LEVEL LOGNORMAL FRAILTY MODEL AND COMPETING RISKS MODEL WITH MISSING CAUSE OF FAILURE

by

Xiongwen Tang

An Abstract

Of a thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Statistics
in the Graduate College of
The University of Iowa

May 2012

Thesis Supervisors: Professor Michael P. Jones
Professor Ying Zhang

ABSTRACT

In clustered survival data, unobservable cluster effects may exert powerful influences on the outcomes and thus induce correlation among subjects within the same cluster. The ordinary partial likelihood approach does not account for this dependence. Frailty models, as an extension to Cox regression, incorporate multiplicative random effects, called frailties, into the hazard model and have become a very popular way to account for the dependence within clusters. We particularly study the two-level nested lognormal frailty model and propose an estimation approach based on the complete data likelihood with frailty terms integrated out. We adopt B-splines to model the baseline hazards and adaptive Gauss-Hermite quadrature to approximate the integrals efficiently. Furthermore, in finding the maximum likelihood estimators, instead of the Newton-Raphson iterative algorithm, Gauss-Seidel and BFGS methods are used to improve the stability and efficiency of the estimation procedure.

We also study competing risks models with missing cause of failure in the context of Cox proportional hazards models. For competing risks data, there exists more than one cause of failure and each observed failure is exclusively linked to one cause. Conceptually, the causes are interpreted as competing risks before the failure is observed. Competing risks models are constructed based on the proportional hazards model specified for each cause of failure respectively, which can be estimated using partial likelihood approach. However, the ordinary partial likelihood is not applicable when the cause of failure could be missing for some reason. We propose a weighted

partial likelihood approach based on complete-case data, where weights are computed as the inverse of selection probability and the selection probability is estimated by a logistic regression model. The asymptotic properties of the regression coefficient estimators are investigated by applying counting process and martingale theory. We further develop a double robust approach based on the full data to improve the efficiency as well as the robustness.

Abstract Approved: _____
Thesis Supervisor

_____
Title and Department

_____
Date

_____
Thesis Supervisor

_____
Title and Department

_____
Date

TWO-LEVEL LOGNORMAL FRAILTY MODEL AND COMPETING RISKS MODEL WITH MISSING CAUSE OF FAILURE

by

Xiongwen Tang

A thesis submitted in partial fulfillment of the
requirements for the Doctor of Philosophy
degree in Statistics
in the Graduate College of
The University of Iowa

May 2012

Thesis Supervisors: Professor Michael P. Jones
Professor Ying Zhang

Graduate College
The University of Iowa
Iowa City, Iowa

CERTIFICATE OF APPROVAL

PH.D. THESIS

This is to certify that the Ph.D. thesis of

Xiongwen Tang

has been approved by the Examining Committee for the thesis requirement for the Doctor of Philosophy degree in Statistics at the May 2012 graduation.

Thesis Committee: _____
Michael P. Jones, Thesis Supervisor

_____
Ying Zhang, Thesis Supervisor

_____
Kung-Sik Chan

_____
Jiang Huang

_____
Russell V. Lenth

To Jing and Alyssa

ACKNOWLEDGEMENTS

I owe my deepest gratitude to my advisors, Professor Michael P. Jones and Professor Ying Zhang, for their guidance throughout my dissertation work. I have been greatly impressed by their inspiration, enthusiasm and immense knowledge, which help me move forward. I have learned from them not only in academic research, but also in American culture, life experience and future career development. I extremely appreciate their time and patience with me.

It is also my big pleasure to thank my committee members, Professors Kung-Sik Chan, Jiang Huang, and Russell V. Lenth, for sharing their insightful comments and suggestions on my research work.

I shall give my immense thanks to my beloved wife Jing Pan for her continuous support and encouragement, and especially for the birth of our daughter Alyssa S. Tang. Though the baby's birth inevitably distracted me quite a bit, I have been really enjoying the new life. The happiness and fun from family is always bringing me more energy and motivation. I am also imdebted to my parents and my parents-in-law for their care and support.

Moreover, I would like to thank the statistics department, every professors I took courses from, and the staff members Tammy, Margie and Dena for all help and convenience they provided.

Last but not the least, I shall thank Eastern Cooperative Oncology Group (ECOG) for providing the data from study E1178.

ABSTRACT

In clustered survival data, unobservable cluster effects may exert powerful influences on the outcomes and thus induce correlation among subjects within the same cluster. The ordinary partial likelihood approach does not account for this dependence. Frailty models, as an extension to Cox regression, incorporate multiplicative random effects, called frailties, into the hazard model and have become a very popular way to account for the dependence within clusters. We particularly study the two-level nested lognormal frailty model and propose an estimation approach based on the complete data likelihood with frailty terms integrated out. We adopt B-splines to model the baseline hazards and adaptive Gauss-Hermite quadrature to approximate the integrals efficiently. Furthermore, in finding the maximum likelihood estimators, instead of the Newton-Raphson iterative algorithm, Gauss-Seidel and BFGS methods are used to improve the stability and efficiency of the estimation procedure.

We also study competing risks models with missing cause of failure in the context of Cox proportional hazards models. For competing risks data, there exists more than one cause of failure and each observed failure is exclusively linked to one cause. Conceptually, the causes are interpreted as competing risks before the failure is observed. Competing risks models are constructed based on the proportional hazards model specified for each cause of failure respectively, which can be estimated using partial likelihood approach. However, the ordinary partial likelihood is not applicable when the cause of failure could be missing for some reason. We propose a weighted

partial likelihood approach based on complete-case data, where weights are computed as the inverse of selection probability and the selection probability is estimated by a logistic regression model. The asymptotic properties of the regression coefficient estimators are investigated by applying counting process and martingale theory. We further develop a double robust approach based on the full data to improve the efficiency as well as the robustness.

TABLE OF CONTENTS

APPENDIX

LIST OF TABLES

Table

LIST OF FIGURES

Figure

PART I

TWO-LEVEL LOGNORMAL FRAILTY MODEL

CHAPTER 1
INTRODUCTION

In survival analysis, Cox's proportional hazards regression models are used as commonly as simple (or multiple) linear models are with uncensored data. However, in some statistical studies, the Cox model may not be applied directly. It assumes the univariate failure times are independent of each other given all relevant covariates. However, some covariates, which can not be observed for some reason, will cause heterogeneity among subjects. For another example, when subjects are recruited in clusters, there may exist unobservable cluster effects that have powerful influences on failure times and thus result in association among subjects within the same cluster or heterogeneity among clusters. More generally, the clustered survival data are called multivariate survival data. In both examples, the unobserved information brings dependence of failure times among subjects that violates the independence assumption of the regular Cox model.

There have been lots of work devoted to extending the Cox's model to handle the dependence in survival data, among which frailty models have become very popular as a way of incorporating heterogeneity among all subjects or between clusters, or dependence among subjects within a cluster. We will first introduce frailty models and then review some alternative approaches.

1.1 The Frailty Model

The term frailty was first introduced by Vaupel et al. (1979) in a univariate survival model. Its meaning can be generalized as the susceptibility to "death". But the idea of frailty used in multivariate survival data dates back to Clayton (1978), where the frailty (not defined in the paper though) was modeled by one fixed parameter to measure the association of bivariate failure times.

Frailty models can be treated as an extension to the Cox model (Cox, 1972) by adding frailty terms that have multiplicative effects on the baseline hazards. Conceptually, they are similar to linear mixed models, which are an extension to simple (or multiple) linear models. The term frailty model appears to be specific to the survival data with censored or truncated observed times. Mathematically, this model can be written as

$$\lambda_{ij}(t) = \lambda_0(t)\, u_i \exp\left(\boldsymbol{\beta}^T \boldsymbol{z}_{ij}\right), \tag{1.1}$$

where $\lambda_{ij}(t)$ is the conditional hazard function for the j^{th} subject from the i^{th} cluster, $\lambda_0(t)$ is the baseline hazard, $\boldsymbol{\beta}$ is the fixed effects vector, $\boldsymbol{z}_{ij}$ is the vector of covariates, and u_i is the frailty term for the i^{th} cluster. Let $\omega_i = \log(u_i)$, then the model can be rewritten as

$$\lambda_{ij}(t) = \lambda_0(t) \exp\left(\boldsymbol{\beta}^T \boldsymbol{z}_{ij} + \omega_i\right), \tag{1.2}$$

where ω_i is called the random effect for the i^{th} cluster.

Like the random effects in mixed models, the frailty terms u_i's (or the random effects ω_i's) in frailty models are assumed to be random following a certain distribution. To avoid unidentifiability issues, some restriction must be put on the

distribution, $E[U] = 1$ or $E[\Omega] = 0$, because we always can choose a baseline hazard function to make the mean of frailties equal to one, or to make the mean of random effects equal to zero. Different frailty distributions have been proposed: gamma distribution, inverse Gaussian distribution, positive stable distribution, power variance function (PVF) distribution, compound Poisson distribution and lognormal distribution. The gamma distribution is most commonly used since the frailty terms in the conditional likelihood can be integrated out analytically and therefore the full likelihood has a closed form. The positive stable distribution, first introduced by Hougaard (1986a) in multivariate survival analysis, has a nice property that the proportionality can be inherited from the conditional hazard to the marginal or population hazard. The Normal and Weibull distributions belong to this family. The PVF distribution (Hougaard, 1986a) is an extension to positive stable distribution. It contains the gamma, inverse Gaussian and positive stable distributions that exhibit different strength of dependence among subjects within clusters at early or late times (Duchateau and Janssen, 2008, Section 4.2-4.4). In the lognormal frailty model, the random effects follow a multivariate normal distribution similar to classical mixed models. We study the lognormal frailty model in this dissertation.

1.2 Alternatives to the Frailty Model

We review some alternative approaches to the clustered survival data: the fixed effects model, the stratified model, the copula model and the marginal model.

In the *fixed effects* model, the cluster effect is assumed fixed rather than ran-

dom. Then the model can be written as

$$\lambda_{ij}(t) = \lambda_0(t) \exp\left(\boldsymbol{\beta}^T \boldsymbol{z}_{ij} + c_i\right),$$

where c_i is the fixed effect for the i^{th} cluster. It further requires a restriction $c_1 = 0$ so that the fixed effects can be uniquely identified separate from the baseline hazards. There are some obvious drawbacks for this approach. First, the estimates of the fixed cluster effects are poor when the cluster size is small, and thus the model interpretation is less persuasive. Furthermore, significant bias may be caused in situations where there are many small clusters, censoring is present, or the subjects within one cluster share similar covariate information (McGilchrist and Aisbett, 1991; Glidden and Vittinghoff, 2004; Duchateau and Janssen, 2008).

The *stratified* model adopts the proportional hazards model within each stratum, assuming the same regression coefficients but different baseline hazards for each cluster (Kalbfleisch and Prentice, 2002). It has the following form

$$\lambda_{ij}(t) = \lambda_{i0}(t) \exp\left(\boldsymbol{\beta}^T \boldsymbol{z}_{ij}\right),$$

where $\lambda_{i0}(t)$ is baseline hazard function for cluster i. Note that the baseline hazard in the fixed effects model for the i^{th} cluster can be expressed as $\lambda_0(t)\exp(c_i)$. Therefore the stratified model is more general than the fixed effects model. For bivariate survival data, one subject within a cluster contributes to the partial likelihood of the stratified model only when a failure is observed for that subject and the other subject is still at risk in the meantime. In this way the actual sample size becomes smaller and thus the resulting estimation is less efficient.

Both the fixed effects model and the stratified model can be estimated easily by applying the partial likelihood approach.

The *copula* model uses the copula to specify the joint survival function in terms of the marginal survival functions for each cluster; hence the copula is the function that links the marginal survival functions to the joint survival function. This approach is not appropriate for data with large or imbalanced cluster size. Its primary application is for bivariate survival data (paired observations) (Andersen et al., 2005; Pipper and Martinussen, 2003; Roy and Mukherjee, 1998; Phelps and Weissfeld, 1997; Nelsen, 1997).

The *marginal* model treats the survival times as independent and thereby ignores the cluster effects when estimating regression coefficients. The resulting estimates are consistent for a marginal or population-averaged hazards model under certain assumptions, but do not possess a mixed model interpretation, i.e. the covariate effects conditional on the random effects. The marginal proportional hazards models have been studied in different settings by Lee et al. (1992), Liang et al. (1993), Lin (1994), Spiekerman and Lin (1998). As an extension to the standard semiparametric marginal approach, Huang and Chen (2003) and Cong et al. (2007) study the clustered survival data particularly with informative cluster size.

1.3 Estimation of Frailty Models

1.3.1 Our Approach

In this dissertation, we study a two-level lognormal frailty model, where the two nested frailties are assumed to independently follow lognormal distributions.

For example, in a typical medical study, several patients may be seen by the same physician, and hence share a number of similarities that are not measured as covariates, called the physician effects, which in turn induces a correlation. Patients being treated at the same hospital may share similar facilities, nurses and other care, and hence similar unmeasured random hospital effects independent of the physician effects. In addition, physicians at the same hospital all subject to the same hospital regulation or care, which is a component of this hospital-level effects as well. In this case, patients are nested within physicians and physicians are nested within hospitals, creating two levels of correlation structure, one at the physician level and one at the hospital level. These are modeled by frailties, one for physicians and one for hospitals. These frailties are assumed independent in this dissertation. Patients with the same physician (or hospital) share the same frailty for that physician (or hospital).

In modeling, we parameterize the baseline cumulative hazard function using a B-spline expansion, and then estimate the model in a parametric fashion. In the estimating approach we consider the full likelihood of all regression parameters, parameters of the B-spline approximation to the cumulative hazard function and the (latent) random effect variances, with the frailty terms integrated out. Hence, the full likelihood involves integrals. We adopt the non-adaptive and adaptive Gauss-Hermite

quadrature for those integrals. A Newton-Raphson iterative algorithm is used but with some modification. For example, for likelihood $L(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2)$, where $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ are the two groups of separate parameters, e.g. $\boldsymbol{\xi}_1$ represents regression parameters and B-spline parameters and $\boldsymbol{\xi}_2$ represents variance parameters of random effects, instead of estimating $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ together, we first estimate $\boldsymbol{\xi}_1$ given $\boldsymbol{\xi}_2$ is fixed, and then estimate $\boldsymbol{\xi}_2$ given $\boldsymbol{\xi}_1$ is fixed, and so on, until both $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ converge. This procedure is called the Gauss-Seidel method. The Newton-Raphson method in our case is very sensitive to initial values, while the Gauss-Seidel method seems more robust to the initial values: its intermediate estimates after a few iterations (not necessarily converged) can then be used as good initial values for the Newton-Raphson method in estimating all parameters together. For the purpose of efficiency and stability, we adopt the BFGS algorithm implemented in R function `constrOptim` instead of the Newton-Raphson and Gauss-Seidel methods in the simulation studies.

1.3.2 Alternative Approaches

Next we briefly discuss some alternative approaches to estimating frailty models in a semiparametric setting (the baseline hazards are treated as some unknown function with infinite dimension).

As defined in (1.2), frailties can be rewritten as random effects. Then EM algorithm can be naturally implemented by treating random effects as missing variables, but it is very slow to converge.

Penalized partial likelihood (PPL) approach provides an alternative approach,

where the PPL consists of the conditional likelihood given frailties and the joint density function of frailties. Duchateau and Janssen (2008, Section 5.2) give a detailed introduction to its computational algorithms and prove theoretically that the EM algorithm and PPL approach yield the same results in the case of gamma frailty models. Also, Bayesian techniques can be used to estimate gamma frailty models through Gibbs sampling (Duchateau and Janssen, 2008, Section 5.3).

In the implementation of EM and PPL approaches, the Gauss-Seidel method is often applied, where the regression coefficients and the frailty variances are estimated separately in each iteration.

Ha et al. (2001) and Ha and Lee (2003) proposed hierarchical likelihood approaches for frailty models with gamma or lognormal frailty distribution. Particularly, the profile hierarchical likelihood can be applied to the two-level lognormal frailty model with nonparametric baseline hazards. Laplace approximation is used when the numerical integration is intractable. This is the only approach we have found that has been implemented to estimate the two-level lognormal frailty model (see R package `frailtyHL`). Hence, we will give a more detailed introduction to this approach in Section 3.4 and evaluate it through simulation studies in Section 5.2.3.

1.4 Layout

In Chapter 2, we review background knowledge of splines. Chapter 3 gives a detailed introduction to notation, modeling of the two-level lognormal frailty model, and formulas for the score functions and the Hessian matrix. All numerical analysis

methods (construction of B-splines, Gauss-Hermite quadrature and BFGS algorithm etc.) involved in the model estimation are discussed in Chapter 4. In Chapter 5, we describe our simulation studies and display the results for our proposed approach and other approaches (e.g. the penalized partial likelihood approach and the hierarchical likelihood approach). Finally, some discussion and potential future work conclude the first part of dissertation in Chapter 6.

CHAPTER 2
SPLINE THEORY

For Cox proportional hazards models, the semiparametric approach can be postulated in which the baseline hazard is estimated in a nonparametric fashion, which entails an infinite dimensional estimation problem.

Assuming that the true baseline (cumulative) hazard function is continuous, it can be approximated by a flexible parametric function. Splines can provide a strategy for this alternative approach. Actually splines have already been widely used in survival analysis, mainly to model the hazard function or incorporate covariate effects. Anderson and Senthilselvan (1980) first proposed a penalized likelihood approach, using smoothing quadratic splines to model the baseline hazard function. Ramsay (1988) first introduced M-splines and I-splines as a set of useful basis for monotone splines. Rondeau et al. (2003, 2006) adopted cubic M-splines to model the baseline hazard function with a smoothing parameter in their maximum penalized likelihood approach for one-level and two-level (nested) gamma-frailty model. Gray (1992) allowed a flexible way to incorporate interactions of continuous covariates and time-dependent covariates in terms of B-splines with smoothing parameter in his partial penalized likelihood approach. These works are all based on proportional hazards models. Kooperberg et al. (1995) used polynomial splines to model the log hazard function with or without covariates in a more general setting.

In this chapter, we explore background knowledge of splines to establish an approach to model the baseline hazard function. A key reference throughout is Schu-

maker (2007).

2.1 Polynomials and Polynomial Splines

It is known, by the Weierstrass Approximation theorem, that any continuous function f in $[a, b]$ (i.e. $f \in C[a, b]$) can be approximated arbitrarily well by a polynomial, where the polynomial function with order m is defined as

$$p(x) = \sum_{i=1}^{m} c_i x^{i-1}, \quad c_1, \ldots, c_m \in R.$$

The approximation using a polynomial seems attractive because the calculation of an integral and derivative is straightforward and the estimation of the parameters c_i's in the linear form is easy to manipulate. To find such a polynomial, one natural approach is to choose m points within $[a, b]$ and then do the interpolation at these points. However, for any sequence of these interpolating points $t_1, t_2, \cdots, t_m$, it is not guaranteed that the resulting polynomials converge to the function to be approximated. There is one counter example shown in Schumaker (2007, Section 3.6). This issue is actually a manifestation of the inflexibility of polynomials: polynomials that well approximate one continuous function in $[a, b]$ may oscillate wildly elsewhere especially when the order is high (10 or above). Unfortunately, high order is often needed for accurate approximation. To overcome this inflexibility issue, we consider *piecewise polynomials* and *polynomial splines*.

Let

$$\mathbf{\Delta} = \{x_i\}_1^k \text{ with } a = x_0 < x_1 < \cdots < x_k < x_{k+1} = b$$

be a *partition* of a finite closed interval $[a, b]$ into $k+1$ subintervals

$$I_i = [x_i, x_{i+1}),\ i = 0, 1, \ldots, k-1 \text{ and } I_k = [x_k, x_{k+1}]\,.$$

Given a partition $\boldsymbol{\Delta}$ of $[a, b]$, the *piecewise polynomial* $f(x)$ of order m is defined as

$$f(x) = p_i(x) \text{ for } x \in I_i,\ i = 0, 1, \ldots, k,$$

where $p_i(x)$'s are any polynomials of order m. We call $x_1, \ldots, x_k$ the *knots* of $f(x)$.

With piecewise polynomials, good approximation can be achieved with fixed order m by increasing the number of knots. Furthermore, the piecewise polynomials become the *polynomial splines* when the smoothness at the knots is set according to the *multiplicity vector* of integers $\boldsymbol{M} = (m_1, \ldots, m_k)$ with $1 \leqslant m_i \leqslant m$, $i = 1, 2, \ldots, k$.

Definition 2.1 (Space of Polynomial Splines)**.** *Let* $\mathcal{S}^m(\boldsymbol{M}, \boldsymbol{\Delta}) = \{s$*: there exist polynomials* $p_0^m, \ldots, p_k^m$ *of order* m *such that* $s(x) = p_i^m(x)$ *for* $x \in I_i$, $i = 0, 1, \ldots, k$, *and* $\mathrm{D}_+^j p_{i-1}^m(x_i) = \mathrm{D}_+^j p_i^m(x_i)$ *for* $j = 0, 1, \ldots, m-1-m_i$, $i = 1, \ldots, k\}$ *be the space of polynomial splines of order* m *with knots* $x_1, \cdots, x_k$ *of multiplicities* $m_1, \ldots, m_k$, *where* D_+^j *is the* j^{th} *right derivative. (Schumaker, 2007, pg. 108)*

As shown above, the multiplicity vector $\boldsymbol{M}$ actually measures the degrees of freedom of piecewise polynomial spline s at the knots x_i's ($i = 1, \ldots, k$): a smaller value of m_i means more smoothness and thus less freedom at x_i for s. Note that the polynomial spline may have a jump (i.e. discontinuous) at x_i when $m_i = m$. In this definition, the knot constraint $\mathrm{D}_+^{-1} p_{i-1}^m(x_i) = \mathrm{D}_+^{-1} p_i^m(x_i)$ does not exist for $m_i = m$.

If each m_i is zero (i.e. no freedom at all), then the spline is just a single polynomial of order m. That is why zero is not allowable for m_i's in the definition of splines.

The following theorem tells that polynomial splines can be used to approximate any function in $C[a, b]$.

Theorem 2.1. *Every continuous function on the interval $[a, b]$ can be approximated arbitrarily well by piecewise polynomial splines with the order m fixed, provided a sufficient number of knots are allowed.*

2.2 B-Splines

It can be shown that $\mathcal{S}^m(\boldsymbol{M}, \boldsymbol{\Delta})$ is a linear space of dimension $m + K$ with $K = \sum_{i=1}^{k} m_i$ (Schumaker, 2007, pg. 110). This implies that any polynomial spline in the space $\mathcal{S}^m(\boldsymbol{M}, \boldsymbol{\Delta})$ can be expanded by a linear combination of a set of bases. There is one set of bases called B-splines, through which the computerized algorithm can be easily implemented to calculate derivatives and integrals for any polynomial spline.

Definition 2.2 (Extended Partition)**.** *Let the partition $\boldsymbol{\Delta}$ of $[a, b]$ and the multiplicity vector $\boldsymbol{M}$ be defined as earlier. Suppose $y_1 \le y_2 \le \cdots \le y_{2m+K}$ is such that*

$$y_1 \le \cdots \le y_m \le a, \qquad b \le y_{m+K+1} \le \cdots \le y_{2m+K} \tag{2.1}$$

and

$$y_{m+1} \le \cdots \le y_{m+K} = \overbrace{x_1, \ldots, x_1}^{m_1} \le \cdots \le \overbrace{x_k, \ldots, x_k}^{m_k}.$$

Then $\tilde{\boldsymbol{\Delta}} = \{y_i\}_{i=1}^{2m+K}$ is the extended partition *associated with $\mathcal{S}^m(\boldsymbol{M}, \boldsymbol{\Delta})$. (Schumaker, 2007, pg. 116)*

Note that the first and last m points in $\tilde{\boldsymbol{\Delta}}$ can be chosen arbitrarily subject to (2.1). In practice, a common approach is to set $y_1 = \cdots = y_m = a$, and $y_{m+K+1} = \cdots = y_{2m+K} = b$. The other points in $\tilde{\boldsymbol{\Delta}}$ are uniquely determined by $\boldsymbol{\Delta}$.

With the extended partition $\tilde{\boldsymbol{\Delta}}$ associated with $\mathcal{S}^m(\boldsymbol{M}, \boldsymbol{\Delta})$, we can define the B-splines by induction. Denote $\{B_i^m\}_{i=1}^{m+K}$ as the m^{th} order B-splines associated with the extended partition $\tilde{\boldsymbol{\Delta}}$. When $m = 1$, the B-splines are given by

$$B_i^1(x) = \begin{cases} 1, & y_i \le x < y_{i+1}, \\ 0, & \text{otherwise.} \end{cases} \tag{2.2}$$

Let $m \ge 2$, then $\{B_i^m\}_{i=1}^{m+K}$ and can be computed from $\{B_i^{m-1}\}_{i=1}^{m-1+K}$:

$$B_i^m(x) = \begin{cases} \dfrac{x - y_i}{y_{i+m-1} - y_i} B_i^{m-1}(x) + \dfrac{y_{i+m} - x}{y_{i+m} - y_{i+1}} B_{i+1}^{m-1}(x), & y_i \le x < y_{i+m}, \\ 0, & \text{otherwise.} \end{cases} \tag{2.3}$$

The derivatives (taking the right derivatives if the B-spline is not smooth enough at the knots) of $\{B_i^m\}_{i=1}^{m+K}$ can also be computed recursively:

$$\mathrm{D}_+ B_i^m(x) = \begin{cases} \dfrac{m-1}{y_{i+m-1} - y_i} B_i^{m-1}(x) - \dfrac{m-1}{y_{i+m} - y_{i+1}} B_{i+1}^{m-1}(x), & y_i \le x < y_{i+m}, \\ 0, & \text{otherwise.} \end{cases} \tag{2.4}$$

Note that in (2.3) and (2.4) the forms $\frac{0}{0}$ are deemed to be zero.

The B-splines have the following properties:

- $0 \le B_i^m(x) \le 1$ for any $x \in R$;
- $\sum_{i=1}^{m+K} B_i^m(x) = 1$ for any $x \in [a, b]$.

Theorem 2.2 (B-spline Expansion)**.** *If $\{B_i^m\}_{i=1}^{m+K}$ are the B-splines associated with*

the extended partition $\tilde{\boldsymbol{\Delta}} = \{y_i\}_{i=1}^{2m+K}$*, then every* $s \in \mathcal{S}^m(\boldsymbol{M}, \boldsymbol{\Delta})$ *has a unique expansion of the form (Schumaker, 2007, pg. 189):*

$$s(x) = \sum_{i=1}^{m+K} c_i B_i^m(x), \text{ all } x \in [y_m, y_{m+K+1}) \tag{2.5}$$

where c_i*'s* $(i = 1, \ldots, m+K)$ *are the constant coefficients. Furthermore, the right derivative of* $s(x)$ *is*

$$\begin{aligned} D_+ s(x) &= \sum_{i=1}^{m+K} c_i \, \mathrm{D}_+ B_i^m(x) \qquad \textit{or} \\ &= \sum_{i=2}^{m+K} c_i' \, B_i^{m-1}(x) \end{aligned}$$

where

$$c_i' = \begin{cases} (m-1)\dfrac{(c_i - c_{i-1})}{(y_{i+m-1} - y_i)}, & \textit{if } y_{i+m-1} > y_i \\ 0, & \textit{otherwise} \end{cases} \qquad i = 2, \ldots, n, \tag{2.6}$$

(Schumaker, 2007, pg. 195)

The B-splines are non-negative. So by (2.6), if the coefficients in the expansion are non-decreasing, i.e. $c_1 \leqslant c_2 \leqslant \ldots \leqslant c_{m+K}$, then $s(x)$ is non-decreasing (Schumaker, 2007, pg. 177).

CHAPTER 3
THE MODEL

To describe the frailty model, we first define some notation:

- T: failure time
- C: censoring time
- $X = T \wedge C = \min(T, C)$: observation time
- $\boldsymbol{Z}$: a p-dimensional vector of time-independent covariates including both subject-specific and cluster-specific covariates
- $\delta = 1_{[T \leq C]}$: censoring indicator
- λ: hazard function, Λ: cumulative hazard function
- h: top-level random effect (e.g. the hospital level)
- p: second-level random effect (e.g. the physician level nested in the hospital level)
- $\boldsymbol{a}^{\otimes 2} = \boldsymbol{a} \times \boldsymbol{a}^T$, where $\boldsymbol{a}$ is a number, vector or matrix.

Let n_h be the number of hospitals, n_i be the number of physicians nested in the i^{th} hospital, and n_{ij} be the number of patients of the j^{th} physician working at the i^{th} hospital. Then $n = \sum_{i=1}^{n_h} \sum_{j=1}^{n_i} n_{ij}$ is the sample size (i.e. the total number of patients).

Our frailty model is defined for the hazard and cumulative hazard functions conditioned on the frailties,

$$\lambda_{ijk}(t) = \lambda_0(t)\, \exp\left(\boldsymbol{\beta}^T \boldsymbol{z}_{ijk} + p_{ij} + h_i\right)\,, \tag{3.1}$$

$$\Lambda_{ijk}(t) = \Lambda_0(t)\, \exp\left(\boldsymbol{\beta}^T \boldsymbol{z}_{ijk} + p_{ij} + h_i\right), \tag{3.2}$$

where $\boldsymbol{z}_{ijk}$ is the covariates for the k^{th} patient of the j^{th} physician within the i^{th} hospital, h_i is the random effect from i^{th} hospital, and p_{ij} is the random physician effect for the j^{th} physician in the i^{th} hospital. In this terminology, $\exp(h_i)$ is the hospital frailty and $\exp(p_{ij})$ is the physician frailty.

Assume $h_i \sim \text{Normal}\,(0, \theta_1)$, $p_{ij} \sim \text{Normal}\,(0, \theta_2)$, $i = 1, \ldots, s$, and $j = 1, ..., n_i$. To avoid the unidentifiability issue, the means of the two normal distributions are set as zeroes. It is further assumed that all h_i's and p_{ij}'s are independent of each other. Unlike the gamma distributed frailty model, we cannot derive an explicit form of the likelihood after integrating out the lognormal frailty terms.

We further need general assumptions for survival models. Given the covariates $\boldsymbol{Z}_{ijk}$'s, the random effects h_i's and p_{ij}'s:

- the failure times T_{ijk}'s are conditionally independent of each other,
- the censoring times C_{ijk}'s are also conditionally independent of each other,
- the failure times T_{ijk}'s are conditionally independent of the censoring time C_{ijk}'s.

3.1 Baseline Hazard

As shown in Chapter 2, the baseline (cumulative) hazard can be modeled by B-splines when being treated as continuous function. To define the B-splines, we need to determine the partition $\boldsymbol{\Delta}$ first.

Practically, the finite closed interval $[a, b]$ can be constructed by the minimum and maximum observed times. The internal knots can be selected from the ordered observed times so that each subinterval contains a similar number of failure times (see Section 5.2 for explanation). Through this approach, the ranges of subintervals may vary greatly. Obviously, a larger range leads to a worse approximation of the baseline hazard within that subinterval by B-splines. However, this is unavoidable since we always observe fewer failures near the end of the study. We will explore this issue in simulation studies in the next chapter.

Let m be the order of B-splines, and k be the number of knots. For the multiplicity vector $\boldsymbol{M}$, let $m_\ell = 1$, $\ell = 1, \ldots, k$, i.e. the splines are assumed as smooth as possible at the knots. Then $q = m + k$ is the dimension of $\mathcal{S}^m(\boldsymbol{M}, \boldsymbol{\Delta})$, which is the number of B-spline basis functions needed to model the baseline hazard function.

The extended partition $\tilde{\boldsymbol{\Delta}}$ is then determined when the first m points are chosen as a and the last m points are chosen as b. As shown in Chapter 2, the B-splines and their derivatives at each observed times can be calculated using (2.2), (2.3) and (2.4).

Let B be a $n \times q$ matrix, of which the row vector $\boldsymbol{b}_{ijk}$ stands for a vector of

B-spline basis functions evaluated at the observed failure time of the k^{th} patient of the j^{th} physician in the i^{th} hospital. Let

$$\boldsymbol{b}_{ijk} = \boldsymbol{b}\left(x_{ijk}\right) = \left(B_1^m(x_{ijk}), \ldots, B_q^m(x_{ijk})\right)^T.$$

Let the derivative of $\boldsymbol{b}_{ijk}$ be

$$\dot{\boldsymbol{b}}_{ijk} = \dot{\boldsymbol{b}}\left(x_{ijk}\right) = \left(\frac{d}{dt}B_1^m(t), \ldots, \frac{d}{dt}B_q^m(t)\right)^T \mid_{\, t=x_{ijk}}.$$

Now we can model the baseline cumulative hazard using the B-spline expansion. Define

$$\Lambda_0(x_{ijk}) = \exp\left(\boldsymbol{c}^T\boldsymbol{b}_{ijk}\right), \tag{3.3}$$

where $\boldsymbol{c}$ is the unknown parameter to be estimated. By taking the derivative of Λ_0, we have

$$\lambda_0(x_{ijk}) = \boldsymbol{c}^T\dot{\boldsymbol{b}}_{ijk}\exp\left(\boldsymbol{c}^T\boldsymbol{b}_{ijk}\right).$$

Then the models (3.1) and (3.2) can be rewritten as

$$\Lambda_{ijk} \equiv \Lambda_{ijk}(x_{ijk}) = \exp\left(\boldsymbol{c}^T\boldsymbol{b}_{ijk} + \boldsymbol{\beta}^T\boldsymbol{z}_{ijk} + p_{ij} + h_i\right), \tag{3.4}$$

$$\lambda_{ijk} \equiv \lambda_{ijk}(x_{ijk}) = \boldsymbol{c}^T\dot{\boldsymbol{b}}_{ijk}\exp\left(\boldsymbol{c}^T\boldsymbol{b}_{ijk} + \boldsymbol{\beta}^T\boldsymbol{z}_{ijk} + p_{ij} + h_i\right). \tag{3.5}$$

By introducing the B-splines, we can estimate the semi-parametric model in a parametric way.

3.2 Likelihood

Let $\boldsymbol{h} = (h_1, \ldots, h_s)^T$, $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{in_i})^T$ and $\boldsymbol{p} = \left(\boldsymbol{p}_1^T, \ldots, \boldsymbol{p}_s^T\right)^T$.

Let f_{h_i} and $f_{p_{ij}}$ be the density functions of h_i and p_{ij}, respectively, for $i = 1, \ldots, s$, and $j = 1, ..., n_i$. Let $f_{\boldsymbol{h}}$, $f_{\boldsymbol{p}_i}$ and $f_{\boldsymbol{p}}$ be the joint density function of $\boldsymbol{h}$, $\boldsymbol{p}_i$ and $\boldsymbol{p}$, respectively, i.e. $f_{\boldsymbol{h}} = \prod_{i=1}^{s} f_{h_i}$, $f_{\boldsymbol{p}_i} = \prod_{j=1}^{n_i} f_{p_{ij}}$ and $f_{\boldsymbol{p}} = \prod_{i=1}^{s} f_{\boldsymbol{p}_i}$.

The full likelihood can be written as

$$
\begin{aligned}
& L(\boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{\theta}) \\
&= \int \left\{ \prod_{i=1}^{s} \prod_{j=1}^{n_i} \prod_{k=1}^{n_{ij}} \left(\lambda_{ijk}^{\delta_{ijk}} e^{-\Lambda_{ijk}} \right) \right\} \times f_{\boldsymbol{h}} f_{\boldsymbol{p}} \, d\boldsymbol{p} d\boldsymbol{h} \\
&= \int \prod_{i=1}^{s} \prod_{j=1}^{n_i} \prod_{k=1}^{n_{ij}} \left\{ \boldsymbol{c}^T \dot{\boldsymbol{b}}_{ijk} \exp\left(\boldsymbol{c}^T \boldsymbol{b}_{ijk} + \boldsymbol{\beta}^T \boldsymbol{z}_{ijk} + p_{ij} + h_i \right) \right\}^{\delta_{ijk}} \\
&\quad \times \exp\left\{ - \exp\left(\boldsymbol{c}^T \boldsymbol{b}_{ijk} + \boldsymbol{\beta}^T \boldsymbol{z}_{ijk} + p_{ij} + h_i \right) \right\} \times f_{\boldsymbol{h}} f_{\boldsymbol{p}} \, d\boldsymbol{p} d\boldsymbol{h} \\
&= \prod_{i=1}^{s} \left[\int \prod_{j=1}^{n_i} \prod_{k=1}^{n_{ij}} \left\{ \boldsymbol{c}^T \dot{\boldsymbol{b}}_{ijk} \exp\left(\boldsymbol{c}^T \boldsymbol{b}_{ijk} + \boldsymbol{\beta}^T \boldsymbol{z}_{ijk} + p_{ij} + h_i \right) \right\}^{\delta_{ijk}} \right. \\
&\quad \left. \times \exp\left\{ - \exp\left(\boldsymbol{c}^T \boldsymbol{b}_{ijk} + \boldsymbol{\beta}^T \boldsymbol{z}_{ijk} + p_{ij} + h_i \right) \right\} \times f_{h_i} f_{\boldsymbol{p}_i} \, d\boldsymbol{p}_i dh_i \right] \\
&= \prod_{i=1}^{s} \prod_{j=1}^{n_i} \prod_{k=1}^{n_{ij}} \left\{ \boldsymbol{c}^T \dot{\boldsymbol{b}}_{ijk} \exp\left(\boldsymbol{c}^T \boldsymbol{b}_{ijk} + \boldsymbol{\beta}^T \boldsymbol{z}_{ijk} \right) \right\}^{\delta_{ijk}} \\
&\quad \times \prod_{i=1}^{s} \left[\int \prod_{j=1}^{n_i} \left\{ \int \prod_{k=1}^{n_{ij}} \exp\left\{ \delta_{ijk}(p_{ij} + h_i) - \exp\left(\boldsymbol{c}^T \boldsymbol{b}_{ijk} + \boldsymbol{\beta}^T \boldsymbol{z}_{ijk} + p_{ij} + h_i \right) \right\} \right. \right. \\
&\quad \left. \left. \times f_{p_{ij}} dp_{ij} \right\} f_{h_i} dh_i \right].
\end{aligned}
$$

Correspondingly, the log likelihood is

$$
\begin{aligned}
l(\boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{\theta}) = & \sum_{i=1}^{s} \sum_{j=1}^{n_i} \sum_{k=1}^{n_{ij}} \delta_{ijk} \left\{ \log\left(\boldsymbol{c}^T \dot{\boldsymbol{b}}_{ijk} \right) + \boldsymbol{c}^T \boldsymbol{b}_{ijk} + \boldsymbol{\beta}^T \boldsymbol{z}_{ijk} \right\} \\
& + \sum_{i=1}^{s} \log \left[\int_{h_i} \prod_{j=1}^{n_i} \int_{p_{ij}} \exp\left\{ \delta_{ij\cdot}(p_{ij} + h_i) - \Lambda_{ij\cdot} \right\} \, f_{p_{ij}} dp_{ij} \, f_{h_i} dh_i \right],
\end{aligned}
$$

where $\Lambda_{ij\cdot} = \sum_{k=1}^{n_{ij}} \Lambda_{ijk}$ and Λ_{ijk} is defined in (3.4).

