Bayesian Semi-parametric Modeling of Time-to-Event Data

by

Yuanting Chen

B.S., University of Science and Technology of China, 2007

M.S., University of Colorado at Boulder, 2010

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Applied Mathematics

2014

This thesis entitled:
Bayesian Semi-parametric Modeling of Time-to-Event Data
written by Yuanting Chen
has been approved for the Department of Applied Mathematics

Prof. Vanja M. Dukic (chair)

Prof. David M. Bortz

Prof. Jem N. Corcoran

Prof. James J. Dignam

Prof. William Kleiber

Date

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Chen, Yuanting (Ph.D., Applied Mathematics)

Bayesian Semi-parametric Modeling of Time-to-Event Data

Thesis directed by Prof. Vanja M. Dukic (chair)

The multiresolution estimator, originally a wavelet-based method for density estimation, was recently extended for estimation of hazard functions. The multiresolution hazard (MRH) method's main advantage is its multiscale property, making simultaneous modeling and inference at multiple time scales possible. Additional advantages, stemming from its Bayesian foundation, are its simple computational implementation, estimation and inference procedures, and ability to easily quantify the uncertainty in hazard function estimates (via point-wise or curve-wise credible bands) adjusted for uncertainty in other model parameters, such as covariate effects. In this dissertation, we further extend the MRH methodology to accommodate the case of varying smoothness in the hazard function over time. The proposed pruned multiresolution hazard (PMRH) performs data-driven "fusing" of adjacent hazard intervals, increasing computational efficiency and reducing uncertainty in hazard rate estimation over regions with low event counts. We apply the PMRH method to examine patterns of failure after treatment for prostate cancer, using data from a large-scale randomized clinical trial.

Additionally, one of the main goals of survival analysis centers around how predictors affect the hazard function, and today, more and more datasets have time-varying predictors and biomarkers, which are functions of time. We extend the MRH methodology to handle time-varying covariates. We study several missingness scenarios, and conclude that when there is no missing data our MRH models perform well and efficiently with time-varying covariates as well. When the amount of missing time-varying covariates increases, our results show how increasing L^2 norm of the predictor function minus its mean within an interval is related to the bias and variance in the MRH model parameter estimators.

Dedication

To my Mom and Dad.

Acknowledgements

First, I would like to express my deepest gratitude to my advisor, Prof. Vanja Dukic, for her continuous support and mentoring during my Ph.D study and research. Her immense knowledge and enthusiasm not only introduced me to the world of Bayesian statistics, but also encouraged me to explore my own solutions in research while helping me out when I got stuck. Without her patience, kindness and guidance, I would not have been able to complete this dissertation.

I would also like to thank my committee members, Prof. David M. Bortz, Prof. Jem N. Corcoran, Prof. James J. Dignam and Prof. William Kleiber for their time, encouragement, insightful comments and helpful advice. Additionally, I want to thank Prof. Xiao-Chuan Cai, for teaching me numerical and parallel methods for PDEs, sharing his knowledge and experience in scientific computing, and for his generosity and kindness to me all these years. My research experience in his numerical analysis lab helped immensely during my dissertation work. Moreover, I would like to thank Dr. Yolanda Hagar, who has always shared her experience and knowledge in research with me, for all the helpful discussions, patient explanations of the algorithms, and suggestions to my code.

In addition, I am indebted to many of my friends and colleagues for supporting and taking care of me both in research and life. A non-exhaustive list includes Si Liu, Quan Yuan, Xiao Liu, Yuqi Wu, Zhaochuan Shen, Jason Hammond and Amrik Sen. It is the kindness and help from these wonderful people that has encouraged me finish this journey.

Finally, I would like to thank my beloved parents, for their endless love and support through all of these years.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Introduction

The multiresolution estimator proposed by Kolaczyk (1999) was originally developed for intensity function estimation in astrophysics, where structure may be visible at multiple scales. This relatively new area of statistical multiscale modeling has recently been introduced into survival analysis (Bouman et al. 2005, 2007; Dukic and Dignam 2007; Dignam et al. 2009), where flexible multiresolution modeling of hazard functions can reveal intricate patterns in patient failure risk over time. Questions of interest in those applications may include assessment of hazard structures over multiple time scales, and implied optimal follow-up monitoring schedules during the post-treatment surveillance period. Associated with these questions is the question of whether a single degree of smoothness is enough to adequately describe hazard behavior over long period of time (Dukic and Dignam 2007; Dignam et al. 2009).

The multiresolution hazard (MRH) model partitions the cumulative hazard function in a tree-like manner, over multiple time scales, via the framework of Bouman et al. (2007) and Dukic and Dignam (2007). The Bayesian MRH framework allows specification of any *a priori* desired shape and amount of smoothness in the hazard function. The multiresolution method's main advantages are its self-consistency across multiple scales (Bouman et al. 2007), simple implementation, estimation and inference procedures – namely its ability to provide uncertainty estimates on hazard functions via point-wise or curve-wise credible bands at multiple scales simultaneously. In addition, effects of factors influencing the hazard (covariates), and the hazard function itself are estimated jointly, resulting in adjusted inference about both of these quantities.

In some survival analysis settings, such as clinical trials investigating failure outcomes under different disease treatments in humans, however, we are faced with the challenge of having periods of intense failure activity combined with longer periods of slower failure activity as more time passes after the initial treatment. In this thesis, we thus further extend the MRH methodology to accommodate the case of varying smoothness in the hazard rate functions over time. (Note that "smoothness" here is taken in the sense of Bouman et al. (2005), and pertains to total variation of the hazard function rather than its differentiability properties.) We propose the "pruned" multiresolution hazard (PMRH), which performs data-driven "fusing" of adjacent hazard intervals, increasing computational efficiency and reducing uncertainty in hazard rate estimation over regions with low event counts.

1.1 Survival analysis and hazard function

Survival analysis is analysis over time-to-event data. In survival analysis, we observe subjects of certain time span and study the time when the event of our interest happens to them. Survival analysis is widely used in many fields. The events studied can be any kind, as long as they are of our interest. In biomedical aspect, events include death, cancer recurrence, heart attack and so on. Social scientists show more interest in graduation from school, first employment, marriage divorce and others. Reliability analysis or failure time analysis are more commonly used in engineering, instead of survival analysis. The timing of an assemble line stops working or the lifetime of a bulb before it burns out is more attractive to engineers. In study of those time-to-event data, we in hope to answer questions as what is the survival rate of a population after a certain time when a highly contiguous disease breaks out, how environmental and individual characteristics can affect one's survival probability.

Censoring happens when the time of event is not observed. Drop out from the study when it is still ongoing or the study time span is not long enough, thereby certain subject experiences no event before the termination of the study are two possible causes for censoring. Generally, there are three types of censoring, left censored, interval censored and right censored. A data point is left

censored if it is only known to be smaller than a certain value. A data point is interval censored if only a range of its value is known. A data point is right censored if its value is above a certain value. In survival analysis, many data are right censored. Also, they are non-informative censoring and informative censoring. When a subject drops out from the study due to reasons not statistically related to the study, such as job rotation to a different state or accidental death, we name it non-informative or random censoring. In a study of graduation from college, it is conceivable that students with not so good performance would have a higher risk to quit than that of students who do well in their classes. And this is informative censoring. Correctly handling censoring in data management will ensure more unbias estimates in analysis.

Survival function $S(t) = Pr(T > t)$ is the probability the time T of death(failure) after a certain time t. And as defined, $S(t)$ is apparently non-increasing and $S(0) = 1$. Moreover, as time t goes to infinity, the survival probability $S(t)$ will decline to zero. As contrast to survival function, we introduce lifetime distribution function $F(t)$, which is the probability the time T of death(failure) before or at a certain time t. Therefore $F(t) = Pr(T \leqslant t) = 1 - S(t)$. Hazard function or hazard rate $\lambda(t)$ is the momentary failure rate conditional on survival so far

$$\lambda(t)\ dt = Pr(t \leqslant T < t + dt | T \geqslant t) = \frac{-S'(t)dt}{S(t)}$$

In consideration of aggregation of the hazard over time, we define $\Lambda(t)$, cumulative hazard function as

$$\Lambda(t) = \int_0^t \lambda(u)\ du$$

Some simple calculation reveals that cumulative hazard function and survival function are related through equation

$$\Lambda(t) = -\ln(S(t))$$

In practice, we sometimes study the cumulative hazard function $\Lambda(t)$ directly, since by transform of

$$F(t) = 1 - e^{-\Lambda(t)} = 1 - e^{-\int_0^t \lambda(u)\ du}$$

we can get information about $F(t)$ and $S(t)$ easily. We define hazard increment from time a to time b as:

$$\Lambda(b) - \Lambda(a) = \int_a^b \lambda(u)\ du$$

In most case, we can't get $\lambda(t)$ at all t's. One way we can take is to estimate hazard increments over a partition of the study window, then by assuming constant hazard rate over each time interval respectively, in order to approximate $\Lambda(t)$. Figure 1.1 illustrates an instance how hazard increments can be employed to estimate survival function.

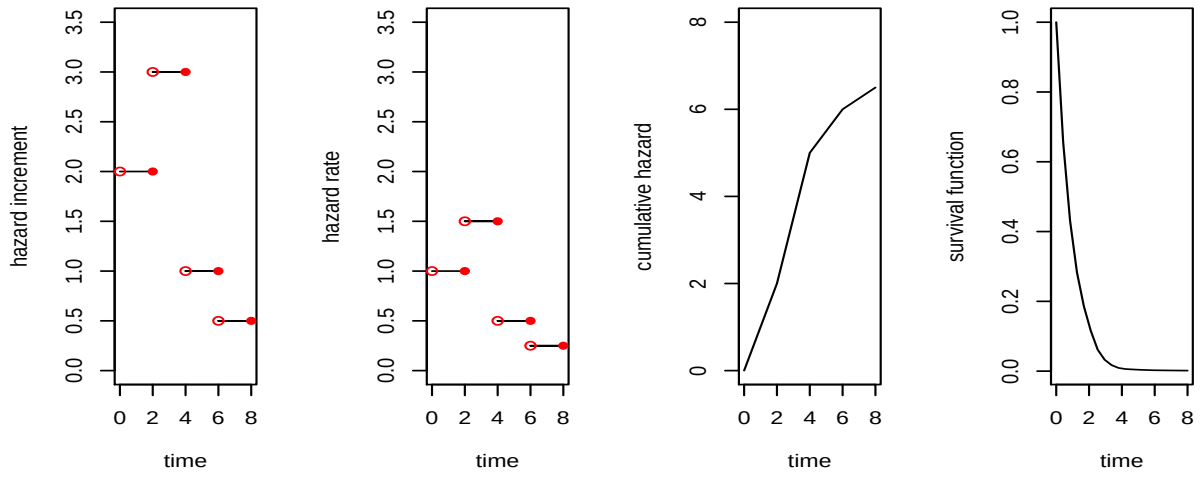

Figure 1.1: Diagram of using hazard increments to estimate survival function

1.2 Hazard estimation

From now on, we will use $h(t)$ to refer hazard function and $H(t)$ as cumulative hazard function. Although in last section we discussed about we may estimate cumulative hazard function instead of hazard function under some situation, the main appeal of hazard functions is that it can visualize details in failure risk patterns not apparent in aggregate summaries such as $S(t)$ or $H(t) = -\ln(S(t))$, the cumulative hazard function, and identify periods of elevated failure risk in a population (Aalen and Gjessing 2001).

There is parametric regression model which is based on exponential distribution.

$$\log(h_i(t)) = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_k X_{i,k} \tag{1.1}$$

In equation (1.1) subjects are indexed with i and $X_{i,j}$ are explanatory covariates of subject i, where $j = 1, \ldots, k$. Once the coefficients β_j can be decided, the hazard function $h_i(t)$ is completely determined. Parametric model can also be built via other distributions, such as Gompertz and Weibull distributions, since these distributions are commonly used in modeling survival analysis data. When we take all the covariates as zero, we would have $h_i(t) = e^{\beta_0} \equiv h_0(t)$, which can be somewhat taken as baseline hazard function and it is a constant. But in most cases, hazard function is always unstable and has a mutable pattern which is not easy to differentiate. Many parametric models have the unimodal assumption, as opposed to possible nonunimodality of hazard function, so we need models allow more flexibility.

Cox (1972) introduces the well-known semiparametric Cox proportional hazard model. Baseline hazard function now can have a flexible shape in any form, since we allow β_0 to vary over time t as a function $\beta_0(t)$ instead of fixing it as we did in equation (1.1)

$$\log(h_i(t)) = \beta_0(t) + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_k X_{i,k} \tag{1.2}$$

In other form,

$$h_i(t) = h_0(t) e^{\beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_k X_{i,k}} \tag{1.3}$$

Cox model estimates the baseline hazard function together with individual covariates effects. By assuming covariates $X_{i,j}$ are not time dependent, Cox model holds its property that any two observations' hazard function ratio is constant. When proportionality is not satisfied, we may consider multiple strata with separate baseline hazard function but all strata share the same covariate effects, example including Dukic and Dignam (2007) model randomized clinical trial data of early stage breast cancer recurrence of different treatment groups jointly with common covariate effects but different hazard baseline among groups.

In order to estimate parameters in Cox model, maximum likelihood can perform this task, but it is difficult. For a general case, we only consider right censored data with indicator $\delta_i = 0$ if

observation t_i is right censored, and $\delta_i = 1$ if not. For notation convenience, we denote explanatory covariates $X_{i,1},\ X_{i,2} \cdots X_{i,k}$ of observation i as X_i and effects $\beta_1,\ \beta_2 \cdots \beta_k$ as β. The likelihood function of Cox model is as following of a size N data set:

$$\prod_{i=1}^{N} [h(t_i|X_i,\beta)]^{\delta_i} S(t_i|X_i,\beta) = \prod_{i=1}^{N} [h_0(t_i) e^{X_i'\beta}]^{\delta_i} S_0(t_i)^{e^{X_i'\beta}}$$

So Cox brings out partial likelihood function and proposes to gain parameter estimates through maximizing the partial likelihood function. Details can be found in Cox (1972). Prentice et al. (1978) modeled cause-specific hazard via this approach in a study of acute leukemia patients in the treatment of bone marrow transplantation. The model is applied to investigate how different factors such as patients age, regimen type can affect the timing of recurrence. Also it models how risk of recurrence and risk of mortality from graft versus host disease are related by patient characteristics. In typical practice, we always take baseline hazard function as a nuisance parameter but put more emphasis studying the covariate effects. However, in this thesis, we will focus on estimating baseline hazard function in a more efficient method besides estimating covariate effects. Particularly, we try to optimize the spread of $t_j's$–time resolutions over the muliresolution model developed in Bouman et al. (2005, 2007), aiming at a more precise estimation hazards and less time-consuming algorithm.

Gray (1990) adopts kernel-based smoothing approach to estimate subgroup baseline hazard over the cumulative hazard function derived from Breslow (1974) and applies it to data from adjuvant therapy for premenopausal breast cancer. It can be used to diagnose the correctness of the fitting model, since when Cox proportionality is satisfied, baseline hazard from subgroups should be the same within sampling variability. Gray (1992) incorporates both linear and spline function to estimate covariate parameters. Estimates from maximizing penalized partial likelihood function summaries more details about covariate effects to breast cancer recurrence risk. Gray (1996) proposes a more general regression approach emphasizing on hazard function estimation by dividing the time axis, and using nonparametric regression smoother. In Andersen et al. (1993), many other methods are discussed such as the non-parametric Nelson-Aalen estimator for cumulative hazard

function, Kaplan-Meier estimator for estimating the survival function. But most of them still pay effort more on covariate effects and model checking, than on hazard function.

Beta process prior and gamma process prior are two highly used stochastic process priors in the context of nonparametric Bayesian hazard estimation. A process is named beta process, provided its cumulative hazard function has independent increments and those increments are approximately beta distributed. Hjort (1990) introduces beta process prior and concludes that the posterior of cumulative hazard function is still a beta process. Lee and Kim (2002) develop an computational algorithm to approximate beta process by generating sample path from a compound Poisson process. Kalbfleisch (1978) and Burridge (1981) model the cumulative hazard function as a gamma process. Correlated process priors are also used in modeling cumulative hazard function. Arjas and Gasbarra (1994) introduce a simple martingale jump process to mode the hazard rate. They assume hazard rate is a constant between two jump times, and the jump times are from a homogeneous Poisson process. Correlated gamma priors are placed to hazard rate to make it a martingale. In Nieto-Barajas and Walker (2002) the hazard priors are correlated by introducing a latent Poisson process between two adjacent hazard increments. More other methods are reviewed in Sinha and Dey (1997).

In the context of Bayesian intensity estimation, multiresolution methods have gained popularity in the recent years. In a now well-known example from astrophysics Kolaczyk (1999), gamma-ray photon counts over equal time intervals were modeled as an inhomogeneous Poisson process, with constant intensity function over the recursive dyadic partitions of the time-axis. In Nowak and Kolaczyk (2000), Bayesian multiscale model was extended to Poisson inverse problems. Non-Bayesian multiresolution modeling examples include the methods based on wavelets, as in Antoniadis et al. (1999).

Bayesian multiresolution models were extended to hazard function estimation, and equipped to deal with censoring and truncation, in the works of Bouman et al. (2005), Bouman et al. (2007), and Dukic and Dignam (2007). Inspired by Kolaczyk (1999), these models assume a binary partition tree structure, with a gamma prior placed on the total cumulative hazard over a finite time

interval, and beta priors describing the recursive partition parameters (the "split" parameters). The smoothness of the resulting piece-wise constant hazard rate is controlled by hyperparameters of these gamma and beta tree priors. The MRH model was used to estimate reporting delay (Bouman et al. 2005), breast cancer recurrence risk (Bouman et al. 2007), and, via its hierarchical multiresolution hazard (HMRH) extension, the multivariate hazard for different subpopulations (Dukic and Dignam 2007).

1.3 Multiresolution hazard(MRH) model

In this section we review the multiresolution hazard model that is used to estimate baseline population survival function. Data for patients from different sources include their failure times, possibly censored and covariate of individual characteristics. We take the Cox proportional hazard model, since it is desirable with censoring failure times and covariates, especially for estimating a overall population survival.

In our analysis, we have the "time resolution" t_j evenly distributed within the whole time span of study and estimate the associated baseline hazard increment d_j from t_{j-1} to t_j. We can transform the posterior estimates of d_j into survival probability estimates via $d_j = \int_{t_{j-1}}^{t_j} h(s)ds$. Here $h(t)$ is the hazard rate function at time t. Hence, we will focus on estimating the cumulative hazard function $H_{\text{base}}(t)$ and the discrete hazard increments $d_j \equiv H_{\text{base}}(t_j) - H_{\text{base}}(t_{j-1})$.

1.3.1 Multiresolution prior for baseline hazard increments

In our model, we firstly choose time points $0 < t_1 < t_2 < \cdots < t_J$ according to clinical interest. The time points don't necessarily have to be evenly spaced. But in our MRH model, we have the them evenly distributed. Then we use Cox proportional hazard model to estimate baseline cumulative hazard $H_{\text{base}}(t)$ at t_j as well as covariates. This differs from a standard Cox model in which the baseline hazard function is treated as a nuisance parameter. We assume $J = 2^M$ where $M > 0$ is the "depth" of the partition tree. M can be chosen in a variety of ways; for example,

using model selection criteria as in Bouman et al. (2007), or using clinical input, as in Dignam et al. (2009). An appropriate M has to ensure that each bin has multiple observations from a statistical convenience point of view.

In our model, $S_{\text{base}}(t)$ is not defined after time point t_J. Therefore times after t_J will be taken as right-censored. In this case, we then have to pick t_J wisely so that our lost in information can be minimized. Our estimation of overall $H_{\text{base}}(t)$ is a piece-wise function. It has constant hazard rate $h_{\text{base}}(t)$ over two neighbored time points. With fixed time points $0 < t_1 < t_2 < \cdots < t_J$, we can transform hazard increments to hazard rate easily. And it is notable that many papers in literature have used the idea of piecewise-constant hazard, for instance, Walker and Mallick (1997). We use this idea to estimate the cumulative hazard function in our model.

The sum of 2^M hazard increments $\{d_j\}_{j=1}^{j=J}$ will always equal the total cumulative hazard over the study interval (t_0, t_J), $H(t_J)$. For simplicity, we will use H to denote $H(t_J)$. Now let $H_{M,0} \equiv d_1, H_{M,1} \equiv d_2, \ldots, H_{M,2^M-1} \equiv d_J$. For each m, $m = 1, 2, \ldots, M-1$, we will recursively define the "level-m" hazard increments as $H_{m,p} \equiv H_{m+1,2p} + H_{m+1,2p+1}$, where $p = 0, 1, 2, \ldots, 2^{m-1} - 1$.