Note that θ_1 and θ_2 must be non-negative. When the true value of θ_1 or θ_2 is

close to zero, which is the boundary of their domains, its MLE may go negative or not converge, and its asymptotic normality may not hold well (which leads to bad coverage probabilities). For improvement, we reparametrize θ_i as e^{τ_i} for $i = 1, 2$. Though the issue of convergence and normality may still exist under some extreme cases since $\theta_i \to 0$ implies $\tau_i \to \infty$, reparametrization brings convenience in computation without having to consider the constraint and gains improvement overall. Then the frailty density functions are modified as follows

$$f_{h_i}(x) = \frac{1}{\sqrt{2\pi}}\theta_1^{-\frac{1}{2}} \exp\left(-\frac{x^2}{2\theta_1}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\tau_1 - \frac{x^2}{2}e^{-\tau_1}\right),$$
$$f_{p_{ij}}(y) = \frac{1}{\sqrt{2\pi}}\theta_2^{-\frac{1}{2}} \exp\left(-\frac{y^2}{2\theta_2}\right) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\tau_2 - \frac{y^2}{2}e^{-\tau_2}\right).$$

Define

$$\tilde{g}_{ij}(h_i, p_{ij}; \boldsymbol{c}, \boldsymbol{\beta}) = \exp\left\{\delta_{ij\cdot}(p_{ij} + h_i) - \Lambda_{ij\cdot}\right\}$$
$$g_{ij}(h_i; \boldsymbol{c}, \boldsymbol{\beta}, \tau_2) = \int \tilde{g}_{ij}(h_i, p_{ij}; \boldsymbol{c}, \boldsymbol{\beta}) \times f_{p_{ij}}\, dp_{ij},$$
$$g_i(\boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{\tau}) = \int \prod_{j=1}^{n_i} g_{ij}(h_i; \boldsymbol{c}, \boldsymbol{\beta}, \tau_2) \times f_{h_i}\, dh_i,$$

then the log likelihood can be rewritten in a simplified way

$$l(\boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{\tau}) = \sum_{i=1}^{s}\sum_{j=1}^{n_i}\sum_{k=1}^{n_{ij}} \delta_{ijk}\left\{\log\left(\boldsymbol{c}^T\dot{\boldsymbol{b}}_{ijk}\right) + \boldsymbol{c}^T\boldsymbol{b}_{ijk} + \boldsymbol{\beta}^T\boldsymbol{z}_{ijk}\right\} + \sum_{i=1}^{s}\log\left\{g_i(\boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{\tau})\right\}. \tag{3.6}$$

3.3 Score Functions and Hessian Matrix

For the sake of brevity, we have defined g_{ij}'s and g_i's ($i = 1, \ldots, s$, and $j = 1, ..., n_i$) to simplify the expression for the log likelihood (3.6). To obtain formulas for

the score functions and the Hessian matrix, we need to compute the first and second partial derivatives of g_{ij}'s with respect to $\boldsymbol{c}, \boldsymbol{\beta}$ and τ_2.

$$\begin{aligned}
\frac{\partial g_{ij}}{\partial \boldsymbol{c}} &= -\int \tilde{g}_{ij} \times \sum_{k=1}^{n_{ij}} (\Lambda_{ijk} \boldsymbol{b}_{ijk}) \times f_{p_{ij}}\ dp_{ij}\,, \\
\frac{\partial g_{ij}}{\partial \boldsymbol{\beta}} &= -\int \tilde{g}_{ij} \times \sum_{k=1}^{n_{ij}} (\Lambda_{ijk} \boldsymbol{z}_{ijk}) \times f_{p_{ij}}\ dp_{ij}\,, \\
\frac{\partial g_{ij}}{\partial \tau_2} &= \int \tilde{g}_{ij} \times \left(-\frac{1}{2} + \frac{p_{ij}^2}{2} e^{-\tau_2}\right) \times f_{p_{ij}}\ dp_{ij}\,, \\
\frac{\partial^2 g_{ij}}{\partial \boldsymbol{c} \partial \boldsymbol{c}^T} &= \int \tilde{g}_{ij} \times \left\{ \left(\sum_{k=1}^{n_{ij}} \Lambda_{ijk} \boldsymbol{b}_{ijk}\right)^{\otimes 2} - \sum_{k=1}^{n_{ij}} \Lambda_{ijk} \boldsymbol{b}_{ijk}^{\otimes 2} \right\} \times f_{p_{ij}}\ dp_{ij}\,, \\
\frac{\partial^2 g_{ij}}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} &= \int \tilde{g}_{ij} \times \left\{ \left(\sum_{k=1}^{n_{ij}} \Lambda_{ijk} \boldsymbol{z}_{ijk}\right)^{\otimes 2} - \sum_{k=1}^{n_{ij}} \Lambda_{ijk} \boldsymbol{z}_{ijk} \right\} \times f_{p_{ij}}\ dp_{ij}\,, \\
\frac{\partial^2 g_{ij}}{\partial \boldsymbol{c} \partial \boldsymbol{\beta}^T} &= \int \tilde{g}_{ij} \times \left\{ \left(\sum_{k=1}^{n_{ij}} \Lambda_{ijk} \boldsymbol{b}_{ijk}\right) \left(\sum_{k=1}^{n_{ij}} \Lambda_{ijk} \boldsymbol{z}_{ijk}\right) - \sum_{k=1}^{n_{ij}} \Lambda_{ijk} \boldsymbol{b}_{ijk} \boldsymbol{z}_{ijk} \right\} \times f_{p_{ij}}\ dp_{ij}\,, \\
\frac{\partial^2 g_{ij}}{\partial \tau_2^2} &= \int \tilde{g}_{ij} \times \left(\frac{1}{4} - p_{ij}^2 e^{-\tau_2} + \frac{1}{4} p_{ij}^4 e^{-2\tau_2}\right) \times f_{p_{ij}}\ dp_{ij}\,, \\
\frac{\partial^2 g_{ij}}{\partial \boldsymbol{c} \partial \tau_2} &= -\int \tilde{g}_{ij} \times \left(\sum_{k=1}^{n_{ij}} \Lambda_{ijk} \boldsymbol{b}_{ijk}\right) \times \left(-\frac{1}{2} + \frac{p_{ij}^2}{2} e^{-\tau_2}\right) \times f_{p_{ij}}\ dp_{ij}\,, \\
\frac{\partial^2 g_{ij}}{\partial \boldsymbol{\beta} \partial \tau_2} &= -\int \tilde{g}_{ij} \times \left(\sum_{k=1}^{n_{ij}} \Lambda_{ijk} \boldsymbol{z}_{ijk}\right) \times \left(-\frac{1}{2} + \frac{p_{ij}^2}{2} e^{-\tau_2}\right) \times f_{p_{ij}} dp_{ij}\,.
\end{aligned}$$

Next we need to compute the first and second derivatives of g_i's with respect to $\boldsymbol{c}, \boldsymbol{\beta}$, τ_1 and τ_2:

$$\begin{aligned}
\frac{\partial g_i}{\partial \boldsymbol{c}} &= \int \left(\prod_{j=1}^{n_i} g_{ij}\right) \left(\sum_{j=1}^{n_i} \frac{1}{g_{ij}} \frac{\partial g_{ij}}{\partial \boldsymbol{c}}\right) \times f_{h_i}\, dh_i\,, \\
\frac{\partial g_i}{\partial \boldsymbol{\beta}} &= \int \left(\prod_{j=1}^{n_i} g_{ij}\right) \left(\sum_{j=1}^{n_i} \frac{1}{g_{ij}} \frac{\partial g_{ij}}{\partial \boldsymbol{\beta}}\right) \times f_{h_i}\ dh_i\,, \\
\frac{\partial g_i}{\partial \tau_1} &= \int \left(\prod_{j=1}^{n_i} g_{ij}\right) \times \left(-\frac{1}{2} + \frac{h_i^2}{2} e^{-\tau_1}\right) \times f_{h_i} dh_i\,, \\
\frac{\partial g_i}{\partial \tau_2} &= \int \prod_{j=1}^{n_i} g_{ij} \times \sum_{j=1}^{n_i} \frac{1}{g_{ij}} \frac{\partial g_{ij}}{\partial \tau_2} \times f_{h_i}\ dh_i\,,
\end{aligned}$$

$$\frac{\partial^2 g_i}{\partial \boldsymbol{c}\partial \boldsymbol{c}^T} = \int \left(\prod_{j=1}^{n_i} g_{ij}\right) \times \left[\sum_{j=1}^{n_i} \left\{ \frac{1}{g_{ij}} \frac{\partial^2 g_{ij}}{\partial \boldsymbol{c}\partial \boldsymbol{c}^T} - \frac{1}{g_{ij}^2} \left(\frac{\partial g_{ij}}{\partial \boldsymbol{c}}\right)^{\otimes 2} \right\} + \left(\sum_{j=1}^{n_i} \frac{1}{g_{ij}} \frac{\partial g_{ij}}{\partial \boldsymbol{c}}\right)^{\otimes 2} \times f_{h_i} \right] \, dh_i \,,$$

$$\frac{\partial^2 g_i}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} = \int \left(\prod_{j=1}^{n_i} g_{ij}\right) \times \left[\sum_{j=1}^{n_i} \left\{ \frac{1}{g_{ij}} \frac{\partial^2 g_{ij}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} - \frac{1}{g_{ij}^2} \left(\frac{\partial g_{ij}}{\partial \boldsymbol{\beta}}\right)^{\otimes 2} \right\} + \left(\sum_{j=1}^{n_i} \frac{1}{g_{ij}} \frac{\partial g_{ij}}{\partial \boldsymbol{\beta}}\right)^{\otimes 2} \times f_{h_i} \right] \, dh_i \,,$$

$$\frac{\partial^2 g_i}{\partial \tau_1^2} = \int \prod_{j=1}^{n_i} g_{ij} \times \left(\frac{1}{4} - h_i^2 e^{-\tau_1} + \frac{1}{4} h_i^4 e^{-2\tau_1}\right) \times f_{h_i} \, dh_i \,,$$

$$\frac{\partial^2 g_i}{\partial \tau_2^2} = \int \prod_{j=1}^{n_i} g_{ij} \times \left[\sum_{j=1}^{n_i} \left\{ \frac{1}{g_{ij}} \frac{\partial^2 g_{ij}}{\partial \tau_2^2} - \frac{1}{g_{ij}^2} \left(\frac{\partial g_{ij}}{\partial \tau_2}\right)^2 \right\} + \left(\sum_{j=1}^{n_i} \frac{1}{g_{ij}} \frac{\partial g_{ij}}{\partial \tau_2}\right)^2 \right] \times f_{h_i} dh_i \,,$$

$$\frac{\partial^2 g_i}{\partial \boldsymbol{c}\partial \boldsymbol{\beta}^T} = \int \prod_{j=1}^{n_i} g_{ij} \times \left[\sum_{j=1}^{n_i} \left\{ \frac{1}{g_{ij}} \frac{\partial^2 g_{ij}}{\partial \boldsymbol{c}\partial \boldsymbol{\beta}^T} - \frac{1}{g_{ij}^2} \frac{\partial g_{ij}}{\partial \boldsymbol{c}} \frac{\partial g_{ij}}{\partial \boldsymbol{\beta}^T} \right\} + \left(\sum_{j=1}^{n_i} \frac{1}{g_{ij}} \frac{\partial g_{ij}}{\partial \boldsymbol{c}}\right) \times \left(\sum_{j=1}^{n_i} \frac{1}{g_{ij}} \frac{\partial g_{ij}}{\partial \boldsymbol{\beta}^T}\right) \right] \times f_{h_i} \, dh_i \,,$$

$$\frac{\partial^2 g_i}{\partial \boldsymbol{c}\partial \tau_1} = \int \prod_{j=1}^{n_i} g_{ij} \times \left(\sum_{j=1}^{n_i} \frac{1}{g_{ij}} \frac{\partial g_{ij}}{\partial \boldsymbol{c}}\right) \times \left(-\frac{1}{2} + \frac{1}{2} h_i^2 e^{-\tau_1}\right) \times f_{h_i} \, \mathrm{d}h_i \,,$$

$$\frac{\partial^2 g_i}{\partial \boldsymbol{c}\partial \tau_2} = \int \prod_{j=1}^{n_i} g_{ij} \times \left[\sum_{j=1}^{n_i} \left\{ \frac{1}{g_{ij}} \frac{\partial^2 g_{ij}}{\partial \boldsymbol{c}\partial \tau_2} - \frac{1}{g_{ij}^2} \frac{\partial g_{ij}}{\partial \tau_2} \frac{\partial g_{ij}}{\partial \boldsymbol{c}} \right\} + \left(\sum_{j=1}^{n_i} \frac{1}{g_{ij}} \frac{\partial g_{ij}}{\partial \tau_2}\right) \times \left(\sum_{j=1}^{n_i} \frac{1}{g_{ij}} \frac{\partial g_{ij}}{\partial \boldsymbol{c}}\right) \right] \times f_{h_i} \, \mathrm{d}h_i \,,$$

$$\frac{\partial^2 g_i}{\partial \tau_1 \partial \tau_2} = \int \prod_{j=1}^{n_i} g_{ij} \times \left(\sum_{j=1}^{n_i} \frac{1}{g_{ij}} \frac{\partial g_{ij}}{\partial \tau_2}\right) \times \left(-\frac{1}{2} + \frac{1}{2} h_i^2 e^{-\tau_1}\right) \times f_{h_i} \, \mathrm{d}h_i \,.$$

With the above formulas about the partial derivatives of g_{ij}'s and g_i's ($i = 1, \ldots, s$, and $j = 1, ..., n_i$), the score functions and Hessian matrix for the log likelihood (3.6) can be computed conveniently as follows. Consequently, the Newton-Raphson and Gauss-Seidel methods can then be implemented to estimate the model

parameters.

$$
\begin{aligned}
\frac{\partial l}{\partial \boldsymbol{c}} &= \sum_{i=1}^{s} \left\{ \sum_{j=1}^{n_i} \sum_{k=1}^{n_{ij}} \delta_{ijk} \left(\frac{\dot{\boldsymbol{b}}_{ijk}}{\boldsymbol{c}^T \dot{\boldsymbol{b}}_{ijk}} + \boldsymbol{b}_{ijk} \right) + \frac{1}{g_i} \frac{\partial g_i}{\partial \boldsymbol{c}} \right\}, \\
\frac{\partial l}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{s} \left(\sum_{j=1}^{n_i} \sum_{k=1}^{n_{ij}} \delta_{ijk} \boldsymbol{z}_{ijk} + \frac{1}{g_i} \frac{\partial g_i}{\partial \boldsymbol{\beta}} \right), \\
\frac{\partial l}{\partial \tau_1} &= \sum_{i=1}^{s} \left(\frac{1}{g_i} \frac{\partial g_i}{\partial \tau_1} \right), \\
\frac{\partial l}{\partial \tau_2} &= \sum_{i=1}^{s} \left(\frac{1}{g_i} \frac{\partial g_i}{\partial \tau_2} \right), \\
\frac{\partial^2 l}{\partial \boldsymbol{c} \partial \boldsymbol{c}^T} &= \sum_{i=1}^{s} \left[\sum_{j=1}^{n_i} \sum_{k=1}^{n_{ij}} \delta_{ijk} \left\{ - \frac{\dot{\boldsymbol{b}}_{ijk}^{\otimes 2}}{\left(\boldsymbol{c}^T \dot{\boldsymbol{b}}_{ijk} \right)^2} \right\} - \frac{1}{g_i^2} \left(\frac{\partial g_i}{\partial \boldsymbol{c}} \right)^{\otimes 2} + \frac{1}{g_i} \frac{\partial^2 g_i}{\partial \boldsymbol{c} \partial \boldsymbol{c}^T} \right], \\
\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} &= \sum_{i=1}^{s} \left\{ - \frac{1}{g_i^2} \left(\frac{\partial g_i}{\partial \boldsymbol{\beta}} \right)^{\otimes 2} + \frac{1}{g_i} \frac{\partial^2 g_i}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right\}, \\
\frac{\partial^2 l}{\partial \tau_j^2} &= \sum_{i=1}^{s} \left\{ - \frac{1}{g_i^2} \left(\frac{\partial g_i}{\partial \tau_j} \right)^2 + \frac{1}{g_i} \frac{\partial^2 g_i}{\partial \tau_j^2} \right\} \text{ for } j = 1, 2, \\
\frac{\partial^2 l}{\partial \boldsymbol{c} \partial \boldsymbol{\beta}^T} &= \sum_{i=1}^{s} \left(- \frac{1}{g_i^2} \frac{\partial g_i}{\partial \boldsymbol{c}} \frac{\partial g_i}{\partial \boldsymbol{\beta}^T} + \frac{1}{g_i} \frac{\partial^2 g_i}{\partial \boldsymbol{c} \partial \boldsymbol{\beta}^T} \right), \\
\frac{\partial^2 l}{\partial \boldsymbol{c} \partial \tau_j} &= \sum_{i=1}^{s} \left(- \frac{1}{g_i^2} \frac{\partial g_i}{\partial \tau_j} \frac{\partial g_i}{\partial \boldsymbol{c}} + \frac{1}{g_i} \frac{\partial^2 g_i}{\partial \boldsymbol{c} \partial \tau_j} \right) \text{ for } j = 1, 2, \\
\frac{\partial^2 l}{\partial \beta \partial \tau_j} &= \sum_{i=1}^{s} \left(- \frac{1}{g_i^2} \frac{\partial g_i}{\partial \tau_j} \frac{\partial g_i}{\partial \boldsymbol{\beta}} + \frac{1}{g_i} \frac{\partial^2 g_i}{\partial \boldsymbol{\beta} \partial \tau_j} \right) \text{ for } j = 1, 2, \\
\frac{\partial^2 l}{\partial \tau_1 \partial \tau_2} &= \sum_{i=1}^{s} \left\{ - \frac{1}{g_i^2} \left(\frac{\partial g_i}{\partial \tau_1} \right) \left(\frac{\partial g_i}{\partial \tau_2} \right)^T + \frac{1}{g_i} \frac{\partial^2 g_i}{\partial \tau_1 \partial \tau_2} \right\}.
\end{aligned}
$$

3.4 Hierarchical Likelihood

In this section, we give a definition of the *hierarchical likelihood* (HL) introduced by Ha et al. (2001) and Ha and Lee (2003) in our setting:

$$
l_h \left(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\Lambda}_0; \boldsymbol{h}, \boldsymbol{p} \right) = \sum_{i,j,k} l_{1ijk} + \sum_{i,j} l_{2ij},
$$

where $l_{1ijk} = l_{1ijk}(\boldsymbol{\beta}, \boldsymbol{\Lambda}_0; y_{ijk}, \delta_{ijk} \mid h_i, p_{ij})$ is the logarithm of the conditional density function for Y_{ijk} and δ_{ijk} given $H_i = h_i$ and $P_{ij} = p_{ij}$ and $l_{2ij} = l_{2i}(\boldsymbol{\theta}; h_i, p_{ij})$ is the logarithm of the joint density function for H_i and P_{ij}. Recall that $var(h_i) = \theta_1$ and $var(p_{ij}) = \theta_2$. Based on models (3.1) and (3.2), l_{1ijk} can be decomposed as follows

$$\begin{aligned} l_{1ijk} &= \delta_{ijk}\left\{\log\lambda_0(y_{ijk}) + \boldsymbol{z}_{ijk}^T\boldsymbol{\beta} + h_i + p_{ij}\right\} - \Lambda_0(y_{ijk})\exp\left(\boldsymbol{z}_{ijk}^T\boldsymbol{\beta} + h_i + p_{ij}\right) \\ &= \delta_{ijk}\left\{\log\Lambda_0(y_{ijk}) + \boldsymbol{z}_{ijk}^T\boldsymbol{\beta} + h_i + p_{ij}\right\} - \Lambda_0(y_{ijk})\exp\left(\boldsymbol{z}_{ijk}^T\boldsymbol{\beta} + h_i + p_{ij}\right) \\ &\quad + \delta_{ijk}\log\left\{\lambda_0(y_{ijk})/\Lambda_0(y_{ijk})\right\} \\ &\equiv l_{10ijk} + l_{11ijk}, \end{aligned}$$

where $l_{10ijk} = \delta_{ijk}\log\mu'_{ijk} - \mu'_{ijk}$, $l_{11ijk} = \delta_{ijk}\log\left\{\lambda_0(y_{ijk})/\Lambda_0(y_{ijk})\right\}$, and $\mu'_{ijk} = \Lambda_0(y_{ijk})\exp\left(\boldsymbol{z}_{ijk}^T\boldsymbol{\beta} + h_i + p_{ij}\right)$. The term l_{10ijk} is identical to kernel of a conditional Poisson likelihood for δ_{ijk} given $H_i = h_i, P_{ij} = p_{ij}$ with mean μ'_{ijk}.

Since the functional form of baseline hazard $\lambda_0(t)$ is unknown, we can compute its nonparametric estimates at the distinct ordered observed failure times $y_{(1)}, \ldots, y_{(D)}$, denoted as the vector $\hat{\boldsymbol{\lambda}}_0 = (\hat{\lambda}_{01}, \ldots, \hat{\lambda}_{0D})$ where

$$\hat{\lambda}_{0l} = \hat{\lambda}_0\left(y_{(l)}\right) = \frac{d_{(l)}}{\sum_{i,j,k\in R_l}\exp\left(\boldsymbol{z}_{ijk}^T\boldsymbol{\beta} + h_i + p_{ij}\right)}.$$

Here $d_{(l)}$ denotes the number of failures at $y_{(l)}$, and R_l denotes the risk set at $y_{(l)}$. With $\hat{\boldsymbol{\lambda}}_0$, the baseline cumulative hazard function $\Lambda_0(t)$ can be estimated as

$$\hat{\Lambda}_0(t) = \sum_{l:y_{(l)}\le t}\hat{\lambda}_{0l}.$$

As a result, the *profile hierarchical likelihood* is defined as

$$l_h^*(\boldsymbol{\beta}, \boldsymbol{\theta}; \boldsymbol{h}, \boldsymbol{p}) = l_h(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\Lambda}_0; \boldsymbol{h}, \boldsymbol{p})\mid_{\Lambda_0=\hat{\Lambda}_0}.$$

Given $\boldsymbol{\theta}$, the estimating equations for $\boldsymbol{\beta}$ and the random effects $\boldsymbol{h}$ and $\boldsymbol{p}$ are identical to the ones from the Poisson hierarchical generalized linear model for δ with offset $\log \hat{\Lambda}_0 (y_{ijk})$.

Consider the *marginal form* of $l_h^* (\boldsymbol{\beta}, \boldsymbol{\theta})$ by integrating out the random effects

$$m_h^* (\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_i \log \left[\int \prod_j \left\{ \int \exp \left\{ l_{h,ij}^* (\boldsymbol{\beta}, \boldsymbol{\theta}) \right\} dp_{ij} \right\} dh_i \right].$$

Given $\boldsymbol{\theta}$, $m_h^* (\boldsymbol{\beta}, \boldsymbol{\theta})$ is replaced by a first-order or second-order Laplace approximation (Lee and Nelder, 2001), denoted as $\hat{m}_h^* (\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{h}, \boldsymbol{p})$, which does not involve integrals. Given $\boldsymbol{\theta}$, the parameters $(\boldsymbol{\beta}, \boldsymbol{h}, \boldsymbol{p})$ are estimated using the estimating equations from l_h^* or $\hat{m}_h^*$. Note that this is a REML-type procedure in which the random components $\boldsymbol{h}$ and $\boldsymbol{p}$ are estimated. $\boldsymbol{\theta}$ is then estimated using the profile likelihood $\hat{m}_h^* \left(\hat{\boldsymbol{\beta}}, \boldsymbol{\theta}, \hat{\boldsymbol{h}} (\boldsymbol{\theta}), \hat{\boldsymbol{p}} (\boldsymbol{\theta}) \right)$ given the most recent estimates of $\boldsymbol{\beta}$ where $\hat{\boldsymbol{\beta}}$ is not treated as a function of $\boldsymbol{\theta}$. This procedure is repeated until both $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ converge. In the R package `frailtyHL`, the orders of Laplace approximation to simply m_h^* need to be specified, where the order for estimating $\boldsymbol{\beta}$ is allowed to be 0 or 1, and the order for estimating $\boldsymbol{\theta}$ is allowed to be 1 or 2. The zero-*th* order means the estimating equations come from the profile hierarchical likelihood l_h^*. The first (or second) order means the estimating equations come from the first (or second) order Laplace approximation of m_h^*.

CHAPTER 4
NUMERICAL ANALYSIS TECHNIQUES

In this chapter, we explore some numerical analysis techniques used in the model estimation. In particular, we will discuss

- knot selection for B-splines,
- constraints on the coefficients in the B-spline expansion of Λ_0,
- non-adaptive and adaptive Gauss-Hermite quadrature,
- the BFGS algorithm.

4.1 Knot Selection for B-Splines

In Section 2.1, we have discussed how to model the hazard functions using B-splines, which reduces the infinite dimension of the space of all continuous non-negative functions to a finite dimension. Equation (2.5) indicates that the dimension q of the B-splines is the order m plus the sum of the multiplicity vector K. In our application, we set each element of the multiplicity vector M as one in order to achieve maximum smoothness of the splines at knots, where K is then reduced to k, the number of knots.

As required in Stone (1986) and Zhou et al. (1998), the following conditions for the selection of internal knots should hold:

- $k = O\left(n^{\upsilon}\right)$ for $0 < \upsilon < 1/2$.

- The maximum spacing of knots, $\delta_{\max} = O\left(n^{-\upsilon}\right)$.

- Let $\delta_{\min}$ be the minimum spacing of knots. For any n, there exists a constant $M > 0$, such that $\delta_{\max}/\delta_{\min} \leq M$.

In our simulation, $q = \left\lceil \sqrt[3]{n'} \right\rceil$, where $\lceil x \rceil$ is the smallest integer that is greater than or equal to x, and n' is the number of distinct observed times. Also, cubic B-splines are used, i.e. the order $m = 4$. The cubic B-spline is twice differentiable at knots as each multiplicity is one as stated previously. Then the number of knots can be determined as $k = \left\lceil \sqrt[3]{n'} \right\rceil - 4$. Furthermore, the minimum and maximum observed times are selected as the boundaries, and the knots are determined by calculating the percentiles at $\left(\frac{1}{k+1} \times 100\%, \ldots, \frac{k}{k+1} \times 100\%\right)$ among the observed failure times. Note that we only use the failure times in the knots selection to avoid some numerical issue while computing the initial values of the B-spline coefficients by forcing positive number of failures in each subinterval (see Section 5.2 for detail).

By doing this, the above conditions for knots selection will then be satisfied.

4.2 Constraints on B-Spline Coefficients

While implementing the computing algorithms, we have to pay extra attention to the B-spline coefficients $\boldsymbol{c}$, since its domain is not the whole R^q. As shown in Chapter 2, the baseline cumulative hazard function $\Lambda_0(t)$ modeled by $\exp\left\{\boldsymbol{c}^T \boldsymbol{b}(t)\right\}$ is non-decreasing if $\boldsymbol{c}$ is non-decreasing. So if we enforce the constraints $c_1 \leq \cdots \leq c_q$ for $\boldsymbol{c}$ in each iteration of Newton-Raphson algorithm, the spline estimate of $\Lambda_0(t)$ is guaranteed non-decreasing. Note that, through this approach, we reduce the dimension

of $\Lambda_0(t)$ since non-decreasing $\Lambda_0(t)$ does not necessarily require $c_1 \leq \cdots \leq c_q$. The *isotonic regression* method (Robertson et al., 1988; Eeden, 1958) is applied towards $\boldsymbol{c}$ when the constraints do not hold for intermediate estimates at some iterations in the estimation. Isotonic here means non-decreasing monotonicity. The idea of isotonic regression for our case can be generalized as the following optimization problem:

$$\min_{\boldsymbol{c}'} \sum_{l=1}^{q} (c'_l - c_l)^2 \text{ subject to } c'_1 \leq c'_2 \leq \cdots \leq c'_q . \tag{4.1}$$

The solution of (4.1) can be found by implementing the pooled adjacent violators algorithm (PAVA) (Barlow, 1972; Robertson et al., 1988). The PAVA starts with $\boldsymbol{c}$ (i.e. $\boldsymbol{c}' = \boldsymbol{c}$). If $\boldsymbol{c}$ is isotonic, then PAVA stops. Otherwise, there exists at least one subscript i (the smallest if there is more than one) such that $c'_{i-1} > c'_i$. Then the adjacent two values $\left(c'_{i-1}, c'_i\right)$ are replaced by their average. If $i > 2$ and $c'_{i-2} > c'_{i-1}$, then the adjacent three values $\left(c'_{i-2}, c'_{i-1}, c'_i\right)$ are replaced by their average. This process is repeated until $(c'_1, \ldots, c'_i)$ is isotonic. Then apply the same steps for the next violator $c'_{j-1} > c'_j$ for $j > i$ until $\left(c'_1, \ldots, c'_q\right)$ is isotonic.

As introduced in Section 5.2, we use an R built-in function `constrOptim`to estimate the model in simulation studies. In `constrOptim`, the log-barrier method (Nocedal and Wright, 2006, Section 19.6) is used to enforce the constraint on the B-spline coefficients. For an optimization problem

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \text{ subject to } \boldsymbol{c}(\boldsymbol{x}) \geq 0, \tag{4.2}$$

the log-barrier function can be defined as follows. Let

$$P(\boldsymbol{x}; \mu) = f(\boldsymbol{x}) - \mu \sum_i \log c_i(\boldsymbol{x}),$$

where $\mu > 0$. It can be shown that the minimizer of $P(\boldsymbol{x}; \mu)$, denoted by $\boldsymbol{x}(\mu)$, approaches a solution to (4.2) as $\mu \searrow 0$ (Forsgren et al., 2003).

4.3 Gauss-Hermite Quadrature

The log likelihood (3.6) involves the integrals over the normally distributed random effects h_i's and p_{ij}'s for $i = 1, \ldots, s$, and $j = 1, ..., n_i$, as do the score functions and the Hessian matrix. Gauss-Hermite quadrature (GHQ) naturally is a simple and efficient way to do the numerical integration. It approximates the integral of a function by a weighted sum over predefined abscissas.

Let $\boldsymbol{a} = (a_1, \ldots, a_d)^T$ be the *abscissas* (or *sampling nodes*) of GHQ, $\boldsymbol{w} = (w_1, \ldots, w_d)^T$ be the corresponding *weights*, where d is the order of GHQ, i.e. the number of the abscissas. The abscissas and weights are widely available for a variety of values of d. The Gauss-Hermite quadrature approximates the integral of $f(x)$ as follows

$$\begin{aligned} \int_{-\infty}^{+\infty} f(x)dx &= \int_{-\infty}^{+\infty} f(x) \exp\left(x^2\right) \times \exp\left(-x^2\right) \, dx \\ &\approx \sum_{l=1}^{d} f(a_l) \exp\left(a_l^2\right) \times w_l \,. \end{aligned} \tag{4.3}$$

We then apply the GHQ algorithm to $g_i(\boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{\tau})$ defined in Section 3.2, $i = 1, \ldots, s$:

$$g_i(\boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{\tau}) = \int \left\{ \prod_{j=1}^{n_i} \left[\int \tilde{g}_{ij}(h_i, p_{ij}; \boldsymbol{c}, \boldsymbol{\beta}) \times f_{p_{ij}} \, \mathsf{d}p_{ij} \right] \times f_{h_i} \right\} dh_i$$

$$= \sum_{l_i=1}^{d} \left\{ \prod_{j=1}^{n_i} \left[\sum_{l_{ij}=1}^{d} \tilde{g}_{ij}(a_{l_i}, a_{l_{ij}}; \boldsymbol{c}, \boldsymbol{\beta})\, f_{p_{ij}}(a_{l_{ij}})\, \exp\left(a_{l_{ij}}^2\right) \times w_{l_{ij}} \right] \right.$$
$$\left. \times\, f_{h_i}(a_{l_i})\, \exp\left(a_{l_i}^2\right) \times w_{l_i} \right\}.$$

Similarly, the partial derivatives of $g_i(\boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{\tau})$, and therefore the score functions and the Hessian matrix (see Section 3.3) can be also approximated by GHQ.

4.4 Adaptive Gauss-Hermite Quadrature

In the GHQ approach, to ensure good approximation through low-order d quadrature, the ratio of the integrand $f(x)$ to a standard normal density must be well approximated by some low-order $(2d+1)$ polynomial in both the *relevant region* of $f(x)$ where $f(x)$ is away from zero and the region where the sampling nodes are taken. When this is not satisfied, low-order GHQ becomes less effective (see Section 5.2.1 for one situation in our problem).

The idea of adaptive Gauss-Hermite quadrature (AGHQ) was first proposed by Liu and Pierce (1994): centralize and standardize the integral variable so that the integrand is sampled in a reasonable range. Assuming $f(x)$ is unimodal, i.e. $f(x)$ has a global maximum, denote $\hat{\mu}$ as the mode of $f(x)$, and $\hat{\sigma}^2 = -\left\{ \frac{d^2}{dx^2} \log f(x) \right\}^{-1} |_{x=\hat{\mu}}$.

Let $t = \frac{x-\hat{\mu}}{\sqrt{2}\hat{\sigma}}$. Then the GHQ approximation procedure is modified as

$$\begin{aligned}
\int_{-\infty}^{+\infty} f(x)dx &= \sqrt{2}\hat{\sigma} \int_{-\infty}^{+\infty} f(\hat{\mu} + \sqrt{2}\hat{\sigma}t)\, dt \\
&= \sqrt{2}\hat{\sigma} \int_{-\infty}^{+\infty} f(\hat{\mu} + \sqrt{2}\hat{\sigma}t) \exp\left(t^2\right) \times \exp\left(-t^2\right)\, dt \\
&= \sum_{l=1}^{d} f(\hat{\mu} + \sqrt{2}\hat{\sigma}a_l) \exp\left(a_l^2\right) \times \sqrt{2}\hat{\sigma}w_l \\
&= \sum_{l=1}^{d} f(a_l^*) \exp\left(a_l^2\right) \times w_l^*,
\end{aligned}$$

where $a_l^* = \hat{\mu} + \sqrt{2}\hat{\sigma}a_l$, $w^* = \sqrt{2}\hat{\sigma}w_l \exp(a_l^2)$.

Compared to the GHQ approximation in (4.3), the sampling nodes and weights in this AGHQ approach are adjusted according to the relevant area of $f(x)$, which is determined by the mode and its standard error. In other words, the integrand is resampled.

Similarly, Pinheiro and Bates (1995) and Pinheiro and Chao (2006) proposed an AGHQ approach for nonlinear mixed models and multilevel generalized linear mixed models: the random effect variables were centralized around their conditional modes given parameters and standardized using the standard errors of the conditional modes. As stated in Chapter 1, frailty models are similar to mixed models dedicated to survival data. Thus we are motivated to develop our *adaptive* GHQ approach for frailty models: in the estimating process, h_i and p_{ij} $(i = 1, ..., s,\ j = 1, \ldots n_i)$ are standardized at each iteration by calculating their conditional modes of h_i and p_{ij} and their standard errors given the data and the current estimates of parameters.

Let $\boldsymbol{h} = (h_1, \ldots, h_s)^T$, $\boldsymbol{p} = (p_{11}, \ldots, p_{1n_1}, \ldots, p_{sn_s})^T$. The calculation of the conditional modes is based on the log conditional likelihood for $(\boldsymbol{h}, \boldsymbol{p})$ given on the

current estimates of $\boldsymbol{c}$, $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$. This is equivalent to (3.6) without the integration over the random effects, namely

$$
\begin{aligned}
& l(\boldsymbol{h}, \boldsymbol{p}; \boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{\tau}) \\
&= \log f_{\boldsymbol{c},\beta}\left(\boldsymbol{c}, \boldsymbol{\beta} \,|\, \boldsymbol{h}, \boldsymbol{p}\right) + \log f_{\boldsymbol{h}}\left(\boldsymbol{h}\right) + \log f_{\boldsymbol{p}}\left(\boldsymbol{p}\right) + \text{constant} \\
&= \sum_{i=1}^{s} \sum_{j=1}^{n_i} \sum_{k=1}^{n_{ij}} \left[\delta_{ijk} \left\{ \log(\boldsymbol{c}^T \dot{\boldsymbol{b}}_{ijk}) + \boldsymbol{c}^T \boldsymbol{b}_{ijk} + \boldsymbol{\beta}^T \boldsymbol{z}_{ijk} + h_i + p_{ij} \right\} - \Lambda_{ijk} \right] \\
&\quad + \sum_{i=1}^{s} \left(-\frac{1}{2} e^{-\tau_1} h_i^2 \right) + \sum_{i=1}^{s} \sum_{j=1}^{n_i} \left(-\frac{1}{2} e^{-\tau_2} p_{ij}^2 \right) + \text{constant} \\
&= \sum_{i=1}^{s} \left\{ \sum_{j=1}^{n_i} \sum_{k=1}^{n_{ij}} \left[\delta_{ijk}(h_i + p_{ij}) - \Lambda_{ijk}\right] - \frac{1}{2} e^{-\tau_1} h_i^2 - \sum_{j=1}^{n_i} \left(\frac{1}{2} e^{-\tau_2} p_{ij}^2 \right) \right\} + \text{constant} \\
&= \sum_{i=1}^{s} l_i + \text{constant},
\end{aligned}
$$

where

$$
l_i = \sum_{j=1}^{n_i} \sum_{k=1}^{n_{ij}} \left[\delta_{ijk}(h_i + p_{ij}) - \Lambda_{ijk}\right] - \frac{1}{2} e^{-\tau_1} h_i^2 - \frac{1}{2} e^{-\tau_2} \sum_{j=1}^{n_i} p_{ij}^2.
$$

Taking the first and secondary derivatives of the log likelihood and recalling (3.2), we have:

$$
\begin{aligned}
\frac{\partial l_i}{\partial h_i} &= \delta_{i\cdot\cdot} - \Lambda_{i\cdot\cdot} - e^{-\tau_1} h_i \,, \\
\frac{\partial l_i}{\partial p_{ij}} &= \delta_{ij\cdot} - \Lambda_{ij\cdot} - e^{-\tau_2} p_{ij} \,, \\
\frac{\partial^2 l_i}{\partial h_i^2} &= -\Lambda_{i\cdot\cdot} - e^{-\tau_1} \,, \\
\frac{\partial^2 l_i}{\partial p_{ij}^2} &= -\Lambda_{ij\cdot} - e^{-\tau_2} \,, \\
\frac{\partial^2 l_i}{\partial h_i \partial p_{ij}} &= -\Lambda_{ij\cdot} \,.
\end{aligned}
$$

Let $\boldsymbol{r}_i = (h_i, p_{i1}, \ldots, p_{in_i})^T$, and $\boldsymbol{r} = (\boldsymbol{r}_1^T, \ldots, \boldsymbol{r}_s^T)^T$. The score functions can

be written as

$$\frac{\partial l(\boldsymbol{h},\boldsymbol{p})}{\partial \boldsymbol{r}_i} = \frac{\partial l_i}{\partial \boldsymbol{r}_i} = \left(\frac{\partial l_i}{\partial h_i}, \frac{\partial l_i}{\partial p_{i1}}, \ldots, \frac{\partial l_i}{\partial p_{in_i}}\right)^T \Rightarrow$$
$$\frac{\partial l(\boldsymbol{h},\boldsymbol{p})}{\partial \boldsymbol{r}} = \left(\left(\frac{\partial l_i}{\partial \boldsymbol{r}_1}\right)^T, \left(\frac{\partial l_i}{\partial \boldsymbol{r}_2}\right)^T, \ldots, \left(\frac{\partial l_i}{\partial \boldsymbol{r}_s}\right)^T\right)^T,$$

and the Hessian matrix is

$$\frac{\partial^2 l_i}{\partial \boldsymbol{r}_i \partial \boldsymbol{r}_i^T} = \begin{pmatrix} \frac{\partial^2 l_i}{\partial h_i^2} & \frac{\partial^2 l_i}{\partial h_i \partial p_{i1}} & \cdots & \frac{\partial^2 l_i}{\partial h_i \partial p_{in_i}} \\ \frac{\partial^2 l_i}{\partial h_i \partial p_{i1}} & \frac{\partial^2 l_i}{\partial p_{i1}^2} & & 0 \\ \vdots & & \ddots & \\ \frac{\partial^2 l_i}{\partial h_i \partial p_{in_i}} & 0 & & \frac{\partial^2 l_i}{\partial p_{in_i}^2} \end{pmatrix} \Rightarrow$$
$$\frac{\partial^2 l(\boldsymbol{h},\boldsymbol{p};\boldsymbol{c},\boldsymbol{\beta},\boldsymbol{\tau})}{\partial \boldsymbol{r} \partial \boldsymbol{r}^T} = \begin{pmatrix} \frac{\partial^2 l_i}{\partial \boldsymbol{r}_1^2} & 0 & \cdots & 0 \\ 0 & \frac{\partial^2 l_i}{\partial \boldsymbol{r}_2^2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \frac{\partial^2 l_i}{\partial \boldsymbol{r}_s^2} \end{pmatrix}.$$

Then we can estimate the conditional modes $\hat{h}_i$'s and $\hat{p}_{ij}$'s of the random effects and their standard errors $\hat{\sigma}_{h_i}$'s and $\hat{\sigma}_{p_{ij}}$ by maximizing $l(\boldsymbol{h},\boldsymbol{p};\boldsymbol{c},\boldsymbol{\beta},\boldsymbol{\tau})$ using Newton-Raphson method. Since the Hessian is a block diagonal matrix, $\boldsymbol{r}_i$'s can be estimated separately and efficiently for each i, $i = 1, \ldots, s$. Let

$$\tilde{h}_i = (h_i - \hat{h}_i) \,/\, (\sqrt{2}\hat{\sigma}_{h_i}), \quad \tilde{p}_{ij} = (p_{ij} - \hat{p}_{ij}) \,/\, (\sqrt{2}\hat{\sigma}_{p_{ij}}),$$

then

$$h_i = \hat{h}_i + \sqrt{2}\hat{\sigma}_{h_i}\,\tilde{h}_i, \quad p_{ij} = \hat{p}_{ij} + \sqrt{2}\hat{\sigma}_{p_{ij}}\,\tilde{p}_{ij}.$$

As an example of applying AGHQ, $g_i(\boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{\tau})$ is approximated as follows:

$$
\begin{aligned}
& g_i(\boldsymbol{c}, \boldsymbol{\beta}, \boldsymbol{\tau}) \\
&= \sqrt{2}\hat{\sigma}_{h_i} \int \left\{ \prod_{j=1}^{n_i} \left[\sqrt{2}\hat{\sigma}_{p_{ij}} \int \tilde{g}_{ij}(\hat{h}_i + \sqrt{2}\hat{\sigma}_{h_i}\,\tilde{h}_i, \quad \hat{p}_{ij} + \sqrt{2}\hat{\sigma}_{p_{ij}}\,\tilde{p}_{ij}; \boldsymbol{c}, \boldsymbol{\beta}) \right. \right. \\
&\quad \left. \left. \times\ f_{p_{ij}}\left(\hat{p}_{ij} + \sqrt{2}\hat{\sigma}_{p_{ij}}\,\tilde{p}_{ij}\right)\, d\tilde{p}_{ij} \right] \times f_{h_i}\left(\hat{h}_i + \sqrt{2}\hat{\sigma}_{h_i}\,\tilde{h}_i\right) \right\} d\tilde{h}_i \\
&\approx \sqrt{2}\hat{\sigma}_{h_i} \sum_{l_i=1}^{d} \left[\prod_{j=1}^{n_i} \left\{ \sqrt{2}\hat{\sigma}_{p_{ij}} \sum_{l_{ij}=1}^{d} \tilde{g}_{ij}(\hat{h}_i + \sqrt{2}\hat{\sigma}_{h_i}\,a_{l_i}, \quad \hat{p}_{ij} + \sqrt{2}\hat{\sigma}_{p_{ij}}\,a_{l_{ij}}; \ \boldsymbol{c}, \boldsymbol{\beta}, \tau_2) \right. \right. \\
&\quad \left. \left. \times\ f_{p_{ij}}(\hat{p}_{ij} + \sqrt{2}\hat{\sigma}_{p_{ij}}\,a_{l_{ij}}) \times e^{a_{l_{ij}}^2} \times w_{l_{ij}} \right\} \times f_{h_i}(\hat{h}_i + \sqrt{2}\hat{\sigma}_{h_i} a_{l_i}) \times e^{a_{l_i}^2} \times w_{l_i} \right].
\end{aligned}
$$

4.5 BFGS Algorithm

Newton-Raphson has a very fast convergence rate. However, in cases where evaluation of the Hessian matrix is impractical or very time-consuming, quasi-Newton methods become appealing since they do not require second derivatives. As discussed in Section 1.3.1, the regular Newton-Raphson method may fail to converge since it is sensitive to initial values. Though this issue can be avoided by using the Gauss-Seidel method in most cases, instability may still exist: the Hessian matrix may be numerically noninvertible, or the algorithm just does not converge. Furthermore, it is time-consuming to compute the Hessian matrix due to the extensive use of Gauss-Hermite quadrature. For these reasons we pursue a quasi-Newton method. Instead of computing the Hessian that requires the second derivative of the log likelihood and taking its inverse, the quasi-Newton method directly computes an approximate Hessian or its inverse, which can be updated from its previous estimate and the two most recent estimates of the parameters. Nocedal and Wright (2006, Chapter 6) give

a detailed introduction to different quasi-Newton methods. We will review the basic idea and focus on one popular quasi-Newton method - the BFGS method.

Let H_{k-1} and B_{k-1} be the previous approximation of the Hessian matrix and its inverse. Let θ_{k-1} and θ_k be the two most recent estimates of the parameters, respectively. Let s_{k-1} and s_k be the score functions evaluated at θ_{k-1} and θ_k, respectively. The parameter can be updated iteratively from the Newton-Raphson formula

$$\theta_k = \theta_{k-1} - B_{k-1}\, s_{k-1}\,.$$

Define $\Delta\theta_k = \theta_k - \theta_{k-1}$, $\Delta s_k = s_k - s_{k-1}$. We first derive the quasi-Newton equation (or secant equation) by Taylor's expansion on the score function:

$$s_k = f'(\theta_k) = f'(\theta_{k-1} + \Delta\theta_k) \approx s_{k-1} + H_k \Delta\theta_k$$

$$\Rightarrow B_k\, \Delta s_k = \Delta\theta_k\,.$$

Then the inverse Hessian can be updated based upon the previous approximation B_{k-1} by solving the following constrained optimization problem:

$$\min_B \|B - B_{k-1}\| \text{ subject to } B = B^T \quad \text{and } B\,\Delta s_k = \Delta\theta_k\,. \tag{4.4}$$

Different norms can be used in (4.4), which yield different quasi-Newton methods.

The BFGS method is the most popular quasi-Newton algorithm. It uses the weighted Frobenius norm and updates the inverse of Hessian at each iteration. The weighted Frobenius norm is defined as

$$\|A\|_W = \left\|W^{1/2}\, A\, W^{1/2}\right\|_F\,,$$

where $\|C\|_F^2 = \sum_i \sum_j c_{ij}^2 = trace(CC^{T)}$. For BFGS method, the weight matrix W can be chosen as any matrix satisfying the equation

$$W\,\Delta\theta_k = \Delta s_k\,. \tag{4.5}$$

For concreteness, let W be the average Hessian

$$\bar{H}_k = \int_0^1 H(\theta_{k-1} + \tau\Delta\theta_k)\,d\tau,$$

where H is the true Hessian matrix. Note that H is positive definite assuming that the objective function has a global minimum.

By Taylor expansion theorem, if the objective function $f(\theta)$ is continuously twice differentiable, then we have

$$\begin{aligned} f'(\theta_k) &= f'(\theta_{k-1}) + \int_0^1 f''(\theta_{k-1} + t\,\Delta\theta_k)\,\Delta\theta_k\,dt \\ &= f'(\theta_{k-1}) + \bar{H}_{k-1}\,\Delta\theta_k\,. \end{aligned}$$

Thus, equation (4.5) holds for $W = \bar{H}_k$.

The unique solution of (4.4) is given by

$$B_k = \left\{I - \rho_k\,\Delta\theta_k\,\Delta s_k^T\right\}\,B_{k-1}\,\left\{I - \rho_k\,\Delta s_k\,\Delta\theta_k^T\right\} + \rho_k\,\Delta s_k\,\Delta s_k^T\,.$$

where $\rho_k = \frac{1}{\Delta s_k^T\,\Delta\theta_k}$ (see Appendix A for the derivation).

The BFGS method has the following appealing properties:

1. Self-correcting: if the approximation of the inverse Hessian B_k gets too far away and slows down the iteration, then the approximation procedure will tend to correct itself within a few steps.

2. Positive definite: if B_{k-1} is positive definite, then B_k will be still positive definite.

3. Efficient: each iteration has lower cost as compared to the Newton-Raphson method though BFGS requires relatively more iterations.

CHAPTER 5
SIMULATION STUDIES

In this chapter, we explore the performance and asymptotic properties of the estimation results through simulation studies.

5.1 Data Generation

First, we need to generate the data for the simulations. Since the model involves nested clusters, sample size includes not only the total number of subjects, but also the number of hospitals, the numbers of physicians within each hospital and the numbers of patients under each physician. In simulation studies, we only consider a balanced design, i.e. the clusters at each level have the same size. The failure times are generated using the cumulative hazard model (3.4), where the baseline hazard function is from a Weibull distributed failure time. The censoring times are generated independently from an exponential distribution with a baseline hazard rate which depends on the same covariates as the failure time model. To control the proportion of censoring (right censoring only) at a certain level in simulations, we need to choose an appropriate constant for the baseline hazard so that the censoring probability has the desired mean.

The data were generated as follows:

1. Determine the sample size $s \times n_i \times n_{ij}$, where

   $s =$ the number of hospitals,

n_i = the number of physicians within each hospital,

n_{ij} = the number of patients per physician.

2. Generate the covariates $\boldsymbol{Z}_{ijk} = (Z_{1,i}, Z_{2,ij}, Z_{3,ijk})^T$ for each patient where

   $Z_{1,i} \sim$ Uniform $(0, 1)$: hospital-specific covariate, $i = 1, \ldots, s$,

   $Z_{2,ij} \sim$ Bernoulli (0.5): physician-specific covariate, $i = 1, \ldots, s$ and $j = 1, \ldots, n_i$,

   $Z_{3,ijk} \sim$ Normal $(0, 1)$: patient-specific covariate, $i = 1, \ldots, s$, $j = 1, \ldots, n_i$ and $k = 1, \ldots, n_{ij}$.

3. Generate the frailty terms h_i and p_{ij} for $i = 1, \ldots, s$ and $j = 1, \ldots, n_i$:

   $h_i \sim$ Normal $(0, \theta_1)$,

   $p_{ij} \sim$ Normal $(0, \theta_2)$.

4. Generate the failure time T_{ijk} with cumulative hazard function $\Lambda(t \mid \boldsymbol{z}_{ijk})$, where $\Lambda_0(t)$ is of Weibull(ρ, α) form, i.e. $\Lambda_0(t) = (t/\alpha)^\rho$. The generating form for T_{ijk} is derived as follows:

$$
\begin{aligned}
S(T_{ijk}) &= \exp\{-\Lambda(T_{ijk} \mid \boldsymbol{Z}_{ijk}, h_i, p_{ij})\} \Rightarrow \\
U_{ijk} &= \exp\left\{-(t/\alpha)^\rho_{ijk} \times e^{\boldsymbol{\beta}^T \boldsymbol{Z}_{ijk} + h_i + p_{ij}}\right\} \Rightarrow \\
T_{ijk} &= \alpha\left\{-\log(U_{ijk}) \times \exp(-\boldsymbol{\beta}^T \boldsymbol{Z}_{ijk} - h_i - p_{ij})\right\}^{1/\rho},
\end{aligned}
$$

   where $U_{ijk} \sim$ Uniform $(0, 1)$ since $S(T) \sim$ Uniform $(0, 1)$.

5. Generate the censoring time C_{ijk} similarly as shown in step (4) with the hazard function

$$\lambda(c_{ijk} \mid \boldsymbol{z}_{ijk}) = \gamma \exp\left(\boldsymbol{\eta}^T \boldsymbol{z}_{ijk}\right), \tag{5.1}$$

where $\boldsymbol{\eta}$ is a vector of predetermined coefficients for the covariates, and γ is the constant baseline hazard to be determined by the desired proportion of censoring (20% for all simulations unless it is specified). The censoring probability $P(T > C)$ can be written as

$$\begin{aligned}
&E_{\boldsymbol{Z},h,p}\left[P\left(T > C \mid \boldsymbol{Z}, h, p\right)\right] \qquad (5.2)\\
&= E_{\boldsymbol{Z},h,p}\left[\int_0^\infty \int_0^t f_T(t) f_C(s)\, ds dt\right]\\
&= E_{\boldsymbol{Z},h,p}\left[\int_0^\infty f_T(t)\left[1 - S_C(t)\right]\, dt\right]\\
&= E_{\boldsymbol{Z},h,p}\Bigg[\int_0^\infty \alpha^{-\rho}\rho t^{\rho-1} \exp\left(\boldsymbol{\beta}^T\boldsymbol{Z} + h + p - \alpha^{-\rho} t^\rho e^{\boldsymbol{\beta}^T\boldsymbol{Z}+h+p}\right)\\
&\quad \times \left\{1 - \exp\left(-\gamma t e^{\boldsymbol{\eta}^T\boldsymbol{Z}}\right)\right\}\, dt\Bigg].
\end{aligned}$$

Define

$$\begin{aligned}
g(\boldsymbol{Z}, h, p; \gamma) = \int_0^\infty &(\rho/\alpha)\,(t/\alpha)^{\rho-1} \exp\left(\boldsymbol{\beta}^T\boldsymbol{Z} + h + p - \alpha^{-\rho} t^\rho e^{\boldsymbol{\beta}^T\boldsymbol{Z}+h+p}\right)\\
&\times \left\{1 - \exp\left(-\gamma t e^{\boldsymbol{\eta}^T\boldsymbol{Z}}\right)\right\}\, dt.
\end{aligned}$$

Recalling that $Z_1 \sim \text{Uniform}\,(0,1)$, $Z_2 \sim \text{Bernoulli}\,(\frac{1}{2})$ and $Z_3 \sim \text{Normal}\,(0,1)$, then the expectation (5.2) can be computed as follows

$$\begin{aligned}
&E_{\boldsymbol{Z},h,p}\left[\Pr\left(T > C \mid \boldsymbol{Z}, h, p\right)\right]\\
&= \int_0^1 \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \frac{1}{2}\left\{g(z_1, 0, z_3; \gamma) + g(z_1, 1, z_3; \gamma)\right\}\\
&\quad \times \theta_1^{-1/2} \varphi\left(h\theta_1^{-1/2}\right) \times \theta_2^{-1/2} \varphi\left(p\theta_2^{-1/2}\right) \times \varphi\left(z_3\right)\, dh\, dp\, dz_3\, dz_1,
\end{aligned}$$

where $\varphi(\cdot)$ is the density function of standard normal distribution. Therefore γ in (5.1) can be solved by the equation $P(T > C) = 0.2$.

6. Determine the observed time $X_{ijk} = \min(T_{ijk}, C_{ijk})$ and the indicator of failure $\delta_{ijk} = 1_{[T_{ijk} \leq C_{ijk}]}$.

5.2 Simulation and Results

We implement the model estimation by applying the techniques discussed in earlier chapters, such as B-splines, AGHQ, and BFGS method.

In the iterative algorithm, initial values of the parameters to be estimated must be provided. The initial values of the frailty variance parameters (θ_1, θ_2) are simply set as one, i.e. $(\tau_1, \tau_2) = (0, 0)$. Initial values for the regression coefficients are obtained by fitting the data to an ordinary Cox proportional-hazards regression model pretending that there is no cluster effect. Finally, the Breslow estimators are obtained for the cumulative hazard function at distinct observed failure times. Subsequently, by taking the log on both sides of (3.3) and replacing the baseline cumulative hazards by the Breslow estimators, we calculate the least square estimators for the spline coefficients as their initial values.

We use cubic B-splines for the good balance of its continuity property (twice differentiable) and computational cost. For the AGHQ method, the number of nodes is fixed at 10 for all simulations. We found 10 is enough for accurate approximation since there is no significant improvement when it increases to 20. To estimate the model using the BFGS algorithm, we adopt the R built-in function `constrOptim`.

Since the Gauss-Seidel method and BFGS method achieve very close results according to pilot simulations, all simulation results presented in this dissertation are obtained using BFGS method for the efficiency purpose. We run 500 simulations for each setting.

5.2.1 Two-Level Frailty Model

In this dissertation, we only consider right censored survival data. To evaluate the model estimation, we first see how the censoring rate affects the results. Table 5.1 shows that the estimates get significantly worse in terms of the bias, the root of mean square error (RMSE) and the coverage as the censoring rate goes up. The censoring rate is fixed at 20% for the following simulation results.

Next we compare the performance of non-adaptive and adaptive GHQ methods. Table 5.2 shows the results from both methods at different values of $\boldsymbol{\theta}$: (0.2, 0.8), (0.4, 0.8) and (0.4, 0.2). When $\boldsymbol{\theta} = (0.2, 0.8)$, there is no significant difference between the non-adaptive and adaptive methods. However, when θ_1 increases to 0.4 and θ_2 is fixed at 0.8, the estimates of standard error of β_1 and θ_1 become worse for the non-adaptive method, in terms of the coverage probability, while the estimates of β_2, β_3, and θ_2 from both methods do not show much difference: the adaptive GHQ performs slightly better. It can be further observed that the relative size of θ_1 compared to θ_2 plays the key role in exhibiting difference between GHQ and AGHQ. When $\boldsymbol{\theta} = (0.4, 0.2)$, compared to $(0.4, 0.8)$, θ_1 is the same while θ_2 is smaller. Therefore we shall expect better results overall for $\boldsymbol{\theta} = (0.4, 0.2)$ (see the

discussion in the next paragraph). However, the non-adaptive method yields even worse estimates of the standard error of β_1 and θ_1. This indicates that the failure of non-adaptive GHQ is caused by the relatively large frailty at the hospital level and the small number of hospitals, which leads to poor estimation of the Hessian matrix and thus poor estimation of the standard errors.

The frailties, which in practice measure the association within any cluster or heterogeneity among clusters, play an important role in model estimation. We are interested in how well the frailty variance parameters (θ_1 and θ_2) can be estimated and how they affect the estimation of the covariate coefficients. Table 5.3 lists the simulation results for different values of θ_1 and θ_2. We first conclude that there exists underestimation of the parameters $(\boldsymbol{\beta}, \boldsymbol{\theta})$ when the frailties at either level are significant (e.g. the corresponding variance is 0.2 or 0.8), because the estimates on average are less than the true value, and the upper bound (UB) error rate is larger than lower bound (LB) error rate in most cases. Furthermore, it is observed that the larger frailty variance (especially at the physician level) impairs the estimation performance of the regression coefficients and θ_2, the variance of physician-level effects, . In Table 5.3, when θ_1 gets smaller from 0.8 to 0.01 and θ_2 is fixed at 0.01 or 0.2, the estimation of β_1 is significantly improved in terms of the bias, SD and the RMSE. There is very little improvement on β_2 and almost no change on β_3. When θ_2 gets smaller from 0.2 to 0.01 and θ_1 is fixed at 0.01 or 0.8, the estimation of all regression coefficients is significantly improved in terms of the bias, SD and the RMSE.

We also study the effect of the sample size. Table 5.4 shows the results for

different numbers of s 5, 10 and 20 (n_i and n_{ij} are fixed at 10). When the number of hospitals doubles, all sizes are doubled. So it is reasonable to see the improvements on the estimates of all parameters. Similarly, Tables 5.5 and 5.6 show the results for different sizes of physician level and patient level respectively. When n_i or n_{ij} increases from 5 to 20 and the other sizes are fixed at 10, the estimates of β_2, β_3 and θ_2 get significantly improved, while the estimates of β_1 and θ_1 do not show a trend of improvement. The latter phenomenon could be interpreted by the large variation due to the small number of hospitals. Overall, the simulation results in Tables 5.4 to 5.6 seem consistent with our expectation based upon the large sample theory.

Table 5.7 shows the simulation results for different combinations of numbers of s and n_i, 20×5, 10×10 and 5×20, while n_{ij} is fixed at 10. As s increases, the estimation of the hospital-level parameters, $beta_1$ and θ_1, gets better in terms of RMSE even though the n_i decreases. It is also reasonable to see almost no change in the estimation of $beta_3$ in terms of RMSE because the total number of patients and the number of patients with each physician are the same for all three settings. As for the estimation of β_2 and θ_2, both the bias and the standard deviation gets larger as s increases and n_i decreases. This may indicate that, given the total number of physicians is fixed, larger number of physicians nested in each hospital leads to better estimation of $beta_2$ and θ_2.

Table 5.8 shows the results for different combinations of n_i and n_{ij}, 20×5, 10×10 and 5×20, while s is fixed at 10.where the total number of patients is the same for the three settings. As n_i increases and n_{ij} decreases, the estimation of β_2 is

improved while there is no trend of change in the estimation of β_1, θ_1 and θ_2 and the estimation of β_3 gets even worse in terms of RMSE. It may indicate that, given the total number of patients is fixed, larger number of patients with each physician leads to better estimation of β_3.

In Tables 5.7 and 5.8, by looking at the change in percent of RMSE, it can be further concluded that the change of s has more significant effect on estimates of β_1 and θ_1, compared to the effect from the change of n_i on estimates of β_2 and θ_2. It is also true that the effect on β_2 from the change of n_i is more significant than the effect on β_3 from the change of n_{ij}. These facts provide some clue to choose the sample size given the total size is fixed: larger size of top level is preferred.

Last we provide a graphical view to examine the estimates of the parameters. Figure 5.1 shows the plots of the true baseline (cumulative) hazard functions and their estimated curves. Note that the B-splines are defined over a finite interval, which is determined by the minimum and maximum observed times. Consequently the finite interval used to define the B-splines for each simulation is different and the plots of hazards were made over the intersection of all the 500 finite intervals. The blue dashed-and-dotted line is the estimated baseline (cumulative) hazard function, which is computed as the average of the estimates from the 500 simulations. The points on the two red dashed lines correspond to the 2.5% and 97.5% percentiles of the 500 estimates respectively. It is observed that the estimated curves of the baseline (cumulative) hazard functions fit the true curves pretty well. However, due to fewer failures observed at late times, it is reasonable to see that the estimated

curves deviate more from the true curve and the lower and upper limits get larger as time increases. Figure 5.2 shows the Q-Q plots for the estimates of all parameters and the estimates of log baseline cumulative hazard function at the midpoint of the time range in the plot. All plots, except the one for τ_2 ($\log \theta_2$), fit a straight line very well, which demonstrates asymptotic normality. However, the estimates of τ_2 exhibit left skewness in the Q-Q plot, which could be due in part to its relative small true value ($\theta_2 = 0.2$).

5.2.2 Comparison with Penalized Partial Likelihood Approach

The R function `coxph` from the `survival` package has an option to estimate lognormal frailty models based on a penalized partial likelihood. But it can only handle a one-level frailty. It does not provide an estimate of its standard error. Besides these limitations of PPL approach, we expect to see some advantage made by the proposed approach, under different choices of the frailty variances (either two-level or one-level). Table 5.9 to 5.11 show the simulation results for $\boldsymbol{\theta} = (0.8, 0), (0, 0.8)$ and $(0.4, 0.4)$, respectively. For each case, we fit the data with a two-level model using the proposed approach, one-level frailty (either the hospital or the physician level) models using both approaches, and an ordinary Cox proportional hazards regression model with no frailties. We will evaluate each approach based upon the estimates of the regression coefficients $\boldsymbol{\beta}$.

In Table 5.9, the data were generated with frailty only at the hospital level, $\boldsymbol{\theta} = (0.8, 0)$. So the one-level (hospital) frailty model is the right model. When

the correct model is specified, both approaches reach very similar results for β_2 and β_3, while our proposed approach yields a much better estimate of β_1 and less biased estimate of θ_1. Actually the results from the PPL approach are only based on the 183 simulations out of the 500 in total, because it failed to estimate β_1 in the remaining simulations, probably due to the small number of hospitals for the sample size is $20 \times 10 \times 5$. The wrong models (no frailty or frailty existed at the physician level), however, yield poor results: larger bias and worse estimates of the standard errors compared to the correct model. Moreover, it is observed that the two-level frailty model performs almost as well as the correct model, the one-level (hospital) frailty model.

In Table 5.10, the data were generated with frailty only at the physician level, $\boldsymbol{\theta} = (0, 0.8)$. Then the one-level (physician) frailty model is the correct one. By comparing the results from the proposed two-level model approach and the correct one-level (physician) model using both approaches, we observe that the three approaches perform similarly in estimating the regression coefficients $\boldsymbol{\beta}$, though substantially biased, while the PPL approach yields slightly better estimates of θ_2 than the other two (0.650 versus 0.602 and 0.593). We also observe that large variation exists in the estimation of β_1: the wrong models (no frailty or frailty existed within the hospital level) somehow yield better estimates of β_1.

Table 5.9 and 5.10 show that the proposed two-level model approach performs almost as well as the correct one-level model approach on the estimation of the regression coefficients. This fact implies an important advantage of our proposed approach

for two-level clustered survival data: if in practice it is hard to tell at which level the cluster effect exists, it is always safe to fit a two-level frailty model.

In Table 5.11, the data were generated with frailties at both levels, $\boldsymbol{\theta} = (0.4, 0.4)$. The two-level frailty model is the only correct model, and it reaches the best estimates of β_2 and β_3 among all approaches. Estimation of β_1 still exhibits large variation, as discussed earlier.

In summary, our proposed two-level approach works reasonably well for two-level clustered survival data even when the cluster effect exists at only one level. The PPL approach was designed only for one-level clustered survival data and failed to converge for roughly 50-60% of the datasets when a cluster effect existed at the hospital level. This may be due in part to the small to moderate number of hospitals (20).

5.2.3 Comparison with Hierarchical Likelihood Approach

The hierarchical likelihood (HL) approach, introduced in Section 3.4, has been implemented in the R package `frailtyHL` , which can be used to estimate the two-level lognormal frailty model. In `frailtyHL`, as mentioned in Section 3.4, the order of Laplace approximation to simply the marginal hierarchical likelihood in estimating the regression coefficients $\boldsymbol{\beta}$ and the frailty variances $\boldsymbol{\theta}$, respectively, needs to be specified. The allowable order is 0 or 1 for $\boldsymbol{\beta}$, and 1 or 2 for $\boldsymbol{\theta}$. Since the preliminary simulations showed no big difference among different choices of orders of Laplace approximation, the first-order Laplace approximation is used for all simulations. The

resulting HL approach is denoted as $HL(1,1)$.

The running time of $HL(1,1)$ for each simulated dataset with sample size $20 \times 10 \times 5$ is around 20 minutes, which is three times longer than our proposed approach. This could be due to the intensive computing in iteratively estimating the 200 random effect variables given the frailty variances.

Table 5.12 shows the simulation results from the proposed approach and the $HL(1,1)$ approach for $\boldsymbol{\theta} = (0.2, 0.8)$ and $\boldsymbol{\theta} = (0.8, 0.2)$. Generally speaking, the two approaches performed closely in terms of RMSE in estimating both the regression coefficients $\boldsymbol{\beta}$ and the frailty variances $\boldsymbol{\theta}$. However, when $\boldsymbol{\theta} = (0.2, 0.8)$, the proposed approach yielded greater estimation bias of β_1 and θ_1. Actually the underestimation of β_1 and θ_1 existed across most simulations presented in this chapter and it became more prominent when the number of hospitals is small or the value of θ_2 is large. This indicates the proposed approach based on maximum likelihood estimation may need some modification in the future to reduce the estimation bias of the hospital-level parameters. We will further discuss this issue in Chapter 6.

Table 5.1: Simulation results for different censoring rates: 10%, 20% and 40%.

Sample Size *	$20\times10\times5$		
Censored (%)	10	20	40
TRUE β	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8
Ave $\hat{\beta}$	-0.236, 0.383, 0.787	-0.251, 0.368, 0.775	-0.296, 0.331, 0.745
SD ($\hat{\beta}$)	0.791, 0.099, 0.047	0.781, 0.103, 0.051	0.757, 0.112, 0.061
Ave SE ($\hat{\beta}$)	0.694, 0.1, 0.047	0.686, 0.104, 0.051	0.664, 0.117, 0.061
RMSE($\hat{\beta}$)	0.792, 0.101, 0.048	0.783, 0.108, 0.057	0.763, 0.132, 0.082
Coverage †	0.916, 0.946, 0.942	0.912, 0.922, 0.902	0.916, 0.916, 0.834
LB Error ‡	0.032, 0.014, 0.014	0.03, 0.016, 0.008	0.02, 0.002, 0
UB Error §	0.052, 0.04, 0.044	0.058, 0.062, 0.09	0.064, 0.082, 0.166
TRUE θ	0.8, 0.2	0.8, 0.2	0.8, 0.2
Ave $\hat{\theta}$	0.706, 0.194	0.683, 0.188	0.622, 0.175
SD ($\hat{\theta}$)	0.255, 0.056	0.248, 0.062	0.235, 0.073
Ave SE ($\hat{\theta}$)	0.242, 0.053	0.238, 0.057	0.227, 0.068
RMSE($\hat{\theta}$)	0.272, 0.057	0.274, 0.063	0.295, 0.077
Coverage	0.91, 0.956	0.916, 0.96	0.848, 0.978
LB Error	0.01, 0.034	0.006, 0.034	0.002, 0.022
UB Error	0.08, 0.01	0.078, 0.006	0.15, 0

*Numbers of hospitals × physicians/hospital × patients/physician.

†Coverage by 95% confidence interval.

‡Proportion of the true value below the lower confidence bounds.

§Proportion of the true value above the upper confidence bounds.

Table 5.2: Simulation Results for non-adaptive and adaptive GHQ methods.

Sample Size *	$20\times10\times5$					
GHQ Type	non-adaptive	adaptive	non-adaptive	adaptive	non-adaptive	adaptive
TRUE β	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8
Ave $\hat{\beta}$	-0.261, 0.328, 0.754	-0.262, 0.328, 0.754	-0.261, 0.325, 0.754	-0.266, 0.326, 0.754	-0.185, 0.365, 0.775	-0.229, 0.366, 0.775
SD ($\hat{\beta}$)	0.434, 0.146, 0.053	0.433, 0.146, 0.053	0.585, 0.146, 0.053	0.551, 0.146, 0.053	0.615, 0.105, 0.051	0.531, 0.104, 0.052
Ave SE ($\hat{\beta}$)	0.405, 0.148, 0.053	0.405, 0.148, 0.053	0.486, 0.149, 0.053	0.516, 0.149, 0.053	0.366, 0.102, 0.051	0.499, 0.103, 0.051
RMSE($\hat{\beta}$)	0.438, 0.163, 0.07	0.438, 0.163, 0.07	0.588, 0.164, 0.07	0.555, 0.164, 0.07	0.615, 0.111, 0.057	0.532, 0.109, 0.058
Coverage *	0.922, 0.914, 0.855	0.918, 0.918, 0.856	0.898, 0.91, 0.841	0.92, 0.92, 0.846	0.749, 0.912, 0.906	0.930, 0.922, 0.902
LB Error *	0.026, 0.004, 0.002	0.026, 0.004, 0.002	0.038, 0.01, 0.002	0.028, 0.006, 0.002	0.138, 0.018, 0.002	0.032, 0.012, 0.006
UB Error *	0.052, 0.082, 0.143	0.056, 0.078, 0.142	0.064, 0.08, 0.157	0.052, 0.074, 0.152	0.113, 0.07, 0.092	0.038, 0.066, 0.092
TRUE θ	0.2, 0.8	0.2, 0.8	0.4, 0.8	0.4, 0.8	0.4, 0.2	0.4, 0.2
Ave $\hat{\theta}$	0.151, 0.691	0.15, 0.691	0.316, 0.688	0.307, 0.69	0.379, 0.182	0.336, 0.182
SD ($\hat{\theta}$)	0.088, 0.133	0.087, 0.132	0.162, 0.132	0.139, 0.132	0.234, 0.058	0.130, 0.057
Ave SE ($\hat{\theta}$)	0.082, 0.128	0.081, 0.127	0.153, 0.128	0.133, 0.127	0.19, 0.057	0.124, 0.057
RMSE($\hat{\theta}$)	0.101, 0.173	0.1, 0.171	0.183, 0.173	0.167, 0.171	0.235, 0.061	0.145, 0.060
Coverage	0.996, 0.869	0.996, 0.868	0.93, 0.873	0.932, 0.87	0.887, 0.971	0.914, 0.984
LB Error	0.004, 0.004	0.004, 0.004	0.006, 0.002	0.002, 0.004	0.035, 0.016	0.002, 0.014
UB Error	0, 0.127	0, 0.128	0.064, 0.124	0.066, 0.126	0.078, 0.012	0.084, 0.002

*See footnotes in Table 5.1 (page 52).

Table 5.3: Simulation results for different values of $\boldsymbol{\theta}$: (0.8, 0.01), (0.8, 0.2), (0.01, 0.2) and (0.01, 0.01).

GHQ Type	$20\times10\times5$			
TRUE β	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8
Ave $\hat{\beta}$	-0.213, 0.391, 0.798	-0.251, 0.368, 0.775	-0.215, 0.368, 0.777	-0.197, 0.398, 0.804
SD ($\hat{\beta}$)	0.736, 0.082, 0.047	0.781, 0.103, 0.051	0.197, 0.099, 0.052	0.16, 0.079, 0.047
Ave SE ($\hat{\beta}$)	0.695, 0.08, 0.048	0.686, 0.104, 0.051	0.192, 0.099, 0.051	0.153, 0.077, 0.048
RMSE($\hat{\beta}$)	0.736, 0.083, 0.047	0.783, 0.108, 0.057	0.198, 0.104, 0.057	0.16, 0.079, 0.048
Coverage *	0.928, 0.932, 0.944	0.912, 0.922, 0.902	0.94, 0.93, 0.907	0.932, 0.937, 0.95
LB Error *	0.036, 0.02, 0.02	0.03, 0.016, 0.008	0.03, 0.01, 0.006	0.033, 0.022, 0.022
UB Error *	0.036, 0.048, 0.036	0.058, 0.062, 0.09	0.03, 0.06, 0.087	0.035, 0.041, 0.028
TRUE θ	0.8, 0.01	0.8, 0.2	0.01, 0.2	0.01, 0.01
Ave $\hat{\theta}$	0.716, 0.015	0.683, 0.188	0.009, 0.177	0.008, 0.015
SD ($\hat{\theta}$)	0.254, 0.021	0.248, 0.062	0.013, 0.054	0.01, 0.02
Ave SE ($\hat{\theta}$)	0.242, 0.019	0.238, 0.057	0.012, 0.055	0.008, 0.02
RMSE($\hat{\theta}$)	0.268, 0.021	0.274, 0.063	0.013, 0.059	0.01, 0.021
Coverage	0.918, 0.906	0.914, 0.964	0.954, 0.99	0.965, 0.904
LB Error	0.004, 0.094	0.006, 0.034	0.046, 0.008	0.035, 0.096
UB Error	0.078, 0	0.08, 0.002	0, 0.002	0, 0

*See footnotes in Table 5.2 (page 52)

Table 5.4: Simulation results for different hospital sizes: 5, 10, 20.

GHQ Type	5×10 × 10	10×10 × 10	20×10 × 10
TRUE β	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8
Ave $\hat{\beta}$	-0.167, 0.379, 0.787	-0.168, 0.384, 0.786	-0.225, 0.38, 0.785
SD ($\hat{\beta}$)	1.875, 0.174, 0.068	1.075, 0.127, 0.049	0.713, 0.087, 0.035
Ave SE ($\hat{\beta}$)	1.281, 0.168, 0.069	0.954, 0.122, 0.048	0.667, 0.086, 0.034
RMSE($\hat{\beta}$)	1.875, 0.176, 0.069	1.076, 0.128, 0.051	0.714, 0.09, 0.039
Coverage *	0.776, 0.932, 0.944	0.9, 0.928, 0.944	0.924, 0.934, 0.918
LB Error *	0.1, 0.02, 0.012	0.048, 0.026, 0.006	0.03, 0.022, 0.012
UB Error *	0.124, 0.048, 0.044	0.052, 0.046, 0.05	0.046, 0.044, 0.07
TRUE θ	0.8, 0.2	0.8, 0.2	0.8, 0.2
Ave $\hat{\theta}$	0.46, 0.18	0.63, 0.189	0.674, 0.189
SD ($\hat{\theta}$)	0.43, 0.074	0.322, 0.055	0.253, 0.038
Ave SE ($\hat{\theta}$)	0.311, 0.072	0.3, 0.053	0.227, 0.037
RMSE($\hat{\theta}$)	0.548, 0.076	0.364, 0.056	0.282, 0.04
Coverage	0.768, 0.982	0.85, 0.96	0.88, 0.938
LB Error	0.006, 0.018	0.004, 0.014	0.006, 0.02
UB Error	0.226, 0	0.146, 0.026	0.114, 0.042

*See footnotes in Table 5.1 (page 52).