If we further define the corresponding split variable $R_{m,p}$ as $H_{m,2p}/H_{m-1,p}$, it follows that the partition tree with depth M can be specified by H, and the set of splits $R_{1,0}, \ldots, R_{M,2^{M-1}-1}$. From there, any d_j can be represented as a product of H and a certain branch of split variables. For example, when $M = 3$ we have the following set of relationships, also depicted in Figure 1.2.

$$\begin{aligned} d_1 &= HR_{1,0}R_{2,0}R_{3,0}, \\ d_2 &= HR_{1,0}R_{2,0}(1 - R_{3,0}), \\ &\vdots \\ d_8 &= H(1 - R_{1,0})(1 - R_{2,1})(1 - R_{3,3}). \end{aligned} \tag{1.4}$$

We adopt Beta priors on the the $R_{m,p}$'s and a Gamma prior on H as in Nowak and Kolaczyk (2000). The prior expectation of each hazard increments are determined by both the shape parameters of each Beta prior for each $R_{m,p}$ and the hyperparameters of H. We adopt a hyperparameter

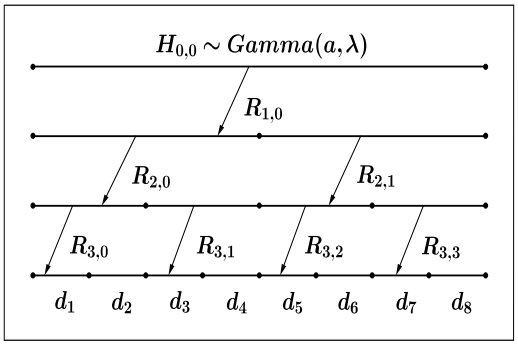

Figure 1.2: Diagram of the 3-level multiresolution prior

k and multiply it to the shape parameter of the Beta priors at each additional level, to gain more smoothness in the multiresolution priors. Follow this fashion, H and $R_{m,p}$ in a $M = 3, J = 8$ model would have the following priors:

$$\begin{aligned} H &\sim \mathcal{G}a(a,\lambda), \\ R_{1,0} &\sim \mathcal{B}e(2\gamma_{1,0}ka, 2(1-\gamma_{1,0})ka), \\ R_{2,p} &\sim \mathcal{B}e(2\gamma_{2,p}k^2a, 2(1-\gamma_{2,p})k^2a), \qquad p = 0,1 \\ R_{3,p} &\sim \mathcal{B}e(2\gamma_{3,p}k^3a, 2(1-\gamma_{3,p})k^3a), \qquad p = 0,1,2,3. \end{aligned} \tag{1.5}$$

Note that the prior expectation of each hazard increment is determined by the shape parameters of the beta priors, and the hyperparameters of H, as each increment d_j can be represented as a product of a set of independent variables on the same branch. For example, in the above, $E(d_1) = E(H)E(R_{1,0})E(R_{2,0})E(R_{3,0})$. The tree hyperparameters are thus directly linked to the distribution of the hazard increment, and control not only their *a-priori* expected values, but also the correlation between them. In that sense, the tree hyperparameters directly relate to the smoothness of the multiresolution prior, as shown in Bouman et al. (2005) and Bouman et al. (2007).

1.3.2 Properties of the MRH prior

Although the MRH prior is a piece-wise constant prior for the hazard rate, the carefully constructed tree-based approach presented above assures the *self-consistency* property. This property means that the joint distribution of hazard increments *at any given level* m does not depend on the total depth of the tree, M. This multiscale property allows researchers to change focus form one resolution to the other, without having to re-derive prior distributions.

To see this and other properties of the generalized multiresolution prior, involving the derivation of the hazard increment distribution, we will first let $0 < \gamma_{m,p} < 1$ and $k = 0.5$. We then consider what happens with the hazard increments at level $m = 1$. As $H \sim \mathcal{G}a(a, \lambda)$ and $R_{1,0} \sim \mathcal{B}e(\gamma_{1,0}a, (1-\gamma_{1,0})a)$, the two level-1 hazard increments can be derived as $H_{1,0} = HR_{1,0}$ and $H_{1,1} = H(1 - R_{1,0})$. Then:

$$P(H_{1,0} \leqslant x) = \int_0^1 \int_0^{x/R_{1,0}} \frac{H^{a-1}e^{-H/\lambda}}{\lambda^a \Gamma(a)} \frac{R_{1,0}^{\gamma_{1,0}a-1}(1-R_{1,0})^{(1-\gamma_{1,0})a-1}}{B(\gamma_{1,0}a, (1-\gamma_{1,0})a)} dH dR_{1,0} \tag{1.6}$$

implying that the density of $H_{1,0}$, which we denote as $f(x)$, is $\mathcal{G}a(\gamma_{1,0}a, \lambda)$, as shown in Equation (1.7):

$$\begin{aligned}
f(x) &= \int_0^1 \frac{d\left(\int_0^{x/R_{1,0}} \frac{H^{a-1}e^{-H/\lambda}}{\lambda^a\Gamma(a)} \frac{R_{1,0}^{\gamma_{1,0}a-1}(1-R_{1,0})^{(1-\gamma_{1,0})a-1}}{B(\gamma_{1,0}a,(1-\gamma_{1,0})a)} dH\right)}{dx} dR_{1,0} \\
&= \int_0^1 \frac{(\frac{x}{R_{1,0}})^{a-1}e^{-\frac{x}{R_{1,0}\lambda}}}{\lambda^a\Gamma(a)} \frac{R_{1,0}^{\gamma_{1,0}a-1}(1-R_{1,0})^{(1-\gamma_{1,0})a-1}}{B(\gamma_{1,0}a,(1-\gamma_{1,0})a)} \frac{1}{R_{1,0}} dR_{1,0} \\
&= \frac{x^{a-1}}{\lambda^a\Gamma(a)B(\gamma_{1,0}a,(1-\gamma_{1,0})a)} \int_1^\infty e^{-xt/\lambda}(t-1)^{(1-\gamma_{1,0})a-1} dt \\
&= \frac{x^{a-1}}{\lambda^a\Gamma(a)B(\gamma_{1,0}a,(1-\gamma_{1,0})a)} \int_0^\infty e^{-x(u+1)/\lambda} u^{(1-\gamma_{1,0})a-1} du \\
&= \frac{x^{\gamma_{1,0}a-1}e^{-x/\lambda}}{\lambda^{\gamma_{1,0}a}\Gamma(\gamma_{1,0}a)} \sim \mathcal{G}a(\gamma_{1,0}a, \lambda)
\end{aligned} \tag{1.7}$$

Following the same reasoning, we can show that $H_{1,1} \sim \mathcal{G}a((1-\gamma_{1,0})a, \lambda)$. We next consider the second-level hazard increments, at $m = 2$, when $R_{2,0} \sim \mathcal{B}e(\gamma_{2,0}a/2, (1-\gamma_{2,0})a/2)$ and $R_{2,1} \sim$

$\mathcal{B}e(\gamma_{2,1}a/2, (1-\gamma_{2,1})a/2)$. Then the prior distribution of $H_{2,0} = HR_{1,0}R_{2,0} = H_{1,0}R_{2,0}$ can be obtained as follows:

$$P(H_{2,0} \leqslant x) = \int_0^1 \int_0^{x/R_{2,0}} \frac{H_{1,0}^{\gamma_{1,0}a-1} e^{-H_{1,0}/\lambda}}{\lambda^{\gamma_{1,0}a}\Gamma(\gamma_{1,0}a)} \frac{R_{2,0}^{\gamma_{2,0}a/2-1}(1-R_{2,0})^{(1-\gamma_{2,0})a/2-1}}{B(\gamma_{2,0}a/2, (1-\gamma_{2,0})a/2)} dH_{1,0} dR_{2,0} \tag{1.8}$$

implying that the form of the prior density of $H_{2,0}$ can be derived as in Equation (1.9):

$$\begin{aligned}
&\int_0^1 \frac{d\left(\int_0^{x/R_{2,0}} \frac{H_{1,0}^{\gamma_{1,0}a-1} e^{-H_{1,0}/\lambda}}{\lambda^{\gamma_{1,0}a}\Gamma(\gamma_{1,0}a)} \frac{R_{2,0}^{\gamma_{2,0}a/2-1}(1-R_{2,0})^{(1-\gamma_{2,0})a/2-1}}{B(\gamma_{2,0}a/2,(1-\gamma_{2,0})a/2)} dH_{1,0}\right)}{dx} dR_{2,0} \\
&= \int_0^1 \frac{(\frac{x}{R_{2,0}})^{\gamma_{1,0}a-1} e^{-\frac{x}{R_{2,0}\lambda}}}{\lambda^{\gamma_{1,0}a}\Gamma(\gamma_{1,0}a)} \frac{R_{2,0}^{\gamma_{2,0}a/2-1}(1-R_{2,0})^{(1-\gamma_{2,0})a/2-1}}{B(\gamma_{2,0}a/2, (1-\gamma_{2,0})a/2)} \frac{1}{R_{2,0}} dR_{2,0} \\
&= \frac{x^{\gamma_{1,0}a-1}\int_1^\infty e^{-xt/\lambda}(t-1)^{(1-\gamma_{2,0})a/2-1} t^{\gamma_{1,0}a-a/2} dt}{\lambda^{\gamma_{1,0}a}\Gamma(\gamma_{1,0}a)B(\gamma_{2,0}a/2, (1-\gamma_{2,0})a/2)}
\end{aligned} \tag{1.9}$$

The expression on the last line in Equation (1.9) is a gamma density only when $\gamma_{1,0} = 0.5$. In that case we have $H_{2,0} \sim \mathcal{G}a(\gamma_{2,0}a/2, \lambda)$. Following the same steps, we can also get the result that $H_{2,1}$, $H_{2,2}$ and $H_{2,3}$ will not have a gamma density unless $\gamma_{1,0} = 0.5$. When $\gamma_{1,0} = 0.5$, $H_{2,1} \sim \mathcal{G}a((1-\gamma_{2,0})a/2, \lambda)$, $H_{2,2} \sim \mathcal{G}a(\gamma_{2,1}a/2, \lambda)$ and $H_{2,3} \sim \mathcal{G}a((1-\gamma_{2,1})a/2, \lambda)$. Thus we have the following properties of multiresolution prior:

- Resolution invariance: the prior of $H_{m,p}$ does not depend on M, the depth of the tree. The prior of $H_{m,p}$ will only contain parameters from the level m itself, and from the upper levels (levels closer to the root). The choice of $\gamma_{m,p}$ will have an impact on the densities of hazard increments.

- When all $\gamma_{m,p} = 0.5$ at level m and the upper levels, *a priori* $H_{m,p} \sim \mathcal{G}a(a/2^m, \lambda)$. Although this is the same gamma distribution for all increments at the same level, they will not in general be independent. The correlation is controlled in part by k, as shown in Bouman et al. (2005), Bouman et al. (2007).

- When all $\gamma_{m,p} = 0.5$, except at the level m, then $H_{m,2p} \sim \mathcal{G}a(\gamma_{m,p}a/2^{m-1}, \lambda)$ and $H_{m,2p+1} \sim \mathcal{G}a((1-\gamma_{m,p})a/2^{m-1}, \lambda)$. This property illustrates the flexibility in using $\gamma_{m,p}$ parameters to specify the hazard shape *a priori*.

1.4 Outline of the thesis

The rest of this dissertation is organized as follows. In Chapter 2, we describe an extension of multiresolution hazard(MRH) models, called pruned multiresolution hazard(PMRH) models and evaluate PMRH model performance with simulated data. As an application of PMRH models, we show the analysis of prostate cancer data in Chapter 3. Next, we give a literature review about hazard models with time-varying covariates and discuss about cumulative hazard function estimation in extended proportional hazards models in Chapter 4. In Chapter 5, we formulate algorithms for MRH models with time-varying covariates. Results of MRH models and PMRH models having time-varying covariates with simulated data are demonstrated as well. In Chapter 6, we review hazard models with missing time-varying covariates and outcomes. In Chapter 7, we discuss about different missing time-varying covariates imputation approaches and give results of MRH models having missing time-varying covariates with simulated data. Finally, we conclude this dissertation along with future work in Chapter 8.

Chapter 2

Pruned multiresolution hazard(PMRH) models

2.1 Pruning the MRH tree

The choice of the level of the maximal resolution in the MRH prior is driven by a compromise between the desire for detail and the amount of data: as the resolution increases (and the number of time bins increases), counts within each bin will decrease. While useful for revealing detailed patterns, large number of bins (and consequently, large number of model parameters) will generally require longer computing times. Similarly, more bins will eventually mean lower event count per bin, and this lower information content will translate into lower efficiency. It would thus make sense to provide an algorithm that could adaptively choose the appropriate number of bins over different time intervals, with an increased number of bins in the regions of high event counts, and fewer bins where the counts are low. Ideally, the method would balance computing time and accuracy. Motivated by these goals, we develop a data-driven "tree pruning" method, which starts with the full MRH prior, and as the end result provides a smaller tree prior with fewer branches.

The idea of MRH tree pruning is simple: two adjacent bins constructed via the same split parameter, $R_{m,p}$, are merged if the the estimated hazard increments in these two bins ($H_{m+1,2p}$ and $H_{m+1,2p+1}$) are statistically similar. Here, the estimate of the hazard increment in a bin is derived as the number of observed failures within the bin divided by the number of patients at risk at the initial time point of the bin. The pruning proceeds as follows: for a given level m (for $m = 1, ..., M$), the null hypothesis $H_0 : R_{m,p} = 0.5$ is tested versus the alternative $H_a : R_{m,p} \neq 0.5$ (with a pre-set type I error α) for each split parameter $R_{m,p}$ ($p = 0, ..., 2^{m-1} - 1$). If the null hypothesis is not

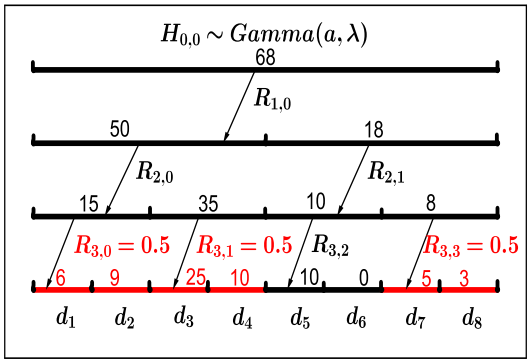

Figure 2.1: Diagram of the 3-level multiresolution pruned prior

rejected, that split $R_{m,p}$ is set to 0.5; the adjacent hazard increments are considered equal and the bins declared "fused". The resulting pruned prior tree is smaller than the full MRH tree used in previous analyses (Bouman et al. 2007; Dukic and Dignam 2007).

The hypothesis testing can be applied to all M levels of the tree, but often in practice only the few bottom levels need be considered. Figure 2.1 presents a hypothetical example diagram with event counts shown at each level of a three-level hierarchy. If all levels are subject to pruning, the split parameters $R_{3,0}$, $R_{3,1}$, and $R_{3,3}$ (shown in red in the figure) will be set to 0.5 after the pruning procedure, and the resulting final resolution will have only 5 bins instead of 8 which the full MRH prior with 3 levels would have.

The information in the counts used for testing the equality of adjacent bin hazard increments is not independent across bins, and small event counts may be frequent in the bins at the bottom levels. For that reason, we perform the pruning hypothesis tests using a modified Fisher's exact test, based on the 2 by 2 table composed of the number of failures within in the bin time interval and at-risk patients at the end of the bin time interval for each pair of adjacent bins sharing a split parameter. It is important to note that the Fisher's exact test provides a simple approximate solution to a fundamentally more complex inference problem, and that other tests could be used instead in future applications. We add a modification to the test in situations where no failures occur in one or both of the bins – the pruning algorithm presented in this thesis fuses the pair of

adjacent bins with no failures, while the bins where only one bin has 0 failures are not fused. Other modifications might be considered instead, though we do not explore them in this thesis.

The pruned (PMRH) model still retains its resolution-invariance under aggregation. Although the pruning is expected to reduce sensitivity of the MRH method in identifying subtle changes in the hazard rate, the pruning method carries several advantages that may be worth considering. With pruning, the posterior hazard increment estimates are expected to be less variable compared to the equivalent non-pruned model. With some split parameters preset to 0.5, fewer total number of model parameters need to be estimated, and, consequently, the estimation time is expected to be reduced as well. Given that the estimation in these models is done using Markov chain Monte Carlo (MCMC) methods, total computing time savings might be substantial.

2.2 Model fitting

The multiresolution prior for the hazard rate can be used in conjunction with any desired likelihood for time-to-event data. One of the more commonly used models in survival analysis is the "proportional-hazards" model (Cox 1972), where the hazard rate for patient i (h_i) is modeled as the product of an unspecified baseline hazard rate (h_0) and the systematic covariate function:

$$h_i(t) = h_0(t)e^{\beta_1 X_{i,1}+\beta_2 X_{i,2}+\cdots+\beta_k X_{i,k}}. \tag{2.1}$$

Here, the parameter vector of log-hazard ratios, $\vec{\beta}$, contains the effects of covariates in $\vec{X}$, the matrix whose p columns correspond to the explanatory variables $\vec{X}_1,\ \vec{X}_2 \cdots \vec{X}_p$. The hazards in different covariate groups are thus assumed to be proportional to each other over the entire study period, giving this model its name.

Unlike the classic Cox proportional hazards model, whose main goal is to estimate the effects of covariates while treating the hazard rate itself as a nuisance parameter, the PMRH model estimates all parameters and the hazard rate jointly. The PMRH prior is placed on the baseline hazard rate h_0, and the posterior distribution for all PMRH tree parameters, as well as the $\vec{\beta}$ parameter, are estimated based on the joint posterior derived using MCMC.

Considering only right censored data, with indicator $\delta_i = 0$ if the patient i's time-to-event, T_i, is right censored, and $\delta_i = 1$ if not (i.e. if that observation was an observed failure event), the likelihood function of the proportional hazards model for a set of N independent observations $(\vec{T}, \vec{\delta}) = \{T_i, \delta_i\}_1^N$ is as follows:

$$\mathcal{L}(\vec{\beta}, h_0 \mid \vec{T}, \vec{\delta}) = \prod_{i=1}^{N} [h(T_i|\vec{X}_i, \vec{\beta})]^{\delta_i} S(T_i|X_i, \vec{\beta}) = \prod_{i=1}^{N} [h_0(T_i) e^{\vec{X}_i'\vec{\beta}}]^{\delta_i} S_0(T_i)^{e^{\vec{X}_i'\vec{\beta}}}$$

The Bayesian model is then completed by specifying prior distributions for $\vec{\beta}$, as well as for the multiresolution hazard tree parameters (a, λ, k, and $\vec{\gamma}$).

The PMRH algorithm proceeds in two steps. The hypothesis testing step is run only once at the beginning, as the prior tree is set. $R_{m,p}$ parameters for which the H_0 is not rejected are set the to 0.5 with probability 1, while those for which the H_0 is rejected are to be estimated in the second step, the MCMC-step. Once the testing of all candidate $R_{m,p}$ is completed, the MCMC step samples the remaining parameters (H, all $R_{m,p}$ that were not set to 0.5, a, λ, k, $\vec{\gamma}$, and $\vec{\beta}$) from their full conditional distributions, following Bouman et al. (2007). We briefly outline each full conditional distribution below (η^- is used to denote the set of all parameters and data except for η):

(1) If k is given an exponential prior distribution with mean μ_k, the full conditional distribution for k is as follows:

$$\pi(k|k^-) \propto \Pi_{m=1}^{M} \Pi_{p=0}^{2^{m-1}-1} \left\{ \frac{R_{m,p}^{2\gamma_{m,p}k^m a}(1-R_{m,p})^{2(1-\gamma_{m,p})k^m a}}{\mathrm{B}(2\gamma_{m,p}k^m a, 2(1-\gamma_{m,p})k^m a)} \right\} e^{-\frac{k}{\mu_k}}$$

(2) If a is given a zero-truncated Poisson prior, $\dfrac{e^{-\mu_a}\mu_a^a}{a!\,(1-e^{-\mu_a})}$ (chosen for computational convenience), the full conditional distribution for a is:

$$\pi(a|a^-) \propto \frac{H^a \mu_a^a}{\lambda^a (a-1)! a!} \Pi_{m=1}^{M} \Pi_{p=0}^{2^{m-1}-1} \left\{ \frac{R_{m,p}^{2\gamma_{m,p}k^m a}(1-R_{m,p})^{2(1-\gamma_{m,p})k^m a}}{\mathrm{B}(2\gamma_{m,p}k^m a, 2(1-\gamma_{m,p})k^m a)} \right\}$$

(3) If the scale parameter λ in the gamma prior for the cumulative hazard function H is given an exponential prior with mean μ_λ, the resulting full conditional is:

$$\pi(\lambda|\lambda^-) \propto \frac{1}{\lambda^a} e^{-\frac{H}{\lambda}} e^{-\frac{\lambda}{\mu_\lambda}}$$

(4) The full conditional for H, $\pi(H|H^-)$ is a gamma density, with the shape parameter $a + \sum_{i=1}^{N} \delta_i$, and rate parameter $\lambda^{-1} + \sum_{i=1}^{N} \exp(X_i'\beta) F(T_i)$, where $F(T_i) = H(\min(T_i, t_J))/H(t_J)$

(5) If a Beta(u, w) prior is placed on each $\gamma_{m,p}$, the full conditional distribution for each $\gamma_{m,p}$ is proportional to:

$$\frac{R_{m,p}^{2\gamma_{m,p}k^m a}(1-R_{m,p})^{2(1-\gamma_{m,p})k^m a}}{\mathrm{B}(2\gamma_{m,p}k^m a, 2(1-\gamma_{m,p})k^m a)}\gamma_{m,p}^{u-1}(1-\gamma_{m,p})^{w-1}$$

(6) A normal $N(0, \sigma_\beta^2)$ prior on each log hazard ratio, β_j, leads to the the following full conditional distribution, where $F(T_i) = H(\min(T_i, t_J))/H(t_J)$:

$$\pi(\beta_j|\beta_j^-) \propto \exp\{-\frac{\beta_j^2}{2\sigma_\beta^2}\}\Pi_{i=1}^{N}\{[\exp(X_{i,j}\beta_j)]^{\delta_i}\exp\left(-\exp\left(X_i'\beta\right)HF(T_i)\right)\}$$

(7) The full conditional for each $R_{m,p}$ for which H_0 was rejected, is proportional to:

$$R_{m,p}^{2\gamma_{m,p}k^m a-1}(1-R_{m,p})^{2(1-\gamma_{m,p})k^m a-1}\Pi_{i=1}^{N}\left\{[h_0(T_i)]^{\delta_i}\exp\left(-\exp(X_i'\beta)HF(T_i)\right)\right\}$$

For most parameters, the full conditionals could be sampled from either directly, or using a Metropolis-Hastings sampler within each iteration of the MCMC sampler. For $R_{m,p}$ and each β_j, we used the adaptive rejection sampling (ARS) algorithm of Gilks and Wild (1992). For parameters with full conditionals that are not log-concave, such as λ and k, adaptive rejection Metropolis sampling (ARMS) of Gilks et al. (1995) can be used.

2.3 Evaluating PMRH performance with simulated data

2.3.1 Simulated data generation

A natural question to ask is how using the data twice – once for the hypothesis testing in the construction of the prior, and once in the likelihood – might impact the results of the PMRH algorithm. The performance of the algorithm is also expected to depend on the number of failures in each bin, the true underlying intensity function, and the type I error of the test, α used in pruning.

In order to assess the overall behavior of PMRH estimators, and compare it to the behavior of the regular MRH estimators, we simulated 200 datasets closely resembling a real clinical trial (Fisher et al. 1989, 1996). In this breast cancer clinical trial, the estimated hazard rate exhibited a mixture of features (bumps and flat regions), which was ideal for evaluating the PMRH method's performance (Dukic and Dignam 2007; Dignam et al. 2009). Each dataset was given 1000 patients (500 in the treatment arm, and 500 in the control arm), and for each patient the failure time was simulated depending only on the treatment indicator (our only covariate). Anyone with the failure time occurring after 5 years was considered right-censored. The resulting data were over 70% censored, matching the amount of censoring in the trial in Fisher et al. (1989). The tamoxifen treatment log-hazard ratio, β_{treat}, was set at -0.44, to match the value estimated in Bouman et al. (2007). In order to see how the model performs in smaller datasets, we then repeated the set of simulations under the same conditions, except with 200 patients instead of 1000 per set, with 100 patients per arm.

We used a rich PMRH model with 5 levels ($M = 5$), with 32 equal length bins over 5 years. The average number of failures in the 32 bins, over all simulated datasets with 200 patients per dataset, is given in Table 2.1 and the average number of failures in the 32 bins, over all simulated datasets with 1000 patients per dataset, is given in Table 2.2. As can be seen, realistically small failure counts per bin were simulated on average. The average failure counts for bins in the larger simulation, where 1000 patients per dataset were used, were approximately 5 times higher on average, as expected.

2.3.2 Evaluating PMRH performance with simulated data

For each set of data, we implemented 4 different PMRH strategies:

- **NPM4:** 4-level model without any pruning,
- **NPM5:** 5-level model without any pruning,
- **PM52:** 5-level model with 4th and 5th level subject to pruning,

bin	1	2	3	4	5	6	7	8	9	10	11
untreated	0.22	0.38	0.76	0.505	1.00	1.165	1.305	1.10	0.92	0.905	1.13
treated	0.13	0.25	0.39	0.365	0.88	0.740	0.765	0.66	0.64	0.680	0.89
pool	0.35	0.63	1.15	0.870	1.88	1.905	2.070	1.76	1.56	1.585	2.02

bin	12	13	14	15	16	17	18	19	20	21	22
untreated	1.360	0.95	1.060	1.02	0.905	0.615	0.765	0.865	1.050	0.675	0.800
treated	0.745	0.69	0.695	0.74	0.620	0.415	0.490	0.675	0.685	0.540	0.585
pool	2.105	1.64	1.755	1.76	1.525	1.030	1.255	1.540	1.735	1.215	1.385

bin	23	24	25	26	27	28	29	30	31	32	censored
untreated	0.67	0.420	1.08	0.975	1.025	0.555	0.49	0.500	0.625	0.635	73.570
treated	0.46	0.375	0.65	0.635	0.630	0.440	0.35	0.355	0.525	0.435	81.875
pool	1.13	0.795	1.73	1.610	1.655	0.995	0.84	0.855	1.150	1.070	155.445

Table 2.1: Average counts in each bin, across 200 simulated datasets with 200 patients each.

bin	1	2	3	4	5	6	7	8	9	10	11
untreated	1.26	2.175	3.420	2.945	4.52	5.45	6.545	5.5	4.825	4.635	6.195
treated	0.85	1.355	2.365	1.910	3.17	3.69	4.285	3.7	3.105	3.465	4.175
all	2.11	3.530	5.785	4.855	7.69	9.14	10.830	9.2	7.930	8.100	10.370

bin	12	13	14	15	16	17	18	19	20	21	22
untreated	6.33	4.685	5.52	5.145	4.92	2.905	4.050	4.475	4.870	3.580	4.220
treated	4.08	3.140	3.86	3.405	3.16	1.990	2.855	3.120	3.435	2.295	2.655
all	10.41	7.825	9.38	8.550	8.08	4.895	6.905	7.595	8.305	5.875	6.875

bin	23	24	25	26	27	28	29	30	31	32	censored
untreated	3.125	2.49	4.49	5.095	4.715	3.15	2.615	2.485	4.00	3.145	366.520
treated	2.365	1.91	3.32	3.455	3.605	2.03	1.830	2.000	2.97	2.085	408.365
all	5.490	4.40	7.81	8.550	8.320	5.18	4.445	4.485	6.97	5.230	774.885

Table 2.2: Average counts in each bin, across 200 simulated datasets with 1000 patients each.

- **PM55:** 5-level model with all levels subject to pruning.

MCMC chains with 1000000 iterations for each of the 200 datasets were run separately, under each of the 4 PMRH strategies. The first half of each MCMC chain was discarded as the burn-in, and every 200th sample from the chain was kept to reduce autocorrelation. In the end, 2500 posterior samples per dataset were used to derive posterior PMRH estimates (posterior means), resulting in 200 sets of estimates. The mean square error (MSE) was computed for each PMRH estimator based on these approximate estimator distributions.

All the simulations were run on a supercomputer with 1368 nodes, each containing two hex-core 2.8Ghz Intel Westmere processors with 12 cores per node and 2GB of RAM per core. For a dataset of size 200, it took about 3 hours for model PM55 to complete 1 million iterations; 5 hours for model PM52; 2.5 hours for model NPM4 and 8 hours for model NPM5. The model NPM5 takes about 2.67 times longer than model PM55, as expected: PMRH method can reduce computing time substantially for a given maximum resolution of the model. Model NPM4 requires the shortest amount of computing time among all 4 models, which is also not surprising: a smaller tree is on average denser (has more events per bin), and the PMRH method may not prune it heavily.

For a dataset with 1000 patients, model PM55 took about 7.5 hours for 1 million iterations; model PM52 took 15 hours; model NPM4 took 13 hours; and model NPM5 took 41 hours. Model NPM5 takes about 5.47 longer than model PM55. In this case, model PM55 has the shortest time among all 4 models. In this case, as the true baseline hazard rate used to generate the datasets was quite flat, it is expected that as the dataset size increases and the number of observations in each bin also increases, the modified Fisher's exact test will result in fewer rejections of the null hypothesis. Eventually, model PM55 contains a tree with fewer branches than the NPM4 model.

Figure 2.2 depicts the square root of MSE of each hazard increment posterior mean in the four PMRH strategies, for simulations with 1000 patients (top) and 200 patients (bottom). PM55 model seems to have the smallest root MSE, while the NPM5 model has the largest, on average over all bins. The first few PM55 hazard increment estimators have larger root MSE than the other increments, which is due to low counts in those first few bins in our simulated data (see Table 2.1),

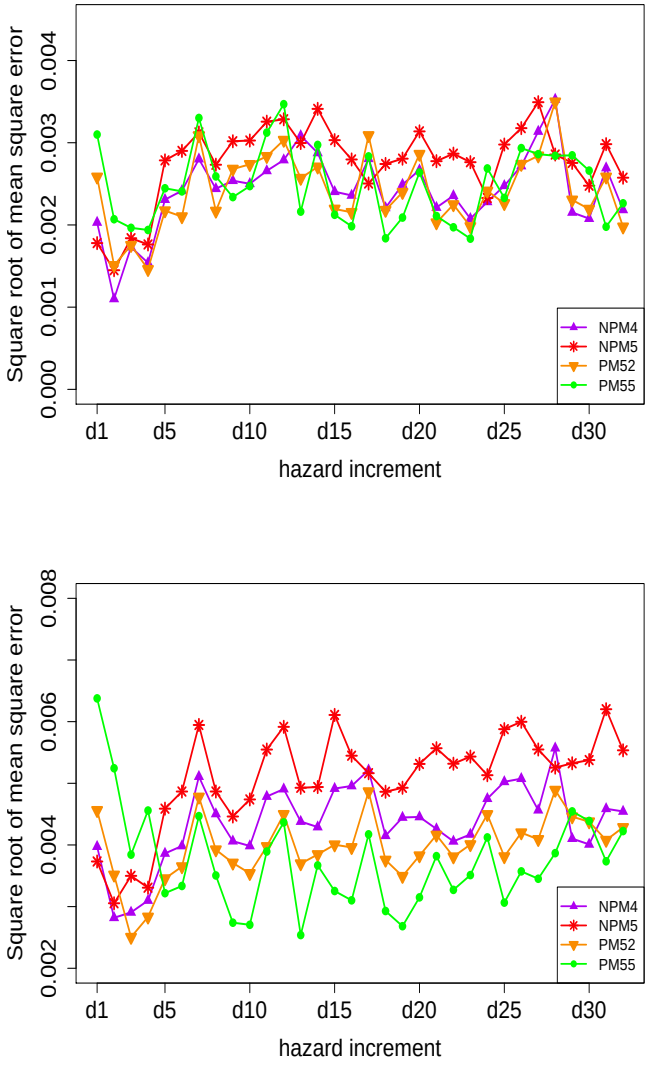

Figure 2.2: Estimated square root of the MSE for 4 PMRH hazard increment estimators (posterior means), based on 200 datasets with 1000 patients per dataset (top figure) and 200 patients per dataset (bottom figure).

and the estimators of hazard increments in these bins are expected to be more variable than the rest. In particular, PM55 performs poorly in those first few bins as these bins are often merged into a single bin under PM55, as the null hypothesis is rarely rejected for bins with low counts.

However, if we examine the square root of the integrated MSE for hazard increments (over all bins), shown in Table 2.3, we see that for the 200-patient simulation the PM55 model has the

smallest square root of the integrated MSE, followed by PM52 model, NPM4 model and NPM5 model. In the 1000-patient simulation, PM52 has the smallest square root of integrated MSE, followed by NPM4 model, PM55 model and NPM5 model. The difference among root-integrated MSEs among models is much smaller in the 1000 patients per-set case than that of 200 patients per-set case. As the pruning only affects the prior, and the prior effect dampens as the dataset size size increases, we can expect that the estimates will be very close among the different 4 models in larger datasets.

Table 2.3: Square root of integrated mean square error of all hazard increments from models PM55, PM52, NPM4 and NPM5, in simulations of 200 datasets with 200 patients, and 200 datasets with 1000 patients.

data set size	PM55	PM52	NPM4	NPM5
200	0.022	0.023	0.025	0.029
1000	0.015	0.013	0.014	0.016

Figure 2.3 shows the 95% probability intervals of posterior means for the 32 individual hazard increments, for the four PMRH strategies, in datasets with 1000 patients (top two plots) and with 200 patients (bottom 2 plots). The left column represents the raw PMRH results, while the right column shows smoothed versions of those using polynomials of degree 7. While more aggressive pruning will generally result in more variation in bins with fewer counts (for example, the first and last bins), it will also tend to reduce the variability over the other bins, resulting in less variable hazard rate estimator.

Figure 2.4 shows the histograms of posterior means of H and β_{treat} over 200 datasets for each of the 4 strategies; the top four shows the performance of H estimator, while the bottom four shows the estimator of β_{treat}. The red line indicates the true value of H and β_{treat} used to simulate the data. Unlike with the hazard increment estimators, it appears that the 4 pruning strategies do not differ much in estimating the baseline H and β_{treat}. This is consistent with the fact that the pruning algorithm only reallocates hazard increments, thus leaving the total of all increments, i.e. the baseline H value, unaffected.

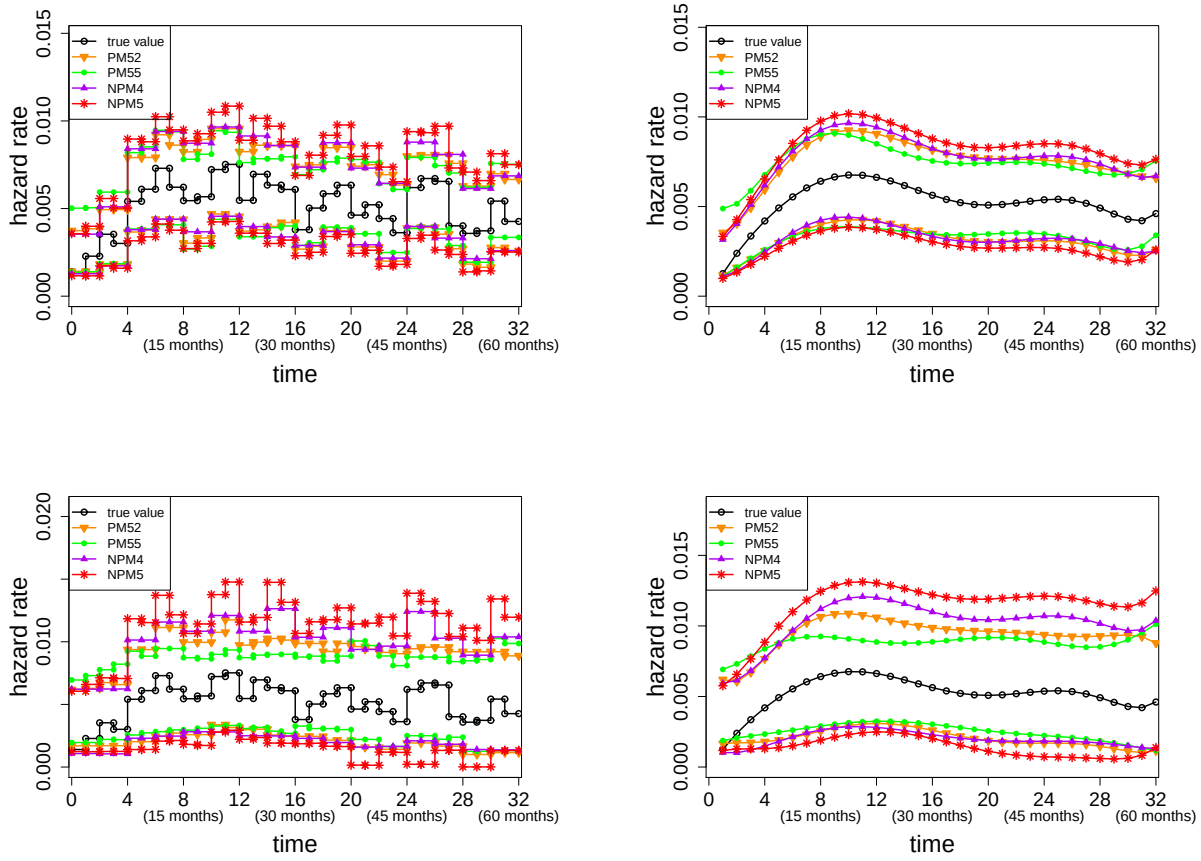

Figure 2.3: 95% probability intervals of individual hazard rate posterior means (left) and their smoothed version (right), for 4 different PMRH estimators. The top row shows the performance over 200 simulated datasets with 1000 patients per dataset, and the bottom row shows the performance over 200 simulated datasets with 200 patients each. (Model NPM5 has the widest 95% probability intervals.)

The effect of number of observations is shown in Figure 2.5, which depicts the distribution of marginal posterior means of H and β_{treat} for two different sample sizes, 1000 (gray) and 200 (white). Both are centered around true values, but the variances of the distributions based on the larger sample size are clearly smaller than their counterparts based on the smaller sample size.

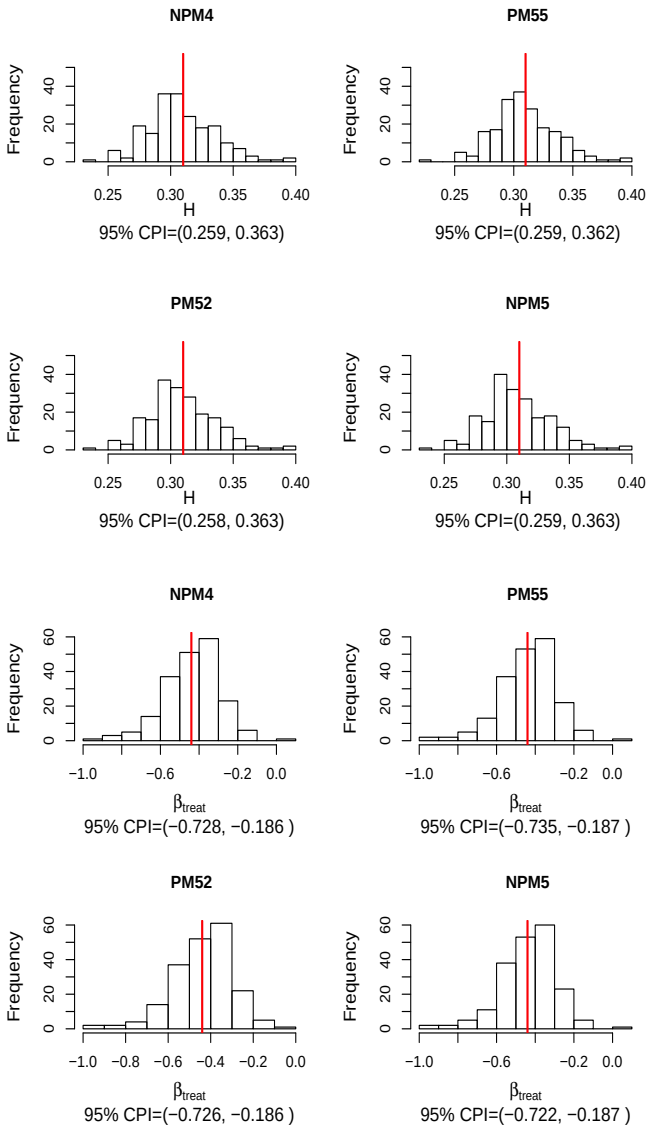

Figure 2.4: Top 4 plots: Histograms of posterior means of H over 200 datasets, with 1000 patients per dataset, for 4 different PMRH estimators. (Red line is the true H value of 0.31, used to simulate the data.) Bottom 4 plots: Histograms of posterior means of β_{treat} over 200 datasets, with 1000 data per dataset, for 4 different PMRH estimators. (Red line is the true β_{treat} value of -0.44, used to simulate the data.) 95% CPI denotes the central 95% probability interval.