Table 5.5: Simulation results for different physician sizes: 5, 10, 20.

GHQ Type	10×5 × 10	10×10 × 10	10×20 × 10
TRUE β	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8
Ave $\hat{\beta}$	-0.219, 0.375, 0.785	-0.168, 0.384, 0.786	-0.18, 0.373, 0.786
SD ($\hat{\beta}$)	1.129, 0.185, 0.07	1.075, 0.127, 0.049	1.102, 0.089, 0.036
Ave SE ($\hat{\beta}$)	0.947, 0.181, 0.069	0.954, 0.122, 0.048	0.924, 0.084, 0.034
RMSE($\hat{\beta}$)	1.129, 0.186, 0.072	1.076, 0.128, 0.051	1.102, 0.093, 0.039
Coverage *	0.862, 0.928, 0.944	0.9, 0.928, 0.944	0.888, 0.91, 0.92
LB Error *	0.062, 0.02, 0.018	0.048, 0.026, 0.006	0.06, 0.022, 0.008
UB Error *	0.076, 0.052, 0.038	0.052, 0.046, 0.05	0.052, 0.068, 0.072
TRUE θ	0.8, 0.2	0.8, 0.2	0.8, 0.2
Ave $\hat{\theta}$	0.596, 0.187	0.63, 0.189	0.618, 0.191
SD ($\hat{\theta}$)	0.342, 0.078	0.322, 0.055	0.317, 0.036
Ave SE ($\hat{\theta}$)	0.304, 0.079	0.3, 0.053	0.286, 0.037
RMSE($\hat{\theta}$)	0.398, 0.079	0.364, 0.056	0.366, 0.037
Coverage	0.882, 0.99	0.85, 0.96	0.854, 0.962
LB Error	0, 0.01	0.004, 0.014	0.006, 0.008
UB Error	0.118, 0	0.146, 0.026	0.14, 0.03

*See footnotes in Table 5.1 (page 52).

Table 5.6: Simulation results for different patient sizes: 5, 10, 20.

GHQ Type	10×10 × 5	10×10 × 10	10×10 × 20
TRUE β	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8
Ave $\hat{\beta}$	-0.226, 0.371, 0.779	-0.168, 0.384, 0.786	-0.265, 0.38, 0.791
SD ($\hat{\beta}$)	1.143, 0.15, 0.071	1.075, 0.127, 0.049	1.17, 0.109, 0.032
Ave SE ($\hat{\beta}$)	0.931, 0.145, 0.072	0.954, 0.122, 0.048	0.941, 0.107, 0.033
RMSE($\hat{\beta}$)	1.144, 0.153, 0.074	1.076, 0.128, 0.051	1.171, 0.111, 0.034
Coverage *	0.872, 0.93, 0.924	0.9, 0.928, 0.944	0.866, 0.934, 0.94
LB Error *	0.06, 0.016, 0.02	0.048, 0.026, 0.006	0.062, 0.022, 0.006
UB Error *	0.068, 0.054, 0.056	0.052, 0.046, 0.05	0.072, 0.044, 0.054
TRUE θ	0.8, 0.2	0.8, 0.2	0.8, 0.2
Ave $\hat{\theta}$	0.581, 0.177	0.63, 0.189	0.617, 0.19
SD ($\hat{\theta}$)	0.323, 0.081	0.322, 0.055	0.331, 0.042
Ave SE ($\hat{\theta}$)	0.289, 0.079	0.3, 0.053	0.289, 0.04
RMSE($\hat{\theta}$)	0.39, 0.084	0.364, 0.056	0.378, 0.043
Coverage	0.824, 0.98	0.85, 0.96	0.844, 0.954
LB Error	0, 0.02	0.004, 0.014	0.002, 0.006
UB Error	0.176, 0	0.146, 0.026	0.154, 0.04

*See footnotes in Table 5.1 (page 52).

Table 5.7: Simulation results for different sizes of hospitals and physicians: 5×20, 10×10, 20×5.

GHQ Type	$5 \times 20 \times 10$	$10 \times 10 \times 10$	$20 \times 5 \times 10$
TRUE β	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8
Ave $\hat{\beta}$	-0.063, 0.394, 0.788	-0.168, 0.384, 0.786	-0.24, 0.382, 0.785
SD ($\hat{\beta}$)	1.967, 0.126, 0.05	1.075, 0.127, 0.049	0.739, 0.128, 0.048
Ave SE ($\hat{\beta}$)	1.246, 0.119, 0.048	0.954, 0.122, 0.048	0.687, 0.127, 0.049
RMSE($\hat{\beta}$)	1.972, 0.126, 0.052	1.076, 0.128, 0.051	0.74, 0.129, 0.05
Coverage *	0.75, 0.932, 0.936	0.9, 0.928, 0.944	0.908, 0.948, 0.944
LB Error *	0.132, 0.034, 0.02	0.048, 0.026, 0.006	0.042, 0.012, 0.004
UB Error *	0.118, 0.034, 0.044	0.052, 0.046, 0.05	0.05, 0.04, 0.052
TRUE θ	0.8, 0.2	0.8, 0.2	0.8, 0.2
Ave $\hat{\theta}$	0.452, 0.193	0.63, 0.189	0.679, 0.185
SD ($\hat{\theta}$)	0.402, 0.051	0.322, 0.055	0.26, 0.058
Ave SE ($\hat{\theta}$)	0.298, 0.052	0.3, 0.053	0.242, 0.055
RMSE($\hat{\theta}$)	0.532, 0.052	0.364, 0.056	0.287, 0.06
Coverage	0.696, 0.966	0.85, 0.96	0.9, 0.952
LB Error	0, 0.014	0.004, 0.014	0.002, 0.024
UB Error	0.304, 0.02	0.146, 0.026	0.098, 0.024

*See footnotes in Table 5.1 (page 52).

Table 5.8: Simulation results for different sizes of physicians and patients: 5×20, 10×10, 20×5.

GHQ Type	$10 \times 5 \times 20$	$10 \times 10 \times 10$	$10 \times 20 \times 5$
TRUE β	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8
Ave $\hat{\beta}$	-0.177, 0.385, 0.794	-0.168, 0.384, 0.786	-0.196, 0.374, 0.776
SD ($\hat{\beta}$)	1.136, 0.174, 0.047	1.075, 0.127, 0.049	1.108, 0.095, 0.052
Ave SE ($\hat{\beta}$)	0.943, 0.16, 0.047	0.954, 0.122, 0.048	0.926, 0.1, 0.05
RMSE($\hat{\beta}$)	1.136, 0.174, 0.048	1.076, 0.128, 0.051	1.108, 0.099, 0.057
Coverage *	0.886, 0.934, 0.948	0.9, 0.928, 0.944	0.874, 0.956, 0.926
LB Error *	0.05, 0.026, 0.014	0.048, 0.026, 0.006	0.056, 0.002, 0.014
UB Error *	0.064, 0.04, 0.038	0.052, 0.046, 0.05	0.07, 0.042, 0.06
TRUE θ	0.8, 0.2	0.8, 0.2	0.8, 0.2
Ave $\hat{\theta}$	0.606, 0.189	0.63, 0.189	0.595, 0.183
SD ($\hat{\theta}$)	0.34, 0.061	0.322, 0.055	0.305, 0.056
Ave SE ($\hat{\theta}$)	0.298, 0.06	0.3, 0.053	0.281, 0.055
RMSE($\hat{\theta}$)	0.391, 0.062	0.364, 0.056	0.367, 0.059
Coverage	0.848, 0.958	0.85, 0.96	0.812, 0.988
LB Error	0.004, 0.018	0.004, 0.014	0.002, 0.006
UB Error	0.148, 0.024	0.146, 0.026	0.186, 0.006

*See footnotes in Table 5.1 (page 52).

Table 5.9: Simulation results from the proposed approach, the PPL and PL approach using the R built-in function `coxph`, by fitting the Cox model with no frailty, the one-level frailty model (either the hospital or the physician level) and the two-level frailty model (only available for the proposed approach) respectively, for $\theta = (0.8, 0)$.

Sample Size *	$20\times10\times5$					
Model (Method)	Two-Level (proposed)	H-Level (proposed)	P-Level (proposed)	H-Level (PPL†) §	P-Level (PPL†)	No Frailty (PL‡)
TRUE β	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8
Ave $\hat{\beta}$	-0.241, 0.397, 0.801	-0.239, 0.393, 0.792	-0.351, 0.263, 0.711	-0.366, 0.391, 0.79	-0.329, 0.280, 0.71	-0.31, 0.188, 0.5
SD ($\hat{\beta}$)	0.741, 0.096, 0.059	0.733, 0.094, 0.057	0.684, 0.15, 0.068	0.788, 0.089, 0.058	0.692, 0.152, 0.068	0.584, 0.118, 0.085
Ave SE ($\hat{\beta}$)	0.691, 0.095, 0.058	0.684, 0.092, 0.056	0.26, 0.145, 0.063	1.519, 0.092, 0.057	0.262, 0.147, 0.058	0.154, 0.085, 0.049
RMSE($\hat{\beta}$)	0.742, 0.096, 0.059	0.734, 0.094, 0.057	0.701, 0.203, 0.112	0.805, 0.089, 0.058	0.704, 0.193, 0.113	0.594, 0.243, 0.312
Coverage *	0.928, 0.948, 0.938	0.93, 0.942, 0.94	0.522, 0.834, 0.664	0.929, 0.967, 0.939	0.518, 0.856, 0.618	0.374, 0.348, 0.01
LB Error *	0.034, 0.022, 0.024	0.034, 0.02, 0.016	0.168, 0.002, 0	0.016, 0.005, 0.017	0.18, 0.004, 0	0.24, 0, 0
UB Error *	0.038, 0.03, 0.038	0.036, 0.038, 0.044	0.31, 0.164, 0.336	0.055, 0.027, 0.044	0.302, 0.140, 0.382	0.386, 0.652, 0.99
TRUE θ	**0.8, 0**	**0.8, 0**	**0.8, 0**	**0.8, 0**	**0.8, 0**	**0.8, 0**
Ave $\hat{\theta}$	0.715, 0.018	0.703	0.55	1.014	0.595	NA
SD ($\hat{\theta}$)	0.27, 0.027	0.263	0.233	2.234	0.249	NA
Ave SE ($\hat{\theta}$)	0.251, 0.024	0.245	0.125	NA	NA	NA
RMSE($\hat{\theta}$)	0.283, 0.033	0.28	0.597	NA	NA	NA
Coverage	0.896, 0	0.882	0	NA	NA	NA
LB Error	0.002, 1	0.002	1	NA	NA	NA
UB Error	0.102, 0	0.116	0	NA	NA	NA

*See footnotes in Table 5.1 (page 52).

†Penalized partial likelihood approach using the R built-in function `coxph`.

‡Ordinary partial likelihood approach for Cox models using the R built-in function `coxph`.

§317 out of the 500 simulations that failed to return the estimates were excluded. Among the remaining 183 simulations, two with extreme estimates of the the hospital-level coefficient (-129.8 and -7.7) were also excluded.

Table 5.10: Simulation results from the proposed approach, the PPL and PL approach using the R built-in function `coxph`, by fitting the Cox model with no frailty, the one-level frailty model (either the hospital or the physician level) and the two-level frailty model (only available for the proposed approach) respectively, for $\theta = (0, 0.8)$.

Sample Size *	$20\times10\times5$					
Model (Method)	Two-Level (proposed)	H-Level (proposed)	P-Level (proposed)	H-Level (PPL†)	P-Level (PPL†)	No Frailty (PL‡)
TRUE β	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8
Ave $\hat{\beta}$	-0.342, 0.248, 0.707	-0.282, 0.17, 0.488	-0.341, 0.248, 0.707	-0.277, 0.169, 0.488	-0.317, 0.267, 0.707	-0.278, 0.164, 0.468
SD ($\hat{\beta}$)	0.287, 0.146, 0.069	0.239, 0.129, 0.058	0.287, 0.145, 0.069	0.234, 0.128, 0.058	0.29, 0.146, 0.069	0.232, 0.125, 0.057
Ave SE ($\hat{\beta}$)	0.279, 0.15, 0.063	0.211, 0.089, 0.049	0.269, 0.15, 0.063	0.222, 0.089, 0.049	0.274, 0.153, 0.058	0.151, 0.085, 0.048
RMSE($\hat{\beta}$)	0.32, 0.211, 0.116	0.252, 0.264, 0.318	0.32, 0.21, 0.116	0.247, 0.264, 0.317	0.313, 0.198, 0.116	0.245, 0.267, 0.337
Coverage *	0.922, 0.82, 0.65	0.87, 0.34, 0	0.91, 0.824, 0.652	0.911, 0.338, 0	0.92, 0.856, 0.602	0.766, 0.304, 0
LB Error *	0.006, 0, 0.002	0.032, 0, 0	0.006, 0, 0.002	0.022, 0, 0	0.006, 0, 0.002	0.052, 0, 0
UB Error *	0.072, 0.18, 0.348	0.098, 0.66, 1	0.084, 0.176, 0.346	0.067, 0.662, 1	0.074, 0.144, 0.396	0.182, 0.696, 1
TRUE θ	**0, 0.8**	**0, 0.8**	**0, 0.8**	**0, 0.8**	**0, 0.8**	**0, 0.8**
Ave $\hat{\theta}$	0.01, 0.593	0.034	0.602	0.464	0.65	NA
SD ($\hat{\theta}$)	0.021, 0.127	0.025	0.127	7.41	0.136	NA
Ave SE ($\hat{\theta}$)	0.015, 0.132	0.022	0.132	NA	NA	NA
RMSE($\hat{\theta}$)	0.023, 0.243	0.042	0.235	NA	NA	NA
Coverage	0, 0.774	0	0.78	NA	NA	NA
LB Error	1, 0.002	1	0.004	NA	NA	NA
UB Error	0, 0.224	0	0.216	NA	NA	NA

*See footnotes in Table 5.1 (page 52).

†Penalized partial likelihood approach using the R built-in function `coxph`.

‡Ordinary partial likelihood approach for Cox models using the R built-in function `coxph`.

Table 5.11: Simulation results from the proposed approach, the PPL and PL approach using the R built-in function `coxph`, by fitting the Cox model with no frailty, the one-level frailty model (either the hospital or the physician level) and the two-level frailty model (only available for the proposed approach) respectively, for $\theta = (0.4, 0.4)$.

Sample Size *	$20 \times 10 \times 5$					
Model (Method)	Two-Level (proposed)	H-Level (proposed)	P-Level (proposed)	H-Level (PPL†) §	P-Level (PPL†)	No Frailty (PL‡)
TRUE β	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8
Ave $\hat{\beta}$	-0.317, 0.29, 0.72	-0.281, 0.237, 0.587	-0.346, 0.248, 0.700	-0.280, 0.238, 0.588	-0.324, 0.266, 0.699	-0.284, 0.168, 0.474
SD ($\hat{\beta}$)	0.553, 0.128, 0.064	0.47, 0.116, 0.056	0.548, 0.138, 0.065	0.458, 0.116, 0.056	0.556, 0.139, 0.065	0.423, 0.116, 0.062
Ave SE ($\hat{\beta}$)	0.485, 0.132, 0.062	0.415, 0.091, 0.052	0.266, 0.148, 0.063	0.577, 0.092, 0.051	0.27, 0.15, 0.058	0.154, 0.085, 0.048
RMSE($\hat{\beta}$)	0.565, 0.169, 0.103	0.477, 0.2, 0.22	0.567, 0.206, 0.119	0.465, 0.200, 0.219	0.569, 0.193, 0.12	0.431, 0.259, 0.332
Coverage *	0.902, 0.87, 0.718	0.92, 0.558, 0.024	0.652, 0.836, 0.644	0.956, 0.556, 0.024	0.65, 0.87, 0.6	0.522, 0.272, 0
LB Error *	0.022, 0.004, 0	0.024, 0.004, 0	0.104, 0.004, 0	0.008, 0.002, 0	0.12, 0.004, 0	0.176, 0.002, 0
UB Error *	0.076, 0.126, 0.282	0.056, 0.438, 0.976	0.244, 0.16, 0.356	0.035, 0.442, 0.976	0.23, 0.126, 0.4	0.302, 0.726, 1
TRUE θ	**0.4, 0.4**	**0.4, 0.4**	**0.4, 0.4**	**0.4, 0.4**	**0.4, 0.4**	**0.4, 0.4**
Ave $\hat{\theta}$	0.286, 0.329	0.229	0.574	1.043	0.62	NA
SD ($\hat{\theta}$)	0.124, 0.096	0.091	0.162	16.135	0.174	NA
Ave SE ($\hat{\theta}$)	0.12, 0.093	0.088	0.128	NA	NA	NA
RMSE($\hat{\theta}$)	0.169, 0.12	0.193	0.238	NA	NA	NA
Coverage	0.874, 0.932	0.65	0.61	NA	NA	NA
LB Error	0, 0.01	0	0.388	NA	NA	NA
UB Error	0.126, 0.058	0.35	0.002	NA	NA	NA

*See footnotes in Table 5.1 (page 52).

†Penalized partial likelihood approach using the R built-in function `coxph`.

‡Ordinary partial likelihood approach for Cox models using the R built-in function `coxph`.

§246 out of the 500 simulations returned warning messages that some inner iterations failed to converge.

Table 5.12: Simulation results from the proposed approach and the hierarchical likelihood approach using the R package `frailtyHL`, for sample $20 \times 10 \times 5$.

Sample Size *	$20 \times 10 \times 5$			
Method	Proposed	$HL(1,1)$ †	Proposed	$HL(1,1)$ †
TRUE β	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8	-0.2, 0.4, 0.8
Ave $\hat{\beta}$	-0.262, 0.328, 0.754	-0.272, 0.328, 0.748	-0.251, 0.368, 0.775	-0.235, 0.365, 0.774
SD ($\hat{\beta}$)	0.433, 0.146, 0.053	0.459, 0.147, 0.053	0.781, 0.103, 0.051	0.73, 0.104, 0.051
Ave SE ($\hat{\beta}$)	0.405, 0.148, 0.053	0.423, 0.146, 0.049	0.686, 0.104, 0.051	0.721, 0.102, 0.048
RMSE($\hat{\beta}$)	0.438, 0.163, 0.07	0.464, 0.164, 0.074	0.783, 0.108, 0.057	0.731, 0.11, 0.058
Coverage *	0.918, 0.918, 0.856	0.909, 0.913, 0.78	0.912, 0.922, 0.902	0.944, 0.924, 0.886
LB Error *	0.026, 0.004, 0.002	0.026, 0.01, 0.002	0.03, 0.016, 0.008	0.026, 0.01, 0.01
UB Error *	0.056, 0.078, 0.142	0.065, 0.077, 0.218	0.058, 0.062, 0.09	0.03, 0.066, 0.104
TRUE θ	0.2, 0.8	0.2, 0.8	0.8, 0.2	0.8, 0.2
Ave $\hat{\theta}$	0.15, 0.691	0.179, 0.685	0.683, 0.188	0.756, 0.178
SD ($\hat{\theta}$)	0.087, 0.132	0.092, 0.128	0.248, 0.062	0.273, 0.055
Ave SE ($\hat{\theta}$)	0.081, 0.127	0.095, 0.118	0.238, 0.057	0.272, 0.052
RMSE($\hat{\theta}$)	0.1, 0.171	0.094, 0.172	0.274, 0.063	0.276, 0.06
Coverage	0.996, 0.868	0.875, 0.772	0.914, 0.964	0.886, 0.89
LB Error	0.004, 0.004	0, 0	0.006, 0.034	0.002, 0.002
UB Error	0, 0.128	0.125, 0.228	0.08, 0.002	0.112, 0.108

*See footnotes in Table 5.1 (page 52).

†Hierarchical likelihood approach using the first order of Laplace approximation.

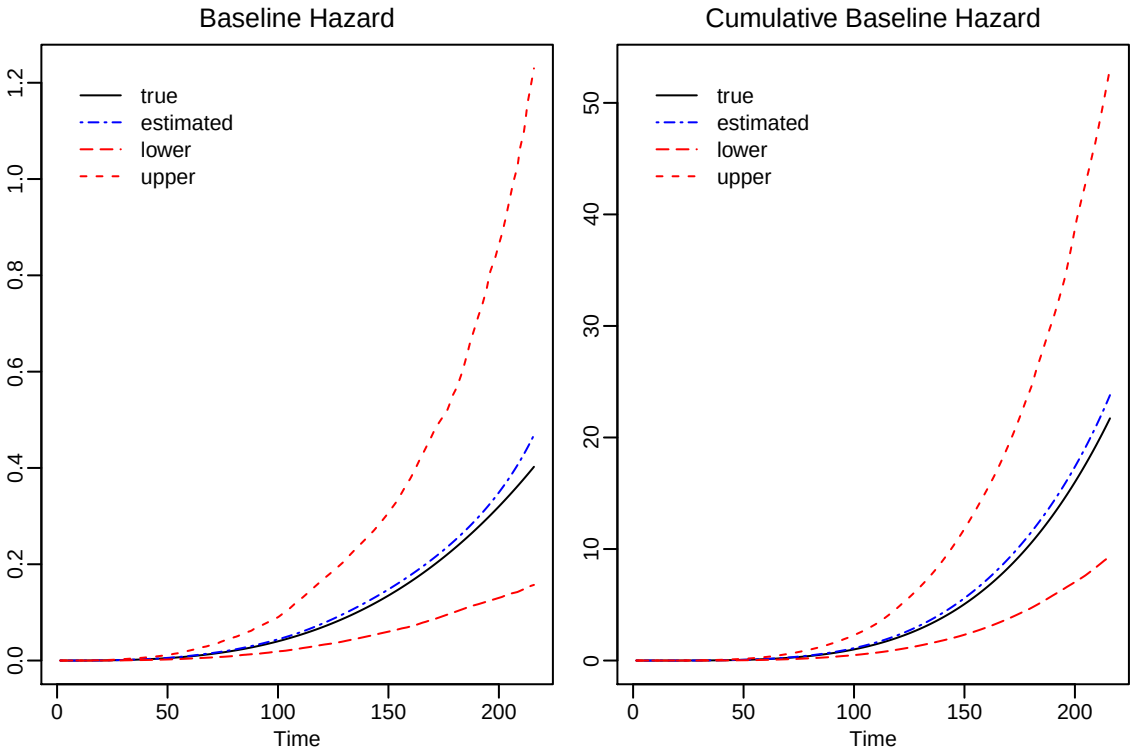

Figure 5.1: Plots for baseline hazard and cumulative hazard functions, with sample size $20 \times 10 \times 5$, $\boldsymbol{\theta} = (0.8, 0.2)$.

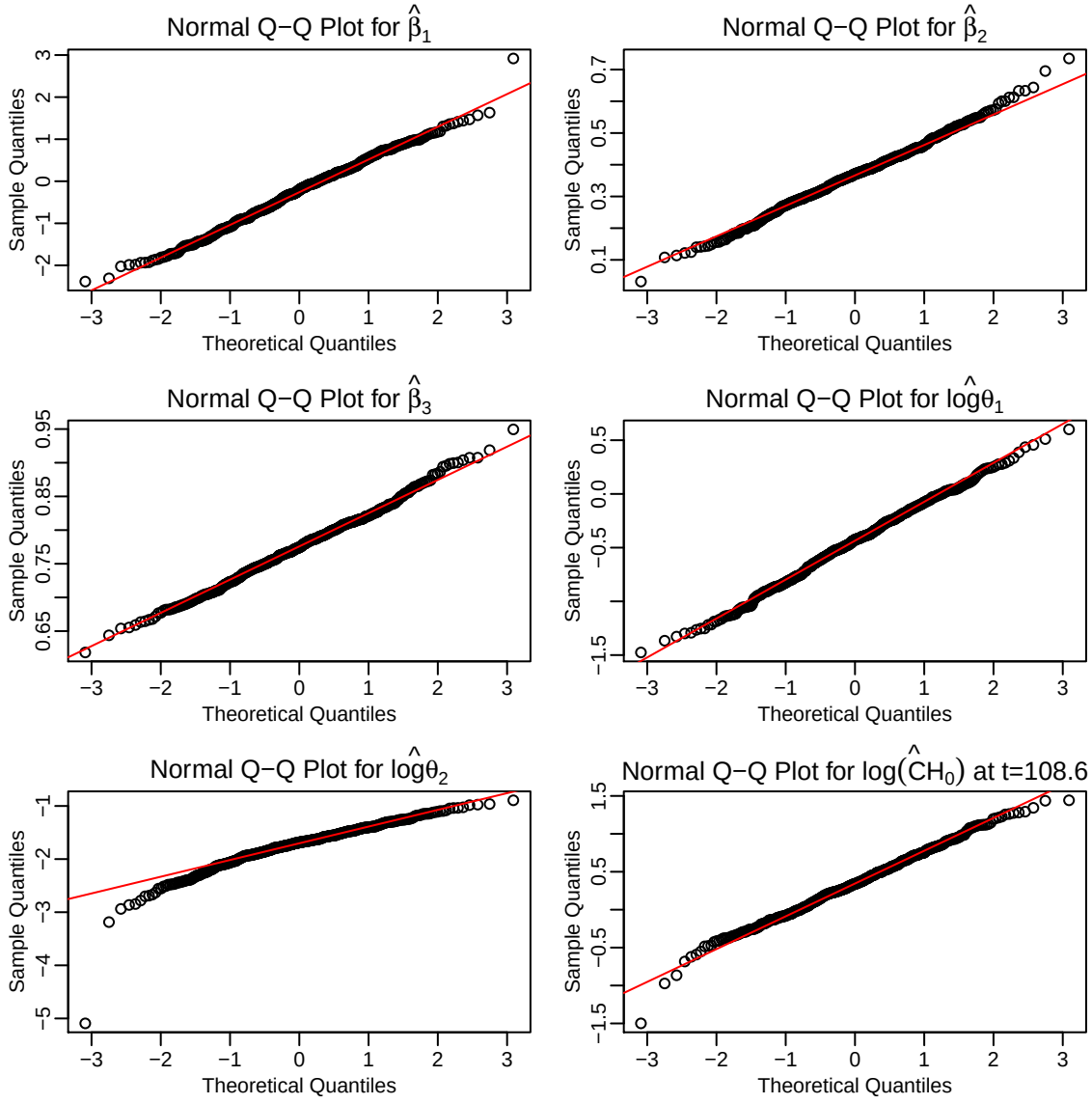

Figure 5.2: Normal Q-Q plots for estimates of the regression coefficients $\boldsymbol{\beta}$, the frailty variances $\boldsymbol{\theta}$ and the log baseline cumulative hazard at the midpoint of observed times. Sample size $= 20 \times 10 \times 5$, $\boldsymbol{\theta} = (0.8, 0.2)$.

CHAPTER 6
DISCUSSION AND FUTURE WORK

In summary, we have proposed an estimation approach to a two-level lognormal frailty model for clustered survival data.

The B-splines method not only allows estimating the baseline hazard function continuously, but also provides a convenient way to estimate the complicated semiparametric model in a parametric way (note that the full likelihood involves multivariate integrals). In our approach, the number of knots for B-splines is determined by the cubic root of the sample size (strictly speaking, the number of distinct observed times) minus the order of B-splines, and the knots were selected according on the percentiles of the observed failure times. However, theoretical justification will need to be further investigated. Moreover, we use cubic splines to model the baseline cumulative hazard function and thus the baseline hazard function is actually modeled by quadratic splines. Quartic splines may be worth a try in future simulations for better estimation of the baseline hazard function.

The adaptive GHQ method provides an easy and efficient way to compute the integrals with respect to the normally distributed random effect variables at both levels (hospital and physician). Attention to the choice of programming language and the implementation of the algorithms are important to reduce the time needed for intensive computation, even though the order of GHQ is chosen at 10. Currently, the AGHQ method has been implemented in R, which is not very efficient in doing loops. For a dataset with sample size $20 \times 10 \times 5$, the analysis of one simulated dataset takes

roughly 5 minutes. Computer time will be significantly reduced if the AGHQ method is recoded with a more efficient language, such as C or Fortran.

We have examined the small sample properties of the parameter estimators using simulation. However, theoretical proof still remains a challenge. Unlike the partial likelihood approach to the Cox model, martingale theory (Andersen and Gill, 1982) can not be applied to our method due to the existence of frailty terms. Furthermore, the adoption of the B-splines method brings more complexity. A detailed study on the asymptotic results presented in Gray (1992) may give us some inspiration relevant for B-splines.

In our simulation studies, we have noticed that the proposed approach led in some cases to significant underestimation of the hospital-level regression coefficient β_1 and frailty variance θ_1 while Ha and Lee's hierarchical likelihood (HL) approach showed much less bias, though still underestimated θ_1. This finding is consistent with simulation results presented in Ha and Lee (2003) that the maximum likelihood estimators (MLEs) had larger bias compared to the HL estimators for one-level lognormal frailty models with nonparametric baseline hazards. Recall that our proposed estimators can be treated as approximate MLEs (due to the use of B-splines). This phenomenon also exists for MLEs and REML estimators in linear mixed models. Ha and Lee (2003) stated that the HL approach actually led to the restricted/residual maximum likelihood (REML) estimator of frailty variance given in McGilchrist (1993). This inspires us in the future to pursue REML estimators rather than MLEs for the frailty variances $\boldsymbol{\theta}$. Then the Gauss-Seidel method can be used to compute the regres-

sion coefficients $\boldsymbol{\beta}$ based on the full likelihood defined in Section 3.2 and the REML estimator of $\boldsymbol{\theta}$ separately.

A much larger assumption of independence between frailties at both levels. The lognormal frailty distribution allows the flexibility in modeling multivariate correlation structure between frailties, e.g. in our case the hospital-level frailty could be correlated with the physician-level frailties within that hospital, or in recurrent event data there exist correlated random patient-specific intercept and slope over time. It is interesting and worthwhile to investigate the identifiability of the dependence parameter and further extend our work to correlated lognormal frailty models. Actually in some case the correlated frailty models are more appropriate than the shared frailty models. To illustrate this, we review some limitations of the shared frailty models, as discussed by Xue and Brookmeyer (1996); Wienke (2009).

First, the subjects within the cluster are forced to have the same frailty. This may not be appropriate, for example, in clusters consisted of family members. Second, the frailty term in the model (1.1) must be positive, which implies one characteristic of frailty's influence: subjects within the same cluster are positively correlated in terms of the risks of "death". However in some situation there exists negative association within the cluster. For example, animals living in one cage with limited food are negatively correlated in terms of their growth or survival rate. Third, the dependence between survival times within the cluster is based on marginal distributions of survival times. For example, when covariates are present in a bivariate proportional hazards model with gamma frailty, the dependence parameter and the regression parameters

are confounded (Clayton and Cuzick, 1985), implying that the joint distribution can be identified from the marginal distributions and that the dependence parameter measures something besides dependence (Hougaard, 1986b). This problem exists for any univariate frailty distribution with a finite mean (Elbers and Ridder, 1982). Positive stable distribution has an infinite mean and thus is recommended for the shared frailty model to overcome this problem (Hougaard, 1986a,b). However, we should be aware that it is rather the data that should guide the choice of the frailty distribution and not an attractive mathematical model property (Duchateau and Janssen, 2008, Section 4.4.1).

However, for clustered survival data where the cluster size is small and the cluster structure is fixed (e.g. each family is a cluster), these limitations can be avoided by developing correlated frailty models: different random variable (frailty) is assigned to each subject within the cluster and the correlation structure is specified between those random variables.

PART II

COMPETING RISKS MODEL WITH MISSING CAUSE OF FAILURE

CHAPTER 7
INTRODUCTION

7.1 Competing Risks Models

In survival analysis, the competing risks model (CRM) is widely used when there exists more than one cause of failure. Conceptually, the causes are defined as risks before the failure occurs. The competing risks are often thought of as if the failure is the result from risks competing with each other. The competing risks are assumed exclusive to each other, in other words, the failure is linked to one cause only.

Failure to take into account the different causes of failure may result in misrepresenting the effects of the covariates on survival. One case is that the analysis should be made based upon one particular failure type rather than the overall failures. For example, as introduced in Gichangi and Vach (2005), in a study of treatments of cancer, the aim is to keep a patient in a state where the patient should neither suffer from a relapse nor die. Consequently, in comparing two treatments the main outcome of interest is the time until death or relapse. Death and relapse can thus be treated as the two causes of failure. For the analyses in which the two causes are not distinguished, one may argue that the death without a prior relapse occurs because the disease has already progressed too far at the time of treatment and only events of relapse can be influenced by a treatment.

Competing risks analysis may also be important in another situation where

people are interested in describing the course of different causes over time with no focus on any specific one. To illustrate this, we look at another example given in Gichangi and Vach (2005). In pharmacoepidemiological studies on drug utilization, treatment of drug therapy is initiated by general practitioner (GP). During the course of the treatment episode, treatment may be discontinued or switched or a patient may die or receive an add-on therapy to boost the initial therapy. If one is interested in analyzing the prescribing behavior of GPs in patients starting a therapy, the outcomes should be the event times of the four cause types.

The Cox proportional hazards model can be specified for each cause type. When all failures have known cause types, the ordinary partial likelihood approach can be applied to each cause-specific hazard model, respectively, to estimate the cause-specific regression coefficients by treating the failures of other causes as the censoring times for the failures due to this specific cause (Kalbfleisch and Prentice, 2002).

In competing risks analysis, one usually uses cumulative incidence functions (CIFs) rather than cause-specific survival functions to evaluate the risks, because cause-specific survival functions are not easily defined or conceptually justified. CIF is conceptually the marginal cumulative function given covariates for a particular cause, which can be determined by the cause-specific hazards without the assumption of independence between competing risks. With CIF, one can compute the probability to fail from a certain cause up to time t or the probability to fail from a certain cause within $[t, t+\delta t]$ given that one survives at time t. If the analysis involves

regression covariates, it should be alerted that the effect of a covariate on the hazard of a certain cause type may not correspond to a similar effect on CIF because CIF is a complicated function of the cause-specific hazards (see Section 8.2 for the definition of cause-specific sub-density functions). To directly assess the effect of covariates on the subdistribution of a particular failure type, Fine and Gray (1999) studied a proportional subdistribution hazards regression model using partial likelihood principle and weighting techniques.

7.2 CRM with Missing Cause of Failure

In this dissertation, we study a special case of competing risks model when the causes of failure could be missing for some reason. Information on cause type should be either completely available (one of the m possible causes) or not available at all (any of the m possible causes). In practice, there also exists the case where information of cause type is not completely missing, e.g. the failure type is not exactly known but limited to a subset of possible causes. We focus on the former case only. First we review existing methods in the literature.

Goetghebeur and Ryan (1990, 1995) proposed a partial likelihood approach for two types of failure based on the proportional hazards for each cause type under the assumption that the cause-specific baseline hazards are proportional as well. A similar partial likelihood approach under the same assumption was developed by Dewanji (1992). Lu and Tsiatis (2005) further showed that the estimator of the regression coefficients based on the Dewanji partial likelihood is not only consistent

and asymptotically normal, but also semiparametric efficient, while the Goetghebeur and Ryan's estimator is more robust in the case of misspecification of proportional baseline hazards.

Lu and Tsiatis (2001) proposed a multiple imputation (MI) approach. The main idea is to impute the missing causes using the complete data under the assumption of missing at random (MAR) and then apply the ordinary partial likelihood approach. Though the imputation requires a parametric model for cause types, e.g. logistic regression model, the MI approach is robust when the parametric model is not correctly specified.

Dewanji and Sengupta (2003) considered a nonparametric estimation of cause-specific hazards without covariates, which used an expectation maximization (EM) algorithm for nonparametric maximum likelihood estimation.

An EM algorithm was studied as well in our preliminary work when covariate information was considered in the cause-specific hazards. But it turned out that the estimates of the regression results were not stable due to the fact that the resulting profile likelihood with missing cause of failure has multiple local maxima (see Figure 7.1). This hinders the success of maximum likelihood method in this particular setting. Though theoretically there exists a consistent and efficient estimator among all possible local maxima in parametric model (Lehmann and Casella, 1998), it appeared not to be the case for the competing risks model under the semiparametric setting.

We propose two estimation approaches for the competing risks data with missing cause of failure: weighted complete-case (WCC) approach and double robust

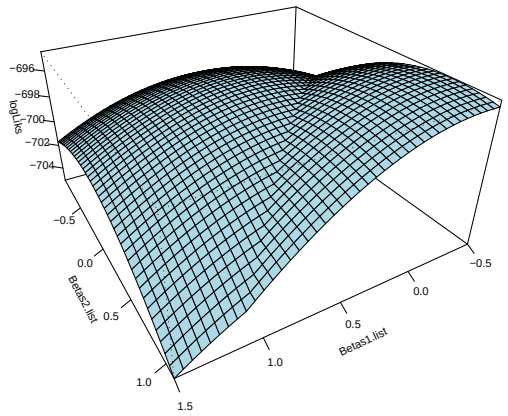

Figure 7.1: 3D plot for the log profile likelihood with respect to the regression parameters β_1 and β_2 for sample size = 200, true $\beta_1 = 0.5$, true $\beta_2 = 0.1$.

approach.

7.2.1 Weighted Complete-Case Approach

In practice, one simple approach for missing data is to not consider data with missing information and use the complete-case data only assuming missingness occurs at random. However, for the complete-case competing risks data, the partial likelihood estimators for the regression coefficients in the cause-specific hazards are biased, because the data include both failed and censored subjects while the missingness occurs only among subjects who fail due to whatever causes. Therefore, the complete-case data, in fact, result in a biased sample even if missingness occurs completely at random (MCAR) for the failures.

We propose a weighted complete-case partial likelihood (WCC, in short) approach based on the idea of bias correction in a biased sample proposed in Pan and Schaubel (2009). The weight is chosen as the inverse of the probability that one observation is selected to construct the likelihood. Note that the weight is always set at one for all censored times because the missingness only occurs for failures and censored data are always included. To obtain the weights for failures with known causes, parametric models are fitted to the event data to predict the selection probability. For simplicity, we assume that $\boldsymbol{W}$, a vector of time-independent covariates, includes all necessary information for missing mechanism. We will show that the resulting estimators of the regression coefficients in the cause-specific hazards are consistent and asymptotically normal.

7.2.2 Double Robust Approach

Though WCC estimators are consistent, they are not efficient because WCC approach does not make full use of missing data. We further propose a double robust (DR) approach to improve the efficiency. Our DR approach was motivated by a similar approach proposed in Lu and Liang (2008) for additive hazards models. The idea to construct DR estimating equations can be first found in Robins et al. (1994). In the competing risks setting with missing cause of failure, the DR approach requires first to estimate the probability of non-missingness (or selection probability) and the probability of failures with certain cause type by specifying parametric models for the non-missingness and cause types, respectively. When at least one of the two models

is correctly specified and the MAR assumption holds, DR estimators are consistent and generally more efficient than WCC estimators.