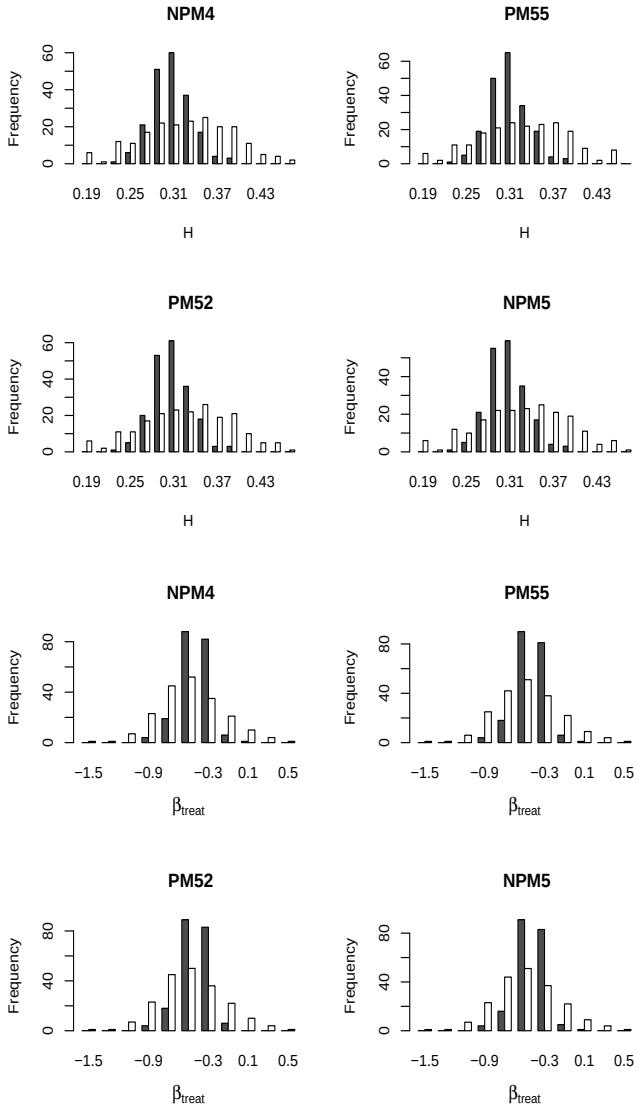

Figure 2.5: Results from two sets of simulations: 1000 observations per dataset (gray) and 200 observations per dataset (white). Top 4 plots: Histogram of 200 posterior means of H, for 4 different PMRH estimators (H was set to 0.31 to simulate the datasets). Bottom 4 plots: Histogram of posterior means of 200 β_{treat}, for 4 different PMRH estimators (β_{treat} was set to -0.44 to simulate the datasets.)

Chapter 3

Applications of PMRH models–analysis of prostate cancer data

In this chapter we employ PMRH method to a landmark clinical trial investigating radiation and hormone treatment for men with localized (e.g., not metastatic) prostate cancer. Early stage prostate cancer presents a valuable example of a complex failure process over time, which, if understood completely, would yield valuable insight and guidance for clinical decisions regarding appropriate degree of treatment and follow-up surveillance for disease recurrence.

3.1 Randomized clinical trials for prostate cancer

In the prostate cancer literature, it is well known that the disease progression and mortality hazard for patients experiencing intermediate clinical events, such as elevated Prostate-Specific Antigen (PSA), is higher than for those that do not experience such events. However, even in men with high-risk disease, the cumulative hazard of death from other causes exceeds that of death from prostate cancer, primarily due to the advanced age at which prostate cancer is often diagnosed. Apart from the common assumption that the disease may be more indolent if diagnosed in late age (or become so as the individual ages), little is still known about how prostate cancer mortality hazard may be changing over time, and what factors may influence it.

The data for this case study come from a large clinical trial performed by the Radiation Therapy Oncology Group (RTOG), which has comprehensively studied prostate cancer treatments in a variety of risk groups. Earlier RTOG trials established that androgen deprivation (AD) after radiation therapy was beneficial for avoidance of recurrence and improvement in prostate cancer-

specific survival (Pilepich et al. 2005, 2001). The study that we will be analyzing in this thesis, RTOG 9202, explored duration of AD among 1521 participants with locally advanced high risk prostate cancer, who underwent external beam radiation therapy and were randomized to either 4 months of AD or additional 24 months of AD (Horwitz et al. 2008).

Our analysis looks at the subset of 1421 patients from RTOG 9202 trial, after excluding data from 100 patients who were missing the Gleason score at baseline (a pathology score assigned to the biopsied tissue to describe the disease severity). Out of the remaining 1421 patients, 716 are treated with long-term AD hormonal therapy, and 705 with short-term AD hormonal therapy. The average age of the 1421 patients at baseline (study entry) is 70 years, with a standard deviation of 6.5 years. The Gleason score distribution was similar in both treatment groups, averaging at 6.7 and ranging from 2 to 10. Primary trial endpoints included time to disease recurrence or death, specific types of recurrence (local/regional recurrence, distant metastasis) and prostate cancer specific death, and overall mortality. Here we model time to death from any cause, which reflects both beneficial and potentially deleterious effects of hormone therapy, and reflects dynamics of patient factors such as age at diagnosis.

3.2 Markov Chain Monte Carlo Bayesian model estimation of PMRH model

The time horizon for our study was set at 160 months, leaving 9 extra patients as administratively right-censored, and 637 patients who were lost follow-up before the study ends. Patients treated with short term AD with age 70 and Gleason score of 2 were taken as the baseline. To address the question of whether different subgroups of men with specific disease and health features at different ages may have different failure hazard patterns over time, our model will examine the effects of Gleason score, linear centered age (age at baseline minus 70), and quadratic centered age (age at baseline minus 70, squared). We also allow for different effects of age before and after the baseline age of 70, to examine the possibility of more aggressive cancers in younger patients. These variables are used as predictors in the proportional hazards setting, according to the following

model:

$$h_i(t) = h_0(t)\exp(\beta_1 X_{i,1} + \beta_2 X_{i,2} + \cdots + \beta_6 X_{i,6})$$

where $X_{i,1}$ to $X_{i,6}$ are the patient i's covariates, as follows:

- $X_{i,1}$ is the treatment indicator: 1 if long-term and 0 if short-term;
- $X_{i,2}$ is the Gleason score minus 2;
- $X_{i,3}$ is the standardized age of the i-th patient, standardized by subtracting the average age of 70 and dividing the age standard deviation of 6.5;
- $X_{i,4}$ equals $X_{i,3}$ if the i-th patient was older than 70 at baseline, and otherwise $X_{i,4} = 0$;
- $X_{i,5}$ is the standardized age squared, $X_{i,3}^2$;
- $X_{i,6}$ equals $X_{i,5}$ if the i-th patient was older than 70 at baseline, and otherwise $X_{i,6} = 0$.

We compare the four PMRH models with 4 and 5 levels (corresponding to time bins with length of 10 months and 5 months, respectively), and consider different levels of pruning: PM52 (5-level MRH with 2 bottom levels subject to pruning), PM55 (5-level MRH with all 5 levels subject to pruning), NPM4 (4-level MRH with no pruning), and NPM5 (5-level MRH with no pruning). For pruning, we set the type I error $\alpha = 0.05$. All $R_{m,p}$s were given symmetric beta priors, $R_{m,p} \sim \mathcal{B}e(a/(2^m), a/(2^m))$. The log hazard ratios (β parameters) were given a flat prior. H was given a $\mathcal{G}a(a, \lambda)$ prior. a was given a zero truncated Poisson prior with $\mu_a = 4$, where μ_a is the mean of the Poisson before truncation. λ was given an exponential prior with mean 100.

For each PMRH model, we run five MCMC chains with one million iterations, using the first half as the burn-in, and taking every 200th sample from the second half to drastically reduce autocorrelation. The resulting 2500 samples from each case were analyzed. Model PM55 took approximately 8 hours for 1 million iterations; model PM52 10.5 hours; model NPM4 11.5 hours, and model NPM5 27 hours. Model NPM5 took about 3 times longer than model PM55, which was expected as PMRH method can reduce computing time substantially.

3.3 Analysis of the death from any cause in prostate cancer

3.3.1 Hazard function estimation

In terms of the baseline hazard, Figure 3.1 shows the smoothed 95% posterior credible intervals of hazard increments from the 4 PMRH models (PM55, NPM4, NPM5, and PM52). Here, the smoothing was done using a degree 7 polynomial of the center points of each bin. The figure indicates that PM55 model yields the narrowest hazard increment posterior intervals, while NPM5 yields the widest. This discrepancy is only notable towards the end of the study, from 11 to 13 years post enrollment, where the number of events and patients remaining under observation are low. All our results indicate however that a linearly increasing hazard function might provide a reasonable simpler model for modeling risk of death due to any cause in prostate studies.

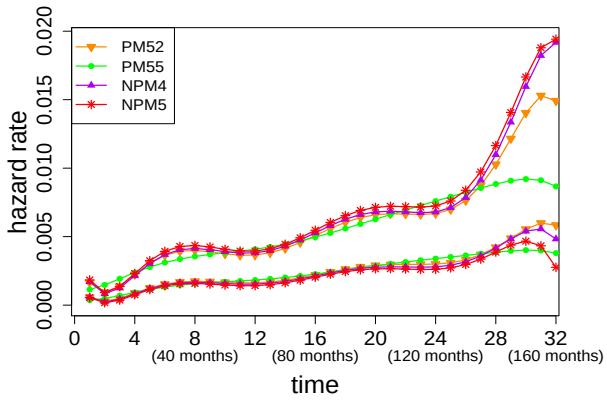

Figure 3.1: Smoothed posterior 95% pointwise credible intervals of individual baseline hazard rate for the prostate cancer data, all 4 PMRH models.

3.3.2 Covariate effect estimation

In terms of covariate effects, Table 3.1 shows the posterior estimates and their 95% credible intervals for each of the 6 covariates used. The posterior estimates appear almost identical across 4 PMRH strategies; as expected, the pruning method did not affect covariate estimation. The

Table 3.1: Estimates for the prostate cancer predictor effects

	Treatment (X_1)	Gleason (X_2)	Age (X_3)	Age >70 (X_4)	Age quad (X_5)	Age quad >70 (X_6)
PM52						
2.5%	-0.253	0.083	0.042	-0.518	-0.005	-0.372
50%	-0.112	0.131	0.353	0.097	0.114	-0.186
mean	-0.113	0.131	0.355	0.096	0.113	-0.186
97.5%	0.030	0.179	0.678	0.696	0.226	-0.004
PM55						
2.5%	-0.252	0.082	0.041	-0.481	-0.004	-0.381
50%	-0.112	0.129	0.361	0.099	0.118	-0.191
mean	-0.112	0.130	0.360	0.102	0.116	-0.194
97.5%	0.029	0.180	0.667	0.707	0.221	-0.017
NPM5						
2.5%	-0.250	0.081	0.032	-0.474	-0.005	-0.379
50%	-0.115	0.130	0.357	0.105	0.115	-0.188
mean	-0.113	0.130	0.353	0.107	0.113	-0.189
97.5%	0.028	0.180	0.662	0.717	0.223	-0.003
NPM4						
2.5%	-0.255	0.079	0.039	-0.493	-0.007	-0.377
50%	-0.114	0.130	0.352	0.104	0.114	-0.185
mean	-0.114	0.130	0.353	0.103	0.113	-0.186
97.5%	0.028	0.180	0.669	0.710	0.221	-0.010

effect of the long-term treatment is estimated at -0.11, with the 95% credible interval of (-0.26, -0.03). This implies that the hazard of death due to any cause for men treated with the long-term treatment is lower by an estimated 11% than the hazard of those treated with the short-term treatment, holding all other covariates constant. Figure 3.2 shows the smoothed estimated hazard rates for patients on short-term treatment versus long-term treatment, aged 70 and with Gleason score of 2 at enrollment, based on model PM55.

The effect of the Gleason score is estimated at 0.13, with the 95% credible interval of (0.08, 0.18), implying that there is an estimated increase of 14% in the hazard with each one-unit increase in the Gleason score at baseline, holding all else constant.

The effect of age has to be examined in more detail. The linear effect of age is estimated

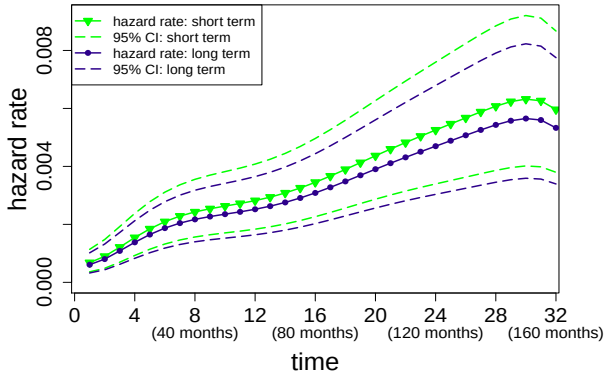

Figure 3.2: The smoothed estimated hazard rates (solid lines) and their pointwise credible intervals (dashed lines) for patients on short-term treatment versus long-term treatment, aged 70 and with Gleason score of 2 at enrollment, based on model PM55.

at 0.35 with the 95% credible interval of (0.04, 0.68). This effect seems to be similar in patients who are younger and older than 70 at baseline, as the estimated effect of $X_{i,4}$ is approximately 0.1 with 95% credible interval of (-0.52,0.70). The quadratic effect of age is estimated at 0.11, with the 95% CI of (-0.0048, 0.226), but this effect appears to differ depending on whether the patient was under or over 70 at baseline: the estimated effect of $X_{i,6}$ is -0.19 with 95% credible interval of (-0.37,-0.004).

The overall effect of age is thus perhaps best interpreted through examples: a 50-year old patient's hazard rate is estimated to be 1.29 times higher than a 60 year old patient's rate, and 2% lower than a 70 year old patient's rate, if patients are on the same treatment and with the same Gleason score. On the other hand, an 80-year old patient's rate is estimated to be 1.68 times higher than a 70-year old patient's rate, 2.23 times higher than a 60-year old's, and 1.72 times higher than a 50-year old patient's rate, for patients who are on the same treatment and have the same Gleason score. Interpreted more broadly, this finding reflects the fact that very aged individuals have high mortality which is expected and mostly due to non-cancer causes, while very young individuals with prostate cancer have high mortality relative to the cases with more typical age at onset, because the disease tends to be most aggressive in early onset cases. Figure 3.3 shows

the smoothed estimated hazard rates for patients on short-term treatment with Gleason score of 2, aged 50, 60, 70 and 80 years at enrollment, based on model PM55. The smoothed estimated hazard rates for patients aged 50 and 70 years at enrollment, and on short-term treatment with Gleason score of 2, are almost identical.

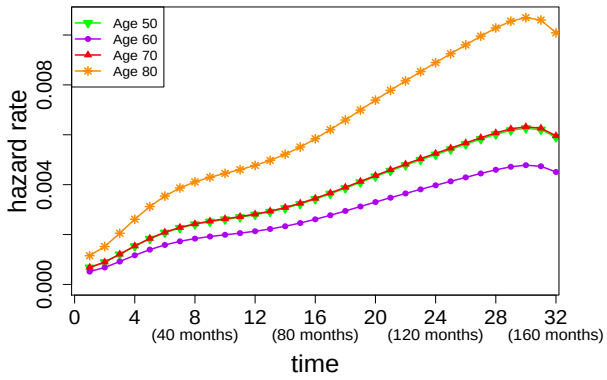

Figure 3.3: The smoothed estimated hazard rates for patients on short-term treatment with Gleason score of 2, aged 50, 60, 70 and 80 years at enrollment, based on model PM55. (The hazard rates for patients on short-term treatment with Gleason score of 2, aged 50 and 70 years at enrollment are almost identical.)

3.4 Comparison to piecewise exponential hazard model

In this section, we run four piecewise exponential hazard models over the same prostate cancer data that we used in Section 3.1:

- **EPEM5:** partition the whole study time evenly to $2^5 = 32$ subintervals, then fit data with piecewise exponential functions based on this partition
- **EPEM4:** partition the whole study time evenly to $2^4 = 16$ subintervals, then fit data with piecewise exponential functions based on this partition

- **QPEM5:** partition the whole study time using $2^5 = 32$ quantiles of available failure times, then fit data with piecewise exponential functions based on this partition
- **QPEM4:** partition the whole study time using $2^5 = 16$ quantiles of available failure times, then fit data with piecewise exponential functions based on this partition

Table 3.2 gives the 95% credible intervals of predictor effects for four multiresolution hazard models discussed Section 3.3.2 and 95% confidence intervals of predictor effects for four piecewise exponential hazard models. The estimates seems very close among all the models and more identical within the same model groups – the multiresolution hazard model group and the piecewise exponential hazard model group, which is consistent with our aforementioned discussions over simulated data. However, using prostate cancer data, the number of subintervals used in model QPEM5 is 32 and in model QPEM4 is 16. This is different from the number of subintervals we get when fitting these two models over simulated data. While we running simulated data sets much fewer subintervals than 16 or 32 were got. Because the prostate cancer data has very different structure which only about 0.6% patients were censored at the end of the study, and also the failure time are measured in a continues scale which means ties are rare. Under this condition, we can expect almost all unique 16 or 32 quantiles yielding 16 or 32 subintervals.

In our PMRH models, we estimate predictor effects together with baseline hazard function and these estimates will affect each other. So when we compare the predictor effects estimated from a PMRH model with a very coarse resolution to one with fine resolution, we can tell there are obviously difference among the estimates. But we also know that when two PMRH models both have although different, but fine enough resolution up to certain level, the predictor effects estimated from these two PMRH models will still be very close. So we claim that pruning will not affect the estimates of predictor effects based on comparing models have fine enough time resolutions.

In Table 3.3 we show the Akaike type information criterion of the four PMRH models and four piecewise exponential hazard models we investigate. The effective number of parameters in

	Treatment (X_1)	Gleason (X_2)	Age (X_3)	Age >70 (X_4)	Age quad (X_5)	Age quad >70 (X_6)
PM52						
2.5%	-0.253	0.083	0.042	-0.518	-0.005	-0.372
50%	-0.112	0.131	0.353	0.097	0.114	-0.186
mean	-0.113	0.131	0.355	0.096	0.113	-0.186
97.5%	0.030	0.179	0.678	0.696	0.226	-0.004
PM55						
2.5%	-0.252	0.082	0.041	-0.481	-0.004	-0.381
50%	-0.112	0.129	0.361	0.099	0.118	-0.191
mean	-0.112	0.130	0.360	0.102	0.116	-0.194
97.5%	0.029	0.180	0.667	0.707	0.221	-0.017
NPM5						
2.5%	-0.250	0.081	0.032	-0.474	-0.005	-0.379
50%	-0.115	0.130	0.357	0.105	0.115	-0.188
mean	-0.113	0.130	0.353	0.107	0.113	-0.189
97.5%	0.028	0.180	0.662	0.717	0.223	-0.003
NPM4						
2.5%	-0.255	0.079	0.039	-0.493	-0.007	-0.377
50%	-0.114	0.130	0.352	0.104	0.114	-0.185
mean	-0.114	0.130	0.353	0.103	0.113	-0.186
97.5%	0.028	0.180	0.669	0.710	0.221	-0.010
EPEM5						
2.5%	-0.254	0.086	0.038	-0.501	0.003	-0.371
MLE	-0.113	0.136	0.356	0.101	0.118	-0.188
97.5%	0.028	0.186	0.673	0.704	0.232	-0.006
QPEM5						
2.5%	-0.255	0.086	0.039	-0.502	0.003	-0.371
MLE	-0.114	0.136	0.356	0.101	0.118	-0.189
97.5%	0.028	0.186	0.67	0.703	0.232	-0.006
EPEM4						
2.5%	-0.254	0.085	0.038	-0.502	0.003	-0.370
MLE	-0.113	0.135	0.355	0.101	0.118	-0.188
97.5%	0.028	0.185	0.673	0.704	0.232	-0.005
QPEM4						
2.5%	-0.254	0.085	0.038	-0.502	0.003	-0.370
MLE	-0.113	0.135	0.355	0.101	0.118	-0.188
97.5%	0.028	0.185	0.673	0.704	0.232	-0.005

Table 3.2: Estimates for the prostate cancer predictor effects (95% credible intervals of predictor effects for four multiresolution hazard models and 95% confidence intervals of predictor effects for four piecewise exponential hazard models)

	-2log(L)	effective number of parameters in the model	AIC calculated
NPM5	9318.092	39	9396.092
NPM4	9331.522	23	9377.522
PM52	9333.798	16	9365.798
PM55	9347.010	13	9373.010
EPEM5	9316.438	38	9392.438
EPEM4	9331.044	22	9375.044
QPEM5	9319.844	38	9395.844
QPEM4	9346.312	22	9390.312

Table 3.3: Akaike type information criterion of pruned multiresolution hazard models and piecewise exponential hazard models

the statistical model is the number of parameters in the statistical model, as in the conventional Akaike information criterion(AIC), when the model is a piecewise exponential hazard model. The baseline cumulative hazard H, the $R_{m,p}$ associated with non-pruned bins and all the covariate effects are counted as effective parameters of a Bayesian PMRH model. And the effective number of parameters is the total number of these parameters plus one. We adjust the sum by one is because the prior of H has two parameters a and λ, where $H = a\lambda$, and if H is known, we just need to know one of a and λ, then we know both of them. And among all these 8 models, model PM52 has the smallest AIC value, which indicates that moderate pruning will definitely balance estimation accuracy and computation cost.