7.3 Outline

Chapter 8 defines the relevant notation and introduces the competing risks model starting from cause-specific hazard models. Chapter 9 proposes the WCC approach for Cox proportional hazards model and provides theoretical proofs for the asymptotic properties of WCC estimators. Chapter 10 discusses the DR estimation approach. Chapter 11 conducts simulation studies to validate the proposed approaches and compares them to other approaches by Goetghebeur and Ryan (1995) and Lu and Tsiatis (2001). The methods are applied to a real life example for further illustration. Finally, Chapter 12 summarizes the findings and discusses the potential future work.

CHAPTER 8
THE MODEL

8.1 Notation

For the ease of notation, we construct the model for two cause types, which can easily be extended to m types for $m \geq 3$.

The weighted partial likelihood contains the following unknown parameters that need to be estimated:

- $\boldsymbol{\beta}_m$: the vector of regression coefficients in the cause-specific hazards model
- $\boldsymbol{\theta}$: the vector of regression coefficients in the logistic regression model for the non-missingness of cause of failure
- $\boldsymbol{\gamma}$: the vector of regression coefficients in the logistic regression model for the cause type of failure.

Denote $\boldsymbol{\beta}_{m0}$, $\boldsymbol{\theta}_0$ and $\boldsymbol{\gamma}_0$ as the true values of $\boldsymbol{\beta}_m$, $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$.

The following notation is defined for a generic subject in a sample, which can be applied to the i^{th} subject of the sample with subscript i, $i = 1, \ldots, n$, where n is the sample size,

- T: the failure time
- C: the censoring time
- $X = T \wedge C = \min(T, C)$: the observed time

- $\boldsymbol{Z}$: the vector of regression variables for cause-specific hazard functions
- $\boldsymbol{W}$: the vector of regression variables for selection probability with the first element the constant being 1
- $\boldsymbol{V}$: the vector of regression variables for cause type with the first element the constant being 1
- δ: the cause type of a failure, 1 for type I, 2 for type II, and 0 for censoring
- π: the missing indicator, 1 for failure with missing cause type, 0 otherwise
- $\lambda_m(\Lambda_m)$: the cause-specific (cumulative) baseline hazard function
- $\lambda_{m0}(\Lambda_{m0})$: the true cause-specific (cumulative) baseline hazard function
- $f\,(f_m)$: the (cause-specific sub) density function of T
- S: the overall survival function of T
- $N_m(t) \equiv 1_{[\delta>0]} \times 1_{[X\leq t,\delta=m,\pi=0|\delta>0]}$

  $N_u(t) \equiv 1_{[\delta>0]} \times 1_{[X\leq t,,\pi=1|\delta>0]}$

  $N(t) \equiv N_1(t) + N_2(t) + N_u(t)$
- $Y(t) \equiv 1_{[X\geq t]}$
- $M_m\,(t) \equiv N_m(t) - \int_0^t Y(u) \exp\left(\boldsymbol{\beta}_{m0}^T \boldsymbol{Z}\right) d\Lambda_{m0}\,(u)$
- $R \equiv I_{[\text{the observed time is censoring time or is a failure time with known cause}]} = I_{[\delta=0]} + I_{[\delta>0]} \times I_{[\pi=0|\delta>0]}$: the indicator function of selection

- $p(\boldsymbol{\theta}; \boldsymbol{W}) \equiv \Pr(\pi = 0 \mid \boldsymbol{W}, \delta > 0)$: the conditional selection probability of observed failures with known cause given covariates $\boldsymbol{W}$

- $p(\boldsymbol{\theta})$: the selection probability of observed failures with known cause

- $p(\boldsymbol{\theta}; \boldsymbol{W}, \delta) \equiv \Pr(R = 1 \mid \boldsymbol{W}, \delta) = I_{[\delta=0]} + I_{[\delta>0]} \times p(\boldsymbol{\theta}; \boldsymbol{W})$: the selection (or non-missingness) probability for either the censored observations or the observations with known cause of failure

- $p(\boldsymbol{\theta}; \delta) \equiv \Pr(R = 1 \mid \delta) = I_{[\delta=0]} + I_{[\delta>0]} \times p(\boldsymbol{\theta})$: the selection (or non-missingness) probability for either the censored observations or the observations with known cause of failure

- $\tau_m(\boldsymbol{\gamma}; \boldsymbol{V}) = \Pr(\delta = m \mid \delta > 0, \pi = 0, \boldsymbol{V})$: the conditional probability that the cause of a failure is the m^{th} type given covariates $\boldsymbol{V}$

- $w(\boldsymbol{\theta}) \equiv \dfrac{R}{p(\boldsymbol{\theta}; \boldsymbol{W}, \delta)}$ or $\dfrac{R}{p(\boldsymbol{\theta}; \delta)}$: the weight, the inverse of selection probability.

The following notation is used for the algebraic convenience in method development:

- $S^{(d)}(\boldsymbol{\beta}, t) \equiv \dfrac{1}{n} \sum_{i=1}^{n} Y_i(t) \exp(\boldsymbol{\beta}^T \boldsymbol{Z}_i) \boldsymbol{Z}_i^{\otimes d}$, for $d = 0, 1, 2$

- $E(\boldsymbol{\beta}, t) \equiv \dfrac{\partial \log S^{(0)}(\boldsymbol{\beta}, t)}{\partial \boldsymbol{\beta}} = \dfrac{S^{(1)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)}$

- $V(\boldsymbol{\beta}, t) \equiv \dfrac{\partial^2 \log S^{(0)}(\boldsymbol{\beta}, t)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = \dfrac{S^{(2)}(\boldsymbol{\beta}, t)}{S^{(0)}(\boldsymbol{\beta}, t)} - E(\boldsymbol{\beta}, t)^{\otimes 2}$
$= \dfrac{1}{S^{(0)}(\boldsymbol{\beta}, t)} \times \dfrac{1}{n} \sum_{i=1}^{n} \{\boldsymbol{Z}_i - E(\boldsymbol{\beta}, t)\}^{\otimes 2} Y_i(t) \exp(\boldsymbol{\beta}^T \boldsymbol{Z}_i)$

- $\bar{\boldsymbol{S}}^{(d)}(\boldsymbol{\beta},\boldsymbol{\theta},t) \equiv \frac{1}{n}\sum_{i=1}^{n} w_i(\boldsymbol{\theta})\, Y_i(t)\exp(\boldsymbol{\beta}^T\boldsymbol{Z}_i)\boldsymbol{Z}_i^{\otimes d}$, for $d = 0, 1, 2$

- $\bar{\boldsymbol{E}}(\boldsymbol{\beta},\boldsymbol{\theta},t) \equiv \dfrac{\partial \log \bar{S}^{(0)}(\boldsymbol{\beta},\boldsymbol{\theta},t)}{\partial \boldsymbol{\beta}} = \dfrac{\bar{\boldsymbol{S}}^{(1)}(\boldsymbol{\beta},\boldsymbol{\theta},t)}{\bar{S}^{(0)}(\boldsymbol{\beta},\boldsymbol{\theta},t)}$

- $\bar{\boldsymbol{V}}(\boldsymbol{\beta},\boldsymbol{\theta},t) \triangleq \dfrac{\partial^2 \log \bar{S}^{(0)}(\boldsymbol{\beta},\boldsymbol{\theta},t)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = \dfrac{\bar{\boldsymbol{S}}^{(2)}(\boldsymbol{\beta},\boldsymbol{\theta},t)}{\bar{S}^{(0)}(\boldsymbol{\beta},\boldsymbol{\theta},t)} - \bar{\boldsymbol{E}}(\boldsymbol{\beta},\boldsymbol{\theta},t)^{\otimes 2}$

  $$= \frac{1}{\bar{S}^{(0)}(\boldsymbol{\beta},\boldsymbol{\theta},t)} \times \frac{1}{n}\sum_{i=1}^{n} w_i(\boldsymbol{\theta}) \left\{\boldsymbol{Z}_i - \bar{\boldsymbol{E}}(\boldsymbol{\beta},\boldsymbol{\theta},t)\right\}^{\otimes 2} Y_i(t)\, \exp(\boldsymbol{\beta}^T\boldsymbol{Z}_i)$$

- $\tilde{\boldsymbol{V}}(\boldsymbol{\beta},\boldsymbol{\theta},t) \equiv \dfrac{1}{\bar{S}^{(0)}(\boldsymbol{\beta},\boldsymbol{\theta},t)} \times \dfrac{1}{n}\sum_{i=1}^{n} w_i^2(\boldsymbol{\theta}) \left\{\boldsymbol{Z}_i - \bar{\boldsymbol{E}}(\boldsymbol{\beta},\boldsymbol{\theta},t)\right\}^{\otimes 2} Y_i(t)\exp(\boldsymbol{\beta}^T\boldsymbol{Z}_i)\,.$

8.2 Competing Risks Model

The proportional hazards competing risks model is defined by the cause-specific hazard functions following Cox proportional hazards model

$$\lim_{\Delta t\to 0}\frac{1}{\Delta t} P\left(T \in (t, t+\Delta t], \tau = m \mid T \geq t, \boldsymbol{Z}\right) \equiv \lambda_{m0}(t)\exp(\boldsymbol{\beta}_m^T\boldsymbol{Z}).$$

That is

$$\lambda_m(t \mid \boldsymbol{Z}) = \lambda_{m0}(t)\exp(\boldsymbol{\beta}_m^T\boldsymbol{Z}), \tag{8.1}$$

for $m = 1, 2$, where $\lambda_{m0}(t)$ is baseline hazard function taking any form of time t. Without loss of generality, we assume that the cause-specific hazards share the same covariates $\boldsymbol{Z}$.

The overall hazard function can be derived as

$$\begin{aligned}
&\lim_{\Delta t\to 0}\frac{1}{\Delta t}P\left(T\in(t,t+\Delta t]\mid T\geq t,\boldsymbol{Z}\right)\\
&=\lim_{\Delta t\to 0}\frac{1}{\Delta t}P\left(T\in(t,t+\Delta t],\,\tau=1\mid T\geq t,\,\boldsymbol{Z}\right)\\
&\quad+\lim_{\Delta t\to 0}\frac{1}{\Delta t}P\left(T\in(t,t+\Delta t],\,\tau=2\mid T\geq t,\boldsymbol{Z}\right)\\
&=\lambda_{10}(t)\exp(\boldsymbol{\beta}_1^T\boldsymbol{Z})+\lambda_{20}(t)\exp(\boldsymbol{\beta}_2^T\boldsymbol{Z}),
\end{aligned}$$

i.e.

$$\lambda(t\mid\boldsymbol{Z})=\lambda_{10}(t)\exp(\boldsymbol{\beta}_1^T\boldsymbol{Z})+\lambda_{20}(t)\exp(\boldsymbol{\beta}_2^T\boldsymbol{Z}) \tag{8.2}$$

Thereby the overall survival function of T is

$$S(t\mid\boldsymbol{Z})=\exp\left\{-\Lambda_{10}(t)\exp(\boldsymbol{\beta}_1^T\boldsymbol{Z})-\Lambda_{20}(t)\exp(\boldsymbol{\beta}_2^T\boldsymbol{Z})\right\} \tag{8.3}$$

where $\Lambda_{m0}(t)=\int_0^t\lambda_{m0}(u)du$ $(m=1,2)$ is the cause-specific baseline cumulative hazard function. The overall survival function in competing risks setting can be mathematically viewed as the survival function of the event $T=T_1\wedge T_2$ with T_1 and T_2 being independent event times having the hazard functions $\lambda_{10}\exp\left(\boldsymbol{\beta}_1^T\boldsymbol{Z}\right)$ and $\lambda_{20}\exp\left(\boldsymbol{\beta}_2^T\boldsymbol{Z}\right)$, respectively. We use this idea to generate competing risks data for simulation studies in Chapter 11.

Consequently, the overall density function of T is then obtained by taking the negative derivative of (8.3)

$$f(t\mid\boldsymbol{Z})=\left\{\lambda_{10}(t)\exp(\boldsymbol{\beta}_1^T\boldsymbol{Z})+\lambda_{20}(t)\exp(\boldsymbol{\beta}_2^T\boldsymbol{Z})\right\}S(t\mid\boldsymbol{Z}) \tag{8.4}$$

The cause-specific sub-density can be directly derived from the definition of

cause-specific hazards

$$f_m(t \mid \boldsymbol{Z}) \equiv f(t, \tau = m \mid \boldsymbol{Z}) = \lambda_{m0}(t) \exp(\boldsymbol{\beta}_m^T \boldsymbol{Z})\, S(t \mid \boldsymbol{Z}) \tag{8.5}$$

for $m = 1, 2$.

When there is no missing cause of failure, the partial likelihood for the competing risks model in counting process notation (Kalbfleisch and Prentice, 2002) has been established by

$$L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2) = \prod_{t \geq 0} \prod_{m=1}^{2} \prod_{i=1}^{n} \left\{ \frac{\exp(\boldsymbol{\beta}_m^T \boldsymbol{Z}_i)}{\sum_{j=1}^{n} Y_j(t) \exp(\boldsymbol{\beta}_m^T \boldsymbol{Z}_j)} \right\}^{dN_{mi}(t)}.$$

However, if there exist missing causes of failure, the partial likelihood approach does not work anymore. The complete likelihood for the observed data can be derived as

$$\begin{aligned}
&L(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2; \lambda_{10}, \lambda_{20}) \\
&= \prod_{i=1}^{n} \prod_{t \geq 0} \left\{ f_1(t \mid \boldsymbol{Z}_i)^{dN_{1i}(t)}\, f_2(t \mid \boldsymbol{Z}_i)^{dN_{2i}(t)}\, f(t \mid \boldsymbol{Z}_i)^{dN_{ui}(t)}\, S(t \mid \boldsymbol{Z}_i)^{-dN_i(t) - dY_i(t)} \right\} \\
&= \prod_{i=1}^{n} \prod_{t \geq 0} \left\{ \lambda_1(t \mid \boldsymbol{Z}_i)^{dN_{1i}(t)}\, \lambda_2(t \mid \boldsymbol{Z}_i)^{dN_{2i}(t)}\, \lambda(t \mid \boldsymbol{Z}_i)^{dN_{ui}(t)}\, S(t \mid \boldsymbol{Z}_i)^{-dY_i(t)} \right\}.
\end{aligned}$$

Then the log complete likelihood can be written as

$$\begin{aligned}
l\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2; \lambda_{10}, \lambda_{20}\right) = \sum_{i=1}^{n} \int_0^{\infty} \Big[&\left\{ \log \lambda_{10}(t) + \boldsymbol{\beta}_1^T \boldsymbol{Z}_i \right\} dN_{1i}(t) + \left\{ \log \lambda_{20}(t) + \boldsymbol{\beta}_2^T \boldsymbol{Z}_i \right\} \\
&\times dN_{2i}(t) - Y_i(t) \exp(\boldsymbol{\beta}_1^T \boldsymbol{Z}_i)\, d\Lambda_{10}(t) - Y_i(t) \exp(\boldsymbol{\beta}_2^T \boldsymbol{Z}_i)\, d\Lambda_{20}(t) \\
&+ \log\{\lambda_{10}(t) \exp(\boldsymbol{\beta}_1^T \boldsymbol{Z}_i) + \lambda_{20}(t) \exp(\boldsymbol{\beta}_2^T \boldsymbol{Z}_i)\}\, dN_{ui}(t) \Big].
\end{aligned} \tag{8.6}$$

The last term in the log complete likelihood (8.6) complicates the estimation as compared to the case with no missing cause type. An attempt with EM algorithm

and profile likelihood approach was experimented but did not yield a satisfactory solution due to the reason stated in Section 7.2.

8.3 Counting Processes for Competing Risks Data

$N_1(t), N_2(t)$ and $Y(t)$, as defined in Section 8.1, are the counting processes for the failure times with known cause type I, II and at risk process, respectively. The filtration $\mathcal{F}_t$ for the competing risks data with no missing causes can be defined as the σ-field generated by these processes by time t, i.e.

$$\mathcal{F}_t = \sigma\left\{N_1(s), N_2(s), Y(s) : 0 \leq s \leq t\right\}.$$

Denote

$$M_m(t) = N_m(t) - \int_0^t Y(u)\exp(\boldsymbol{\beta}_{m0}^T \boldsymbol{Z}) d\Lambda_{m0}(u).$$

$M_m(t)$ is a martingale with respect to $\mathcal{F}_t$ (Fleming and Harrington, 1991, Section 1.3).

CHAPTER 9
WEIGHTED COMPLETE-CASE APPROACH

9.1 Introduction

As pointed out in Chapter 7, the complete-case data is a biased sample. Since the records of observed failures with missing causes are excluded, having weights greater than one applied to the selected records of failures (with known cause) may reduce the bias. If we can predict the probability of failures with known cause, it is reasonable to think that one selected failure with a smaller probability of being included in the analytic sample should have a larger weight because it represents more failures that have missing causes and are thus excluded.

For the competing risks model with no missing cause, the log partial likelihood can be written as

$$l\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2\right) = \frac{1}{n} \sum_{m=1}^{2} \sum_{i=1}^{n} \int_0^{\infty} \left\{ \boldsymbol{\beta}_m^T \boldsymbol{Z}_i - \log S^{(0)}(\boldsymbol{\beta}_m, t) \right\} dN_{mi}\left(t\right).$$

This motivates to define the following stochastic process that mimics the log partial likelihood aforementioned but also takes care of weighting the observations selected for analysis.

$$l^w\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\theta}, t\right) = \frac{1}{n} \sum_{m=1}^{2} \sum_{i=1}^{n} \int_0^{t} w_i\left(\boldsymbol{\theta}\right) \left\{ \boldsymbol{\beta}_m^T \boldsymbol{Z}_i - \log \bar{S}^{(0)}(\boldsymbol{\beta}_m, \boldsymbol{\theta}, u) \right\} dN_{mi}\left(u\right), \quad (9.1)$$

where $w_i\left(\boldsymbol{\theta}\right)$ is the weight for the i^{th} subject. Then the log weighted partial likelihood in counting process form is $l^w\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\theta}, \infty\right)$, i.e. the limit of $l^w\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\theta}, t\right)$ as $t \to \infty$. For a given data from a study, let τ be the termination time, i.e.

$\tau = \inf\{t : dN(t) \equiv 0\}$, then $l^w(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\theta}, \tau) \equiv l^w(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\theta}, \infty)$, and is denoted for the corresponding log weighted partial likelihood.

Similarly, define $\boldsymbol{u}^w(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\theta}, t) = \left(\boldsymbol{u}_1^w(\boldsymbol{\beta}_1, \boldsymbol{\theta}, t)^T, \boldsymbol{u}_2^w(\boldsymbol{\beta}_2, \boldsymbol{\theta}, t)^T\right)^T$, where

$$\begin{aligned}\boldsymbol{u}_m^w(\boldsymbol{\beta}_m, \boldsymbol{\theta}, t) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^t w_i(\boldsymbol{\theta})\left\{\boldsymbol{Z}_i - \bar{\boldsymbol{E}}(\boldsymbol{\beta}_m, \boldsymbol{\theta}, u)\right\} dN_{mi}(u) \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^t w_i(\boldsymbol{\theta})\left\{\boldsymbol{Z}_i - \bar{\boldsymbol{E}}(\boldsymbol{\beta}_m, \boldsymbol{\theta}, u)\right\} dM_{mi}(u). \qquad (9.2)\end{aligned}$$

Then $\boldsymbol{u}^w(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\theta}, \tau)$ is the score functions of the log weighted partial likelihood for the given data.

To compute the weights, we need to first estimate the selection probability of the observed failures with known cause. Thus whether or not the cause type is missing rather than the cause type itself becomes our primary interest. Due to the binary outcome of the non-missingness of failure cause, it is natural to consider the logistic regression model. Let $\boldsymbol{\theta}$ be the regression coefficients to be estimated, and $p(\boldsymbol{\theta}; \boldsymbol{W})$ be the conditional selection probability given $\boldsymbol{W}$, a vector of covariates that are predictive of the missingness. Then the model can be written as

$$\log\frac{p(\theta; \boldsymbol{W})}{1 - p(\theta; \boldsymbol{W})} = \boldsymbol{\theta}^T\boldsymbol{W}. \qquad (9.3)$$

Let $\hat{\boldsymbol{\theta}}$ be the estimator of $\boldsymbol{\theta}$ in (9.3). By the ordinary maximum likelihood theorem, it follows that $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}$ and $p\left(\hat{\boldsymbol{\theta}}; \boldsymbol{W}\right) \xrightarrow{p} p(\boldsymbol{\theta}_0; \boldsymbol{W})$, where $\boldsymbol{\theta}_0$ is the true value of $\boldsymbol{\theta}$. If (9.3) is the correct model for non-missingness, in other words, which requires that $\boldsymbol{W}$ should include all covariates that affect the missingness of failure causes, then $p(\boldsymbol{\theta}_0; \boldsymbol{W})$ is the true conditional selection probability given $\boldsymbol{W}$.

There also exists a simplified way to predict the selection probability without using covariate information. Note that if $\boldsymbol{W}$ is set as a constant ONE and thus $\boldsymbol{\theta}$ is the intercept, then the logistic model yields a constant estimate $p\left(\hat{\boldsymbol{\theta}}\right)$ for all observed failures, which is the proportion of observed failures with known cause from given data. To make a difference, $\boldsymbol{W}$ is omitted in the notation. We call $p\left(\hat{\boldsymbol{\theta}}\right)$ as the empirical estimate of the selection probability. When (9.3) is the correct model, the following is true

$$\hat{p} \xrightarrow{p} E\left[p\left(\boldsymbol{\theta}_0; \boldsymbol{W}\right)\right] \equiv p_0,$$

where p_0 is the true overall selection probability without conditioning on the covariates. The resulting weighting method works well even when the missingness depends on additional covariate information. We will compare the two weighting methods by simulation studies in Chapter 11.

To generalize the above discussion, the weights can be computed as follows

$$\begin{cases} w\left(\boldsymbol{\theta}\right) = 1 & \text{for censored data,} \\ w\left(\boldsymbol{\theta}\right) = 0 & \text{for observed failures with missing cause,} \\ w\left(\boldsymbol{\theta}\right) = \dfrac{1}{p\left(\boldsymbol{\theta}; \boldsymbol{W}\right)} \text{ or } \dfrac{1}{p\left(\boldsymbol{\theta}\right)} & \text{for observed failures with known cause.} \end{cases}$$

By combining the three cases, as defined in Section 8.1, we can rewrite weights as

$$w(\boldsymbol{\theta}) = \frac{I_{[\delta=0]} + I_{[\delta>0]} \times I_{[\pi=0|\delta>0]}}{I_{[\delta=0]} + I_{[\delta>0]} \times p(\boldsymbol{\theta}; \boldsymbol{W})} \tag{9.4}$$

or

$$w(\boldsymbol{\theta}) = \frac{I_{[\delta=0]} + I_{[\delta>0]} \times I_{[\pi=0|\delta>0]}}{I_{[\delta=0]} + I_{[\delta>0]} \times p(\boldsymbol{\theta})}. \tag{9.5}$$

For the sake of generality, only the weighting method with the weight given by (9.4) will be considered in the proofs of asymptotic properties in next section. If the weight is computed by (9.5), we will show later that the assumption of independence between the missingness and the covariates in the cause-specific hazards, i.e. $\boldsymbol{W} \perp \boldsymbol{Z}$, is needed.

9.2 Asymptotic Properties

In this section, we describe and prove the consistency and asymptotic normality for the estimates of the regression coefficients from the weighted partial likelihood approach.

9.2.1 Regularity Conditions

Condition 1. *Let* $0 < \tau < \infty$ *be the termination time in a study, such that*

$$0 < \int_0^\tau d\Lambda_{m0}(t) < \infty.$$

Condition 2. *The covariate vectors* $\boldsymbol{Z}$ *and* $\boldsymbol{W}$ *is bounded, i.e. there exists* $C > 0$ *such that*

$$\Pr\left(\|\boldsymbol{Z}\| < C\right) = 1,$$

$$\Pr\left(\|\boldsymbol{W}\| < C\right) = 1.$$

Condition 3. *For* $\boldsymbol{S}_n^{(d)}(\boldsymbol{\beta}_m, t)$, $d = 0, 1, 2$, *defined in Section 8.1, there exists a function* $\boldsymbol{s}^{(d)}(\boldsymbol{\beta}_m, t)$ *respectively, and a compact and convex set* $\mathcal{B}_m \subset \mathbb{R}^p$ *(*$m = 1, 2$*) that contains* $\boldsymbol{\beta}_{m0}$ *such that*

$$\sup_{\boldsymbol{\beta}_m \in \mathcal{B}_m, t \in [0,\tau]} \|\boldsymbol{S}_n^{(d)}(\boldsymbol{\beta}_m, t) - \boldsymbol{s}^{(d)}(\boldsymbol{\beta}_m, t)\| \xrightarrow{p} 0$$

as $n \to \infty$.

Remark 1. *By law of large numbers,* $\boldsymbol{s}^{(d)}(\boldsymbol{\beta}_m, t)$ *(*$d = 1, 2$*), defined in Condition 3, can be written as* $E\left[Y(t)\exp(\boldsymbol{\beta}^T\boldsymbol{Z})\boldsymbol{Z}^{\otimes d}\right]$.

Then Condition 2 implies that $\boldsymbol{s}^{(d)}(\boldsymbol{\beta}_m, t)$ *(*$d = 1, 2$*) are bounded and* $\boldsymbol{s}^{(0)}(\boldsymbol{\beta_m}, t)$ *is bounded away from zero uniformly on* $\mathcal{B}_m \times [0, \tau]$.

Let

$$\begin{aligned}
\boldsymbol{e}(\boldsymbol{\beta}, t) &= \frac{\boldsymbol{s}^{(1)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)}, \\
\boldsymbol{v}(\boldsymbol{\beta}, t) &= \frac{\boldsymbol{s}^{(2)}(\boldsymbol{\beta}, t)}{s^{(0)}(\boldsymbol{\beta}, t)} - \boldsymbol{e}(\boldsymbol{\beta}, t)^{\otimes 2} \\
&= \frac{1}{s^{(0)}(\boldsymbol{\beta}, t)} E\left[\{\boldsymbol{Z} - \boldsymbol{e}(\boldsymbol{\beta}, t)\}^{\otimes 2} Y(t)\exp(\boldsymbol{\beta}^T\boldsymbol{Z})\right].
\end{aligned}$$

by continuous mapping theorem, it immediately follows that

$$\boldsymbol{E}(\boldsymbol{\beta}, t) \xrightarrow{p} \boldsymbol{e}(\boldsymbol{\beta}, t), \tag{9.6}$$

$$\boldsymbol{V}(\boldsymbol{\beta}, t) \xrightarrow{p} \boldsymbol{v}(\boldsymbol{\beta}, t). \tag{9.7}$$

Remark 2. *In the logistic regression model (9.3), the selection probability can be written as*

$$p(\boldsymbol{\theta}; \boldsymbol{W}) = \frac{\exp\left(\boldsymbol{\theta}^T\boldsymbol{W}\right)}{1 + \exp\left(\boldsymbol{\theta}^T\boldsymbol{W}\right)}.$$

Hence, Condition 2 implies that $p(\boldsymbol{\theta}; \boldsymbol{W})$ *is bounded away from zero uniformly on a compact and convex set* $\mathcal{S} \subset \mathbb{R}^p$ *that contains* $\boldsymbol{\theta}_0$.

Remark 3. *Since*

$$w(\boldsymbol{\theta}) = \frac{I}{1_{[\delta=0]} + 1_{[\delta>0]}\, p(\boldsymbol{\theta}; \boldsymbol{W})} \leq \frac{1}{p(\boldsymbol{\theta}; \boldsymbol{W})},$$

by Remark 2, $w(\boldsymbol{\theta})$ is bounded uniformly on $\mathcal{S}$, i.e. there exists a constant $C > 0$ such that

$$\Pr\left(\sup_{\boldsymbol{\theta}\in\mathcal{S}} w(\boldsymbol{\theta}) < C\right) = 1. \tag{9.8}$$

Condition 4. *Let*

$$\tilde{\boldsymbol{v}}(\boldsymbol{\beta},\boldsymbol{\theta},t) = \frac{1}{s^{(0)}(\boldsymbol{\beta},t)} E\left[w^2(\boldsymbol{\theta})\left\{\boldsymbol{Z} - \boldsymbol{e}(\boldsymbol{\beta},t)\right\}^{\otimes 2} Y(t)\exp(\boldsymbol{\beta}^T\boldsymbol{Z})\right].$$

The matrices

$$\boldsymbol{\Sigma}_m = \int_0^\tau s^{(0)}(\boldsymbol{\beta}_{m0},t)\ \boldsymbol{v}(\boldsymbol{\beta}_{m0},t)\ d\Lambda_{m0}(t),$$
$$\tilde{\boldsymbol{\Sigma}}_m = \int_0^\tau s^{(0)}(\boldsymbol{\beta}_{m0},t)\ \tilde{\boldsymbol{v}}(\boldsymbol{\beta}_{m0},\boldsymbol{\theta}_{\mathbf{0}},t)\ d\Lambda_{m0}(t)$$

are positive definite.

9.2.2 Asymptotic Properties

First we introduce some technical lemmas needed for the proofs of consistency and asymptotic normality. Lemma 9.1 generalizes the result of Andersen and Gill (1982) on finding extrema in sequences of random concave functions. Lemma 9.2 is the Martingale version of multivariate central limit theorem, which is useful in the proof of normality for martingale sequences. Lemma 9.3 shows the special case of Lenglart's Inequality (Lenglart, 1977; Fleming and Harrington, 1991). In Lemma 9.4, we further provide a convergence property that is heavily used in the proofs.

Lemma 9.1. *Let E be an open convex subset of $\mathbb{R}^p$, and let $F_1, F_2, \cdots$ be a sequence of random concave functions on E and f be real-valued function on E, such that, for*

all $x \in E$,

$$F_n \xrightarrow{p} f(x) \text{ as } n \to \infty.$$

Then

a. *The function* f *is concave.*

b. *For all compact subjects* A *of* E

$$\sup_{x \in A} |F_n(x) - f(x)| \xrightarrow{p} 0 \text{ as } n \to \infty.$$

c. *If* F_n *has a unique maximum at* X_n *and* f *has one at* x, *then* $X_n \xrightarrow{p} x$ *as* $n \to \infty$.

Lemma 9.2 (Rebolledo's Martingale Central Limit Theorem)**.** *Let* $\{N_1(t), \ldots, N_n(t)\}$ *be a multivariate counting process,* $A_i(t)$ *be the (continuous) compensator of* $N_i(t)$ $(i = 1, \ldots, n)$. *Let* $u_l(t) = \sum_{i=1}^{n} \int_0^t H_{i,l}(u)\, dM_i(u)$, *where* $H_{i,l}(t)$*'s* $(i = 1, \ldots, n, l = 1, \ldots, r)$ *are bounded* $\mathcal{F}_t$*-predictable processes. If*

i. $\langle u_{l_1}, u_{l_2} \rangle(t) = \sum_{i=1}^{n} \int_0^t H_{i,l_1}(u) H_{i,l_2}(u)\, dA_i(u) \xrightarrow{p} C_{l_1,l_2}(t)$ *as* $n \to \infty$,

ii. *(Lindeberg condition)* $\langle u_{l,\varepsilon}, u_{l,\varepsilon} \rangle(t) = \sum_{i=1}^{n} \int_0^t H_{i,l}^2(u)\, 1_{\left[|H_{i,l}(u)| > \varepsilon\right]} dA_i(u) \xrightarrow{p} 0$,

then $(u_1, \ldots, u_r) \xrightarrow{d} (w_1, \ldots, w_r)$ *as* $n \to \infty$, *where* $(w_1, \ldots, w_r)$ *is a* r*-variate continuous Gaussian process with* $w_l(0) = 0$, $E[w_l(t)] = 0$, *and* $E[w_{l_1}(s), w_{l_2}(t)] = C_{l_1,l_2}(s \wedge t)$*(Fleming and Harrington, 1991, Theorem 5.3.5).*

Lemma 9.3 (Lenglart's Inequality)**.** *Let* N_i *be a univariate counting process with a continuous compensator* A_i, *let* $M_i = N_i - A_i$, *and let* H_i *be a locally bounded, predictable process,* $i = 1, \ldots n$. *Let* $N = \sum_{i=1}^{n} N_i$, $A = \sum_{i=1}^{n} A_i$. *Then for all* $\delta, \rho > 0$ *and any* $t \geq 0$,

a. $\Pr\{N(t) \geq \rho\} \leq \dfrac{\delta}{\rho} + \Pr\{A(t) \geq \delta\}$,

b. $\Pr\left\{ \sup_{0\leq y\leq t} \left| \sum_{i=1}^{n} \int_0^y H_i(x) dM_i(x) \right| \geq \rho \right\} \leq \dfrac{\delta}{\rho^2} + \Pr\left\{ \sum_{i=1}^{n} \int_0^t H_i^2(x) dA_i(x) \geq \delta \right\}$.

Lemma 9.4. *Let* $\hat{\boldsymbol{\beta}}_{mn} \xrightarrow{p} \boldsymbol{\beta}_{m0}$, *and* $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$. *Suppose Condition (2) and (3) hold, then*

$$\bar{\boldsymbol{S}}^{(d)}(\hat{\boldsymbol{\beta}}_{mn}, \hat{\boldsymbol{\theta}}_n, t) \xrightarrow{p} \boldsymbol{s}^{(d)}(\boldsymbol{\beta}_{m0}, t)$$

for $0 \leq t \leq \tau$, $m = 1, 2$, $d = 0, 1, 2$.