Figure 3.4 shows the MLEs of baseline hazard rates and its smoothed version from the 4 different piecewise exponential hazard models(EPEM5, EPEM4, QPEM5, QPEM4), and the posterior medians of baseline hazard rates and it smoothed version from the 4 different PMRH models (PM55, NPM4, NPM5, and PM52). We can see that the hazard rate is obviously decreasing near the endpoint of the study from the result of model EPEM5 and NPM5. And the estimation result of predictor effects and baseline hazard rate from these two models are more similar in comparison with other models, as showed in Table 3.2 and Figure 3.4. Different from all the other models we are comparing here, they don't have such notable decrements in hazard rate around the termination point of the study. All this is because both EPEM5 and NPM5 models are having 32

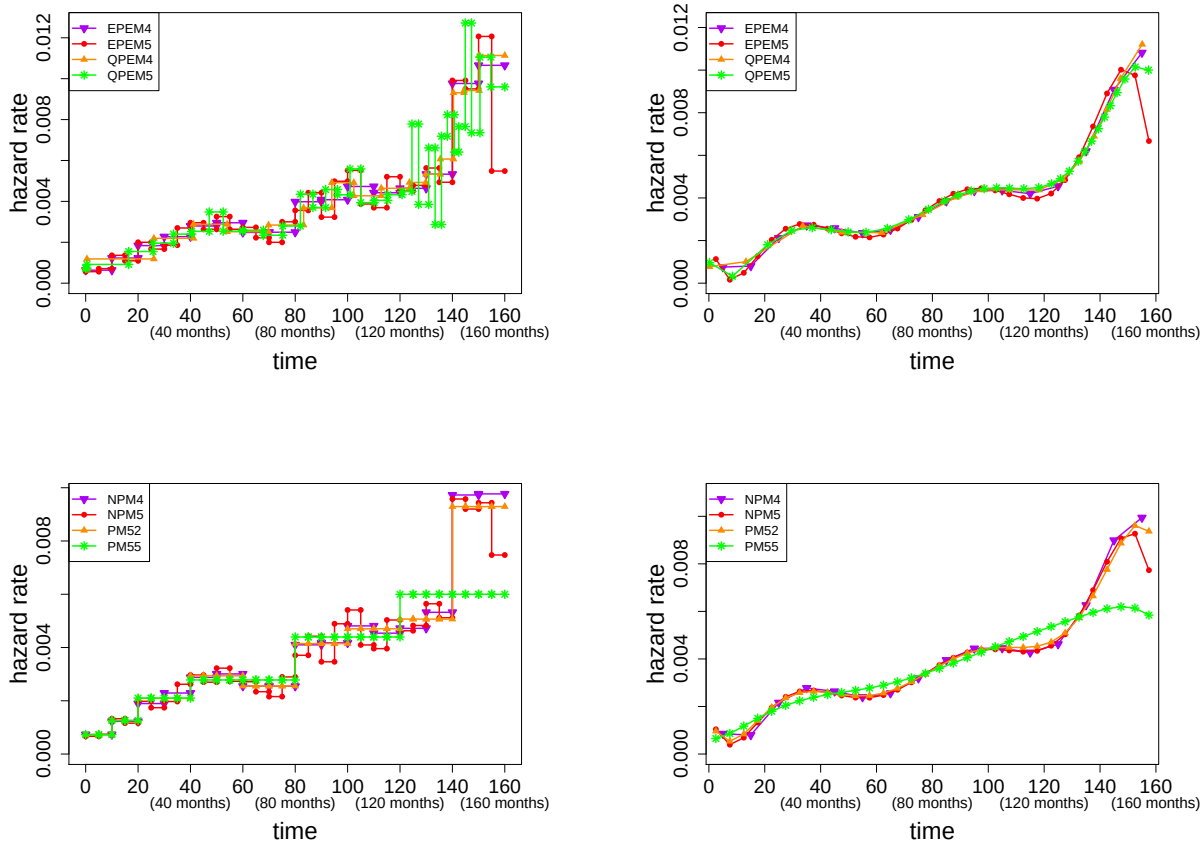

Figure 3.4: MLEs of hazard rates for patients on short-term treatment with Gleason score of 2, aged 70 years at enrollment (top-left) and its smoothed version (top-right), for 4 different piecewise exponential hazard models. Posterior medians of hazard rates for patients on short-term treatment with Gleason score of 2, aged 70 years at enrollment (bottom-left) and its smoothed version (bottom-right), for 4 different PMRH models.

equal-size subintervals and there is only a few failures in the last a few intervals with no pruning is performed over these two models, the hazard rate would drop according to its definition. For the other models, we either have a coarser grid, subintervals partitioned by the failure counts quantiles , or a pruned MRH model. Any of these can resulting in averaging the hazard rate near the endpoint to that from the subinterval next to it. Hence, we won't have apparent changes in hazard rates in the end for those models.

Chapter 4

Hazard models with time-varying covariates

In survival analysis, we always try to estimate how covariates affect hazard function, jointly with the study of hazard function. Nowadays, more and more analysis need to handle time-varying covariates and biomarkers. Covariates are often no longer constants, they are functions of time, and can be both discrete and continuous. In this chapter, we give an overview of hazard models with time-varying covariates, discuss cumulative hazard function estimation in hazard models and talk about different standardization approaches in handling time-varying covariates.

4.1 Hazard models with time-varying covariates

Treatment is one commonly used variable of interest in survival models. Throughout a study, subjects may change treatments due to changes in their physical condition or other factors. If the change in treatment is not incorporated into the survival model, biased parameter estimates may result. Another type of time-varying covariates are biomarkers, which are measured repeatedly or periodically over time. Examples include glucose levels for patients with diabetes, which are measured and recorded routinely in their treatment. Depending on the purpose of study, sometimes only measures at enrollment would be enough for covariate values. But in many cases, the time-span of study will last a few years, such as studies of chronic diseases, then considering lab result changes in modeling fitting, has its necessity.

Hazard function always reflects the underlying process in a study. When covariates influence the process, failure risk associated with process will have corresponding response. In this sense,

some covariates may have instant effect on the risk function. For instance, if a patient's oxygen saturation suddenly drops below normal, this will immediately affect the risk of death. On the other hand, some covariates only indicate how far a process has developed, such as CD4 counts in HIV infection.

Time-varying covariates can also be categorized as internal covariates or external covariates. Internal covariates can only be measured when people are alive, such as lab results and body temperature, to name a few. If a variable changes in a way independent of all individuals, then it is an external covariate, for instance, humidity in the air.

The classic Cox proportional-hazards model can be extended to handle time-varying covariates, and in this way, the model no longer has proportional-hazards as the hazard ratio depends on time-varying covariates now. Advantages and disadvantages of using extended Cox model are reviewed in Fisher and Lin (1999). The Cox model including time-dependent covariates will provide more opportunities for investigating uncertain associations and mechanisms of covariates and survival, such as surrogate outcome analysis. Tumor response is always an ideal outcome in cancer clinical trials, since in many cases the primary endpoint of cancer study, such as death, is undesired. If we investigate the relationship between a tumor shrinking and survival time, we can better know how survival time can be predicted based on tumor response. One drawback of the extended Cox model is that it may not be useful for individual predictive analysis, since we may not have future values of a time-varying covariate. Meanwhile, choosing an inappropriate form for a time-varying covariate can also lead to incorrect estimates. Some biological understanding of the treatment mechanism and the clinical trial can help in choosing functional forms of time-varying covariates and checking the parameter estimates. With a formatted dataset, proportional hazard model with time-varying covariates can directly be processed by SAS, Stata and R.

Fisher and Lin (1999) propose partial likelihood score function to estimate unknown regression parameters of proportional hazard model using time-varying covariates. In this approach, T is the failure time of our interest and Z is the set of possible time-varying covariates. Covariate value at time t is denoted as $Z(t)$ and notation $\overline{Z}(t)$ represents the history of covariate values up

to time t, where $\overline{Z}(t) = \{Z(s) : 0 \leqslant s \leqslant t\}$. As usual, let β be the vector of unknown regression coefficents and $\lambda_0(t)$ be the baseline hazard function. The conditional hazard function of T given $\overline{Z}$ can be rewritten as

$$\lambda(t|\overline{Z}) = \lambda_0(t)e^{\beta' Z(t)} \tag{4.1}$$

Then based on this hazard function, the partial likelihood is

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{e^{\beta' Z_i(X_i)}}{\sum_{j \in R_i} e^{\beta' Z_j(X_i)}} \right)^{\delta_i} \tag{4.2}$$

and the corresponding log partial likelihood is

$$l(\beta) = \sum_{i=1}^{n} \delta_i \left(\beta' Z_i(X_i) - \log \sum_{j \in R_i} e^{\beta' Z_j(X_i)} \right) \tag{4.3}$$

Taking the first derivative of Equation (4.3) with respective to β, we have the partial likelihood score function $U(\beta)$

$$U(\beta) = \sum_{i=1}^{n} \delta_i \left(Z_i(X_i) - \frac{\sum_{j \in R_i} e^{\beta' Z_j(X_i)} Z_j(X_i)}{\sum_{j \in R_i} e^{\beta' Z_j(X_i)}} \right) \tag{4.4}$$

The solution of $U(\beta) = 0$, which serves as the maximum partial likelihood estimator $\hat{\beta}$, is the partial likelihood estimator of β.

Now we talk about some notation used in Equation (4.2), Equation (4.3) and Equation (4.4). n is the total number of units under observation in the study; X_i is the last follow-up time of the i-th unit; indicator $\delta_i = 0$ if the i-th unit is right censored at time X_i, and $\delta_i = 1$ if not; $\overline{Z}_i(t)$ is the history of covariates of the i-th unit up to time t; R_i is the set of units who are at risk at time X_i. Examining Equation (4.4) carefully, we notice that $Z_j(X_i)$ may not have a value if time X_j is less than X_i. In this scenario, data imputation need to be considered.

Although age is time-varying, when using the Cox model to study the effect of age on failure time, we use the age at enrollment as a fixed covariate instead of a time-varying covariate. Because of the property of partial likelihood structure both methods will yield the same estimates in age effects. Let us have a look at the partial likelihood when taking age as a time-varying covariate

and a fixed covariate. Still use the notation in Equation (4.3), for the convenience of illustration, we consider age as the only covariate in the model. For a general case, the derivation steps will be similar. For the i-th subject in the study, we assume the age at enrollment is W_i. And we also assume that all the subjects enroll the study at the same time B. When taking age as a fixed covariate, the partial likelihood Equation (4.2) can be rewritten as

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{e^{\beta Z_i(X_i)}}{\sum_{j \in R_i} e^{\beta Z_j(X_i)}} \right)^{\delta_i} = \prod_{i=1}^{n} \left(\frac{e^{\beta W_i}}{\sum_{j \in R_i} e^{\beta W_j}} \right)^{\delta_i} \tag{4.5}$$

With all the assumptions made above, we have $Z_j(X_i) = W_j + X_i - B$. Therefore, when taking age as a time-varying covariate, the partial likelihood Equation (4.2) can be rewritten as

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{e^{\beta Z_i(X_i)}}{\sum_{j \in R_i} e^{\beta Z_j(X_i)}} \right)^{\delta_i} = \prod_{i=1}^{n} \left(\frac{e^{\beta (W_i + X_i - B)}}{\sum_{j \in R_i} e^{\beta (W_j + X_i - B)}} \right)^{\delta_i} = \prod_{i=1}^{n} \left(\frac{e^{\beta W_i}}{\sum_{j \in R_i} e^{\beta W_j}} \right)^{\delta_i} \tag{4.6}$$

We can see that no matter age is considered as a fixed or a time-varying covariate, the survival model ends up having the same partial likelihood function as showed in Equation (4.5) and Equation (4.6). Therefore, the estimates for age effect will be the same.

In survival analysis, not only covariates can change over time but also covariate effects can depend on time. Motivated by the decay of predictive effect as time goes by, Anderson and Senthilselvan (1982) propose a two-step hazard model extended from Cox model. And this two-step model can be considered as a fixed effects Cox model with time-varying covariates. To allow more feasibility, Gore et al. (1984) introduce a step function proportional hazards model to survival data with fixed covariates and time-varying covariate effects. They are straightforward methods but may cause inefficient parameter estimates if the chosen form of step function is inappropriate. This thesis does not consider time-varying effects.

4.2 Cumulative hazard estimation in extended proportional hazards models

In survival analysis, one of the goals is to study the cumulative hazard function. The cumulative hazard function of a subject is determined as

$$H(t) = \int_0^t \exp(\mathbf{x}(\tau)\boldsymbol{\beta})\lambda_0(\tau)d\tau \tag{4.7}$$

where $\mathbf{x}(t)$ is a vector of momentary covariate value functions along the time, $\boldsymbol{\beta}$ is a vector of corresponding predictor effects, and $\lambda_0(t)$ is the hazard rate function. Here we still assume that all the covariate effects are constant. One of the issues we have is how to evaluate Equation (4.7). Most of the time, we can't evaluate it analytically due to the mathematical form of $\mathbf{x}(t)$ and $\lambda_0(t)$. In our MRH and PMRH models so far we assume that covariates $\mathbf{x}(t)$ are constants through the whole study and hazard function $\lambda_0(t)$ is a piecewise function having constant value of each piece. In this section, we are going to discuss how to approximate Equation (4.7), when covariates $\mathbf{x}(t)$ are time-varying and $\lambda_0(t)$ is still a piecewise constant function.

4.2.1 Cumulative hazard function estimation in models with one predictor

4.2.1.1 Approximation using first order Taylor expansion

First, we consider the simplest case with only one time-varying predictor. Given $x_i(t)$ is the covariate function of the i-th subject, we expand the exponential function at $x_i(t) = a_{is}$, over the interval $[t_{s-1}, t_s]$, where $0 = t_0 < t_1 < t_2 < \ldots < t_{J-1} < t_J$ is a partition of the whole time axis with constant hazard function within each interval and censoring time as t_J.

$$\begin{aligned} I_{is} &= \int_{t_{s-1}}^{t_s} \exp\{x_i(\tau)\beta\}\lambda_0(\tau)d\tau \\ &= \int_{t_{s-1}}^{t_s} \{1 + \beta(x_i(\tau) - a_{is}) + \beta^2(x_i(\tau) - a_{is})^2/2 + \cdots\}e^{a_{is}\beta}\lambda_0(\tau)d\tau \end{aligned} \tag{4.8}$$

For $y_i = \min(T_i, T_J)$, where T_i is the observed failure time or censoring time of the i-th subject, we still assume that we have $t_{j-1} < y_i \leqslant t_j$, for some j, then Equation (4.7) becomes:

$$\begin{aligned} H_i(y_i) = &\sum_{s=1}^{j-1} \int_{t_{s-1}}^{t_s} \{1 + \beta(x_i(\tau) - a_{is}) + \beta^2(x_i(\tau) - a_{is})^2/2 + \cdots\} e^{a_{is}\beta} \lambda_0(\tau) d\tau \\ &- \int_{t_{j-1}}^{y_i} \{1 + \beta(x_i(\tau) - a_{ij}) + \beta^2(x_i(\tau) - a_{ij})^2/2 + \cdots\} e^{a_{ij}\beta} \lambda_0(\tau) d\tau \end{aligned} \tag{4.9}$$

Using first order Taylor expansion,

$$I_{is} \approx \int_{t_{s-1}}^{t_s} \{1 + \beta(x_i(\tau) - a_{is})\} e^{a_{is}\beta} \lambda_0(\tau) d\tau \tag{4.10}$$

and error can be written as

$$\text{error}_{is} = \int_{t_{s-1}}^{t_s} \frac{1}{2!} \beta^2 (x_i(\tau) - a_{is})^2 e^{c(\tau)\beta} \lambda_0(\tau) d\tau \tag{4.11}$$

where $c_{is}(\tau)$ is between $x_i(\tau)$ and a_{is}. Denote m_{is} as the minimal value of $x_i(\tau)$ over interval $[t_{s-1}, t_s]$ and M_{is} as the maximal value of $x_i(\tau)$ over interval $[t_{s-1}, t_s]$. In order to minimize $|\text{error}_{is}|$, it is obvious that constant a should take some value between m_{is} and M_{is}. Consequently, $c_{is}(\tau)$ is bounded by m_{is} and M_{is}. With the assumption of constant hazard rate λ_s on time interval $[t_{s-1}, t_s]$, and predictor effect β positive we have

$$|\text{error}_{is}| \leqslant \frac{1}{2!} \beta^2 \lambda_s e^{M_{is}\beta} \int_{t_{s-1}}^{t_s} (x_i(\tau) - a_{is})^2 d\tau \tag{4.12}$$

When β is negative, the result will be similar. Now we want to minimize $\int_{t_{s-1}}^{t_s} (x_i(\tau) - a_{is})^2 d\tau$ to choose the appropriate constant a_{is} between m_{is} and M_{is}. Then the overall error using first order of Taylor expansion to estimate integral part in Equation (4.9)is bounded as:

$$|\text{error}| \leqslant \frac{1}{2!} \beta^2 \max(\lambda_s) e^{\max(x_i(\tau))\beta} \left\{ \sum_{s=1}^{j-1} \int_{t_{s-1}}^{t_s} (x_i(\tau) - a_{is})^2 d\tau + \int_{t_{j-1}}^{y_i} (x_i(\tau) - a_{ij})^2 d\tau \right\} \tag{4.13}$$

Let

$$G(a_{is}) = \int_{t_{s-1}}^{t_s} (x_i(\tau) - a_{is})^2 d\tau \tag{4.14}$$

then our goal is to find some constant a_{is}, such that the value of function $G(a_{is})$ is as small as possible. As

$$G'(a_{is}) = \int_{t_{s-1}}^{t_s} -2(x_i(\tau) - a_{is})d\tau \tag{4.15}$$

it can be verified that

$$a_{is} = \frac{1}{t_s - t_{s-1}} \int_{t_{s-1}}^{t_s} x_i(\tau)d\tau \tag{4.16}$$

makes $G'(a_{is})$ equal to 0. And

$$G''(a_{is}) = 2(t_s - t_{s-1}) > 0 \tag{4.17}$$

so Equation (4.14) reaches its minimal value

$$\begin{aligned} G(\frac{1}{t_s - t_{s-1}} \int_{t_{s-1}}^{t_s} x_i(\tau)d\tau) &= \int_{t_{s-1}}^{t_s} (x_i(\tau) - \frac{1}{t_s - t_{s-1}} \int_{t_{s-1}}^{t_s} x_i(\tau)d\tau)^2 d\tau \\ &= \int_{t_{s-1}}^{t_s} x_i^2(\tau)d\tau - \frac{1}{t_s - t_{s-1}} \left(\int_{t_{s-1}}^{t_s} x_i(\tau)d\tau \right)^2 \end{aligned} \tag{4.18}$$

at $a_{is} = \dfrac{1}{t_s - t_{s-1}} \displaystyle\int_{t_{s-1}}^{t_s} x_i(\tau)d\tau$.

Then overall error bound Equation (4.13) becomes

$$|\text{error}| \leqslant \frac{1}{2!}\beta^2 \max(\lambda_0(\tau)) e^{\max(x_i(\tau))\beta} \left\{ \int_0^{y_i} x_i^2(\tau)d\tau - \sum_{s=1}^{j-1} \frac{\left(\int_{t_{s-1}}^{t_s} x_i(\tau)d\tau \right)^2}{t_s - t_{s-1}} - \frac{\left(\int_{t_{j-1}}^{y_i} x_i(\tau)d\tau \right)^2}{y_i - t_{j-1}} \right\} \tag{4.19}$$

From Cauchy-Schwartz inequality, we know that the right side of Equation (4.19) is always non-negative. As $a_{is} = \dfrac{1}{t_s - t_{s-1}} \displaystyle\int_{t_{s-1}}^{t_s} x_i(\tau)d\tau$ is the average of $x_i(t)$ over interval $[t_{s-1}, t_s]$, we can see in order to make the right side of Equation (4.13) small, we need the L^2-norm of the difference of $x_i(t)$ and its mean over interval $[t_{s-1}, t_s]$ to be small.

Now we are going to illustrate two examples. First one is when covariate function $x_i(t)$ is a piecewise linear function. Assume $x_i(t)$ is linear over interval $[t_{s-1}, t_s]$, denoting as $x_i(t) = p_{is}t + q_{is}$, when take

$$a_{is} = \frac{1}{t_s - t_{s-1}} \int_{t_{s-1}}^{t_s} x_i(\tau)d\tau = \frac{1}{t_s - t_{s-1}} \int_{t_{s-1}}^{t_s} p_{is}\tau + q_{is}d\tau \tag{4.20}$$

Using Equation (4.9), the cumulative hazard function for the i-th subject can be approximated as

$$\begin{aligned} H_i(y_i) = &\sum_{s=1}^{j-1}(t_s - t_{s-1})\exp\{(\frac{p_{is}(t_s + t_{s-1})}{2} + q_{is})\beta\}\lambda_s \\ &- (y_i - t_{j-1})\exp\{(\frac{p_{ij}(y_i + t_{j-1})}{2} + q_{ij})\beta\}\lambda_j \end{aligned} \tag{4.21}$$

the over all error bound Equation (4.19) becomes

$$|\text{error}| \leqslant \frac{1}{2!}\beta^2 \max(\lambda_s)e^{\max(x_i(\tau))\beta}\left\{\sum_{s=1}^{j-1}\frac{p_{is}^2(t_s - t_{s-1})^3}{6} + \frac{p_{ij}^2(y_i - t_{j-1})^3}{6}\right\} \tag{4.22}$$

We can see that the closer to zero of the slope p_{is} within each interval $[t_{s-1}, t_s]$, the better approximation we can get.

Second example is when the covariate is a piecewise quadratic function. Assume $x_i(t)$ is quadratic over interval $[t_{s-1}, t_s]$, denoting as $x_i(t) = p_{is}t^2 + q_{is}t + w_{is}$, when take

$$a_{is} = \frac{1}{t_s - t_{s-1}}\int_{t_{s-1}}^{t_s} x_i(\tau)d\tau = \frac{1}{t_s - t_{s-1}}\int_{t_{s-1}}^{t_s} p_{is}\tau^2 + q_{is}\tau + w_{is}d\tau \tag{4.23}$$

This time, the cumulative hazard function for the i-th subject can be approximated as

$$\begin{aligned} H_i(y_i) = &\sum_{s=1}^{j-1}(t_s - t_{s-1})\exp\{(\frac{p_{is}(t_s^2 + t_s t_{s-1} + t_{s-1}^2)}{3} + \frac{q_{is}(t_s + t_{s-1})}{2} + w_{is})\beta\}\lambda_s \\ &- (y_i - t_{j-1})\exp\{(\frac{p_{ij}(y_i^2 + y_i t_{j-1} + t_{j-1}^2)}{3} + \frac{q_{ij}(y_i + t_{j-1})}{2} + w_{ij})\beta\}\lambda_j \end{aligned} \tag{4.24}$$

4.2.1.2 Approximation using second order Taylor expansion

Following the same fashion from last section, now we use second order Taylor expansion to approximate the cumulative hazard function. With all the same assumptions, Equation (4.10) now becomes

$$I_{is} \approx \int_{t_{s-1}}^{t_s}\{1 + \beta(x_i(\tau) - a_{is}) + \beta^2(x_i(\tau) - a_{is})^2/2\}e^{a_{is}\beta}\lambda_0(\tau)d\tau \tag{4.25}$$

and error can be written as

$$\text{error}_{is} = \int_{t_{s-1}}^{t_s}\frac{1}{3!}\beta^3(x_i(\tau) - a_{is})^3 e^{c_{is}(\tau)\beta}\lambda_0(\tau)d\tau \tag{4.26}$$

where $c_{is}(t)$ is between $x_i(t)$ and a_{is}. Denote m_{is} as the minimal value of $x_i(t)$ over interval $[t_{s-1}, t_s]$ and M_{is} as the maximal value of $x_i(t)$ over interval $[t_{s-1}, t_s]$. In order to minimize $|\text{error}_{is}|$, it is obvious that constant a_{is} should take some value between m_{is} and M_{is}. Consequently, $c_{is}(t)$ is bounded by m_{is} and M_{is}. Still assuming constant hazard rate λ_s on time interval $[t_{s-1}, t_s]$, then

$$|\text{error}_{is}| \leqslant \frac{1}{3!}\beta^3\lambda_s e^{M_{is}\beta}\int_{t_{s-1}}^{t_s}|(x_i(\tau)-a_{is})^3|d\tau \tag{4.27}$$

Here we assume β to be positive(the negative case will be very similar). Now we want to minimize $\int_{t_{s-1}}^{t_s}|(x_i(\tau)-a_{is})^3|d\tau$ by choosing the appropriate constant a_{is} between m_{is} and M_{is}. Then the overall error using first order Taylor expansion to estimate integral part in Equation (4.9) is bounded as:

$$|\text{error}| \leqslant \frac{1}{3!}\beta^3 \max(\lambda_s)e^{\max(x_i(\tau))\beta}\left\{\sum_{s=1}^{j-1}\int_{t_{s-1}}^{t_s}|(x_i(\tau)-a_{is})^3|d\tau + \int_{t_{j-1}}^{y_i}|(x_i(\tau)-a_{ij})^3|d\tau\right\} \tag{4.28}$$

Let $D_1 = \{t \in [t_{s-1}, t_s]|x_i(t) \geqslant a_{is}\}$ and $D_2 = \{t \in [t_{s-1}, t_s]|x_i(t) < a_{is}\}$, then

$$\int_{t_{s-1}}^{t_s}|(x_i(\tau)-a_{is})^3|d\tau = \int_{D_1}(x_i(\tau)-a_{is})^3 d\tau + \int_{D_2}(a_{is}-x_i(\tau))^3 d\tau \tag{4.29}$$

Denote Equation (4.29) as $G(a_{is})$, then

$$G'(a_{is}) = \int_{D_1} -3(x_i(\tau)-a_{is})^2 d\tau + \int_{D_2} 3(a_{is}-x_i(\tau))^2 d\tau \tag{4.30}$$

In order to have $G'(a_{is}) = 0$, a needs to be the solution of

$$\int_{D_1}(x_i(\tau)-a_{is})^2 d\tau = \int_{D_2}(a_{is}-x_i(\tau))^2 d\tau \tag{4.31}$$

When we have enough assumptions about $x_i(\tau)$, such as monotonic, or continuous, Equation (4.31) can be solved numerically to minimize the difference between the left side and right side. And as

$$G''(a_{is}) = \int_{D_1} 6(x_i(\tau)-a_{is})d\tau + \int_{D_2} 6(a_{is}-x_i(\tau))d\tau > 0, \tag{4.32}$$

the critical point we find from Equation (4.31) will serve as a minimizer of function $G(a_{is})$.

4.2.2 Cumulative hazard function estimation in models with multiple predictors

In this section, we will discuss the cumulative hazard function estimation in models with multiple predictors. Recall that Taylor expansion for scalar-valued function of more than one variable can be written compactly as:

$$f(\mathbf{x}) = f(\mathbf{a}) + Df(\mathbf{a})^T(\mathbf{x}-\mathbf{a}) + \frac{1}{2!}(\mathbf{x}-\mathbf{a})^T D^2 f(\mathbf{a})(\mathbf{x}-\mathbf{a}) + \cdots \tag{4.33}$$

where $Df(\mathbf{a})$ is the gradient of f evaluated at $\mathbf{x}=\mathbf{a}$ and $D^2f(\mathbf{a})$ is the Hessian matrix. If we let $\mathbf{x} = (x_1, \ldots, x_p)^T$ and $\mathbf{a} = (a_1, a_2, \ldots, a_p)^T$, then we can rewrite Equation (4.33) in Σ notation as

$$f(\mathbf{x}) = f(\mathbf{a}) + \sum_{j=1}^{p} \frac{\partial f(\mathbf{a})}{\partial x_j}(x_j - a_j) + \frac{1}{2!}\sum_{j=1}^{p}\sum_{k=1}^{p} \frac{\partial^2 f(\mathbf{a})}{\partial x_j \partial x_k}(x_j - a_j)(x_k - a_k) + \cdots \tag{4.34}$$

When we have multiple covariates in our extended proportional hazard model, with vector $\mathbf{x}_i(t) = (x_{i1}(t), \ldots, x_{ip}(t))^T$ as the covariates for the i-th subject and vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ as the associated constant predictor effect, we again expand, I_{is}, the cumulative hazard increment for the i-th subject over interval $[t_{s-1}, t_s]$ at $\mathbf{a_{is}} = (a_{is1}, a_{isp}, \ldots, a_{isp})^T$, the we have

$$\begin{aligned} I_{is} &= \int_{t_{s-1}}^{t_s} \exp\{\mathbf{x}_i^T(\tau)\boldsymbol{\beta}\}\lambda_0(\tau)d\tau \\ &= \int_{t_{s-1}}^{t_s} [e^{\mathbf{a_{is}}'\boldsymbol{\beta}} + e^{\mathbf{a_{is}}'\boldsymbol{\beta}}(\mathbf{x}_i(\tau) - \mathbf{a_{is}})'\boldsymbol{\beta} + \frac{e^{\mathbf{a_{is}}'\boldsymbol{\beta}}}{2}(\mathbf{x}_i(\tau) - \mathbf{a_{is}})'\mathbf{W}(\mathbf{x}_i(\tau) - \mathbf{a_{is}}) + \cdots]\lambda_0(\tau)d\tau \\ &= \int_{t_{s-1}}^{t_s} [1 + (\mathbf{x}_i(\tau) - \mathbf{a_{is}})'\boldsymbol{\beta} + \frac{1}{2}(\mathbf{x}_i(\tau) - \mathbf{a_{is}})'\mathbf{W}(\mathbf{x}_i(\tau) - \mathbf{a_{is}}) + \cdots]e^{\mathbf{a_{is}}'\boldsymbol{\beta}}\lambda_0(\tau)d\tau \end{aligned} \tag{4.35}$$

where

$$\mathbf{W} = \begin{bmatrix} \beta_1\beta_1 & \beta_1\beta_2 & \cdots & \beta_1\beta_p \\ \beta_2\beta_1 & \beta_2\beta_2 & \cdots & \beta_2\beta_p \\ \vdots & \vdots & \ddots & \vdots \\ \beta_p\beta_1 & \beta_p\beta_2 & \cdots & \beta_p\beta_p \end{bmatrix} \tag{4.36}$$

Using the assumptions and notations as in Equation (4.9), the cumulative hazard function for the

i-th subject becomes:

$$\begin{aligned} H_i(y_i) =& \sum_{s=1}^{j-1} \int_{t_{s-1}}^{t_s} [1 + (\mathbf{x}_i(\tau) - \mathbf{a_{is}})^T \boldsymbol{\beta} + \frac{1}{2}(\mathbf{x}_i(\tau) - \mathbf{a_{is}})^T \mathbf{W}(\mathbf{x}_i(\tau) - \mathbf{a_{is}}) + \cdots] e^{\mathbf{a_{is}}^T \boldsymbol{\beta}} \lambda_0(\tau) d\tau \\ & - \int_{t_{j-1}}^{y_i} [1 + (\mathbf{x}_i(\tau) - \mathbf{a_{ij}})^T \boldsymbol{\beta} + \frac{1}{2}(\mathbf{x}_i(\tau) - \mathbf{a_{ij}})^T \mathbf{W}(\mathbf{x}_i(\tau) - \mathbf{a_{ij}}) + \cdots] e^{\mathbf{a_{ij}}^T \boldsymbol{\beta}} \lambda_0(\tau) d\tau \end{aligned} \tag{4.37}$$

As we have done before, with the same assumption that hazard rate is constant λ_s over interval $[t_{s-1}, t_s]$, if we only use the first order Taylor expansion in Equation (4.35), then it will yield

$$\begin{aligned} H_i(y_i) &\approx \sum_{s=1}^{j-1} \int_{t_{s-1}}^{t_s} [1 + (\mathbf{x}_i(\tau) - \mathbf{a_{is}})^T \boldsymbol{\beta}] e^{\mathbf{a_{is}}^T \boldsymbol{\beta}} \lambda_0(\tau) d\tau - \int_{t_{j-1}}^{y_i} [1 + (\mathbf{x}_i(\tau) - \mathbf{a_{ij}})^T \boldsymbol{\beta}] e^{\mathbf{a_{ij}}^T \boldsymbol{\beta}} \lambda_0(\tau) d\tau \\ &= \sum_{s=1}^{j-1} e^{\mathbf{a_{is}}^T \boldsymbol{\beta}} \lambda_s \int_{t_{s-1}}^{t_s} [1 + \sum_{l=1}^{p} (x_{il}(\tau) - a_{isl}) \beta_l] d\tau - e^{\mathbf{a_{ij}}^T \boldsymbol{\beta}} \lambda_j \int_{t_{j-1}}^{y_i} [1 + \sum_{l=1}^{p} (x_{il}(\tau) - a_{ijl}) \beta_l] d\tau \end{aligned} \tag{4.38}$$

Moreover, for $l = 1, 2, \ldots, p$, when we take $a_{isl} = \dfrac{1}{t_s - t_{s-1}} \displaystyle\int_{t_{s-1}}^{t_s} x_{il}(\tau) d\tau$, $s = 1, 2, \ldots, j-1$ and $a_{ijl} = \dfrac{1}{y_i - t_{j-1}} \displaystyle\int_{t_{s-1}}^{y_i} x_{il}(\tau) d\tau$, Equation (4.38) becomes

$$H_i(y_i) \approx \sum_{s=1}^{j-1} e^{\mathbf{a}'_{is} \boldsymbol{\beta}} \lambda_s (t_s - t_{s-1}) - e^{\mathbf{a}'_{ij} \boldsymbol{\beta}} \lambda_j (y_i - t_{j-1}) \tag{4.39}$$

Equation (4.39) gives us a more general formula for cumulative hazard function estimation using first order Taylor expansion. When a covariate is time-varying, we can use its average value over a interval to represent the value for this covariate within that interval in the process of approximating cumulative hazard function.

4.3 Standardization of time-varying covariates

Sometimes, covariate values in a dataset can vary from widely. If we don't standardize them before using them in models, numerical issues, such as blow-up or running out of memory

during model fitting procedure, may occur. When a covariate is time-varying, there are several ways of standardizing. In this section, we give a brief discussion of different ways of standardizing time-varying covariates.

For the convenience of discussion, let us introduce some notation. For the i-th subject, we denote the observed failure time as T_i, the corresponding censoring time as C_i and y_i as the smaller one between T_i and C_i, i.e. $y_i = \min(T_i, C_i)$. We partition the whole time axis into J intervals, $0 = t_0 < t_1 < t_2 < \ldots < t_{J-1} < t_J$, with t_J no less than the maximum of y_i. $\lambda_k(t)$ is the hazard rate function within each interval $(t_{k-1}, t_k]$. As usual, the covariate effects is denoted as $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$. And vector $\mathbf{x}_{ik} = (x_{i1k}, \ldots, x_{ipk})'$, $k = 1, 2, \ldots, J$ represents the observed covariate values in the k-th time interval for i-th subject. To make it clear enough, x_{ipk} is the observed value associated to p-th covariate in the k-th time interval for the i-th subject. The same as other chapters in this dissertation, we only consider constant covariate effects, which means $\boldsymbol{\beta}$ doesn't change over time.

First, we could standardize all the covariates related to the l-th predictor with the same mean μ_l and sample standard deviation σ_l for all the subjects, where μ_l and σ_l are calculated all units in the data set. Then we have the standardized covariate values for the i-th unit in the k-th interval associated to all predictors as

$$\mathbf{x}^*_{ik} = (\frac{x_{i1k} - \mu_1}{\sigma_1}, \ldots, \frac{x_{ipk} - \mu_p}{\sigma_p}) \tag{4.40}$$

In extended proportional hazard models, when we directly use original(non-standardized) data $\mathbf{x}_{ik}$, we denote the corresponding baseline hazard function as $\lambda(t)$ and covariate effects as $\boldsymbol{\beta}$; and when we use the standardized $\mathbf{x}^*_{ik}$ as showed in Equation (4.40), we denote the corresponding baseline hazard function as $\lambda^*(t)$ and covariate effects as $\boldsymbol{\beta}^*$. Considering the hazard function as a given time t, we have the following equation

$$\lambda(t)\exp(\mathbf{x}'_{ik}\boldsymbol{\beta}) = \lambda^*(t)\exp((\mathbf{x}^*_{ik})'\boldsymbol{\beta}^*) \tag{4.41}$$

Substitute $\mathbf{x}^*_{ik}$ in Equation (4.41) with Equation (4.40), then we have

$$\begin{cases} \beta_l^* = \beta_l \sigma_l \\ \lambda^*(t) = \lambda(t) \exp(\mu_1 \beta_1 + \ldots + \mu_p \beta_p) \end{cases} \tag{4.42}$$

where $l = 1, 2, \ldots, p$.

From Equation (4.42), we can tell that if covariates are standardized in this way, the baseline resulting from using standardized data will remain the same along the whole study and we can easily figure out the baseline hazard function and predictor effects from using standardized data provided the original baseline hazard function, predictor effects and standardization information are given, or the vice versa. More generally, the baseline will always keep unchanged over time, as long as for a covariate, all the values associated to it from all the subjects in the data set are standardized by subtracting the same number and dividing by the same number. This method of standardization can provide us a time independent baseline, which is mostly used in survival analysis. Moreover, we know that the baseline is only determined by the covariate means used in standardization procedure. As long as these means keep the same, any selection of sample standard deviations in performing standardization won't change the baseline. So we can always scale the data such that extreme values exceeding the the computer memory capacity will not be produced in numerical simulations. This will greatly reduce the numerical issues while running simulations and in analyzing real data. With this advantage, no matter the covariates are time-varying or not in a data set, we always use this way to standardize data.

Secondly, we could standardize all the covariates related to the l-th predictor in the k-th time interval with the same mean μ_{lk} and sample standard deviation σ_{lk} for all the subjects. In this way, we have the covariates for the i-th unit in the k-th time interval after standardization as

$$\hat{\mathbf{x}}_{ik} = (\frac{x_{i1k} - \mu_{1k}}{\sigma_{1k}}, \ldots, \frac{x_{ipk} - \mu_{pk}}{\sigma_{pk}}) \tag{4.43}$$

Again, we still denote baseline hazard function as $\lambda(t)$ and covariate effects as $\boldsymbol{\beta}$ when we directly use $\mathbf{x}_{ik}$ in the extended proportional hazard model; and when we use the standardized $\hat{\mathbf{x}}_{ik}$, we denote baseline hazard function as $\hat{\lambda}(t)$ and covariate effects as $\hat{\boldsymbol{\beta}}$. As defined already, $\lambda(t) = \lambda_k(t)$

and $\hat{\lambda}(t) = \hat{\lambda}_k(t)$ within each interval $(t_{k-1}, t_k]$. Similarly, we have equation

$$\lambda(t)\exp(\mathbf{x}_{ik}'\boldsymbol{\beta}) = \hat{\lambda}(t)\exp(\hat{\mathbf{x}}_{ik})'\hat{\boldsymbol{\beta}}) \tag{4.44}$$

Substitute $\hat{\mathbf{x}}_{ik}$ in Equation (4.44) with Equation (4.43), then we find that in order to have $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ be constant over time, σ_{lk} must equal to 1, for $l = 1, 2, \ldots, p$ and $k = 1, 2, \ldots, J$. The baseline hazard function and predictor effects before and after standardizing the data are related to each other as following:

$$\begin{cases} \hat{\beta}_l = \beta_l \\ \hat{\lambda}_k(t) = \lambda_k(t)\exp(\mu_{1k}\beta_1 + \ldots + \mu_{pk}\beta_p) \end{cases} \tag{4.45}$$

over interval $(t_{k-1}, t_k]$, where $l = 1, 2, \ldots, p$ and $k = 1, 2, \ldots, J$. In this case, we can note that instead of having a time independent baseline along the whole study, we have a piecewise baseline resulting from standardization. Baseline for the standardized data only remain unchanged within each time interval. This may cause us trouble in interpreting our results. Also, with the restriction of having all the sample standard deviation used in standardization step to be 1, we then can only center the data using their corresponding means. The ranges of the covariates will remain unchanged. As discussed in the first approach, this may cause numerical issues if the original covariates have wide ranges. But this way of standardization may possibly be useful in handling piecewise covariates, whose value doesn't change much within each interval, but change a lot among different intervals over the time.

Chapter 5

MRH models and PMRH models with time-varying covariates

In this chapter, we first talk about how MRH models can be extended to handling time-varying covariates. Then we evaluate MRH models and PMRH models with time-varying covariates using simulated datasets. The simulated datasets that we use here consist of three different types. In each observation, we include a fixed covariate and a time-varying covariate. The time-varying covariates are generated from linear functions, five degree polynomials or cosine like functions. Our results show that the MRH models perform quite well and efficiently with time-varying covariates. In the end of this chapter, we demonstrate some results of fitting the same datasets to piecewise exponential hazard function.

5.1 MRH model with time-varying covariates

5.1.1 Time-varying covariates in MRH models

In MRH models, we need to have at least one observation of the covariate of interest in order to investigate how that covariate affects the hazard function. MRH models have already been studied with constant covariates in Bouman et al. (2005), Bouman et al. (2007), Dukic and Dignam (2007), and Dignam et al. (2009). When it comes to time-varying covariates, there are different scenarios. One scenario is, for a time-varying covariate, we have one and only one observation in each bin of the finest level of a MRH model. In Section 5.1.2, and Section 5.1.3 we will implement MRH model with time-varying covariates in this scenario. The other scenario is we have no observation or multiple observations of a time-varying covariate within a bin. Unlike constant covariates, time-

varying covariates can be observed at any time to some extent. For example, the red blood cell counts of a patient with leukemia is repeatedly tested when the disease is diagnosed. But when a patient progresses to a later stage, we can expect that blood tests will be performed more frequently than when he is in an early stage. So we may have no observation of red blood cell counts for a period a time we picked, and also we may have more than one result for a time period. In Chapter 7 we are going to discuss approaches to analyse data sets with missing time-varying covariate values.