Proof. For any i, $1 \leq i \leq n$,

$$\begin{aligned}
\left| w_i\big(\hat{\boldsymbol{\theta}}_n\big) - w_i\left(\boldsymbol{\theta}_0\right) \right| &\leq \left| \frac{1}{p_i\big(\hat{\boldsymbol{\theta}}_n\big)} - \frac{1}{p_i\left(\boldsymbol{\theta}_0\right)} \right| \\
&\leq \left| \exp\big(-\hat{\boldsymbol{\theta}}_n^T \boldsymbol{Z}_i \big) - \exp\big(-\boldsymbol{\theta}_0^T \boldsymbol{Z}_i \big) \right| \\
&\leq \left\| \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \right\| \times \left\| \exp\big(-\boldsymbol{\theta}_0^T \boldsymbol{Z}_i \big)\, \boldsymbol{Z}_i \right\| + o_p(\left\| \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \right\|) \\
&\leq K \left\| \hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \right\| + o_P(1),
\end{aligned}$$

where $K > 0$ is some constant due to Condition 2. This results in

$$\sup_{1\leq i\leq n} \left| w_i\big(\hat{\boldsymbol{\theta}}_n\big) - w_i\big(\boldsymbol{\theta}_0\big) \right| = o_P(1). \tag{9.9}$$

Hence, by conclusion (9.8), there exists a $C > 0$ such that

$$P\left(\sup_{1\leq i\leq n} \left| w_i\big(\hat{\boldsymbol{\theta}}_n\big) \right| < C \right) = 1.$$

Then it follows that

$$
\begin{aligned}
&\left\|\bar{\boldsymbol{S}}^{(d)}(\hat{\boldsymbol{\beta}}_{mn},\hat{\boldsymbol{\theta}}_n,t)-\bar{\boldsymbol{S}}^{(d)}\left(\boldsymbol{\beta}_{m0},\boldsymbol{\theta}_0,t\right)\right\| \\
&\leq \left\|\bar{\boldsymbol{S}}^{(d)}(\hat{\boldsymbol{\beta}}_{mn},\hat{\boldsymbol{\theta}}_n,t)-\bar{\boldsymbol{S}}^{(d)}\left(\boldsymbol{\beta}_{m0},\hat{\boldsymbol{\theta}}_n,t\right)\right\|+\left\|\bar{\boldsymbol{S}}^{(d)}\left(\boldsymbol{\beta}_{m0},\hat{\boldsymbol{\theta}}_n,t\right)-\bar{\boldsymbol{S}}^{(d)}\left(\boldsymbol{\beta}_{m0},\boldsymbol{\theta}_0,t\right)\right\| \\
&= \left\|\frac{1}{n}\sum_{i=1}^{n} w_i\left(\hat{\boldsymbol{\theta}}_n\right) Y_i\left(t\right)\left\{\exp\left(\hat{\boldsymbol{\beta}}_{mn}^T\boldsymbol{Z}_i\right)-\exp(\boldsymbol{\beta}_{m0}^T\boldsymbol{Z}_i)\right\}\boldsymbol{Z}_i^{\otimes d}\right\| \\
&\quad+\left\|\frac{1}{n}\sum_{i=1}^{n}\left\{w_i\big(\hat{\boldsymbol{\theta}}_n\big)-w_i\big(\boldsymbol{\theta}_0\big)\right\}Y_i(t)\exp(\boldsymbol{\beta}_{m0}^T\boldsymbol{Z}_i)\boldsymbol{Z}_i^{\otimes d}\right\| \\
&\leq \sup_{1\leq i\leq n} w_i\big(\hat{\boldsymbol{\theta}}_n\big)\times\left\|\frac{1}{n}\sum_{i=1}^{n} Y_i\left(t\right)\left\{\exp\left(\boldsymbol{\beta}_{m0}^T\boldsymbol{Z}_i\right)\boldsymbol{Z}_i^T\left(\hat{\boldsymbol{\beta}}_{mn}-\boldsymbol{\beta}_{m0}\right)+o_P(1)\right\}\boldsymbol{Z}_i^{\otimes d}\right\| \\
&\quad+\sup_{1\leq i\leq n}\left|w_i\big(\hat{\boldsymbol{\theta}}_n\big)-w_i\left(\boldsymbol{\theta}_0\right)\right|\times\left|\frac{1}{n}\sum_{i=1}^{n} Y_i(t)\exp(\boldsymbol{\beta}_{m0}^T\boldsymbol{Z}_i)\boldsymbol{Z}_i^{\otimes d}\right| \\
&\leq C\times\left\{C\,\boldsymbol{S}^{(d)}(\boldsymbol{\beta}_{m0},t)\times\left\|\hat{\boldsymbol{\beta}}_{mn}-\boldsymbol{\beta}_{m0}\right\|+o_P(1)\right\}+o_P(1)\times\boldsymbol{S}^{(d)}(\boldsymbol{\beta}_{m0},t) \\
&\quad\text{in probability} \qquad (9.10) \\
&= o_P(1). \qquad (9.11)
\end{aligned}
$$

If the weight is computed by (9.4), then $E\left[w\left(\boldsymbol{\theta}_0\right)\mid\delta,\boldsymbol{W}\right]=1$. Since W includes all covariates relevant to the missingness, $E\left[w\left(\boldsymbol{\theta}_0\right)\mid\delta,\boldsymbol{Z},\boldsymbol{W}\right]=1$. This leads to

$$
\begin{aligned}
&\bar{\boldsymbol{S}}^{(d)}(\boldsymbol{\beta}_{m0},\boldsymbol{\theta}_0,t) \\
&= \frac{1}{n}\sum_{i=1}^{n} w_i\left(\boldsymbol{\theta}_0\right) Y_i(t)\exp(\boldsymbol{\beta}_{m0}^T\boldsymbol{Z}_i)\boldsymbol{Z}_i^{\otimes d} \\
&\xrightarrow{p} E\Big[w\left(\boldsymbol{\theta}_0\right) Y(t)\exp(\boldsymbol{\beta}_{m0}^T\boldsymbol{Z})\boldsymbol{Z}^{\otimes d}\Big] \\
&= E\Big[Y(t)\,\exp(\boldsymbol{\beta}_{m0}^T\boldsymbol{Z})\,\boldsymbol{Z}^{\otimes d}\,E\left\{w\left(\boldsymbol{\theta}_0\right)\mid\delta,\boldsymbol{Z},\boldsymbol{W}\right\}\Big] \\
&= E\Big[Y(t)\,\exp(\boldsymbol{\beta}_{m0}^T\boldsymbol{Z})\,\boldsymbol{Z}^{\otimes d}\Big] \\
&= \boldsymbol{s}^{(d)}(\boldsymbol{\beta}_{m0},t). \qquad (9.12)
\end{aligned}
$$

If the weight is computed by (9.5), then $E\left[w\left(\boldsymbol{\theta}_0\right) \mid \delta\right] = 1$, and hence the above derivation needs a little modification. Assuming that $\boldsymbol{W}$ is independent of $\boldsymbol{Z}$, it follows that $E\left[w\left(\boldsymbol{\theta}_0\right) \mid \delta, \boldsymbol{Z}\right] = 1$. Therefore

$$\bar{\boldsymbol{S}}^{(d)}(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_0, t) \xrightarrow{p} E\Big[Y(t)\, \exp(\boldsymbol{\beta}_{m0}^T \boldsymbol{Z})\, \boldsymbol{Z}^{\otimes d}\, E\left\{w\left(\boldsymbol{\theta}_0\right) \mid \delta, \boldsymbol{Z}\right\}\Big] = \boldsymbol{s}^{(d)}(\boldsymbol{\beta}_{m0}, t).$$

Consequently,

$$\begin{aligned}
&\left\|\bar{\boldsymbol{S}}^{(d)}(\hat{\boldsymbol{\beta}}_{mn}, \hat{\boldsymbol{\theta}}_n, t) - \boldsymbol{s}^{(d)}(\boldsymbol{\beta}_{m0}, t)\right\| \\
&\leq \left\|\bar{\boldsymbol{S}}^{(d)}(\hat{\boldsymbol{\beta}}_{mn}, \hat{\boldsymbol{\theta}}_n, t) - \bar{\boldsymbol{S}}^{(d)}(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_0, t)\right\| + \left\|\bar{\boldsymbol{S}}^{(d)}(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_0, t) - \boldsymbol{s}^{(d)}(\boldsymbol{\beta}_{m0}, t)\right\| \\
&= o_P(1),
\end{aligned}$$

and it concludes that

$$\bar{\boldsymbol{S}}^{(d)}(\hat{\boldsymbol{\beta}}_{mn}, \hat{\boldsymbol{\theta}}_n, t) \xrightarrow{p} \boldsymbol{s}^{(d)}(\boldsymbol{\beta}_{m0}, t)$$

for $d = 0, 1, 2$. □

Remark 4. *By the continuous mapping theorem, it follows immediately from Lemma 9.4 that*

a. $\bar{\boldsymbol{E}}(\hat{\boldsymbol{\beta}}_{mn}, \hat{\boldsymbol{\theta}}_n, t) \xrightarrow{p} e(\boldsymbol{\beta}_{m0}, t)$

b. $\bar{\boldsymbol{V}}(\hat{\boldsymbol{\beta}}_{mn}, \hat{\boldsymbol{\theta}}_n, t) \xrightarrow{p} v(\boldsymbol{\beta}_{m0}, t)$

c. $\tilde{\boldsymbol{V}}\left(\boldsymbol{\beta}, \boldsymbol{\theta}, t\right) \xrightarrow{p} \tilde{\boldsymbol{v}}\left(\boldsymbol{\beta}, \boldsymbol{\theta}, t\right)$

Theorem 9.5 (Consistency)**.** *Suppose the regularity conditions given in Section 9.2.1 hold. Let* $\hat{\boldsymbol{\beta}}_{mn}^{w}$ *be the maximum weighted partial likelihood estimator (MWPLE) of* $\boldsymbol{\beta}_{m0}$*, the true value of* $\boldsymbol{\beta}_m$ *in the cause-specific hazard model (8.1),* $m = 1, 2$*. Then*

$$\hat{\boldsymbol{\beta}}_{mn}^{w} \xrightarrow{p} \boldsymbol{\beta}_{m0} \text{ as } n \to \infty,\ m = 1, 2.$$

Proof. Let $\hat{\boldsymbol{\theta}}_n$ be the MLE of the regression coefficient $\boldsymbol{\theta}_0$ in the logistic regression model (9.3). Then $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$ by the ordinary MLE theory. Denote

$$\begin{aligned}
&X_n\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \hat{\boldsymbol{\theta}}_n, t\right) \\
&= l^w\left(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \hat{\boldsymbol{\theta}}_n, t\right) - l^w\left(\boldsymbol{\beta}_{10}, \boldsymbol{\beta}_{20}, \hat{\boldsymbol{\theta}}_n, t\right) \\
&= \frac{1}{n}\sum_{m=1}^{2}\sum_{i=1}^{n}\int_0^t w_i\left(\hat{\boldsymbol{\theta}}_n\right)\left\{\left(\boldsymbol{\beta}_m - \boldsymbol{\beta}_{m0}\right)^T \boldsymbol{Z}_i - \log\frac{\bar{S}^{(0)}(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, u)}{\bar{S}^{(0)}(\boldsymbol{\beta}_{m0}, \hat{\boldsymbol{\theta}}_n, u)}\right\} dN_{mi}(u) \\
&= X_n\left(\boldsymbol{\beta}_1, \hat{\boldsymbol{\theta}}_n, t\right) + X_n\left(\boldsymbol{\beta}_2, \hat{\boldsymbol{\theta}}_n, t\right),
\end{aligned}$$

where

$$\begin{aligned}
&X_n\left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, t\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^t w_i\left(\hat{\boldsymbol{\theta}}_n\right)\left\{\left(\boldsymbol{\beta}_m - \boldsymbol{\beta}_{m0}\right)^T \boldsymbol{Z}_i - \log\frac{\bar{S}^{(0)}(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, u)}{\bar{S}^{(0)}(\boldsymbol{\beta}_{m0}, \hat{\boldsymbol{\theta}}_n, u)}\right\} dN_{mi}(u).
\end{aligned}$$

Because the log partial likelihood of $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ is the sum of two individual functions of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ alone, the parameter spaces of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are orthogonal with each other. The study of MLE of $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ can be simplified to the study of the maximum of those individual functions. In $X_n\left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, t\right)$, for $m = 1, 2$, replacing $dN_{mi}(t)$ with $Y_i(t)\exp\left(\boldsymbol{\beta}_{m0}^T \boldsymbol{Z}_i\right) d\Lambda_{m0}(t)$, we denote

$$\begin{aligned}
&A_n\left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, t\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^t w_i\left(\hat{\boldsymbol{\theta}}_n\right)\left\{\left(\boldsymbol{\beta}_m - \boldsymbol{\beta}_{m0}\right)^T \boldsymbol{Z}_i - \log\frac{\bar{S}^{(0)}(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, u)}{\bar{S}^{(0)}(\boldsymbol{\beta}_{m0}, \hat{\boldsymbol{\theta}}_n, u)}\right\} \\
&\quad \times Y_i(u)\exp\left(\boldsymbol{\beta}_{m0}^T \boldsymbol{Z}_i\right) d\Lambda_{m0}(u) \\
&= \int_0^t \left\{\left(\boldsymbol{\beta}_m - \boldsymbol{\beta}_{m0}\right)^T \bar{\boldsymbol{S}}^{(1)}(\boldsymbol{\beta}_{m0}, \hat{\boldsymbol{\theta}}_n, u) - \bar{S}^{(0)}(\boldsymbol{\beta}_{m0}, \hat{\boldsymbol{\theta}}_n, u)\log\frac{\bar{S}^{(0)}(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, u)}{\bar{S}^{(0)}(\boldsymbol{\beta}_{m0}, \hat{\boldsymbol{\theta}}_n, u)}\right\} \\
&\quad \times d\Lambda_{m0}(u)
\end{aligned}$$

Let

$$A\left(\boldsymbol{\beta}_m, t\right) = \int_0^t \left\{ \left(\boldsymbol{\beta}_m - \boldsymbol{\beta}_{m0}\right)^T \boldsymbol{s}^{(1)}(\boldsymbol{\beta}_{m0}, u) - s^{(0)}(\boldsymbol{\beta}_{m0}, u) \log \frac{s^{(0)}(\boldsymbol{\beta}_m, u)}{s^{(0)}(\boldsymbol{\beta}_{m0}, u)} \right\} d\Lambda_{m0}\left(u\right)$$

for $m = 1, 2$. Since $\bar{\boldsymbol{S}}^{(d)}(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, t) \xrightarrow{p} \boldsymbol{s}^{(d)}(\boldsymbol{\beta}_m, t)$ for $d = 0, 1, 2$, by continuous mapping theorem and dominant convergence theorem, we have

$$A_n\left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, t\right) \xrightarrow{p} A\left(\boldsymbol{\beta}_m, t\right) \tag{9.13}$$

as $n \to \infty$ for any $t \in [0, \tau]$, $m = 1, 2$. Moreover, for $m = 1, 2$,

$$\begin{aligned}
& X_n\left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, t\right) - A_n\left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, t\right) \\
= \quad & \frac{1}{n} \sum_{i=1}^n \int_0^t w_i\left(\hat{\boldsymbol{\theta}}_n\right) \left\{ \left(\boldsymbol{\beta}_m - \boldsymbol{\beta}_{m0}\right)^T \boldsymbol{Z}_i - \log \frac{\bar{S}^{(0)}(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, u)}{\bar{S}^{(0)}(\boldsymbol{\beta}_{m0}, \hat{\boldsymbol{\theta}}_n, u)} \right\} dM_{mi}\left(\boldsymbol{\beta}_m, u\right),
\end{aligned}$$

and for $0 \le t \le \tau$, it follows that

$$\begin{aligned}
& \left\langle X_n\left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, \cdot\right) - A_n\left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, \cdot\right), X_n\left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, \cdot\right) - A_n\left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, \cdot\right) \right\rangle (t) \\
& = \frac{1}{n^2} \sum_{i=1}^n \int_0^t w_i^2\left(\hat{\boldsymbol{\theta}}_n\right) \left\{ \left(\boldsymbol{\beta}_m - \boldsymbol{\beta}_{m0}\right)^T \boldsymbol{Z}_i - \log \frac{\bar{S}^{(0)}(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, u)}{\bar{S}^{(0)}(\boldsymbol{\beta}_{m0}, \hat{\boldsymbol{\theta}}_n, u)} \right\}^2 \\
& \quad \times Y_i(t) \exp(\boldsymbol{\beta}_{m0}^T \boldsymbol{Z}_i) d\Lambda_{m0}\left(u\right) \\
& \le \frac{C}{n^2} \sum_{i=1}^n \int_0^t \left\{ \left(\boldsymbol{\beta}_m - \boldsymbol{\beta}_{m0}\right)^T \boldsymbol{Z}_i - \log \frac{\bar{S}^{(0)}(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, u)}{\bar{S}^{(0)}(\boldsymbol{\beta}_{m0}, \hat{\boldsymbol{\theta}}_n, u)} \right\}^2 \\
& \quad \times Y_i(t) \exp\left(\boldsymbol{\beta}_{m0}^T \boldsymbol{Z}_i\right) d\Lambda_{m0}\left(u\right) \text{ in probability} \\
& = \frac{C}{n} \int_0^t \Bigg[\left(\boldsymbol{\beta}_m - \boldsymbol{\beta}_{m0}\right)^T \boldsymbol{S}^{(2)}\left(\boldsymbol{\beta}_{m0}, u\right) \left(\boldsymbol{\beta}_m - \boldsymbol{\beta}_{m0}\right) - 2 \left(\boldsymbol{\beta}_m - \boldsymbol{\beta}_{m0}\right)^T \boldsymbol{S}^{(1)}\left(\boldsymbol{\beta}_{m0}, u\right) \\
& \quad \times \log \frac{\bar{S}^{(0)}(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, u)}{\bar{S}^{(0)}(\boldsymbol{\beta}_{m0}, \hat{\boldsymbol{\theta}}_n, u)} + \left\{ \log \frac{\bar{S}^{(0)}(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, u)}{\bar{S}^{(0)}(\boldsymbol{\beta}_{m0}, \hat{\boldsymbol{\theta}}_n, u)} \right\}^2 S^{(0)}\left(\boldsymbol{\beta}_{m0}, u\right) \Bigg] d\Lambda_{m0}\left(u\right)
\end{aligned}$$

$$= \frac{C}{n}\int_0^t \Bigg[\left(\boldsymbol{\beta}_m - \boldsymbol{\beta}_{m0}\right)^T \boldsymbol{s}^{(2)}\left(\boldsymbol{\beta}_{m0}, u\right)\left(\boldsymbol{\beta}_m - \boldsymbol{\beta}_{m0}\right) - 2\left(\boldsymbol{\beta}_m - \boldsymbol{\beta}_{m0}\right)^T \boldsymbol{s}^{(1)}\left(\boldsymbol{\beta}_{m0}, u\right)$$
$$\times \log\frac{s^{(0)}(\boldsymbol{\beta}_m, u)}{s^{(0)}(\boldsymbol{\beta}_{m0}, u)} + \left\{\log\frac{s^{(0)}(\boldsymbol{\beta}_m, u)}{s^{(0)}(\boldsymbol{\beta}_{m0}, u)}\right\}^2 s^{(0)}\left(\boldsymbol{\beta}_{m0}, u\right)\Bigg] d\Lambda_{m0}\left(u\right) + o_p(1)$$
$$\xrightarrow{p} 0,$$

by Part (b) of Lemma 9.3, it can be concluded

$$X_n\left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, \tau\right) - A_n\left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, \tau\right) \xrightarrow{p} 0 \text{ as } n \to \infty.$$

Therefore the convergence of $A_n\left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, \tau\right)$ in (9.13) results in

$$X_n\left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, \tau\right) \xrightarrow{p} A\left(\boldsymbol{\beta}_m, \tau\right) \text{ as } n \to \infty.$$

Note that

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\beta}_m} A\left(\boldsymbol{\beta}_m, \tau\right) &= \int_0^\tau \left\{\boldsymbol{s}^{(1)}(\boldsymbol{\beta}_{m0}, t) - s^{(0)}(\boldsymbol{\beta}_{m0}, t)\frac{\boldsymbol{s}^{(1)}(\boldsymbol{\beta}_m, t)}{s^{(0)}(\boldsymbol{\beta}_m, t)}\right\} d\Lambda_{m0}\left(t\right) \\
&= \int_0^\tau s^{(0)}\left(\boldsymbol{\beta}_{m0}, t\right)\left\{\boldsymbol{e}\left(\boldsymbol{\beta}_{m0}, t\right) - \boldsymbol{e}\left(\boldsymbol{\beta}_m, t\right)\right\} d\Lambda_{m0}\left(t\right), \\
\frac{\partial^2}{\partial \boldsymbol{\beta}_m \partial \boldsymbol{\beta}_m^T} A\left(\boldsymbol{\beta}_m, \tau\right) &= -\int_0^\tau s^{(0)}(\boldsymbol{\beta}_{m0}, t)\left[\frac{\boldsymbol{s}^{(2)}\left(\boldsymbol{\beta}_m, t\right)}{s^{(0)}\left(\boldsymbol{\beta}_m, t\right)} - \left\{\frac{\boldsymbol{s}^{(1)}\left(\boldsymbol{\beta}_m, t\right)}{s^{(0)}\left(\boldsymbol{\beta}_m, t\right)}\right\}^{\otimes 2}\right] d\Lambda_{m0}\left(t\right) \\
&= -\int_0^\tau s^{(0)}(\boldsymbol{\beta}_{m0}, t)\, \boldsymbol{v}\left(\boldsymbol{\beta}_m, t\right) d\Lambda_{m0}\left(t\right) \\
&= -\boldsymbol{\Sigma}_m.
\end{aligned}$$

It is easily seen that $\frac{\partial}{\partial \boldsymbol{\beta}_m} A\left(\boldsymbol{\beta}_m, \tau\right) |_{\boldsymbol{\beta}_m = \boldsymbol{\beta}_{m0}} = 0$. So by the positive definiteness of $\boldsymbol{\Sigma}_m$ given in Condition 4, $A\left(\boldsymbol{\beta}_m, \tau\right)$ has a unique maximum at $\boldsymbol{\beta}_m = \boldsymbol{\beta}_{m0}$, for $m = 1, 2$.

Similarly,

$$\begin{aligned}
&\frac{\partial^2}{\partial \boldsymbol{\beta}_m \, \partial \boldsymbol{\beta}_m^T} X_n \left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, \tau \right) \\
&= -\frac{1}{n} \sum_{i=1}^{n} \int_0^{\tau} w_i \left(\hat{\boldsymbol{\theta}}_n \right) \frac{\partial^2 \log \bar{\boldsymbol{S}}^{(0)} (\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, t)}{\partial \boldsymbol{\beta}_m \, \partial \boldsymbol{\beta}_m^T} dN_{mi} (t) \\
&= -\frac{1}{n} \sum_{i=1}^{n} \int_0^{\tau} w_i \left(\hat{\boldsymbol{\theta}}_n \right) \bar{\boldsymbol{V}} (\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, t) \, dN_{mi} (t) \, ,
\end{aligned}$$

by (9.14) and (9.24) shown in the proof of Theorem 9.6,

$$\frac{1}{n} \sum_{i=1}^{n} \int_0^{t} w_i \left(\hat{\boldsymbol{\theta}}_n \right) \bar{\boldsymbol{V}} (\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, t) \, dN_{mi} (t) \xrightarrow{p} \boldsymbol{\Sigma}_m,$$

$\frac{\partial^2}{\partial \boldsymbol{\beta}_m \, \partial \boldsymbol{\beta}_m^T} X_n \left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, \tau \right)$ is negative definite as $n \to \infty$. Let $\hat{\boldsymbol{\beta}}_{mn}^w$ be the unique maximum of $X_n \left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}_n, \tau \right)$. Then it follows from Lemma 9.1 that

$$\hat{\boldsymbol{\beta}}_n^w \xrightarrow{p} \boldsymbol{\beta}_0 \text{ as } n \to \infty.$$

□

Theorem 9.6 (Asymptotic Normality)**.** *Suppose the regularity conditions given in Section 9.2.1 hold. Let* $\hat{\boldsymbol{\beta}}_{mn}^w$ *be the MWPLE of* $\boldsymbol{\beta}_{m0}$*,* $m = 1, 2$*. Then*

$$\sqrt{n} \left(\hat{\boldsymbol{\beta}}_{mn}^w - \boldsymbol{\beta}_{m0} \right) \xrightarrow{d} Normal \left(0, \boldsymbol{\Sigma}_m^{-1} \, \tilde{\boldsymbol{\Sigma}}_m \, \boldsymbol{\Sigma}_m^{-1} \right) \text{ as } n \to \infty \, ,$$

where $\boldsymbol{\Sigma}_m$ *and* $\tilde{\boldsymbol{\Sigma}}_m$ *are given in Condition 4.*

Proof. Apply Taylor expansion to the score function $u_m^w \left(\hat{\boldsymbol{\beta}}_{mn}^w, \hat{\boldsymbol{\theta}}_n, \tau \right)$ defined in (9.2) at $\boldsymbol{\beta}_m = \boldsymbol{\beta}_{m0}$,

$$\boldsymbol{u}_m^w \left(\hat{\boldsymbol{\beta}}_{mn}^w, \hat{\boldsymbol{\theta}}_n, \tau \right) = \boldsymbol{u}_m^w \left(\boldsymbol{\beta}_{m0}, \hat{\boldsymbol{\theta}}_n, \tau \right) - \boldsymbol{i}_m^w \left(\boldsymbol{\beta}_m^*, \hat{\boldsymbol{\theta}}_n, \tau \right) \left(\hat{\boldsymbol{\beta}}_{mn}^w - \boldsymbol{\beta}_0 \right) ,$$

where $\boldsymbol{\beta}^*_{mn} \in \left\{\boldsymbol{\beta}_m \in \mathcal{B}_m : \|\boldsymbol{\beta}_m - \boldsymbol{\beta}_{m0}\| \leq \left\|\hat{\boldsymbol{\beta}}^w_{mn} - \boldsymbol{\beta}_{m0}\right\|\right\}$, and $i^w_m(\boldsymbol{\beta}_m, \boldsymbol{\theta}, t)$ is the negative Hessian matrix given by

$$\boldsymbol{i}^w_m(\boldsymbol{\beta}_m, \boldsymbol{\theta}, t) = -\frac{\partial^2 l^w(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\theta}, t)}{\partial \boldsymbol{\beta}_m \partial \boldsymbol{\beta}_m^T} \tag{9.14}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\int_0^t w_i(\boldsymbol{\theta})\,\bar{\boldsymbol{V}}(\boldsymbol{\beta}_m, \boldsymbol{\theta}, u)\,dN_{mi}(u)\,. \tag{9.15}$$

Then

$$\begin{aligned}
&i^w_m\left(\boldsymbol{\beta}^*_{mn}, \hat{\boldsymbol{\theta}}_n, \tau\right)\sqrt{n}\left(\hat{\boldsymbol{\beta}}^w_{mn} - \boldsymbol{\beta}_{m0}\right)\\
&= \sqrt{n}u^w_m\left(\boldsymbol{\beta}_{m0}, \hat{\boldsymbol{\theta}}_n, \tau\right)\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau w_i\left(\hat{\boldsymbol{\theta}}_n\right)\left\{\boldsymbol{Z}_i - \bar{\boldsymbol{E}}(\boldsymbol{\beta}_{m0}, \hat{\boldsymbol{\theta}}_n, t)\right\}dM_{mi}(t)\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau w_i(\boldsymbol{\theta}_0)\left\{\boldsymbol{Z}_i - \bar{\boldsymbol{E}}(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_0, t)\right\}dM_{mi}(t)\\
&\quad - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau w_i(\boldsymbol{\theta}_0)\left\{\bar{\boldsymbol{E}}(\boldsymbol{\beta}_{m0}, \hat{\boldsymbol{\theta}}_n, t) - \bar{\boldsymbol{E}}(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_0, t)\right\}dM_{mi}(t)\\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\int_0^\tau \left\{w_i\left(\hat{\boldsymbol{\theta}}_n\right) - w_i(\boldsymbol{\theta}_0)\right\}\left\{\boldsymbol{Z}_i - \bar{\boldsymbol{E}}(\boldsymbol{\beta}_{m0}, \hat{\boldsymbol{\theta}}_n, t)\right\}dM_{mi}(t)\\
&\equiv \sqrt{n}\left\{\boldsymbol{u}^w_{m1} + \boldsymbol{u}^w_{m2} + \boldsymbol{u}^w_{m3}\right\},
\end{aligned} \tag{9.16}$$

where

$$\begin{aligned}
\boldsymbol{u}^w_{m1} &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau w_i(\boldsymbol{\theta}_0)\left\{\boldsymbol{Z}_i - \bar{\boldsymbol{E}}(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_0, t)\right\}dM_{mi}(t)\,,\\
\boldsymbol{u}^w_{m2} &= -\frac{1}{n}\sum_{i=1}^{n}\int_0^\tau w_i(\boldsymbol{\theta}_0)\left\{\bar{\boldsymbol{E}}(\boldsymbol{\beta}_{m0}, \hat{\boldsymbol{\theta}}_n, t) - \bar{\boldsymbol{E}}(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_0, t)\right\}dM_{mi}(t)\,,\\
\boldsymbol{u}^w_{m3} &= \frac{1}{n}\sum_{i=1}^{n}\int_0^\tau \left\{w_i\left(\hat{\boldsymbol{\theta}}_n\right) - w_i(\boldsymbol{\theta}_0)\right\}\left\{\boldsymbol{Z}_i - \bar{\boldsymbol{E}}(\boldsymbol{\beta}_{m0}, \hat{\boldsymbol{\theta}}_n, t)\right\}dM_{mi}(t)\,.
\end{aligned}$$

We will show that the followings are true

(i). $\sqrt{n}\boldsymbol{u}_{m1}^{w} \xrightarrow{p} N\left(0, \tilde{\boldsymbol{\Sigma}}_m\right)$ as $n \to \infty$,

(ii). $\sqrt{n}\boldsymbol{u}_{m2}^{w} \xrightarrow{p} 0$ as $n \to \infty$,

(iii). $\sqrt{n}\boldsymbol{u}_{m3}^{w} \xrightarrow{p} 0$ as $n \to \infty$.

(i).

It follows from Lemma 9.4 and conclusion (4) that

$$\begin{aligned}
&\left\langle n^{-1/2}\boldsymbol{u}_{m1}^{w}, n^{-1/2}\boldsymbol{u}_{m1}^{w}\right\rangle \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} w_i^2\left(\boldsymbol{\theta}_0\right)\left\{\boldsymbol{Z}_i - \bar{\boldsymbol{E}}(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_{\mathbf{0}}, t)\right\}^{\otimes 2} Y_i(u)\exp\left(\boldsymbol{\beta}_0^T\boldsymbol{Z}_i\right) d\Lambda_{m0}\left(t\right) \\
&= \int_0^{\tau} \bar{S}^{(0)}\left(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_{\mathbf{0}}, t\right)\, \tilde{V}\left(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_{\mathbf{0}}, t\right) d\Lambda_{m0}(t) \\
&\xrightarrow{p} \int_0^{\tau} \boldsymbol{s}^{(0)}\left(\boldsymbol{\beta}_{m0}, t\right)\, \tilde{\boldsymbol{v}}\left(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_0, t\right)\, d\Lambda_{m0}(t) \equiv \tilde{\boldsymbol{\Sigma}}_m.
\end{aligned}$$

Define

$$u_{m1l,\varepsilon}^{w} = \sum_{i=1}^{n}\int_0^{\tau} w_i\left(\boldsymbol{\theta}_0\right)\left\{Z_{i,l} - \bar{E}_l(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_{\mathbf{0}}, t)\right\} 1_{\left[n^{-1/2}\left|Z_{i,l} - \bar{E}_l(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_{\mathbf{0}}, t)\right| > \varepsilon\right]} dM_{mi}\left(t\right),$$

where $Z_{i,l}$ is the l^{th} element of $\boldsymbol{Z}_i$, and $\bar{E}_l(\boldsymbol{\beta}_{m0}, u)$ is the l^{th} element of $\bar{E}(\boldsymbol{\beta}_{m0}, u)$, $l = 1, \ldots, p$. Then

$$\begin{aligned}
&\left\langle n^{-1/2}u_{m1l,\varepsilon}^{w}, n^{-1/2}u_{m1l,\varepsilon}^{w}\right\rangle \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^{\tau} w_i^2\left(\boldsymbol{\theta}_0\right)\left\{Z_{i,l} - \bar{E}_l(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_{\mathbf{0}}, t)\right\}^2 \\
&\quad \times 1_{\left[n^{-1/2}\left|Z_{i,l} - \bar{E}_l(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_{\mathbf{0}}, t)\right| > \varepsilon\right]} Y_i(t)\exp\left(\boldsymbol{\beta}_{m0}^T\boldsymbol{Z}_i\right) d\Lambda_{m0}\left(t\right).
\end{aligned}$$

From Condition 2 and its follow-up conclusion (9.8), together with (9.6), there exists

a constant $0 < C < \infty$, such that

$$\Pr\Big(\sup_{1\le i\le n, t\in[0,\tau]} \left|Z_{i,l} - \bar{E}_l(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_{\mathbf{0}}, t)\right| \le C\Big) = 1,$$

$$\Pr\Big(\sup_{1\le i\le n} \left|w_i\left(\boldsymbol{\theta}_0\right)\right| \le C\Big) = 1.$$

Therefore,

$$\begin{aligned}
&\left\langle n^{-1/2} u^w_{m1l,\varepsilon}, n^{-1/2} u^w_{m1l,\varepsilon}\right\rangle \\
&\le \frac{1}{n}\sum_{i=1}^{n} \int_0^\tau C^3 Y_i(t) \exp\left(\boldsymbol{\beta}^T_{m0}\boldsymbol{Z}_i\right) 1_{\left[n^{-1/2}C>\varepsilon\right]} d\Lambda_{m0}\left(t\right) \text{ in probability} \\
&= C^4 \times 1_{\left[n^{-1/2}C>\varepsilon\right]} \int_0^\tau S^{(0)}\left(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_{\mathbf{0}}, t\right) d\Lambda_{m0}\left(t\right) \\
&\xrightarrow{p} 0\,.
\end{aligned}$$

Then it follows directly from Lemma 9.2 that

$$\sqrt{n} u^w_{m1}\left(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_{\mathbf{0}}, \tau\right) \xrightarrow{d} N\left(0, \tilde{\boldsymbol{\Sigma}}_m\right) \text{ as } n \to \infty. \tag{9.17}$$

(ii).

For $\sqrt{n}\boldsymbol{u}^w_{m2}$, we have

$$\begin{aligned}
\left\|\sqrt{n}\boldsymbol{u}^w_{m2}\right\| \le \sup_{1\le i\le n} \left\|w_i(\boldsymbol{\theta}_0)\right\| \times \sup_{t\in[0,\tau]} \left\|\bar{\boldsymbol{E}}(\boldsymbol{\beta}_{m0}, \hat{\boldsymbol{\theta}}_n, t) - \bar{\boldsymbol{E}}(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_0, t)\right\| \\
\times \left| n^{-1/2} \sum_{i=1}^{n} \int_0^\tau dM_{mi}\left(t\right)\right|,
\end{aligned}$$

where

$$\sup_{t\in[0,\tau]} \left\|\bar{\boldsymbol{E}}(\boldsymbol{\beta}_{m0}, \hat{\boldsymbol{\theta}}_n, t) - \bar{\boldsymbol{E}}(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_0, t)\right\| = o_P(1)$$

by Remark 4a, and

$$\sup_{1\le i\le n} \left\|w_i(\boldsymbol{\theta}_0)\right\| \le C$$

in probability by (9.8).

By Part (b) of Lemma 9.3, for any $\delta, \rho > 0$,

$$\begin{aligned}
&\Pr\left\{\sup_{0\le t\le\tau}\left|\sum_{i=1}^{n}\int_0^t n^{-1/2}dM_{mi}(u)\right|\ge\rho\right\}\\
&\le\frac{\delta}{\rho^2}+\Pr\left\{\int_0^\tau\frac{1}{n}\sum_{i=1}^{n}Y_i\left(t\right)\exp\left(\boldsymbol{\beta}_{m0}^T\boldsymbol{Z}_i\right)d\Lambda_{m0}\left(t\right)\ge\delta\right\}.
\end{aligned}$$

Note that $\int_0^\tau s^{(0)}\left(\boldsymbol{\beta}_{m0},t\right)$ is finite by Condition 1. Let $\delta > \int_0^\tau s^{(0)}\left(\boldsymbol{\beta}_{m0},t\right)d\Lambda_{m0}\left(t\right)$, then $\left|n^{-1/2}\sum_{i=1}^{n}\int_0^\tau dM_{mi}(t)\right|$ is bounded in probability, i.e.

$$n^{-1/2}\sum_{i=1}^{n}\int_0^\tau dM_i(t)=O_P(1). \tag{9.18}$$

This concludes that

$$\sqrt{n}\boldsymbol{u}_{m2}^w=\boldsymbol{o}_P(1). \tag{9.19}$$

(iii).

For $\boldsymbol{u}_{m3}^w$, we have

$$\begin{aligned}
\left\|\sqrt{n}\boldsymbol{u}_{m3}^w\right\|\le&\sup_{1\le i\le n}\left|\left\{w_i\left(\hat{\boldsymbol{\theta}}_n\right)-w_i\left(\boldsymbol{\theta}_0\right)\right\}\right|\times\sup_{\substack{1\le i\le n,\\ t\in[0,\tau]}}\left\|\boldsymbol{Z}_i-\bar{\boldsymbol{E}}(\boldsymbol{\beta}_{m0},\hat{\boldsymbol{\theta}}_n,t)\right\|\\
&\times\left|n^{-1/2}\sum_{i=1}^{n}\int_0^\tau dM_{mi}\left(t\right)\right|
\end{aligned}$$

Again, by Part (b) of Lemma 9.3, together with conclusions (9.8) and (9.18), we get

$$\sqrt{n}\boldsymbol{u}_{m3}^w\left(\boldsymbol{\beta}_{m0},\hat{\boldsymbol{\theta}}_n,\tau\right)=\boldsymbol{o}_P(1). \tag{9.20}$$

Then it follows from conclusions (9.17), (9.19) and (9.20)

$$\sqrt{n}i_m^w\left(\boldsymbol{\beta}_m^*,\hat{\boldsymbol{\theta}}_n,\tau\right)\left(\hat{\boldsymbol{\beta}}_{mn}^w-\boldsymbol{\beta}_{m0}\right)\xrightarrow{d}N\left(0,\tilde{\boldsymbol{\Sigma}}_m\right)\text{ as }n\to\infty. \tag{9.21}$$

Next to prove

$$i_m^w\left(\boldsymbol{\beta}_m^*, \hat{\boldsymbol{\theta}}_n, \tau\right) \xrightarrow{p} \boldsymbol{\Sigma}_m \text{ as } n \to \infty.$$

Let $\bar{N}_m(t) = n^{-1}\sum_{i=1}^n N_{mi}(t)$. By Part (a) of Lemma 9.3, together with Condition 3, for any $\delta, \rho > 0$,

$$\begin{aligned}
\Pr\left\{\bar{N}_m(t) > \rho\right\} &\leq \frac{\delta}{\rho} + \Pr\left\{\int_0^\tau n^{-1}\sum_{i=1}^n Y_i(t)\exp\left(\boldsymbol{\beta}_{m0}^T\boldsymbol{Z}_i\right)\, d\Lambda_{m0}(t) > \delta\right\} \\
&= \frac{\delta}{\rho} + \Pr\left\{\int_0^\tau S^{(0)}\left(\boldsymbol{\beta}_{m0}, t\right)\, d\Lambda_{m0}(t) > \delta\right\} \\
&\xrightarrow{p} \frac{\delta}{\rho} + \Pr\left\{\int_0^\tau s^{(0)}\left(\boldsymbol{\beta}_{m0}, t\right)\, d\Lambda_{m0}(t) > \delta\right\}.
\end{aligned}$$

Taking $\delta > \int_0^\tau s^{(0)}\left(\boldsymbol{\beta}_{m0}, t\right)\, d\Lambda_{m0}(t)$ gives

$$\lim_{\rho\to\infty}\lim_{n\to\infty}\Pr\left\{\bar{N}_m(\tau) > \rho\right\} = 0. \tag{9.22}$$

Since $E\left[\int_0^\tau dM_m(t)\right] = 0$, by the law of large numbers,

$$n^{-1}\sum_{i=1}^n\int_0^\tau dM_{mi}(t) = o_P(1). \tag{9.23}$$

Note that $\boldsymbol{\beta}_{mn}^* \xrightarrow{p} \boldsymbol{\beta}_{m0}$, $\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0$, by applying Lemma 9.4 and Condition 1, together with conclusions (9.9), (9.8), (9.22) and (9.23),

$$\begin{aligned}
&\left\|i_m^w\left(\boldsymbol{\beta}_{mn}^*, \hat{\boldsymbol{\theta}}_n, \tau\right) - \boldsymbol{\Sigma}_m\right\| \\
&= \left\|n^{-1}\sum_{i=1}^n\int_0^\tau w_i\left(\hat{\boldsymbol{\theta}}_n\right)\bar{\boldsymbol{V}}\left(\boldsymbol{\beta}_m^*, \hat{\boldsymbol{\theta}}_n, t\right) dN_{mi}(t) - \int_0^\tau s^{(0)}\left(\boldsymbol{\beta}_{m0}, t\right)\boldsymbol{v}\left(\boldsymbol{\beta}_{m0}, t\right)\, d\Lambda_{m0}(t)\right\| \\
&\leq \sup_{1\leq i\leq n}\left|w_i\left(\hat{\boldsymbol{\theta}}_n\right) - w_i\left(\boldsymbol{\theta}_0\right)\right| \times \left\|\int_0^\tau \bar{\boldsymbol{V}}\left(\boldsymbol{\beta}_m^*, \hat{\boldsymbol{\theta}}_n, t\right) d\bar{N}_m(t)\right\| \\
&\quad + \sup_{1\leq i\leq n}\left\{w_i\left(\boldsymbol{\theta}_0\right)\right\} \times \left\|\int_0^\tau\left\{\bar{\boldsymbol{V}}\left(\boldsymbol{\beta}_m^*, \hat{\boldsymbol{\theta}}_n, t\right) - \boldsymbol{v}\left(\boldsymbol{\beta}_{m0}, t\right)\right\} d\bar{N}_m(t)\right\| \\
&\quad + \sup_{1\leq i\leq n}\left\{w_i\left(\boldsymbol{\theta}_0\right)\right\} \times \sup_{t\in[0,\tau]}\left\|\boldsymbol{v}\left(\boldsymbol{\beta}_{m0}, t\right)\right\| \times \left|n^{-1}\sum_{i=1}^n\int_0^\tau dM_{mi}(t)\right|
\end{aligned}$$

$$
\begin{aligned}
&+ \left\| \int_0^\tau \left\{ \bar{S}^{(0)} \left(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_0, t \right) - s^{(0)} \left(\boldsymbol{\beta}_{m0}, t \right) \right\} \boldsymbol{v} \left(\boldsymbol{\beta}_{m0}, t \right) \, d\Lambda_{m0}(t) \right\| \\
&\leq o_P(1) \times \left\{ \sup_{t \in [0,\tau]} \left\| \boldsymbol{v} \left(\boldsymbol{\beta}_{m0}, t \right) \right\| + o_P(1) \right\} \times \bar{N}_m \left(\tau \right) + C \times o_P(1) \times \bar{N}_m \left(\tau \right) \\
&\quad + C \times o_P(1) + o_P(1) \times \sup_{t \in [0,\tau]} \left\| \boldsymbol{v} \left(\boldsymbol{\beta}_{m0}, t \right) \right\| \times \int_0^\tau d\Lambda_{m0}(t) \\
&= o_P(1).
\end{aligned}
$$

Then it follows that

$$
i_m^w \left(\boldsymbol{\beta}_m^*, \hat{\boldsymbol{\theta}}_n, \tau \right) \xrightarrow{p} \boldsymbol{\Sigma}_m \text{ as } n \to \infty. \tag{9.24}
$$

Applying Slutsky's theorem to (9.21) leads to

$$
\sqrt{n} \left(\hat{\boldsymbol{\beta}}_{mn}^w - \boldsymbol{\beta}_{m0} \right) \xrightarrow{d} N \left(0, \boldsymbol{\Sigma}_m^{-1} \, \tilde{\boldsymbol{\Sigma}}_m \, \boldsymbol{\Sigma}_m^{-1} \right)
$$

as $n \to \infty$. □

CHAPTER 10
DOUBLE ROBUST APPROACH

As we have shown in Chapter 9, the WCC approach yields consistent and asymptotically normal estimators. However, it partially uses the information from the missing data via a parametric model for the non-missingness, while the relationship between observed failure times with missing cause and the corresponding covariates via the cause-specific hazard models is not accounted. Motivated by a double robust approach for additive hazards models proposed by Lu and Liang (2008), we further make use of the missing data by weighting the counting processes of failures with the selection probability as well as the probability of failures with a certain cause type. The former probability can be estimated using the parametric model introduced for the WCC approach, while the latter one requires to specify another parametric model for the cause type. We use the logistic regression model proposed in Lu and Tsiatis (2001),

$$\log \frac{\tau_1(\boldsymbol{\gamma}; \boldsymbol{V})}{1 - \tau_1(\boldsymbol{\gamma}; \boldsymbol{V})} = \boldsymbol{\gamma}^T \boldsymbol{V}, \tag{10.1}$$

where $\tau_m(\boldsymbol{\gamma}; \boldsymbol{V}) = P\left(\delta = m \mid \delta > 0, \pi = 0, \boldsymbol{V}\right)$, is the conditional probability that the cause of a failure is the m^{th} type given relevant covariates $\boldsymbol{V}$. In the case of m ($m > 2$) cause types, multinomial logistic regression model can be adopted.