5.1.2 Posteriors of parameters of MRH models with time-varying covariates

As discussed in Section 5.1.1, now we derive the posteriors of parameters of a five level MRH model with N_β time-varying covariates, provided that for any time-varying covariate, each bin in the finest level contains one observation of it. If a covariate is a constant, we just need to assign the same value over all the bins. For a $M = 5$ level model, we have $2^M = 32$ time intervals(bins) at the finest level, based on our model structure. Intuitively, we index those bins with integers starting from 1 and ending in 2^M. We format covariate matrix of the i-th subject, X_i as following:

$$X_i = \begin{array}{c} \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{N_\beta} \end{array} \begin{array}{c} \begin{array}{cccc} \text{bin1} & \text{bin2} & \cdots & \text{bin32} \end{array} \\ \begin{pmatrix} * & * & \cdots & * \\ * & * & \cdots & * \\ \cdots & \cdots & \cdots & \cdots \\ * & * & \cdots & * \end{pmatrix} \end{array}$$

X_i is a N_β by 2^M matrix with entry $X_i[j, l]$ denoting the observed value associated to the j-th covariate in the l-th bin for the i-th unit. Some more notations are introduced for the convenience of discussion:

N_i : the number of bin which the i-th subject falls into, i.e. the bin contains failure time T_i. If the patient is censored when the study ends, $N_i = 2^M$.

ratio_i : the portion of hazard increment the i-th subject takes in bin N_i. For instance, with bin width as 1, $M = 2$, $T_i = 3.75$, we will have $N_i = 4$ and $\text{ratio}_i = \dfrac{3.75 - (N_i - 1) \times 1}{1} =$

$\frac{3.75-(4-1)\times 1}{1} = 0.75$. When the i-th subject is censored as the study ends, we will have $\text{ratio}_i = 1$.

δ_i : censoring indicator, $\delta_i = 1$ means the i-th subject is not censored and $\delta_i = 0$ means the i-th subject is censored.

F_l : F_l is defined as $F_l = \frac{d_l}{H}$ or $d_l = HF_l$, where d_l is the l-th hazard increment of the finest level, and from our model structure, we know F_l is a product of some $R_{m,p}$ and $1 - R_{m,p}$

β: $\beta = (\beta_1, \beta_2, \ldots, \beta_{N_\beta})$ is the vector of predictor effects.

When we consider the the baseline cumulative hazard function H with gamma prior $\mathcal{G}a(a, \lambda)$, the posterior of H is then proportional to a gamma density:

$$\mathcal{G}a(\sum_{i=1}^{N}\delta_i + a, \frac{1}{\frac{1}{\lambda} + \sum_{i=1}^{N}(\sum_{l=1}^{N_i-1}\exp(\beta X_i[\ ,l])F_l + \exp(\beta X_i[\ ,N_i])F_{N_i}\text{ratio}_i)}) \tag{5.1}$$

$$\text{with mean}= \frac{\sum_{i=1}^{N}\delta_i + a}{\frac{1}{\lambda} + \sum_{i=1}^{N}(\sum_{l=1}^{N_i-1}\exp(\beta X_i[\ ,l])F_l + \exp(\beta X_i[\ ,N_i])F_{N_i}\text{ratio}_i)}$$

The following is the log full conditional distribution for $R_{m,p}$ (conditioning on all other model parameter $R^-_{m,p}$), and notation $R^-_{m,p}$ is used to denote all other parameters and data except for $R_{m,p}$ itself.

$$\begin{aligned}&\sum_{i=1}^{N}\delta_i\log(F_{N_i}) - \sum_{i=1}^{N}(\sum_{l=1}^{N_i-1}\exp(\beta X_i[\ ,l])HF_l + \exp(\beta X_i[\ ,N_i])HF_{N_i}\text{ratio}_i)\\ &+(2\gamma_{m,p}k^m a - 1)\log(R_{m,p}) + (2(1-\gamma_{m,p})k^m a - 1)\log(1-R_{m,p})\end{aligned} \tag{5.2}$$

We place a normal prior with mean μ_{β_j} and variance σ_j^2 on β_j, yielding the log full conditional distribution for β_j as

$$\sum_{i=1}^{N}\delta_i(\beta X_i[\ ,N_i]) - \sum_{i=1}^{N}(\sum_{l=1}^{N_i-1}\exp(\beta X_i[\ ,l])HF_l + \exp(\beta X_i[\ ,N_i])HF_{N_i}\text{ratio}_i) - \frac{(\beta_j-\mu_{\beta_j})^2}{2\sigma_j^2} \tag{5.3}$$

When we consider the hyperprior of k is exponential distributed with mean μ_k, we have the

full conditional distribution for k.

$$\pi(k|k^-) \propto \Pi_{m=1}^{M}\Pi_{p=0}^{2^{m-1}-1}\left\{\frac{R_{m,p}^{2\gamma_{m,p}k^m a}(1-R_{m,p})^{2(1-\gamma_{m,p})k^m a}}{\mathcal{B}e(2\gamma_{m,p}k^m a, 2(1-\gamma_{m,p})k^m a)}\right\} e^{-\frac{k}{\mu_k}} \tag{5.4}$$

A zero-truncated Poisson hyperprior $\frac{e^{-\mu_a}\mu_a^a}{a!\,(1-e^{-\mu_a})}$ is chosen for a for computational convenience. In most practical cases, integer shape parameters can serve the goal. Hence, the full conditional distribution for a is

$$\pi(a|a^-) \propto \left\{\frac{1}{\lambda^a\Gamma(a)}H^a\right\}\Pi_{m=1}^{M}\Pi_{p=0}^{2^{m-1}-1}\left\{\frac{R_{m,p}^{2\gamma_{m,p}k^m a}(1-R_{m,p})^{2(1-\gamma_{m,p})k^m a}}{\mathcal{B}e(2\gamma_{m,p}k^m a, 2(1-\gamma_{m,p})k^m a)}\right\}\frac{\mu_a^a}{a!} \tag{5.5}$$

For the scale parameter λ of our cumulative hazard function H, we choose an exponential distribution with mean μ_λ, resulting in

$$\pi(\lambda|\lambda^-) \propto \frac{1}{\lambda^a}e^{-\frac{H}{\lambda}}e^{-\frac{\lambda}{\mu_\lambda}} \tag{5.6}$$

A Beta prior with shape u and w is placed to $\gamma_{m,p}$. So the full conditional distribution for $\gamma_{m,p}$ is proportional to

$$\frac{R_{m,p}^{2\gamma_{m,p}k^m a}(1-R_{m,p})^{2(1-\gamma_{m,p})k^m a}}{\mathcal{B}e(2\gamma_{m,p}k^m a, 2(1-\gamma_{m,p})k^m a)}\gamma_{m,p}^{u-1}(1-\gamma_{m,p})^{w-1} \tag{5.7}$$

5.1.3 Model fitting

Our algorithm is implemented via Gibbs sampler steps (G-step). And the whole procedure is very similar to model fitting of PMRH models as we discussed in Chapter 2. The Gibbs sampler steps (Geman and Geman 1984) for the parameters H, $R_{m,p}$, a, λ, β's, $\gamma_{m,p}$ and k:

(1) Draw H from its full conditional posterior density (5.1)

(2) Draw the $R_{m,p}$ only those are rejected in H-step, (in any order) from its density(5.2)

(3) Draw λ from $\pi(\lambda|\lambda^-)$, a from $\pi(a|a^-)$, β_j from $\pi(\beta_j|\beta_j^-)$, $\gamma_{m,p}$ from $\pi(\gamma_{m,p}|\gamma_{m,p}^-)$ and k from $\pi(k|k^-)$, as described in Section 5.1.2.

In our simulations, we fix $k = 0.5$ and $\gamma_{m,p} = 0.5$. The conditional posterior distributions for H, is Gamma. The full conditionals for $R_{m,p}$, and each β_j are log-concave and are therefore sampled via the adaptive rejection sampling (ARS) algorithm of Gilks and Wild (1992). Since the hyperparameters λ is in general not log-concave, we use adaptive rejection Metropolis sampling (ARMS) of Gilks et al. (1995). ARMS is known as an extension of ARS.

5.2 Simulation of time-varying predictors

In this section we present the methodology of simulating failure time T_i with time-varying covariates. As a reminder, our covariate effects β_j's are all constants here. We use a two-level multiresolution hazard model to illustrate the idea. Without loss of generality, we only consider two covariates, one is a constant and the other is time-varying. For the sake of discussion, we let the constant covariate be the indicator of gender, and let the time-varying one be glucose values. If a subject is a male, we will set the indicator as 1. Otherwise, we will set the indicator as 0 for a female subject. We denote the i-th subject's gender indicator as Gen_i. As described in Section 5.1.2, for a subject, we need to simulate 4 glucose values in a two-level MRH model, one value per bin, and we use $G_{i,j}$ to denote the glucose value of i-th patient observed in the j-th bin. The baseline group here consists of patients with all $G_{i,j} = 0$ and $\text{Gen}_i = 0$. Here, all these $G_{i,j}$'s are values after standardization. In practice, we always standardize the original data of a covariate with its sample mean and sample standard deviation before use them in a MRH model.

First we pick a set of values for the baseline cumulative hazard H and the splits $R_{m,p}$'s. Then we can calculate hazard increments d_1 to d_4 via equations discussed in Section 1.3.1. And this step only need to be performed once. All the values here are taken as true values of parameters of our MRH model and will be used to simulate failure times from now on. And we also pick the terminate time of the study, the covariate effect of gender–β_{gen} and the covariate effect of glucose–β_{glu}.

Secondly, we simulate the covariates of a patient. If a patient is male, we set the gender indicator as 1, otherwise 0. In a simulated data set, half of the patients are male and half are

female. We generate the timing-varying glucose values $G_{i,j}$'s using equation

$$G_{i,j} = a_{0,i} + a_{1,i}t_{i,j} \tag{5.8}$$

where $a_{0,i}$ is from a normal distribution with mean 100 and standard deviation 0.5, and $a_{1,i}$ has a normal distribution with mean 5 and standard deviation 0.1. For each patient, we draw a vector of $(a_{0,i}, a_{1,i})$ from the above normal distributions, and use it to simulate this patient's $G_{i,j}$ values. $t_{i,j}$ is picked randomly within bin_j following a uniform distribution.

The third step is to standardize the simulated $G_{i,j}$ values. Provided we are trying to generate data for a data set with size 200, after the first two steps, for each patient we then already have one gender covariate and 4 glucose values. In all we have $4 \times 200 = 800$ glucose values. Then we standardize each simulated glucose value by subtracting the mean of all these available glucose values and dividing by the standard deviation of all these available glucose values. For simplicity, we still denote the glucose values after standardization $G_{i,j}$.

The fourth step is to simulate a failure time using the generated gender covariate from step two and standardized glucose values from step three. Now, hazard increments of patient i over four bins are :

$$\begin{aligned} &d_1 \exp(\text{Gen}_i\beta_{\text{gen}} + G_{i,1}\beta_{\text{glu}}), \quad d_2 \exp(\text{Gen}_i\beta_{\text{gen}} + G_{i,2}\beta_{\text{glu}}) \\ &d_3 \exp(\text{Gen}_i\beta_{\text{gen}} + G_{i,3}\beta_{\text{glu}}), \quad d_4 \exp(\text{Gen}_i\beta_{\text{gen}} + G_{i,4}\beta_{\text{glu}}) \end{aligned} \tag{5.9}$$

Still using the assumption that hazard function is a constant within each interval, we can get life time distribution function $F_i(t)$ using $F_i(t) = 1 - \exp(-H_i(t))$, noticing that $H_i(t)$ is a step function fully determined by hazard increments above. Then we use inverse probability to generate one failure time T_i from $F_i(t)$. For our case, we can't have an analytic function to generate failure time. Given the baseline hazard rate is piecewise constant, our simulation is conducted via numerical methods. With some assumptions, we can have a closed analytic form to sample failure time too. Austin (2012) discusses generating failure time for Cox proportional hazards model with time-varying covariates and derives closed-form expression to simulate failure time for some

special cases. Examples are illustrated over three kinds of distribution for baseline survival time: Exponential, Weibull and Gompertz distribution with three types of time-varying covariate: a continuous liner covariate, a dichotomous covariate with one change before the study ends, and a dichotomous covariate with more than one change before the study ends.

When we have to generate a data set with certain size, first we run step one once, then run step two as many times as we meet the data set size, then run step three once and run step four for each patient in the set. In the end, for each patient in this data set, we generated his/her observed failure time T_i, constant covariate Gen_i, and four observations of time-varying glucose values. This methodology can be easily extended to a multiresolution hazard model with any level and also with more than one time-varying covariate, and the function chosen for generating covariates can be in any form not limited to linear as in Equation (5.8).

Following the above steps, we simulated 200 data sets, each of them containing 200 patients, 200 data sets of 500 patients per set, and 200 data sets of 1000 patients per set, all with two covariates of a three-level multiresolution hazard model. In each data set, half of the patients are male with time-varying glucose values and half of them are female with time-varying glucose values. Assume the whole study takes 20 years, so that failure time after 20 would be taken as right-censored. We have $M = 3$ in our model, so there are 8 equal length bins before the censored time 20. For the true parameters, we set $H = 1$, $a = 10$, $\lambda = 0.03$ and $R_{m,p}$'s as in the table 5.1. The gender effect β_{gen} is set as 0.48 and the glucose effect β_{glu} is set as 0.7.

$R_{1,0}$	$R_{2,0}$	$R_{2,1}$	$R_{3,0}$	$R_{3,1}$	$R_{3,2}$	$R_{3,3}$
0.368	0.550	0.547	0.650	0.496	0.396	0.289

Table 5.1: True $R_{m,p}$ values used in generating the simulated data set

For each data set, we count the number of people with failure time falling in each bin, and

then we average those counts over 200 data set for a same bin. Table 5.2, Table 5.3 and Table 5.4 are the average failures over each bins and average censored subjects in the end, over all simulated datasets with 200 patients per set, 500 patients per set and 1000 patients per set. For the picked censoring time 20, a total right censoring rate is about 16.5%, where the female group has a right censoring rate at about 11.7% and the male group has a right censoring rate at about 4.8%. These rates differ very subtly with different data set sizes.

	1	2	3	4	5	6	7	8	censored
female	4.4	3.2	4.7	5.9	11.9	19.6	8.1	18.7	23.4
male	7.3	4.9	7.1	8.7	16.3	23.1	8.0	15.0	9.6
pool	11.8	8.1	11.8	14.6	28.2	42.7	16.2	33.6	32.9

Table 5.2: Average counts in each bin of 200 simulated data sets with one constant covariate and one time-varying covariate (200 data per set, time-varying covariate is generated from linear function)

	1	2	3	4	5	6	7	8	censored
female	11.2	7.5	11.9	14.9	30.2	49.2	20.3	46.2	58.6
male	17.8	12.6	18.1	22.4	40.7	56.6	20.3	37.8	23.7
pool	29.0	20.1	30.0	37.3	70.9	105.9	40.6	84.0	82.2

Table 5.3: Average counts in each bin of 200 simulated data sets with one constant covariate and one time-varying covariate (500 data per set, time-varying covariate is generated from linear function)

Apart from using linear function to generate time-varying covariates, we also use a cosine shaped function and a five degree polynomial to generate time-varying covariates. Following exact the same fashion as aforementioned in this section, and keep everything the same, but just replace

	1	2	3	4	5	6	7	8	censored
female	21.9	15.6	23.8	30.2	62.2	95.5	40.6	92.5	117.8
male	35.7	24.3	35.8	43.0	82.8	115.0	40.4	75.0	48.0
pool	57.6	39.9	59.6	73.2	144.9	210.4	81.0	167.5	165.8

Table 5.4: Average counts in each bin of 200 simulated data sets with one constant covariate and one time-varying covariate (1000 data per set, time-varying covariate is generated from linear function)

Equation (5.8) with

$$G(t) = a_0 + a_1 \cos(a_2 t) \tag{5.10}$$

where a_0 has a normal distribution with mean 100 and standard deviation 0.5; a_1 has a normal distribution with mean 5 and standard deviation 0.1; a_2 has a normal distribution with mean 2 and standard deviation 0.1. We generate 200 datasets with 200 subjects per set.

Similarly, keeping all other parameters the same, the time-varying covariate values are generated using a five degree polynomial

$$G(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + a_5 t^5 \tag{5.11}$$

where a_0 has a normal distribution with mean 100 and standard deviation 0.5; a_1 has a normal distribution with mean 5 and standard deviation 1; a_2 has a normal distribution with mean 3 and standard deviation 1; a_3 has a normal distribution with mean 2 and standard deviation 1; a_4 has a normal distribution with mean 1 and standard deviation 1; a_5 has a normal distribution with mean 0.5 and standard deviation 1. We again generate 200 datasets with 200 subjects per set.

We count the average of failures in each bin and subjects censored when the study terminated over 200 datasets for these two cases as well. When the time-varying covariates are generated from cosine shape function as Equation (5.10), as showed in Table 5.5, the average censoring rate for female is 15.3%, for male is 7.4% and overall censoring rate is 22.7%. When the time-varying covariates are generated from five degree polynomials as Equation (5.11), from Table 5.6, we can

see the average censoring rate for female is 15.5%, for male is 8.7% and overall censoring rate is 24.1%. The total censoring rate for these two scenarios are very close.

	1	2	3	4	5	6	7	8	censored
female	12.6	8.1	8.9	6.7	8.4	11.9	4.4	8.5	30.6
male	19.3	11.9	12.0	9.0	9.8	11.8	4.0	7.5	14.8
pool	31.9	19.9	20.9	15.7	18.2	23.7	8.3	16.0	45.3

Table 5.5: Average counts in each bin of 200 simulated data sets with one constant covariate and one time-varying covariate (200 data per set, time-varying covariate is generated from cosine like functions)

	1	2	3	4	5	6	7	8	censored
female	10.3	5.0	5.4	5.0	8.6	12.4	6.5	15.9	30.9
male	16.1	7.6	8.0	7.4	10.9	14.0	5.9	12.8	17.3
pool	26.4	12.6	13.4	12.4	19.5	26.4	12.4	28.7	48.1

Table 5.6: Average counts in each bin of 200 simulated data sets with one constant covariate and one time-varying covariate (200 data per set, time-varying covariate is generated from five degree polynomials)

Finally, we have generated the following datasets and we will use them to evaluate our MRH models and PRMH models later in this chapter.

- 200 datasets with size 200, including a time-varying covariate generated from linear functions and a fixed covariate.
- 200 datasets with size 500, including a time-varying covariate generated from linear functions and a fixed covariate.

- 200 datasets with size 1000, including a time-varying covariate generated from linear functions and a fixed covariate.
- 200 datasets with size 200, including a time-varying covariate generated from five degree polynomials and a fixed covariate.
- 200 datasets with size 200, including a time-varying covariate generated from cosine shape functions and a fixed covariate.

5.3 Evaluating MRH models with time-varying covariates with simulated data

First, for each set of data from Section 5.2, we implemented MRH strategy TVC-NPM3 (3-level model with time-varying covariates without any pruning, no missing data). MCMC chains with 200000 iterations for each of the 200 datasets were run separately. The first 50000 iterations of each MCMC chain was discarded as the burn-in, and every 10th sample from the chain was kept to reduce autocorrelation. In the end, 15000 posterior samples per dataset were used to derive posterior PMRH estimates (posterior means), resulting in 200 sets of estimates. And we use these means to calculate the corresponding 95% probability intervals for each parameter of our interest.

All the simulations were coded in R and run on a supercomputer with 1368 nodes, each containing two hex-core 2.8Ghz Intel Westmere processors with 12 cores per node and 2GB of RAM per core. When the time-varying covariates are generated from linear functions, for a dataset of size 200, it took about 2.1 hours for model TVC-NPM3 to complete 200000 iterations; 3.2 hours for a dataset of size 500; 5.5 hours for a dataset of size 1000. For the dataset of size 200, with time-varying covariates are generated using Equation (5.10), it took about 2.1 hours for model TVC-NPM3 to complete 200000 iterations. For the dataset of size 200, with time-varying covariates are generated using Equation (5.11), it took about 2.1 hours for model TVC-NPM3 to complete 200000 iterations. With the same dataset size, the model TVC-NPM3 takes about the same time to finish a certain number of iterations, no matter how the covariates are generated. As dataset size

increases, it requires longer time to complete the same number of iterations for the same model.

In Table 5.7, we give the 95% probability intervals for all parameters of model TVC-NPM3 over 200 datasets with 200 subjects per set. The time-varying covariates in these datasets are generated from linear function Equation (5.8), cosine shape function Equation (5.10) and five degree polynomial Equation (5.11) separately. We can see, for each model parameter, all the 95% probability intervals are centered around its true value. For the group that time-varying covariates are generated from cosine shape functions and five degree polynomials, the posterior mean seems to be a very good estimate for all model parameters repectively. But the overall result from the group with time-varying covariates generated from linear functions sounds to be more variant when comparing with the other two groups. In Table 5.8 we show that variations will be reduced and better estimates will be attained as dataset size increases for this group.