To correctly estimate $\tau_m(\boldsymbol{\gamma}; \boldsymbol{V})$ ($m = 1, 2$) using the model (10.1), we need the assumption of *missing at random* (MAR), i.e. the missingness of failure cause does

not depend on the cause type itself,

$$P\left(\pi = 0 \mid \delta, \delta > 0, \boldsymbol{W}\right) = P\left(\pi = 0 \mid \delta > 0, \boldsymbol{W}\right).$$

The MAR assumption stipulates that π and δ are independent given $\{\delta > 0, \boldsymbol{W}\}$, which can be expressed equivalently as

$$P\left(\delta = m \mid \pi = 0, \delta > 0, \boldsymbol{W}\right) = P\left(\delta = m \mid \delta > 0, \boldsymbol{W}\right).$$

Therefore, with the MAR assumption, the probability $\tau_m(\boldsymbol{\gamma}; \boldsymbol{V})$ estimated using the complete-case data is guaranteed to be consistent with the one from the full data.

Let $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\gamma}}$ be the estimators of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ from the logistic regression models (9.3) and (10.1) respectively.

The double robust approach is based on the following estimating equations

$$\begin{aligned}
&\boldsymbol{u}_m^{dr}\left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, t\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^t \left\{\boldsymbol{Z}_i - \boldsymbol{E}(\boldsymbol{\beta}_m, u)\right\}\left\{\frac{R_i}{p_i\left(\hat{\boldsymbol{\theta}}; \boldsymbol{W}, \delta\right)} dN_{mi}(u) - \frac{R_i - p_i\left(\hat{\boldsymbol{\theta}}; \boldsymbol{W}, \delta\right)}{p_i\left(\hat{\boldsymbol{\theta}}; \boldsymbol{W}, \delta\right)}\right. \\
&\quad \left.\times\, \tau_{mi}\left(\hat{\boldsymbol{\gamma}}; \boldsymbol{V}\right)\, dN_i(u)\right\} \equiv 0
\end{aligned}$$

for $m = 1, 2$. Define

$$N_i^*\left(t; \boldsymbol{\theta}, \boldsymbol{\gamma}\right) = \frac{R_i}{p_i\left(\boldsymbol{\theta}; \boldsymbol{W}, \delta\right)} N_{mi}(t) - \frac{R_i - p_i\left(\boldsymbol{\theta}; \boldsymbol{W}, \delta\right)}{p_i\left(\boldsymbol{\theta}; \boldsymbol{W}, \delta\right)} \tau_{mi}\left(\boldsymbol{\gamma}; \boldsymbol{V}\right)\, N_i(t).$$

Then the double robust estimating equations can be rewritten as

$$\boldsymbol{u}_m^{dr}\left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, t\right) = \frac{1}{n}\sum_{i=1}^{n}\int_0^t \left\{\boldsymbol{Z}_i - \boldsymbol{E}(\boldsymbol{\beta}_m, u)\right\} dN_i^*\left(u; \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}\right) \equiv 0 \tag{10.2}$$

The double robust estimator $\hat{\boldsymbol{\beta}}_m^{dr}$ of $\boldsymbol{\beta}_m$ can be solved using Newton-Raphson method based upon (10.2). According to the result given by Lu and Liang (2008), we conjecture the following asymptotic normality result

$$\sqrt{n}\left(\hat{\boldsymbol{\beta}}_m^{dr} - \boldsymbol{\beta}_{m0}\right) \to N\left(0, \left(\boldsymbol{\Sigma}_m^{dr}\right)^{-1} \tilde{\boldsymbol{\Sigma}}_m^{dr} \left\{\left(\boldsymbol{\Sigma}_m^{dr}\right)^{-1}\right\}^T\right),$$

where $\boldsymbol{\Sigma}_m^{dr}$ can be estimated by $\frac{1}{n}\sum_{i=1}^n \int_0^\infty \boldsymbol{V}(\hat{\boldsymbol{\beta}}_m^{dr}, t) dN_i^*\left(t; \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}\right)$, and $\tilde{\boldsymbol{\Sigma}}_m^{dr}$ can be estimated by $n^{-1}\sum_{i=1}^n \hat{\boldsymbol{\phi}}_i \hat{\boldsymbol{\phi}}_i^T$, with $\hat{\boldsymbol{\phi}}_i$ calculated by

$$\int_0^\infty \left\{\boldsymbol{Z}_i - \boldsymbol{E}(\hat{\boldsymbol{\beta}}_m^{dr}, t)\right\} dN_i^*\left(t; \hat{\boldsymbol{\theta}}, \boldsymbol{\gamma}\right) - \hat{\boldsymbol{B}}_\theta \hat{\boldsymbol{I}}_\theta^{-1} \hat{\boldsymbol{S}}_{\theta i} - \hat{\boldsymbol{B}}_\gamma \hat{\boldsymbol{I}}_\gamma^{-1} \hat{\boldsymbol{S}}_{\gamma i}.$$

Here $\hat{\boldsymbol{B}}_\theta$ and $\hat{\boldsymbol{B}}_\gamma$ are the derivatives of $\boldsymbol{u}_m^{dr}(\boldsymbol{\beta}_m, \boldsymbol{\theta}, \boldsymbol{\gamma}, t)$ with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ respectively with all parameters replaced by their estimates. $\hat{\boldsymbol{I}}_\theta$ and $\hat{\boldsymbol{S}}_{\theta i}$ are the estimate of Fisher information matrix and the score function contributed by the i^{th} subject, respectively, for the parametric model (9.3). $\hat{\boldsymbol{I}}_\gamma$ and $\hat{\boldsymbol{S}}_{\gamma i}$ are similarly defined for the parametric model (10.1).

While the complete proof for the asymptotic properties of $\hat{\boldsymbol{\beta}}_m^{dr}$ is more involved and left for future work, we would like to justify the double robustness by showing that one necessary condition for the consistency of the double robust estimator holds

$$E\left[\boldsymbol{u}_m^{dr}(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0, t) \mid \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{V}\right] = 0, \tag{10.3}$$

where $\boldsymbol{\beta}_{m0}$, $\boldsymbol{\gamma}_0$ and $\boldsymbol{\theta}_0$ are the true values of parameters assuming at least one of the parametric models (9.3) and (10.1) is correctly specified.

Following the definition of $\boldsymbol{E}(\boldsymbol{\beta}_m, t)$ in Section 8.1, we have

$$\sum_{i=1}^n \left\{\boldsymbol{Z}_i - \boldsymbol{E}(\boldsymbol{\beta}_m, t)\right\} Y_i(t) \exp\left(\boldsymbol{\beta}_m^T \boldsymbol{Z}_i\right) = 0$$

for any $t \geq 0$, then the double robust estimating equations (10.2) can be rewritten as

$$
\begin{aligned}
&\boldsymbol{u}_m^{dr}\left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, t\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^t \left\{\boldsymbol{Z}_i - \boldsymbol{E}(\boldsymbol{\beta}_m, u)\right\}\left\{\frac{R_i}{p_i\left(\hat{\boldsymbol{\theta}}; \boldsymbol{W}, \delta\right)} dN_{mi}(u) - \frac{R_i - p_i\left(\hat{\boldsymbol{\theta}}; \boldsymbol{W}, \delta\right)}{p_i\left(\hat{\boldsymbol{\theta}}; \boldsymbol{W}, \delta\right)}\right. \\
&\quad \left.\times \tau_{mi}\left(\hat{\boldsymbol{\gamma}}; \boldsymbol{V}\right) dN_i(u)\right\} - \frac{1}{n}\sum_{i=1}^{n}\int_0^t \left\{\boldsymbol{Z}_i - \boldsymbol{E}(\boldsymbol{\beta}_m, u)\right\} Y_i(u)\exp\left(\boldsymbol{\beta}_m^T \boldsymbol{Z}_i\right) d\Lambda_{m0}(u) \\
&= \frac{1}{n}\sum_{i=1}^{n}\int_0^t \left\{\boldsymbol{Z}_i - \boldsymbol{E}(\boldsymbol{\beta}_m, u)\right\}\left[dM_{mi}(u) + \frac{R_i - p_i\left(\hat{\boldsymbol{\theta}}; \boldsymbol{W}, \delta\right)}{p_i\left(\hat{\boldsymbol{\theta}}; \boldsymbol{W}, \delta\right)}\right. \\
&\quad \left.\times \left\{I_{[\delta_i = m | \delta_i > 0]} - \tau_{mi}\left(\hat{\boldsymbol{\gamma}}; \boldsymbol{V}\right)\right\}\right] dN_i(u) \\
&\equiv \boldsymbol{u}_{m1}^{dr}\left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, t\right) + \boldsymbol{u}_{m2}^{dr}\left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, t\right),
\end{aligned}
$$

where $\Lambda_{m0}(t)$ is the true cause-specific cumulative hazard, $M_{mi}(t)$ is the cause-specific martingale defined in Section 8.1, $\boldsymbol{u}_{m1}^{dr}\left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, t\right)$ and $\boldsymbol{u}_{m2}^{dr}\left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, t\right)$ are given by

$$
\begin{aligned}
\boldsymbol{u}_{m1}^{dr}\left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, t\right) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^t \left\{\boldsymbol{Z}_i - \boldsymbol{E}(\boldsymbol{\beta}_m, u)\right\} dM_{mi}(u), \\
\boldsymbol{u}_{m2}^{dr}\left(\boldsymbol{\beta}_m, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}, t\right) &= \frac{1}{n}\sum_{i=1}^{n}\int_0^t \frac{1}{p_i\left(\hat{\boldsymbol{\theta}}; \boldsymbol{W}, \delta\right)}\left\{\boldsymbol{Z}_i - \boldsymbol{E}(\boldsymbol{\beta}_m, u)\right\} \\
&\qquad \times \left\{R_i - p_i\left(\hat{\boldsymbol{\theta}}; \boldsymbol{W}, \delta\right)\right\}\left\{I_{[\delta_i = m | \delta_i > 0]} - \tau_{mi}\left(\hat{\boldsymbol{\gamma}}; \boldsymbol{V}\right)\right\} dN_i(u).
\end{aligned}
$$

It is obvious that

$$
E\left[\boldsymbol{u}_{m1}^{dr}\left(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0, t\right) \mid \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{V}\right] = 0,
$$

due to the fact that $M_{mi}(t)$ is a martingale.

As stated earlier, MAR implies that π and δ are independent given $\{\delta > 0, \boldsymbol{W}\}$.

Therefore, assuming MAR holds, we have

$$
\begin{aligned}
& E\left[\boldsymbol{u}_{m2}^{dr}\left(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0, t\right) \mid \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{V}\right] \\
& = \frac{1}{n} \sum_{i=1}^{n} \int_0^t \frac{1}{p_i\left(\boldsymbol{\theta}_0; \boldsymbol{W}\right)} \left\{\boldsymbol{Z}_i - \boldsymbol{E}(\boldsymbol{\beta}_m, u)\right\} \\
& \quad \times E\Big[\left\{I_{[\pi_i=0|\delta_i>0]} - p_i\left(\boldsymbol{\theta}_0; \boldsymbol{W}\right)\right\} \left\{I_{[\delta_i=m|\delta_i>0]} - \tau_{mi}\left(\boldsymbol{\gamma}_0; \boldsymbol{V}\right)\right\} dN_i\left(u\right) \Big] \\
& = \frac{1}{n} \sum_{i=1}^{n} \int_0^t \frac{1}{p_i\left(\boldsymbol{\theta}_0; \boldsymbol{W}\right)} \left\{\boldsymbol{Z}_i - \boldsymbol{E}(\boldsymbol{\beta}_m, u)\right\} \times E\Big[E[\left\{I_{[\pi_i=0|\delta_i>0]} - p_i\left(\boldsymbol{\theta}_0; \boldsymbol{W}\right)\right\} \\
& \quad \times \left\{I_{[\delta_i=m|\delta_i>0]} - \tau_{mi}\left(\boldsymbol{\gamma}_0; \boldsymbol{V}\right)\right\} dN_i\left(t\right) \mid \mathcal{F}_t, \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{V}] \mid \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{V} \Big] \\
& = \frac{1}{n} \sum_{i=1}^{n} \int_0^t \frac{1}{p_i\left(\boldsymbol{\theta}_0; \boldsymbol{W}\right)} \left\{\boldsymbol{Z}_i - \boldsymbol{E}(\boldsymbol{\beta}_m, u)\right\} E\Big[E\left[I_{[\pi_i=0|\delta_i>0]} - p_i\left(\boldsymbol{\theta}_0; \boldsymbol{W}\right) \mid \mathcal{F}_t, \boldsymbol{W}\right] \\
& \quad \times E\left[I_{[\delta_i=m|\delta_i>0]} - \tau_{mi}\left(\boldsymbol{\gamma}_0; \boldsymbol{V}\right) \mid \mathcal{F}_t, \boldsymbol{V}\right] dN_i\left(t\right) \mid \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{V} \Big].
\end{aligned}
$$

When either of the parametric models (9.3) and (10.1) is correct,

$$
E\left[\boldsymbol{u}_{m2}^{dr}\left(\boldsymbol{\beta}_{m0}, \boldsymbol{\theta}_0, \boldsymbol{\gamma}_0, t\right) \mid \boldsymbol{Z}, \boldsymbol{W}, \boldsymbol{V}\right] = 0,
$$

and therefore (10.3) holds. This indicates that the double robust approach is valid as long as the MAR assumption holds and at least one of the parametric models (9.3) and (10.1) is correctly specified.

CHAPTER 11
SIMULATION STUDIES

11.1 Data Generation

In simulations, we assume the cause-specific baseline hazard $\lambda_{m0}(t)$ and baseline cumulative hazard $\Lambda_{m0}(t)$ are given by

$$\lambda_{m0}(t) = (\rho_m/\alpha_m)(t/\alpha_m)^{\rho_m - 1}, \tag{11.1}$$

$$\Lambda_{m0}(t) = (t/\alpha_m)^{\rho_m} \tag{11.2}$$

for $m = 1, 2$. They are actually the hazard functions of event times that follow Weibull(ρ_m, α_m), respectively. Then the cause-specific hazards (8.1), the cause-specific sub-densities (8.5), the overall density (8.4) and the overall survival function (8.3) can be explicitly given according to definitions (11.1) and (11.2).

Assume that the censoring time C has the hazard function

$$\lambda_C(t \mid \boldsymbol{Z}) = \gamma e^{\boldsymbol{\eta}^T \boldsymbol{Z}},$$

where $\boldsymbol{\eta}$ is a vector of fixed coefficients for the covariates, and γ is the constant baseline hazard.

Let f_C and S_C be the conditional density function and the conditional survival function for the censoring time C respectively given the covariates $\boldsymbol{Z}$. Then we have

$$S_C(t \mid \boldsymbol{Z}) = \exp\left(-\gamma t e^{\boldsymbol{\eta}^T \boldsymbol{Z}}\right),$$

$$f_C(t \mid \boldsymbol{Z}) = \gamma e^{\boldsymbol{\eta}^T \boldsymbol{Z}} \exp\left(-\gamma t e^{\boldsymbol{\eta}^T \boldsymbol{Z}}\right).$$

We generate the data by taking the following steps:

1. Determine the censoring rate r_c and the proportion r_1 of non-censored failure times from cause type I:

   - The censoring rate is determined by

$$\begin{aligned} r_c &= E\left[P\left(T > C \mid \boldsymbol{Z}\right)\right] \\ &= E\left[\int_0^{\infty} S_{T|\boldsymbol{Z}}\left(t\right) \times f_{C|\boldsymbol{Z}}\left(t\right)\, dt\right] \\ &\equiv h_1(\alpha_1, \alpha_2, \rho_1, \rho_2, \gamma, \boldsymbol{\eta}). \end{aligned} \tag{11.3}$$

   - Similarly, the proportion of type I failures is determined by

$$\begin{aligned} r_1 &= E\left[P\left(\delta = 1 \mid T < C, \boldsymbol{Z}\right)\right] \\ &= E\left[\frac{\int_0^{+\infty} P\left(\delta = 1, C > T \mid T = t, \boldsymbol{Z}\right) \times f_{T|\boldsymbol{Z}}\left(t\right)\, dt}{\int_0^{+\infty} P\left(C > T \mid T = t, \boldsymbol{Z}\right) \times f_{T|\boldsymbol{Z}}\left(t\right)\, dt}\right] \\ &= E\left[\frac{\int_0^{+\infty} \frac{\lambda_1(t|\boldsymbol{Z})}{\lambda_1(t|\boldsymbol{Z})+\lambda_2(t|\boldsymbol{Z})} \times S_{C|\boldsymbol{Z}}\left(t\right) \times f_{T|\boldsymbol{Z}}\left(t\right)\, dt}{\int_0^{+\infty} S_{C|\boldsymbol{Z}}(t) \times f_{T|\boldsymbol{Z}}(t)\, dt}\right] \\ &= E\left[\frac{\int_0^{+\infty} S_{C|\boldsymbol{Z}}\left(t\right) \times f_{T_1|\boldsymbol{Z}}\left(t\right)\, dt}{\int_0^{+\infty} S_{C|\boldsymbol{Z}}(t) \times f_{T|\boldsymbol{Z}}(t)\, dt}\right] \\ &\equiv h_2(\alpha_1, \alpha_2, \rho_1, \rho_2, \gamma, \boldsymbol{\eta}). \end{aligned} \tag{11.4}$$

   - Let α_1, ρ_1, ρ_2 and $\boldsymbol{\eta}$ be predetermined. Then equations (11.3) and (11.4) can be solved jointly for α_2 and γ in order to achieve the desired censoring rate r_c and proportion of type I failures r_1. In our simulation studies, we set $\alpha_1 = 0.1$, $\rho_1 = 4$, $\rho_2 = 2$, $\boldsymbol{\eta} = \mathbf{1}$, $r_c = 30\%$, and $\gamma_1 = 60\%$.

2. Generate failure times T_1 and T_2 independently from the distributions with

hazard functions $\lambda_1(t \mid \boldsymbol{Z})$ and $\lambda_2(t \mid \boldsymbol{Z})$, respectively. That is

$$\exp\{-\Lambda(T_m \mid \boldsymbol{Z})\} = S_m(T_m) \equiv U_m \Rightarrow$$
$$T_m = \frac{1}{\alpha_m}\left[-\log(U_m) \times \exp\left(-\boldsymbol{\beta}^T\boldsymbol{Z}\right)\right]^{\frac{1}{\rho_m}}$$

for $U_m \sim Uniform(0,1)$.

3. Generate the censoring time C from the distribution with hazard function $\lambda_C(t \mid \boldsymbol{Z})$, i.e.

$$C = -\frac{1}{\gamma}\log(U_C) \times \exp\left(-\boldsymbol{\eta}^T\boldsymbol{Z}\right)$$

for $U_C \sim Uniform(0,1)$.

4. Let $X = (T_1 \wedge T_2 \wedge C)$ be the observed time. Determine its status as follows:

$$\begin{cases} \delta = 0 & \text{if } T_1 > C \text{ and } T_2 > C, \\ \delta = 1 & \text{if } T_1 \leq C \text{ and } T_1 \leq T_2, \\ \delta = 2 & \text{if } T_2 \leq C \text{ and } T_2 < T_1. \end{cases}$$

5. Generate the missing indicator π for those failures from a Bernoulli distribution with parameter $1-p(\boldsymbol{\theta};\boldsymbol{W})$, where $p(\boldsymbol{\theta};\boldsymbol{W})$, defined in Section 8.1, is the selection probability among failure times given covariates $\boldsymbol{W}$ that is modeled using the logistic regression model given by (9.3). The covariate matrix $\boldsymbol{W}$ actually determines the missing mechanism. In our simulation studies, we consider the following four cases:

   - $\boldsymbol{W}$ is one, which yields the case of *missing completely at random* (MCAR).

- $\boldsymbol{W}$ is the vector that includes constant one and the covariates $\boldsymbol{Z}$ given in the cause-specific hazards models, which yields the case of *missing at random* (MAR1).
- The missing probability depends on the failure time via a linear function. This is the case of *missing at random* as well (MAR2), but there is no corresponding correct logistic model.
- $\boldsymbol{W}$ is the vector that includes constant one and the dummy variable for cause type, which is the case of *missing not at random* (MNAR).

Note that the missingness is applicable only for the event data. We will compare different estimation approaches for data generated using each of the four missing mechanisms.

11.2 Comparison of Weighted and Non-Weighted Complete-Case Approaches

We first compare the complete-case (CC) approaches with different weighting methods: (i) non-weighted method, (ii) weighted method 1 (WCC1) with weight being the sample proportion of non-missing failures, constant for all censored subjects, are computed using the empirical proportion of missingness, (iii) weighted method 2 (WCC2) with weight computed using the logistic regression model, which is the true model for the missing mechanism in the simulation setting. Tables 11.1 and 11.2 show simulation results from the three CC approaches with sample size 100 and 200 respectively. For the comparison purpose, we also add the results in each table based

on the full data with no missing cause. Obviously, the full-data approach yields the best results because it makes use of the most information.

The results are for the estimates of the regression coefficient for each cause type. Note that the proportion of cause type I failure is larger than that of cause type II, 60% vs 40%, which results in better estimation of β_1 than estimation of β_2 in terms of bias and coverage.

The CC approach without using weights results in biased estimators in all cases (MCAR, MAR1, MAR2 and MNAR). When the sample size increases from 100 to 200, there is no improvement on the bias. This reinforces our argument in Section 7.2.1 that the CC data is actually a biased sample from the full data and hence the ordinary estimation methods of Cox proportional hazards model introduce estimation bias.

Results of the two weighted complete-case (WCC) approaches show that the bias has been corrected very well in the cases of MCAR, MAR1 and even MNAR, where the true logistic regression model for the missingness exists and is adopted by the WCC2 approach. The consistency observed for the WCC1 estimators in the cases of MCAR and MNAR demonstrates the theoretical argument discussed in Section 9.1 that using the sample proportion of failures with known cause as the estimate of the selection probability can also yield consistent estimation of the regression coefficients in cause-specific hazards models under the assumption that $\boldsymbol{W}$ does not include $\boldsymbol{Z}$. However, in the case of MAR1 where $\boldsymbol{W} \equiv \boldsymbol{Z}$ that violates that assumption, WCC1 estimators do show the bias. It is further observed that WCC1 estimators

have slightly smaller standard error than WCC2 estimators and thus have smaller root mean square error (RMSE). The same phenomenon was also observed in the simulation results presented in Lu and Liang (2008). This is somehow against our intuition as WCC2 approach utilizes the true logistic regression model designed in the simulation setting and it requires further investigation.

In the case of MNAR, the missingness of failure cause depends on the cause itself, consequently, the MAR assumption is violated. Results show that the WCC approaches work well in bias correction, which is consistent with the theoretical proof of consistency and asymptotic normality shown in Chapter 9 where the MAR assumption is not needed.

In the case of MAR2 where the missingness depends on the observed times via a linear function and there is no correct logistic regression model for the missing mechanism, the estimation bias is more prominent for both WCC1 and WCC2, and it does not show any decreasing trend as sample size increase from 100 to 200. But the results are still comparable between the two approaches.

In summary, WCC1 and WCC2 appear to be comparable in terms of estimation results and it may indicate that the selection of parametric model for computing the weight is not critical.

Compared to the full-data estimators, the WCC estimators, while removing the estimation bias presented in the CC approach, are still not very efficient as the standard error of the estimates is significantly larger. We next study the improved double robust approach.

Table 11.1: Comparison of three complete-case approaches with different weighting methods, for sample size 100, censoring 30%, proportion of missingness 21% and repetitions 1000.

Methods		Full Data *	CC †	WCC1 ‡	WCC2 §
MCAR	True β	0.5, 0.2	0.5, 0.2	0.5, 0.2	0.5, 0.2
	Ave $\hat{\beta}$	0.523, 0.22	0.511, 0.181	0.539, 0.221	0.541, 0.222
	Emp. SD ($\hat{\beta}$)	0.592, 0.717	0.722, 0.875	0.712, 0.868	0.718, 0.87
	Ave SE ($\hat{\beta}$)	0.576, 0.687	0.704, 0.836	0.703, 0.802	0.706, 0.805
	RMSE ($\hat{\beta}$)	0.592, 0.717	0.721, 0.875	0.713, 0.868	0.719, 0.87
	Emp. Coverage	0.943, 0.955	0.952, 0.964	0.953, 0.954	0.954, 0.954
MAR1¶	True β	0.5, 0.2	0.5, 0.2	0.5, 0.2	0.5, 0.2
	Ave $\hat{\beta}$	0.523, 0.22	0.477, 0.143	0.514, 0.195	0.53, 0.215
	Emp. SD ($\hat{\beta}$)	0.592, 0.717	0.73, 0.892	0.719, 0.885	0.727, 0.89
	Ave SE ($\hat{\beta}$)	0.576, 0.687	0.716, 0.851	0.716, 0.814	0.717, 0.817
	RMSE ($\hat{\beta}$)	0.592, 0.717	0.73, 0.893	0.719, 0.885	0.728, 0.889
	Emp. Coverage	0.943, 0.955	0.952, 0.957	0.956, 0.951	0.956, 0.95
MAR2‖	True β	0.5, 0.2	0.5, 0.2	0.5, 0.2	0.5, 0.2
	Ave $\hat{\beta}$	0.523, 0.22	0.474, 0.145	0.502, 0.185	0.498, 0.177
	Emp. SD ($\hat{\beta}$)	0.592, 0.717	0.73, 0.853	0.722, 0.846	0.726, 0.85
	Ave SE ($\hat{\beta}$)	0.576, 0.687	0.715, 0.815	0.713, 0.78	0.716, 0.782
	RMSE ($\hat{\beta}$)	0.592, 0.717	0.73, 0.855	0.721, 0.846	0.726, 0.849
	Emp. Coverage	0.943, 0.955	0.952, 0.963	0.956, 0.956	0.954, 0.954
MNAR**	True β	0.5, 0.2	0.5, 0.2	0.5, 0.2	0.5, 0.2
	Ave $\hat{\beta}$	0.523, 0.22	0.5, 0.188	0.526, 0.227	0.549, 0.227
	Emp. SD ($\hat{\beta}$)	0.592, 0.717	0.657, 0.983	0.647, 0.978	0.759, 0.84
	Ave SE ($\hat{\beta}$)	0.576, 0.687	0.652, 0.951	0.649, 0.912	0.784, 0.707
	RMSE($\hat{\beta}$)	0.592, 0.717	0.657, 0.983	0.647, 0.978	0.761, 0.84
	Emp. Coverage	0.943, 0.955	0.953, 0.957	0.959, 0.948	0.96, 0.923

*Partial likelihood approach for full data given all causes are known.

†Complete-case approach.

‡WCC approach with weights computed from sample proportion of failures with known cause.

§WCC approach with weights computed from a logistic regression model.

¶Missingness depends on the covariates in the cause-specific hazards.

‖Missingness depends on the failure time.

**Missingness depends on the cause type.

Table 11.2: Comparison of three complete-case approaches with different weighting methods, for sample size 200, censoring 30%, proportion of missingness 21% and repetitions 1000.

Methods		Full Data *	CC *	WCC1 *	WCC2 *
MCAR	True β	0.5, 0.2	0.5, 0.2	0.5, 0.2	0.5, 0.2
	Ave $\hat{\beta}$	0.509, 0.192	0.471, 0.158	0.501, 0.200	0.504, 0.202
	Emp. SD ($\hat{\beta}$)	0.411, 0.468	0.494, 0.561	0.487, 0.556	0.487, 0.558
	Ave SE ($\hat{\beta}$)	0.397, 0.478	0.480, 0.576	0.477, 0.556	0.478, 0.557
	RMSE ($\hat{\beta}$)	0.411, 0.468	0.495, 0.563	0.487, 0.556	0.487, 0.558
	Emp. Coverage	0.942, 0.964	0.946, 0.963	0.951, 0.961	0.951, 0.961
MAR1*	True β	0.5, 0.2	0.5, 0.2	0.5, 0.2	0.5, 0.2
	Ave $\hat{\beta}$	0.509, 0.192	0.448, 0.123	0.487, 0.177	0.508, 0.202
	Emp. SD ($\hat{\beta}$)	0.411, 0.468	0.497, 0.581	0.489, 0.576	0.489, 0.581
	Ave SE ($\hat{\beta}$)	0.397, 0.478	0.489, 0.587	0.486, 0.565	0.486, 0.567
	RMSE ($\hat{\beta}$)	0.411, 0.468	0.499, 0.586	0.489, 0.577	0.489, 0.581
	Emp. Coverage	0.942, 0.964	0.949, 0.963	0.956, 0.959	0.957, 0.959
MAR2*	True β	0.5, 0.2	0.5, 0.2	0.5, 0.2	0.5, 0.2
	Ave $\hat{\beta}$	0.509, 0.192	0.451, 0.125	0.483, 0.168	0.481, 0.163
	Emp. SD ($\hat{\beta}$)	0.411, 0.468	0.5, 0.545	0.491, 0.539	0.49, 0.541
	Ave SE ($\hat{\beta}$)	0.397, 0.478	0.488, 0.563	0.484, 0.543	0.485, 0.544
	RMSE ($\hat{\beta}$)	0.411, 0.468	0.502, 0.55	0.491, 0.539	0.491, 0.542
	Emp. Coverage	0.942, 0.964	0.943, 0.959	0.945, 0.961	0.948, 0.963
MNAR*	True β	0.5, 0.2	0.5, 0.2	0.5, 0.2	0.5, 0.2
	Ave $\hat{\beta}$	0.509, 0.192	0.474, 0.158	0.503, 0.199	0.521, 0.202
	Emp. SD ($\hat{\beta}$)	0.411, 0.468	0.464, 0.651	0.457, 0.647	0.524, 0.558
	Ave SE ($\hat{\beta}$)	0.397, 0.478	0.448, 0.655	0.444, 0.631	0.537, 0.49
	RMSE($\hat{\beta}$)	0.411, 0.468	0.464, 0.652	0.457, 0.646	0.525, 0.558
	Emp. Coverage	0.942, 0.964	0.94, 0.954	0.942, 0.95	0.956, 0.924

*See footnotes in Table 11.1 (page 116).

11.3 Comparison of Double Robust Approaches and Two Existing Approaches

Table 11.3 and 11.4 show the results from the double robust (DR) approaches with the two weighting methods used for the CC approach, the multiple imputation (MI) approach proposed by Lu and Tsiatis (2001) and the GR approach proposed by Goetghebeur and Ryan (1995).

In this simulation setting, the covariates of the logistic model for cause type used in the DR approaches includes the constant one, the observed failure time and the covariates in the cause-specific hazards. Strictly speaking, this actually is not the correct model. For further illustration, we first look at the following relationship between the probability of cause type and the cause-specific hazards

$$\frac{\tau_1(\boldsymbol{\gamma}; \boldsymbol{V})}{1 - \tau_1(\boldsymbol{\gamma}; \boldsymbol{V})} = \frac{\lambda_1(t \mid \boldsymbol{Z})}{\lambda_2(t \mid \boldsymbol{Z})},$$

which can be further rewritten as

$$\begin{aligned} \frac{\tau_1(\boldsymbol{\gamma}; \boldsymbol{V})}{1 - \tau_1(\boldsymbol{\gamma}; \boldsymbol{V})} &= \frac{\lambda_{10}(t) \exp(\boldsymbol{\beta}_1^T \boldsymbol{Z})}{\lambda_{20}(t) \exp(\boldsymbol{\beta}_2^T \boldsymbol{Z})} \\ &= \exp\left\{ \log \frac{\lambda_{10}(t)}{\lambda_{20}(t)} + (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2)^T \boldsymbol{Z} \right\}. \end{aligned} \tag{11.5}$$

Since the baseline hazards take arbitrary functional form, the first exponential term in (11.5) is not necessarily a linear function of time. Lu and Tsiatis (2001) showed the misspecified model did not cause substantial bias or loss of efficiency when the baseline hazards come from survival distributions such as Weibull, gamma, lognormal and log logistic.

Compared to the WCC approaches, the DR, MI and GR approaches not only reduce the bias, but also result in significant smaller estimation of standard error in the cases of MCAR, MAR1 and MAR2. Moreover, the estimation bias for MAR2 is also largely removed with those methods. But in the case of MNAR, the estimation bias exists for those methods: it becomes even more prominent when the sample size increases from 100 to 200. This supports the theoretical argument that the MAR assumption is required for DR, MI and GR approaches.

Generally speaking, those methods are quite comparable. But, in terms of RMSE, the two DR approaches perform slightly better than the others and the MI approach performs slightly better than the GR approach. Furthermore, of the two DR approaches, DR1 yields smaller standard error and thus smaller RMSE in the cases of MCAR, MAR2 and even MAR1. This fact seems consistent with the corresponding WCC approaches. Lastly, since the proportionality of the cause-specific baseline hazards does not hold in all cases, the results from the GR approach demonstrate the robustness of this method as mentioned in Goetghebeur and Ryan (1995, Section 5.1).

11.4 An Example

Eastern Cooperative Oncology Group (ECOG) Study E1178 compared a two-year tamoxifen therapy to placebo in breast cancer patients with age ≥ 65 and positive axillary nodes for the survival. This is a classical example of competing risks data, used by many researchers studying competing risks models, for example, Goetghebeur

Table 11.3: Comparison of the double robust approaches and the two existing approaches, for sample size 100, censoring 30%, proportion of missingness 21% and repetitions 1000.

Methods		Full Data *	GR †	MI ‡	DR1 §	DR2 ¶
MCAR	True β	0.5, 0.2	0.5, 0.2	0.5, 0.2	0.5, 0.2	0.5, 0.2
	Ave $\hat{\beta}$	0.523, 0.22	0.524, 0.209	0.532, 0.207	0.527, 0.216	0.527, 0.215
	Emp. SD ($\hat{\beta}$)	0.592, 0.717	0.649, 0.838	0.648, 0.830	0.639, 0.826	0.644, 0.828
	Ave SE ($\hat{\beta}$)	0.576, 0.687	0.623, 0.787	0.665, 0.828	0.615, 0.748	0.618, 0.752
	RMSE ($\hat{\beta}$)	0.592, 0.717	0.649, 0.838	0.648, 0.829	0.639, 0.825	0.644, 0.828
	Emp. Coverage	0.943, 0.955	0.946, 0.952	0.955, 0.944	0.942, 0.932	0.938, 0.933
MAR1 ‖	True β	0.5, 0.2	0.5, 0.2	0.5, 0.2	0.5, 0.2	0.5, 0.2
	Ave $\hat{\beta}$	0.523, 0.22	0.533, 0.191	0.527, 0.203	0.522, 0.215	0.523, 0.213
	Emp. SD ($\hat{\beta}$)	0.592, 0.717	0.688, 0.917	0.651, 0.851	0.643, 0.841	0.646, 0.844
	Ave SE ($\hat{\beta}$)	0.576, 0.687	0.641, 0.821	0.672, 0.841	0.624, 0.762	0.623, 0.761
	RMSE ($\hat{\beta}$)	0.592, 0.717	0.689, 0.916	0.651, 0.850	0.643, 0.841	0.647, 0.844
	Emp. Coverage	0.943, 0.955	0.939, 0.94	0.959, 0.940	0.942, 0.93	0.941, 0.93
MAR2 ‖	True β	0.5, 0.2	0.5, 0.2	0.5, 0.2	0.5, 0.2	0.5, 0.2
	Ave $\hat{\beta}$	0.523, 0.22	0.534, 0.214	0.537, 0.204	0.529, 0.215	0.53, 0.214
	Emp. SD ($\hat{\beta}$)	0.592, 0.717	0.662, 0.816	0.646, 0.826	0.636, 0.819	0.639, 0.82
	Ave SE ($\hat{\beta}$)	0.576, 0.687	0.635, 0.765	0.669, 0.811	0.616, 0.748	0.621, 0.755
	RMSE ($\hat{\beta}$)	0.592, 0.717	0.662, 0.815	0.647, 0.825	0.636, 0.819	0.639, 0.819
	Emp. Coverage	0.943, 0.955	0.946, 0.946	0.956, 0.946	0.95, 0.933	0.948, 0.934
MNAR ‖	True β	0.5, 0.2	0.5, 0.2	0.5, 0.2	0.5, 0.2	0.5, 0.2
	Ave $\hat{\beta}$	0.523, 0.22	0.489, 0.201	0.504, 0.194	0.503, 0.198	0.559, 0.24
	Emp. SD ($\hat{\beta}$)	0.592, 0.717	0.583, 0.951	0.593, 0.938	0.588, 0.927	0.817, 0.829
	Ave SE ($\hat{\beta}$)	0.576, 0.687	0.57, 0.906	0.57, 0.849	0.57, 0.836	0.777, 0.721
	RMSE($\hat{\beta}$)	0.592, 0.717	0.582, 0.951	0.593, 0.937	0.588, 0.927	0.819, 0.83
	Emp. Coverage	0.943, 0.955	0.945, 0.952	0.943, 0.941	0.947, 0.923	0.952, 0.915

*Ordinary partial likelihood approach for the full data given all causes are known.

†Goethebeur and Ryan's proportional baseline hazards approach.

‡Lu and Tsiatis's multiple imputation approach with 10 repeats.

§DR approach with selection probabilities estimated by the sample proportion.

¶DR approach with selection probabilities estimated using the logistic regression model.

‖See footnotes in Table 11.1 (page 116).

Table 11.4: Comparison of the double robust approaches and the two existing approaches, for sample size 200, censoring 30%, proportion of missingness 21% and repetitions 1000.

Methods		Full Data *	GR *	MI *	DR1 *	DR2 *
MCAR	True β	0.5, 0.2	0.5, 0.2	0.5, 0.2	0.5, 0.2	0.5, 0.2
	Ave $\hat{\beta}$	0.509, 0.192	0.503, 0.195	0.508, 0.190	0.505, 0.195	0.506, 0.194
	Emp. SD ($\hat{\beta}$)	0.411, 0.468	0.450, 0.528	0.446, 0.510	0.442, 0.506	0.443, 0.507
	Ave SE ($\hat{\beta}$)	0.397, 0.478	0.430, 0.546	0.440, 0.547	0.426, 0.526	0.427, 0.527
	RMSE ($\hat{\beta}$)	0.411, 0.468	0.450, 0.528	0.446, 0.510	0.442, 0.506	0.443, 0.507
	Emp. Coverage	0.942, 0.964	0.941, 0.959	0.936, 0.963	0.941, 0.958	0.942, 0.952
MAR1 *	True β	0.5, 0.2	0.5, 0.2	0.5, 0.2	0.5, 0.2	0.5, 0.2
	Ave $\hat{\beta}$	0.509, 0.192	0.516, 0.173	0.508, 0.185	0.504, 0.192	0.505, 0.190
	Emp. SD ($\hat{\beta}$)	0.411, 0.468	0.471, 0.575	0.445, 0.521	0.443, 0.519	0.444, 0.523
	Ave SE ($\hat{\beta}$)	0.397, 0.478	0.441, 0.571	0.444, 0.555	0.432, 0.537	0.430, 0.534
	RMSE ($\hat{\beta}$)	0.411, 0.468	0.471, 0.576	0.445, 0.521	0.443, 0.520	0.444, 0.523
	Emp. Coverage	0.942, 0.964	0.934, 0.951	0.944, 0.957	0.937, 0.956	0.937, 0.951
MAR2 *	True β	0.5, 0.2	0.5, 0.2	0.5, 0.2	0.5, 0.2	0.5, 0.2
	Ave $\hat{\beta}$	0.509, 0.192	0.511, 0.197	0.506, 0.192	0.506, 0.193	0.507, 0.192
	Emp. SD ($\hat{\beta}$)	0.411, 0.468	0.459, 0.516	0.446, 0.51	0.442, 0.508	0.443, 0.509
	Ave SE ($\hat{\beta}$)	0.397, 0.478	0.437, 0.532	0.44, 0.542	0.426, 0.528	0.428, 0.53
	RMSE ($\hat{\beta}$)	0.411, 0.468	0.459, 0.515	0.446, 0.51	0.442, 0.508	0.443, 0.509
	Emp. Coverage	0.942, 0.964	0.941, 0.963	0.94, 0.958	0.943, 0.963	0.943, 0.959
MNAR *	True β	0.5, 0.2	0.5, 0.2	0.5, 0.2	0.5, 0.2	0.5, 0.2
	Ave $\hat{\beta}$	0.509, 0.192	0.476, 0.166	0.49, 0.154	0.486, 0.163	0.552, 0.209
	Emp. SD ($\hat{\beta}$)	0.411, 0.468	0.411, 0.617	0.418, 0.595	0.416, 0.588	0.558, 0.533
	Ave SE ($\hat{\beta}$)	0.397, 0.478	0.393, 0.625	0.393, 0.583	0.395, 0.586	0.531, 0.504
	RMSE($\hat{\beta}$)	0.411, 0.468	0.411, 0.618	0.418, 0.596	0.416, 0.589	0.56, 0.533
	Emp. Coverage	0.942, 0.964	0.94, 0.962	0.938, 0.952	0.937, 0.95	0.934, 0.939

*See footnotes in Table 11.3 (page 120).

and Ryan (1995), Lu and Tsiatis (2001), Lu and Liang (2008) and others.

We use the data presented in Cummings et al. (1986), indicating two covariates presence of 4 or more positive axillary nodes and having an ER-negative primary, that were reported significantly associated with overall survival. The dataset is summarized as follows

- 169 patients were enrolled in this study,
- 79 patients had died by the end of the study (thus 90 were treated as censored),
- among the 79 patients who died, there were 18 patients had unknown cause of death and 61 patients had known cause of death,
- among the 61 patients with known cause of death, there were 44 of breast cancer and 17 of other causes.

The cause of death can be categorized as breast cancer and other diseases. We are more interested in breast cancer mortality. The CRM with missing cause is fitted to this dataset in order to ascertain how the two covariates were associated with death caused by breast cancer only. Therefore the cause-specific hazard model includes the two covariates. In addition to the two covariates, we assume that the missingness and the cause type may depend on the tumor size and treatment (tamoxifen) as well. Due to the fact that of the five patients who died and had ER-negative status, all died from breast cancer, the covariates in the logistic regression model for the cause type does not include the ER status.

Table 11.5 presents the results from different estimation approaches. It is observed that all CRM estimates of the coefficient of ER-negative status increase significantly, compared to the *overall* estimate without distinguishing failure types, while there is no trend for the coefficient of presence of 4 or more positive nodes. The larger estimate of ER-negative status indicates that ER-negative status is particularly a mortality marker for breast cancer death. Since the simulation study shows that DR1 approach has the best performance with respect to RMSE, the DR1 estimates is generally the most reliable to be reported.

Table 11.5: Comparison of different approaches for the breast cancer data from ECOG study E1178.

	Overall *	CC †	WCC1 †	WCC2 †	GR ‡	MI ‡	DR1 ‡	DR2 ‡
$\hat{\beta}_1$ ($\geq$ 4 nodes)	0.59	0.71	0.65	0.57	0.57	0.60	0.55	0.54
(SE($\hat{\beta}_1$))	(0.2294)	(0.3065)	(0.2779)	(0.2759)	(0.2803)	(0.2618)	(0.2695)	(0.2714)
$\hat{\beta}_2$ (ER-negative)	1.19	1.70	1.62	1.59	1.59	1.61	1.72	1.70
(SE($\hat{\beta}_2$))	(0.4688)	(0.4861)	(0.4727)	(0.4746)	(.4822)	(0.4794)	(0.3677)	(0.3811)

*The overall estimates were computed based upon the ordinary Cox proportional hazards model without distinguishing the types of mortality.

†See footnotes in Table 11.1 (page 116).

‡See footnotes in Table 11.3 (page 120).

CHAPTER 12
DISCUSSION AND FUTURE WORK

In summary, we have proposed two approaches to estimate the competing risks models with missing cause of failure: the weighted complete-case (WCC) approach and the double robust (DR) approach. Simulation studies show that the DR estimator is superior to the WCC estimator in terms of efficiency and robustness under the assumption of missing at random while only the WCC estimator is consistent in the case of *missing not at random*.

We have studied the theoretical properties for the WCC method in Chapter 9 using counting process and martingale theory. The asymptotic properties of the DR estimators are open for future investigation. Furthermore, the estimation of baseline hazard functions is not fully explored in the context of weighted approaches for CRM with missing cause of failure and remains an open question for future research.

Although the proposed DR approaches improve the estimation efficiency, it is not known if the estimation method is indeed semiparametric efficient. The maximum likelihood method is generally desired for this purpose. As motivated by the EM algorithm proposed in Dewanji and Sengupta (2003) for estimating the cumulative hazards without covariates in the context of CRM with missing cause of failure, a similar EM algorithm along with profile likelihood technique can be developed. But our experiment shows that the ordinary EM algorithm does not work well due to multiple local maxima in the profile likelihood. A modified EM algorithm is desired and remains another open problem for future investigation.

APPENDIX A
DERIVATION OF BFGS ITERATIVE FORMULA

As discussed in Section 4.5, in BFGS algorithm, the inverse Hessian matrix can be updated based upon the previous approximation B_{k-1} by solving the following optimization problem

$$\min_{B} trace\left(\left\{W^{1/2}(B - B_{k-1})W^{1/2}\right\}^{\otimes 2}\right) \text{ subject to}$$

$$B = B^T,$$

$$B\,\Delta s_k = \Delta\theta_k\,,$$

$$W\,\Delta\theta_k = \Delta s_k\,. \tag{A.1}$$

Let $\boldsymbol{\lambda}$ be a $n \times 1$ vector, and $\boldsymbol{\Lambda} = (\lambda_{ij})$ be a $n \times n$ vector. Let M be any matrix that satisfies (A.1). Without including the constraint (A.1) on the weight matrix, the Lagrange function for the above optimization problem can be written as

$$f\,(B) = trace\left(\left\{W^{1/2}(B - B_k)W^{1/2}\right\}^{\otimes 2}\right) - 2\boldsymbol{\lambda}^T(B\Delta s_k - \Delta\theta_k) - 2\sum_{i=1}^{n}\sum_{j=i+1}^{n}\lambda_{ij}(b_{ij} - b_{ji}).$$

Take the partial derivative on $f\,(B)$ with respect to B, and rewrite the problem as

$$WBW - WB_kW - \boldsymbol{\lambda}\Delta s_k^T - \boldsymbol{\Lambda} = 0 \text{ with} \tag{A.2}$$

$$\boldsymbol{\Lambda}^T = -\boldsymbol{\Lambda},$$

$$B = B^T, \tag{A.3}$$

$$B\Delta s_k = \Delta\theta_k, \tag{A.4}$$

$$W\,\Delta\theta_k = \Delta s_k \text{ or } W^{-1}\Delta s_k = \Delta\theta_k\,. \tag{A.5}$$

Next we show how to solve the above problem.

Multiplying (A.2) with $\Delta\theta_k$, we first get

$$WBW\Delta\theta_k - WB_kW\Delta\theta_k - \Delta s_k^T\Delta\theta_k\,\boldsymbol{\lambda} - \Lambda\Delta\theta_k = \mathbf{0}. \tag{A.6}$$

Then, using the conditions (A.4) and (A.5), the equation (A.6) can be solved to obtain

$$\boldsymbol{\lambda} = \rho_k(\Delta s_k - WB_k\Delta\theta_k - \boldsymbol{\Lambda}\Delta\theta_k), \tag{A.7}$$

where $\rho_k = \frac{1}{\Delta s_k^T\Delta\theta_k}$.

Next, substitute (A.7) into the equation (A.2)

$$WBW - WB_kW - \rho_k(\Delta s_k\Delta s_k^T - WB_k\Delta s_k\Delta s_k^T) = \boldsymbol{\Lambda}(I - \rho_k\Delta\theta_k\Delta s_k^T), \tag{A.8}$$

and left multiply both sides of the equation (A.8) with $(I - \rho_k\Delta s_k\Delta\theta_k^T)$

$$\begin{aligned} &WBW - WB_kW - \rho_k\Delta s_k\Delta s_k^T + \rho_kWB_k\Delta s_k\Delta s_k^T + \rho_k\Delta s_k\Delta s_k^TB_kW \\ &- \rho_k^2\Delta s_k\Delta s_k^TB_k\Delta s_k\Delta s_k^T = (I - \rho_k\Delta s_k\Delta\theta_k^T)\boldsymbol{\Lambda}(I - \rho_k\Delta\theta_k\Delta s_k^T), \end{aligned} \tag{A.9}$$

furthermore, left multiply and right multiply both sides of the equation (A.9) with W^{-1},

$$\begin{aligned} &B - B_k - \rho_k\Delta\theta_k\Delta\theta_k^T + \rho_kB_k\Delta s_k\Delta\theta_k^T + \rho_k\Delta\theta_k\Delta s_k^TB_k \\ &- \rho_k^2\Delta\theta_k\Delta s_k^TB_k\Delta s_k\Delta\theta_k^T = W^{-1}(I - \rho_k\Delta s_k\Delta\theta_k^T)\Lambda(I - \rho_k\Delta\theta_k\Delta s_k^T)W^{-1}. \end{aligned} \tag{A.10}$$

Since $\boldsymbol{\Lambda}^T = -\boldsymbol{\Lambda}$, the transpose of the left side of the equation (A.10) is its negative. So add (A.10) to its the transpose on both sides:

$$\begin{aligned} &2 \times (B - B_k - \rho_k\Delta\theta_k\Delta\theta_k^T + \rho_kB_k\Delta s_k\Delta\theta_k^T + \rho_k\Delta\theta_k\Delta s_k^TB_k \\ &- \rho_k^2\Delta\theta_k\Delta s_k^TB_k\Delta s_k\Delta\theta_k^T) = \mathbf{0}. \end{aligned}$$

Finally, we have

$$B = B_k + \rho_k \Delta\theta_k \Delta\theta_k^T - \rho_k B_k \Delta s_k \Delta\theta_k^T - \rho_k \Delta\theta_k \Delta s_k^T B_k + \rho_k^2 \Delta\theta_k \Delta s_k^T B_k \Delta s_k \Delta\theta_k^T$$

$$= \left\{ I - \rho_k \, \Delta\theta_k \, \Delta s_k^T \right\} \, B_{k-1} \, \left\{ I - \rho_k \, \Delta s_k \, \Delta\theta_k^T \right\} + \rho_k \, \Delta s_k \, \Delta s_k^T \, .$$

APPENDIX B
ONE-LEVEL LOGNORMAL FRAILTY MODEL

Estimation of a one-level lognormal frailty model requires a simplified approach compared to the proposed two-level model approach. Just for completeness, we generalize the results in this appendix.

B.1 Likelihood

The one-level frailty model can be similarly defined for the hazard function and the cumulative hazard function respectively, as shown in (3.5) and (3.4),

$$\lambda_{ij}(t_{ij}) = \lambda_0(t_{ij}) \exp\left(\boldsymbol{\beta}^T \boldsymbol{x}_{ij} + h_i\right),$$

$$\Lambda_{ij}(t_{ij}) = \Lambda_0(t_{ij}) \exp\left(\boldsymbol{\beta}^T \boldsymbol{x}_{ij} + h_i\right),$$

where, in the one-level setting, t_{ij} and $\boldsymbol{x}_{ij}$ are the observed failure time and covariates for the j^{th} patient of the i^{th} hospital.

With B-splines applied to model the (cumulative) hazard function, we have

$$\Lambda_{ij} \triangleq \Lambda_{ij}(t_{ij}) = \exp\left(\boldsymbol{c}^T \boldsymbol{b}_{ij} + \boldsymbol{\beta}^T \boldsymbol{x}_{ij} + h_i\right), \tag{B.1}$$

$$\lambda_{ij} \triangleq \lambda_{ij}(t_{ij}) = \boldsymbol{c}^T \dot{\boldsymbol{b}}_{ij} \exp\left(\boldsymbol{c}^T \boldsymbol{b}_{ij} + \boldsymbol{\beta}^T \boldsymbol{x}_{ij} + h_i\right). \tag{B.2}$$

Let f_{h_i} be the density function of h_i for $i = 1, \ldots, s$. The marginal likelihood can be written as

$$\begin{aligned} L(\boldsymbol{c}, \boldsymbol{\beta}, \theta) &= \int \prod_{i=1}^{s} \prod_{j=1}^{n_i} \left\{ \lambda_{ij}^{\delta_{ij}} e^{-\Lambda_{ij}} f_{h_i} \right\} dh_i \\ &= \int \prod_{i=1}^{s} \prod_{j=1}^{n_i} \left\{ \boldsymbol{c}^T \dot{\boldsymbol{b}}_{ij} \exp\left(\boldsymbol{c}^T \boldsymbol{b}_{ij} + \boldsymbol{\beta}^T \boldsymbol{x}_{ij} + h_i\right) \right\}^{\delta_{ij}} \end{aligned}$$

$$
\begin{aligned}
&\times \exp\left\{-\exp\left(\boldsymbol{c}^T\boldsymbol{b}_{ij}+\boldsymbol{\beta}^T\boldsymbol{x}_{ij}+h_i\right)\right\}\ f(h_i)\ dh_i \\
&= \prod_{i=1}^{s}\prod_{j=1}^{n_i}\left\{\boldsymbol{c}^T\dot{\boldsymbol{b}}_{ij}\exp\left(\boldsymbol{c}^T\boldsymbol{b}_{ij}+\boldsymbol{\beta}^T\boldsymbol{x}_{ij}\right)\right\}^{\delta_{ij}} \\
&\quad\times\prod_{i=1}^{s}\int\exp\left\{\delta_{i\cdot}h_i-\sum_{j=1}^{n_i}\exp\left(\boldsymbol{c}^T\boldsymbol{b}_{ij}+\boldsymbol{\beta}^T\boldsymbol{x}_{ij}+h_i\right)\right\}\ f_{h_i}\,dh_i\,.
\end{aligned}
$$

Then the log likelihood is

$$
\begin{aligned}
l(\boldsymbol{c},\boldsymbol{\beta},\theta) &= \sum_{i=1}^{s}\sum_{j=1}^{n_i}\delta_{ij}\left\{\log\left(\boldsymbol{c}^T\dot{\boldsymbol{b}}_{ij}\right)+\boldsymbol{c}^T\boldsymbol{b}_{ij}+\boldsymbol{\beta}^T\boldsymbol{x}_{ij}\right\} \\
&\quad+\sum_{i=1}^{s}\log\left\{\int\exp\left(\delta_{i\cdot}h_i-\Lambda_{i\cdot}\right)\ f_{h_i}\,dh_i\right\}.
\end{aligned}
$$

Considering the constraint $\theta \geq 0$, we reparametrize θ as e^{τ}, $\tau \in (-\infty, +\infty)$. Then the frailty density function is modified as follows

$$
f_{h_i}(x) = \frac{1}{\sqrt{2\pi}}\theta^{-\frac{1}{2}}\exp\left(-\frac{x^2}{2\theta}\right) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}\tau-\frac{x^2}{2}e^{-\tau}\right).
$$

Define

$$
g_i(\boldsymbol{c},\boldsymbol{\beta},\tau) = \int\exp\left(\delta_{i\cdot}h_i-\Lambda_{i\cdot}\right)\ f_{h_i}\,dh_i\,.
$$

Then the log likelihood can be rewritten as

$$
l(\boldsymbol{c},\boldsymbol{\beta},\tau) = \sum_{i=1}^{s}\sum_{j=1}^{n_i}\delta_{ij}\left\{\log\left(\boldsymbol{c}^T\dot{\boldsymbol{b}}_{ij}\right)+\boldsymbol{c}^T\boldsymbol{b}_{ij}+\boldsymbol{\beta}^T\boldsymbol{x}_{ij}\right\}+\sum_{i=1}^{s}\log\left\{g_i\left(\boldsymbol{c},\boldsymbol{\beta},\tau\right)\right\}. \quad \text{(B.3)}
$$

The first and second derivatives of g_i's with respect to $\boldsymbol{c}, \boldsymbol{\beta}$ and τ are computed as follows

$$
\begin{aligned}
\frac{\partial g_i}{\partial \boldsymbol{c}} &= -\int\exp\left(\delta_{i\cdot}h_i-\Lambda_{i\cdot}\right)\times\sum_{j=1}^{n_i}\Lambda_{ij}\boldsymbol{b}_{ij}\times f_{h_i}\ dh_i\,, \\
\frac{\partial g_i}{\partial \boldsymbol{\beta}} &= -\int\exp\left(\delta_{i\cdot}h_i-\Lambda_{i\cdot}\right)\times\sum_{j=1}^{n_i}\Lambda_{ij}\boldsymbol{x}_{ij}\times f_{h_i}\ dh_i\,,
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial g_i}{\partial \tau} &= \int \exp\left(\delta_{i\cdot} h_i - \Lambda_{i\cdot}\right) \times \left(-\frac{1}{2} + \frac{h_i^2}{2} e^{-\tau}\right) \times f_{h_i}\, dh_i\,, \\
\frac{\partial^2 g_i}{\partial \boldsymbol{c} \partial \boldsymbol{c}^T} &= \int \exp\left(\delta_{i\cdot} h_i - \Lambda_{i\cdot}\right) \times \left\{ \left(\sum_{j=1}^{n_i} \Lambda_{ij} \boldsymbol{b}_{ij}\right)^{\otimes 2} - \sum_{j=1}^{n_i} \Lambda_{ij} \boldsymbol{b}_{ij}^{\otimes 2} \right\} \times f_{h_i}\, dh_i\,, \\
\frac{\partial^2 g_i}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} &= \int \exp\left(\delta_{i\cdot} h_i - \Lambda_{i\cdot}\right) \times \left\{ \left(\sum_{j=1}^{n_i} \Lambda_{ij} \boldsymbol{x}_{ij}\right)^{\otimes 2} - \sum_{j=1}^{n_i} \Lambda_{ij} \boldsymbol{x}_{ij}^{\otimes 2} \right\} \times f_{h_i}\, dh_i\,, \\
\frac{\partial^2 g_i}{\partial \boldsymbol{c} \partial \boldsymbol{\beta}^T} &= \int \exp\left(\delta_{i\cdot} h_i - \Lambda_{i\cdot}\right) \times \left\{ \left(\sum_{j=1}^{n_i} \Lambda_{ij} \boldsymbol{b}_{ij}\right) \left(\sum_{j=1}^{n_i} \Lambda_{ij} \boldsymbol{x}_{ij}^T\right) - \sum_{j=1}^{n_i} \Lambda_{ij} \boldsymbol{b}_{ij} \boldsymbol{x}_{ij}^T \right\} \\
&\quad \times f_{h_i}\, dh_i\,, \\
\frac{\partial^2 g_i}{\partial \tau^2} &= \int \exp\left(\delta_{i\cdot} h_i - \Lambda_{i\cdot}\right) \times \left(\frac{1}{4} - h_i^2 e^{-\tau} + \frac{1}{4} h_i^4 e^{-2\tau}\right) \times f_{h_i}\, dh_i\,, \\
\frac{\partial^2 g_i}{\partial \boldsymbol{c} \partial \tau} &= \int \exp\left(\delta_{i\cdot} h_i - \Lambda_{i\cdot}\right) \times \left(\sum_{j=1}^{n_i} \Lambda_{ij} \boldsymbol{b}_{ij}\right) \times \left(-\frac{1}{2} + \frac{h_i^2}{2} e^{-\tau}\right) \times f_{h_i}\, dh_i\,, \\
\frac{\partial^2 g_i}{\partial \boldsymbol{\beta} \partial \tau} &= \int \exp\left(\delta_{i\cdot} h_i - \Lambda_{i\cdot}\right) \times \left(\sum_{j=1}^{n_i} \Lambda_{ij} \boldsymbol{x}_{ij}\right) \times \left(-\frac{1}{2} + \frac{h_i^2}{2} e^{-\tau}\right) \times f_{h_i}\, dh_i\,.
\end{aligned}
$$

Then we can write the fomulas for the score function and the Hessian matrix conveniently

$$
\begin{aligned}
\frac{\partial l}{\partial \boldsymbol{c}} &= \sum_{i=1}^{s} \left\{ \sum_{j=1}^{n_i} \delta_{ij} \left(\frac{\dot{\boldsymbol{b}}_{ij}}{\boldsymbol{c}^T \dot{\boldsymbol{b}}_{ij}} + \boldsymbol{b}_{ij} \right) + \frac{1}{g_i} \frac{\partial g_i}{\partial \boldsymbol{c}} \right\}\,, \\
\frac{\partial l}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{s} \left(\sum_{j=1}^{n_i} \delta_{ij} \boldsymbol{x}_{ij} + \frac{1}{g_i} \frac{\partial g_i}{\partial \boldsymbol{\beta}} \right)\,, \\
\frac{\partial l}{\partial \tau} &= \sum_{i=1}^{s} \left(\frac{1}{g_i} \frac{\partial g_i}{\partial \tau} \right)\,, \\
\frac{\partial^2 l}{\partial \boldsymbol{c} \partial \boldsymbol{c}^T} &= \sum_{i=1}^{s} \left[\sum_{j=1}^{n_i} \delta_{ij} \left\{ -\frac{\dot{\boldsymbol{b}}_{ij}^{\otimes 2}}{\left(\boldsymbol{c}^T \dot{\boldsymbol{b}}_{ij}\right)^2} \right\} - \frac{1}{g_i^2} \frac{\partial g_i}{\partial \boldsymbol{c}} \frac{\partial g_i}{\partial \boldsymbol{c}^T} + \frac{1}{g_i} \frac{\partial^2 g_i}{\partial \boldsymbol{c} \partial \boldsymbol{c}^T} \right]\,, \\
\frac{\partial^2 l}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} &= \sum_{i=1}^{s} \left\{ -\frac{1}{g_i^2} \left(\frac{\partial g_i}{\partial \boldsymbol{\beta}} \right)^{\otimes 2} + \frac{1}{g_i} \frac{\partial^2 g_i}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right\}\,, \\
\frac{\partial^2 l}{\partial \tau^2} &= \sum_{i=1}^{s} \left\{ -\frac{1}{g_i^2} \left(\frac{\partial g_i}{\partial \tau} \right)^2 + \frac{1}{g_i} \frac{\partial^2 g_i}{\partial \tau^2} \right\}\,,
\end{aligned}
$$

$$\frac{\partial^2 l}{\partial \boldsymbol{c}\partial \boldsymbol{\beta}^T} = \sum_{i=1}^{s}\left(-\frac{1}{g_i^2}\frac{\partial g_i}{\partial \boldsymbol{c}}\frac{\partial g_i}{\partial \boldsymbol{\beta}^T} + \frac{1}{g_i}\frac{\partial^2 g_i}{\partial \boldsymbol{c}\partial \boldsymbol{\beta}^T}\right),$$
$$\frac{\partial^2 l}{\partial \boldsymbol{c}\partial \tau} = \sum_{i=1}^{s}\left(-\frac{1}{g_i^2}\frac{\partial g_i}{\partial \tau}\frac{\partial g_i}{\partial \boldsymbol{c}} + \frac{1}{g_i}\frac{\partial^2 g_i}{\partial \boldsymbol{c}\partial \tau}\right),$$
$$\frac{\partial^2 l}{\partial \beta\partial \tau} = \sum_{i=1}^{s}\left(-\frac{1}{g_i^2}\frac{\partial g_i}{\partial \tau}\frac{\partial g_i}{\partial \boldsymbol{\beta}} + \frac{1}{g_i}\frac{\partial^2 g_i}{\partial \boldsymbol{\beta}\partial \tau}\right).$$

B.2 Adaptive Gauss-Hermite Quadrature

The idea of AGHQ for one-level case is the same as discussed in Section 4.4. Here we rewrite the formulas to compute the conditional mode of frailties h_i ($i = 1, ..., s$) and approximate functions g_i ($i = 1, ..., s$). The approximation of the partial derivatives of g_i's and thus the score functions and the Hessian matrix can then be similarly performed.

The log conditional likelihood for $\boldsymbol{h}$ given the current estimates of $\boldsymbol{c}$ and $\boldsymbol{\beta}$ can be written as

$$\begin{aligned} l(\boldsymbol{h}; \boldsymbol{c}, \boldsymbol{\beta}, \tau) &= \log f_{\boldsymbol{c},\boldsymbol{\beta}}(\boldsymbol{c}, \boldsymbol{\beta} \mid \boldsymbol{h}) + \log f_{\boldsymbol{h}}(\boldsymbol{h}) + \text{constant} \\ &= \sum_{i=1}^{s}\sum_{j=1}^{n_i}\left[\delta_{ij}\left\{\log(\boldsymbol{c}^T\dot{\boldsymbol{b}}_{ij}) + \boldsymbol{c}^T\boldsymbol{b}_{ij} + \boldsymbol{\beta}^T\boldsymbol{x}_{ij} + h_i\right\} - \Lambda_{ij}\right] \\ &\quad + \sum_{i=1}^{s}\left(-\frac{1}{2}e^{-\tau}h_i^2\right) + \text{constant} \\ &= \sum_{i=1}^{s}\left(\delta_{i\cdot}h_i - \Lambda_{i\cdot} - \frac{1}{2}e^{-\tau}h_i^2\right) + \text{constant}. \end{aligned}$$

Taking the first and second derivatives of the log likelihood with respect to $\boldsymbol{h}$

$$\frac{\partial l}{\partial h_i} = \delta_{i\cdot} - \Lambda_{i\cdot} - e^{-\tau}h_i, i = 1, \ldots, s,$$
$$\frac{\partial^2 l}{\partial h_i^2} = -\Lambda_{i\cdot} - e^{-\tau}, i = 1, \ldots, s,$$

$$\frac{\partial^2 l}{\partial h_{i_1} \partial h_{i_2}} = 0\,, \text{ for } i_1 \neq i_2,$$

we can compute the MLE (i.e. conditional mode) $\hat{h}_i$ and its standard error $\hat{\sigma}_{h_i}$, $i = 1, \ldots, s$.

Let

$$\tilde{h}_i = (h_i - \hat{h}_i) \,/\, (\sqrt{2}\hat{\sigma}_{h_i})\,,\ i = 1, \ldots, s,$$

then

$$h_i = \hat{h}_i + \sqrt{2}\hat{\sigma}_{h_i}\,\tilde{h}_i,\ i = 1, \ldots, s.$$

The function $g_i(\boldsymbol{c}, \boldsymbol{\beta}, \tau)$ can be approximated using the AGHQ algorithm for $i = 1, \ldots, s$:

$$\begin{aligned}
&g_i(\boldsymbol{c}, \boldsymbol{\beta}, \tau) \\
&= \int \exp\left(\delta_{i\cdot} h_i - \sum_{j=1}^{n_i} \exp\left(\boldsymbol{c}^T \boldsymbol{b}_{ij} + \boldsymbol{\beta}^T \boldsymbol{x}_{ij} + h_i\right)\right) f_{h_i}\, dh_i \\
&= \sqrt{2}\hat{\sigma}_{h_i} \int \exp\left\{\delta_{i\cdot}\left(\hat{h}_i + \sqrt{2}\hat{\sigma}_{h_i}\,\tilde{h}_i\right) - \sum_{j=1}^{n_i} \exp\left(\boldsymbol{c}^T \boldsymbol{b}_{ij} + \boldsymbol{\beta}^T \boldsymbol{x}_{ij} + \hat{h}_i + \sqrt{2}\hat{\sigma}_{h_i}\,\tilde{h}_i\right)\right\} \\
&\quad \times\ f_{h_i}(\hat{h}_i + \sqrt{2}\hat{\sigma}_{h_i}\tilde{h}_i)\, d\tilde{h}_i \\
&\approx \sqrt{2}\hat{\sigma}_{h_i} \sum_{l=1}^{d} \exp\left(\delta_{i\cdot}\left(\hat{h}_i + \sqrt{2}\hat{\sigma}_{h_i}\, a_l\right) - \sum_{j=1}^{n_i} \exp\left(\boldsymbol{c}^T \boldsymbol{b}_{ij} + \boldsymbol{\beta}^T \boldsymbol{x}_{ij} + \hat{h}_i + \sqrt{2}\hat{\sigma}_{h_i}\, a_l\right)\right) \\
&\quad \times\ f_{h_i}(\hat{h}_i + \sqrt{2}\hat{\sigma}_{h_i} a_l) \times e^{a_l^2} \times w_l,
\end{aligned}$$

where $\boldsymbol{a} = (a_1, \ldots, a_d)^T$ is the vector of abscissas needed for GHQ, $\boldsymbol{w} = (w_1, \ldots, w_d)^T$ is the vector of corresponding weights, and d is the order of GHQ.

REFERENCES

Andersen, P. K., Ekstrom, C. T., Klein, J. P., Shu, Y. and Zhang, M. J. (2005). A class of goodness of fit tests for a copula based on bivariate right-censored data. *Biometrical Journal*, **47** 815–24.

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

Anderson, J. A. and Senthilselvan, A. (1980). Smooth estimates for the hazard function. *Journal of the Royal Statistical Society, Series B*, **42** 322–327.

Barlow, R. E. (1972). *Statistical Inference Under Order Restrictions*. John Wiley and Sons Ltd.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, **65** 141–151.

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

Cong, X. J., Yin, G. and Shen, Y. (2007). Marginal analysis of correlated failure time data with informative cluster sizes. *Biometrics*, **63** 663–672.

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

Cummings, F. J., Gray, R., Davis, T. E., Tormey, D. C., Harris, J. E., Falkson, G. G. and Arseneau, J. (1986). Tamoxifen versus placebo: double-blind adjuvant trial in elderly women with stage II breast cancer. *NCI Monographs* 119–123.

Dewanji, A. (1992). A note on a test for competing risks with missing failure type. *Biometrika*, **79** 855–857.

Dewanji, A. and Sengupta, D. (2003). Estimation of competing risks with general missing pattern in failure types. *Biometrics*, **59** 1063–1070.

Duchateau, L. and Janssen, P. (2008). *The Frailty Model*. Springer.

Eeden, C. (1958). *Testing and estimating ordered parameters of probability distributions*. Ph.D. thesis, University of Amsterdam.

Elbers, C. and Ridder, G. (1982). True and spurious duration dependence: The identifiability of the proportional hazard model. *The Review of Economic Studies*, **49** 403–409.

Fine, P. J. and Gray, R. J. (1999). A proportional hazards model for the subdistribution of a competing risk. *Journal of the American Statistical Association*, **94** 496–509.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. John Wiley & Sons, Ltd.

Forsgren, A., Gill, P. E. and Wright, M. H. (2003). Interior methods for nonlinear optimization. *SIAM Review*, **44** 525–597.

Gichangi, A. and Vach, W. (2005). *The analysis of competing risks data: A guided tour*. Unpublished manuscript, Odense.

Glidden, D. V. and Vittinghoff, E. (2004). Modelling clustered survival data from multicentre clinical trials. *Statistics in Medicine*, **23** 369–388.

Goetghebeur, E. and Ryan, L. (1990). A modified log rank test for competing risks with missing failure type. *Biometrika*, **77** 207–211.

Goetghebeur, E. and Ryan, L. (1995). Analysis of competing risks survival data when some failure types are missing. *Biometrika*, **82** 821–833.

Gray, R. J. (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association*, **87** 942–951.

Ha, I. D. and Lee, Y. (2003). Estimating frailty models via Poisson hierarchical generalized linear models. *Journal of Computational and Graphical Statistics*, **12** 663–681.

Ha, I. D., Lee, Y. and Song, J.-K. (2001). Hierarchical likelihood approach for frailty models. *Biometrika*, **88** 233–243.

Hougaard, P. (1986a). Survival models for heterogeneous populations derived from stable distributions. *Biometrika*, **73** 387–396.

Hougaard, P. (1986b). A class of multivariate failure time distributions. *Biometrika*, **73** 671–678.

Huang, Y. and Chen, Y. Q. (2003). Marginal regression of gaps between recurrent events. *Lifetime Data Analysis*, **9** 293–303.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. 2nd ed. Wiley-Interscience.

Kooperberg, C., Stone, C. J. and Truong, Y. K. (1995). Hazard regression. *Journal of the American Statistical Association*, **90** 78–94.

Lee, E. W., Wei, L. J. and Amato, D. A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. *Survival Analysis: State of the Art, Dordrecht: Kluwer Academic* 237–247.

Lee, Y. and Nelder, J. A. (2001). Hierarchical generalised linear models: A synthesis of generalised linear models, random-effect models and structured dispersions. *Biometrika*, **88** pp. 987–1006.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. 2nd ed. Springer.

Lenglart, E. (1977). Relation de domination entre deux processus. *Ann. Inst. H. Poincaré Sect. B*, **13** 171–179.

Liang, K. Y., Self, S. G. and Chang, Y. C. (1993). Modeling marginal hazards in multivariate failure time data. *Journal of the Royal Statistical Society, Series B (Methodological)*, **55** 441–453.

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

Liu, Q. and Pierce, D. (1994). A note on Gauss-Hermite quadrature. *Biometrika*, **81** 624–629.

Lu, K. and Tsiatis, A. (2005). Comparison between two partial likelihood approaches for the competing risks model with missing cause of failure. *Lifetime Data Analysis*, **11** 29–40.

Lu, K. and Tsiatis, A. A. (2001). Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. *Biometrics*, **57** 1191–1197.

Lu, W. and Liang, Y. (2008). Analysis of competing risks data with missing cause of failure under additive hazards model. *Statistica Sinica* 219–234.

McGilchrist, C. A. (1993). REML estimation for survival models with frailty. *Biometrics*, **49** pp. 221–225.

McGilchrist, C. A. and Aisbett, C. W. (1991). Regression with frailty in survival analysis. *Biometrics*, **47** 461–466.

Nelsen, R. B. (1997). Dependence and order in families of Archimedean copulas. *Journal of Multivariate Analysis*, **60** 111–122.

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer.

Pan, Q. and Schaubel, D. (2009). Evaluating bias correction in weighted proportional hazards regression. *Lifetime Data Analysis*, **15** 120–146.

Phelps, A. L. and Weissfeld, L. A. (1997). A comparison of dependence estimators in bivariate copula models. *Communications in Statistics — Simulation and Computation*, **26** 1583–1597.

Pinheiro, J. C. and Bates, D. M. (1995). Approximations to the log-likelihood function in the nonlinear mixed-effects model. *Journal of Computational and Graphical Statistics*, **4** 12–35.

Pinheiro, J. C. and Chao, E. C. (2006). Efficient Laplacian and adaptive Gaussian quadrature algorithms for multilevel generalized linear mixed models. *Journal of Computational and Graphical Statistics*, **15** 58–81.

Pipper, C. B. and Martinussen, T. (2003). A likelihood based estimating equation for the Clayton-Oakes model with marginal proportional hazards. *Scandinavian Journal of Statistics*, **30** 509–521.

Ramsay, J. O. (1988). Monotone regression splines in action. *Statistical Science*, **3** 425–461.

Robertson, T., Wright, F. T. and Dykstra, R. (1988). *Order Restricted Statistical Inference*. Wiley.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, **89** 846–866.

Rondeau, V., Commenges, D. and Joly, P. (2003). Maximum penalized likelihood estimation in a Gamma-frailty model. *Lifetime Data Analysis*, **9** 139–153.

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

Roy, D. and Mukherjee, S. P. (1998). Multivariate extensions of univariate life distributions. *Journal of Multivariate Analysis*, **67** 72–79.

Schumaker, L. (2007). *Spline Functions: Basic Theory*. Cambridge University Press.

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

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, **14**.

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

Wienke, A. (2009). *Frailty Models In Survival Analysis*. Chapman & Hall/CRC, Taylor & Francis Group.

Xue, X. and Brookmeyer, R. (1996). Bivariate frailty model for the analysis of multivariate survival time. *Lifetime Data Analysis*, **2** 277–290.

Zhou, S., Shen, X. and Wolfe, D. A. (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics*, **26** 1760–1782.