In Table 5.8, we show the 95% probability intervals for all parameters of model TVC-NPM3 running over 200 datasets with 200, 500, and 1000 patients per set with time-varying covariates generated from linear function Equation (5.8). We can see, for each model parameter, all the 95% probability intervals are still centered around its true value. The posterior mean can be a very good estimate for baseline cumulative hazard function H_0 and gender effect. Although the posterior mean of glucose effect is not so close to the true glucose effect β_{glu} 0.7, when we look at the results from dataset of size 200 and dataset of size 500, the 95% probability intervals still contain the true β_{glu}. However, if we look at these results together with the probability interval of β_{glu} from dataset of 1000, we can see we do have better estimates as dataset size increases. Combining the result for glucose effect β_{glu} estimates in Table 5.7 and Table 5.8, it definitely indicates that for time-varying covariate effects, in order to get an estimate with similar tolerance to the truth, a larger dataset size will probably be required comparing with that needed for estimating other constant covaraites associated effects in some scenario. In all, Table 5.7 and Table 5.8 verify that our MRH model with time-varying covariates without missing data works well.

	H_0	Gender	Glucose	$R_{1,0}$	$R_{2,0}$	$R_{2,1}$	$R_{3,0}$	$R_{3,1}$	$R_{3,2}$	$R_{3,3}$
True:	1.00	0.48	0.70	0.37	0.55	0.55	0.65	0.50	0.40	0.29
Time-varying covariates are generated from linear functions										
2.5%	0.79	0.19	0.51	0.30	0.45	0.48	0.43	0.35	0.31	0.20
50%	1.04	0.49	0.86	0.44	0.57	0.58	0.65	0.51	0.41	0.30
mean	1.03	0.50	0.87	0.44	0.58	0.58	0.65	0.51	0.41	0.31
97.5%	1.27	0.84	1.22	0.56	0.72	0.68	0.82	0.68	0.53	0.43
Time-varying covariates are generated from cosine shape functions										
2.5%	0.73	0.21	0.54	0.30	0.46	0.43	0.54	0.35	0.28	0.15
50%	0.97	0.52	0.71	0.38	0.55	0.54	0.64	0.50	0.39	0.30
mean	0.98	0.51	0.72	0.38	0.55	0.54	0.65	0.51	0.39	0.30
97.5%	1.26	0.82	0.91	0.47	0.66	0.66	0.77	0.66	0.52	0.47
Time-varying covariates are generated from five degree polynomials										
2.5%	0.75	0.19	0.54	0.30	0.43	0.44	0.51	0.32	0.29	0.18
50%	0.96	0.51	0.73	0.38	0.55	0.55	0.65	0.50	0.41	0.30
mean	0.97	0.50	0.73	0.38	0.55	0.55	0.65	0.50	0.41	0.30
97.5%	1.26	0.80	0.87	0.47	0.66	0.66	0.78	0.66	0.56	0.43

Table 5.7: Estimates and 95% probability intervals for all parameters of model TVC-NPM3 over 200 datasets with 200 subjects per set (time-varying covariates are generated from linear functions, cosine shape functions and five degree polynomials separately, no missing data)

5.4 Evaluating PMRH models with time-varying covariates with simulated data

In Section 5.3, we already applied regular MRH models to datasets with time-varying covariates. In this section, we are going to investigate the performance of PMRH models over datasets with time-varying covariates. As discussed in Chapter 2 and Chapter 3, PMRH models are applied to baseline hazards only. For each set of data, we implemented 3 different PMRH strategies:

- **TVC-NPM3:** 3-level model with time-varying covariates without any pruning, no missing data

	H_0	Gender	Glucose	$R_{1,0}$	$R_{2,0}$	$R_{2,1}$	$R_{3,0}$	$R_{3,1}$	$R_{3,2}$	$R_{3,3}$
True:	1.00	0.48	0.70	0.37	0.55	0.55	0.65	0.50	0.40	0.29
200 datasets with 200 patients per set										
2.5%	0.79	0.19	0.51	0.30	0.45	0.48	0.43	0.35	0.31	0.20
50%	1.04	0.49	0.86	0.44	0.57	0.58	0.65	0.51	0.41	0.30
mean	1.03	0.50	0.87	0.44	0.58	0.58	0.65	0.51	0.41	0.31
97.5%	1.27	0.84	1.22	0.56	0.72	0.68	0.82	0.68	0.53	0.43
200 datasets with 500 patients per set										
2.5%	0.88	0.29	0.39	0.27	0.45	0.46	0.52	0.37	0.34	0.22
50%	1.02	0.49	0.81	0.42	0.57	0.57	0.65	0.51	0.41	0.30
mean	1.03	0.49	0.80	0.42	0.57	0.57	0.65	0.51	0.41	0.30
97.5%	1.25	0.67	1.18	0.57	0.66	0.66	0.78	0.63	0.49	0.38
200 datasets with 1000 patients per set										
2.5%	0.91	0.36	0.41	0.27	0.46	0.49	0.55	0.42	0.35	0.24
50%	1.02	0.48	0.73	0.39	0.56	0.56	0.65	0.50	0.41	0.29
mean	1.02	0.48	0.74	0.39	0.56	0.56	0.65	0.51	0.41	0.30
97.5%	1.16	0.62	1.05	0.53	0.64	0.64	0.75	0.60	0.47	0.36

Table 5.8: Estimates and 95% probability intervals for all parameters of model TVC-NPM3 over 200 datasets with 200, 500, and 1000 patients per set (time-varying covariates are generated from linear functions, no missing data)

- **TVC-PM31:** 3-level model with time-varying covariates with the 3rd level subject to pruning, no missing data
- **TVC-PM33:** 3-level model with time-varying covariates with all 3 levels subject to pruning, no missing data

First, for each dataset with size 200 from Section 5.2, we implemented PMRH strategy TVC-PM33 (3-level model with time-varying covariates with all 3 levels subject to pruning, no missing data). To fresh our memory, there are three groups of datasets with size 200, and each group contains 200 datasets. One group consists of 200 datasets whose time-varying covariates are generated from linear function as Equation (5.8), one consists of 200 datasets having time-varying

covariates generated from cosine shape function as Equation (5.10), and one consists of 200 datasets with time-varying covariates from five degree polynomial as Equation (5.11).

MCMC chains with 200000 iterations for each of the 200 datasets were run separately. The first 50000 iterations of each MCMC chain was discarded as the burn-in, and every 20th sample from the chain was kept to reduce autocorrelation. In the end, 7500 posterior samples per dataset were used to derive posterior PMRH estimates (posterior means), resulting in 200 sets of estimates. And we use these means to calculate the corresponding 95% probability intervals for each parameter of our interest.

The same as in Section 5.3, all the simulations were coded in R and run on a supercomputer with 1368 nodes, each containing two hex-core 2.8Ghz Intel Westmere processors with 12 cores per node and 2GB of RAM per core. On average, for a dataset of size 200, when the time-varying covariates were generated from linear functions, it took about 1.4 hours for model TVC-PM33 to complete 200000 iterations; when the time-varying covariates were generated from cosine shape functions, it took about 1.2 hours; when the time-varying covariates were generated from five degree polynomials, it took about 1.2 hours. And results from last section tell us it took about 2.1 hours for model TVC-NPM3 to complete 200000 iterations, for a dataset of size 200. Pruning saved computation time about 50% in this case.

In Table 5.9, we give the 95% probability intervals for all parameters of model TVC-PM33 over 200 datasets with 200 subjects per set. The time-varying covariates in these datasets are generated from linear function Equation (5.8), cosine shape function Equation (5.10) and five degree polynomial Equation (5.11) separately. We can see, for each model parameter, all the 95% probability intervals are centered around its true value. For the group that time-varying covariates were generated from cosine shape functions and five degree polynomials, the posterior mean seems to be a very good estimate for all model parameters repectively even we made all the levels in our MRH models subject to pruning. The result from cosine shape function group tells us that, for each dataset, except the bins associated with R_{10}, all the other bins have been merged in most of the 200 datasets. When it comes to the five degree polynomial group, for each dataset, except the bins

	H_0	Gender	Glucose	$R_{1,0}$	$R_{2,0}$	$R_{2,1}$	$R_{3,0}$	$R_{3,1}$	$R_{3,2}$	$R_{3,3}$
True:	1.00	0.48	0.70	0.37	0.55	0.55	0.65	0.50	0.40	0.29
Time-varying covariates are generated from linear functions										
2.5%	0.82	0.19	0.23	0.21	0.39	0.41	0.50	0.31	0.28	0.18
50%	1.06	0.49	0.50	0.30	0.50	0.50	0.50	0.50	0.38	0.28
mean	1.07	0.50	0.54	0.31	0.48	0.50	0.52	0.48	0.38	0.28
97.5%	1.32	0.83	1.01	0.45	0.51	0.58	0.80	0.50	0.50	0.44
Time-varying covariates are generated from cosine shape functions										
2.5%	0.72	0.20	0.51	0.30	0.50	0.50	0.50	0.50	0.28	0.15
50%	0.97	0.52	0.68	0.38	0.50	0.50	0.50	0.50	0.50	0.50
mean	0.98	0.51	0.68	0.40	0.50	0.51	0.54	0.51	0.44	0.37
97.5%	1.26	0.80	0.87	0.50	0.61	0.65	0.77	0.65	0.50	0.50
Time-varying covariates are generated from five degree polynomials										
2.5%	0.76	0.20	0.53	0.28	0.50	0.44	0.50	0.50	0.29	0.18
50%	0.96	0.51	0.70	0.38	0.50	0.50	0.50	0.50	0.50	0.30
mean	0.99	0.51	0.69	0.38	0.52	0.51	0.59	0.50	0.44	0.30
97.5%	1.26	0.81	0.85	0.47	0.66	0.58	0.78	0.50	0.50	0.50

Table 5.9: Estimates and 95% probability intervals for all parameters of model TVC-PM33 over 200 datasets with 200 subjects per set (time-varying covariates are generated from linear functions, cosine shape functions and five degree polynomials separately, no missing data)

associated with R_{10} and R_{33}, all the other bins have been merged in most of the 200 datasets. For the linear function group, for each dataset, except the bins associated with R_{10}, R_{33} and R_{32}, all the other bins have been merged in most of the 200 datasets. We could say that all the models applied to these three groups have been pruned heavily. But the estimates for cumulative baseline hazard function H_0 and time-varying covariate glucose effect β_{glu} are not as close to the true values as those from the other two groups. As we already discussed in Section 2.3, since the estimates for each parameter counting on all the other parameters, when a model is pruned too heavily, it is possible that the accuracy for estimating model parameter, such as predictor effects and cumulative baseline hazard function, will be affected. One advantage of PMRH models is to balance computation cost and estimates accuracy. Apart from the extent of pruning, the performance of PMRH models also

depends on the dataset itself. For instance, here we have MRH models for all three groups be heavily pruned, but only the results from group having linear function generating time-varying covariates are not so good. And we also can see that even under such extensive pruned MRH models, the estimates for the constant covariate effect are always very close to the corresponding true values.

Taking the same group of datasets whose time-varying covariates were generated from linear functions as in Equation (5.8), for each dataset with size 200, 500 and 1000, we implemented PMRH strategy TVC-PM31 (3-level model with time-varying covariates with only the third level subject to pruning, no missing data) and TVC-PM33 (3-level model with time-varying covariates with all 3 levels subject to pruning, no missing data).

MCMC chains with 200000 iterations for each of the 200 datasets were run separately. The first 50000 iterations of each MCMC chain was discarded as the burn-in, and every 50th sample from the chain was kept to reduce autocorrelation. In the end, 3000 posterior samples per dataset were used to derive posterior PMRH estimates (posterior means), resulting in 200 sets of estimates. And we use these means to calculate the corresponding 95% probability intervals for each parameter of our interest.

The same as in Section 5.3, all the simulations were coded in R and run on a supercomputer with 1368 nodes, each containing two hex-core 2.8Ghz Intel Westmere processors with 12 cores per node and 2GB of RAM per core. When PMRH model TVC-PM31 is used, for a dataset of size 200, it took about 1.7 hours to complete 200000 iterations; 3 hours for a dataset of size 500; 4.1 hours for a dataset of size 1000, on average. When PMRH model TVC-PM33 is used, for a dataset of size 200, it took about 1.4 hours to complete 200000 iterations; 2.2 hours for a dataset of size 500; 4.1 hours for a dataset of size 1000, on average.

Considering how the pruning procedure is implemented of PMRH models, we know that it is highly possible that a TVC-PM31 model and a TVC-PM33 model will end up be the same when applied to the same dataset with relative big size. We only have about 14% TVC-PM31 model and TVC-PM33 model be identical, when they were applied to datasets of size 200; this rate becomes

	H_0	Gender	Glucose	$R_{1,0}$	$R_{2,0}$	$R_{2,1}$	$R_{3,0}$	$R_{3,1}$	$R_{3,2}$	$R_{3,3}$
True:	1.00	0.48	0.70	0.37	0.55	0.55	0.65	0.50	0.40	0.29
TVC-PM31 (200 datasets with 200 patients per set)										
2.5%	0.79	0.19	0.36	0.26	0.42	0.46	0.50	0.34	0.30	0.19
50%	1.03	0.49	0.75	0.39	0.55	0.56	0.50	0.50	0.40	0.29
mean	1.03	0.50	0.75	0.39	0.55	0.56	0.52	0.49	0.40	0.30
97.5%	1.29	0.83	1.16	0.51	0.69	0.65	0.82	0.50	0.50	0.44
TVC-PM33 (200 datasets with 200 patients per set)										
2.5%	0.82	0.19	0.23	0.21	0.39	0.41	0.50	0.31	0.28	0.18
50%	1.06	0.49	0.50	0.30	0.50	0.50	0.50	0.50	0.38	0.28
mean	1.07	0.50	0.54	0.31	0.48	0.50	0.52	0.48	0.38	0.28
97.5%	1.32	0.83	1.01	0.45	0.51	0.58	0.80	0.50	0.50	0.44
TVC-PM31 (200 datasets with 500 patients per set)										
2.5%	0.89	0.29	0.29	0.23	0.43	0.43	0.50	0.37	0.33	0.21
50%	1.04	0.49	0.65	0.35	0.53	0.53	0.50	0.50	0.39	0.29
mean	1.05	0.49	0.64	0.35	0.53	0.53	0.54	0.48	0.39	0.29
97.5%	1.28	0.66	0.96	0.48	0.62	0.62	0.78	0.50	0.47	0.37
TVC-PM33 (200 datasets with 500 patients per set)										
2.5%	0.89	0.29	0.23	0.20	0.41	0.42	0.50	0.36	0.32	0.21
50%	1.07	0.49	0.53	0.31	0.50	0.51	0.50	0.50	0.38	0.28
mean	1.08	0.49	0.56	0.32	0.50	0.51	0.54	0.48	0.38	0.28
97.5%	1.33	0.66	0.94	0.48	0.60	0.61	0.77	0.50	0.46	0.36
TVC-PM31 (200 datasets with 1000 patients per set)										
2.5%	0.92	0.36	0.21	0.20	0.43	0.44	0.50	0.40	0.33	0.23
50%	1.06	0.48	0.55	0.31	0.51	0.52	0.50	0.50	0.39	0.28
mean	1.07	0.48	0.57	0.32	0.52	0.52	0.56	0.48	0.39	0.28
97.5%	1.27	0.62	0.88	0.46	0.60	0.60	0.74	0.50	0.45	0.36
TVC-PM33 (200 datasets with 1000 patients per set)										
2.5%	0.92	0.36	0.21	0.20	0.43	0.44	0.50	0.40	0.33	0.22
50%	1.06	0.48	0.54	0.31	0.50	0.51	0.50	0.50	0.39	0.28
mean	1.07	0.48	0.55	0.31	0.51	0.52	0.56	0.48	0.39	0.28
97.5%	1.27	0.62	0.88	0.46	0.60	0.60	0.74	0.50	0.45	0.36

Table 5.10: Estimates and 95% probability intervals for all parameters of model TVC-PM31 and TVC-PM33 over 200 datasets with 200, 500, and 1000 patients per set (time-varying covariates are generated from linear functions, no missing data)

60%, when TVC-PM31 model and TVC-PM33 model were applied to datasets of size 500; when it comes to datasets of size 1000, we have 91% identical models. This is not surprising. Because the modified fisher's exact test used in pruning steps have a higher power when larger number of counts are used in the test, then the chance of rejecting the null increases, resulting in keeping more branches. In this case, the counts of failures in each bin are completely proportional to the dataset size, so when dataset size increases, even we have all 3 levels of a MRH model subject to pruning, the branches in level 1 and level 2 will still be kept rather than pruned. This then ends up with the two PMRH models TVC-PM31 and TVC-PM33 being the same. In this case, with datasets of size 1000, we have 91% of the TVC-PM31 and TVC-PM33 models be identical for a dataset. Then on average, surely the computation time for model TVC-PM31 and TVC-PM33 are about the same.

In Table 5.10, we show the 95% probability intervals for all parameters of model TVC-PM31 and TVC-PM33 running over 200 datasets with 200, 500, and 1000 patients per set with time-varying covariates generated from linear function Equation (5.8). We can see, for each model parameter, all the 95% probability intervals still contain its true value. First, we can see no matter how dataset size changes and the pruning strategy changes, the estimates for constant covariate(gender) effect can always be estimated very accurately. Second, in this case, it seems the more intensively the MRH models have been pruned, the mean and median of the estimates for cumulative baseline hazard function H_0 is getting further and further away from its true value 1, in the positive direction. Since all the parameter values affect each other, in the mean time, the mean and median of the estimates for time-varying covariate(glucose) effect is getting further and further away from its true value 0.7, in the negative direction. It is also not surprising that the results for datasets with size 1000 over model TVC-PM31 and TVC-PM33 are almost identical, since we already know that 90% of these two models are the same over a same dataset in this case. In all, from Table 5.9 and Table 5.10, we can conclude that the performance of PMRH models depends on the extent of pruning, the time-varying covariate values and the dataset size and constant covariates associated effects can always be well estimated almost in all PRMH models with and without time-varying covariates.

5.5 Comparison to piecewise exponential hazard model

In this section, we compare the results from applying MRH models and piecewise exponential hazard models to the same simulated dataset that we used in Section 5.3. For each dataset, we implemented the following two strategies:

- **TVC-NPM3:** 3-level model with time-varying covariates without any pruning, no missing data
- **TVC-EPEM3:** "TVC" means the dataset used having time-varying covariates and no missing data; the first "E" means the time axis is partitioned with equal width intervals; "PE" refers to piecewise exponential hazard model; "M3" means the time axis is partitioned into $2^3 = 8$ intervals

As discussed in Section 5.3, for each dataset implemented by strategy TVC-NPM3, MCMC chains with 200000 iterations for each of the 200 datasets were run separately. The first 50000 iterations of each MCMC chain was discarded as the burn-in, and every 10th sample from the chain was kept to reduce autocorrelation. In the end, 15000 posterior samples per dataset were used to derive posterior PMRH estimates (posterior means), resulting in 200 sets of estimates. And we use these means to calculate the corresponding 95% probability intervals for each parameter of our interest. For each dataset implemented by strategy TVC-EPEM3, the MLEs were considered as the piecewise exponential hazards model estimates, also resulting in 200 sets of estimates. Following the same fashion, these MLEs were used to calculate the corresponding 95% probability intervals for each parameter of our interest.

Table 5.11 gives the estimates and 95% probability intervals for all parameters and hazard increments of model TVC-NPM3 and model TVC-EPEM3 over 200 datasets with 200 subjects per set. All the time-varying covariates in the datasets are generated from cosine shape functions and there is no missing data. We can see that the results from model TVC-NPM3 and model TVC-EPEM3 are very close to each other, and all are centered around their true values respectively.

In Table 5.12 we demonstrate the estimates and 95% probability intervals for all parameters and hazard increments of model TVC-NPM3 and model TVC-EPEM3 over 200 datasets with 200 subjects per set. All the time-varying covariates in the datasets are generated from five degree polynomials and there is no missing data. We reach the same conclusion that the results from model TVC-NPM3 and model TVC-EPEM3 are very close to each other, and all are centered around their true values respectively.

In Table 5.13 we show the estimates and 95% probability intervals for all parameters and hazard increments of model TVC-NPM3 and model TVC-EPEM3 over 200 datasets with 200, 500, and 1000 subjects per set. In each dataset, the time-varying covariates are generated from linear functions and there is no missing data. When dataset size is the same, for a parameter, we can see that the estimates from model TVC-EPEM3 are more variant comparing with that from model TVC-NPM3. As the dataset size increases, for model TVC-EPEM3, we get narrower 95% probability intervals for all parameters and hazard increments. We also have narrower 95% probability intervals for all parameters and hazard increments when examine the result from applying model TVC-NPM3 to larger datasets. But the changes in probability interval width are not as obvious as that from applying model TVC-EPEM3.

Based on the results we illustrate in Table 5.11, Table 5.12 and Table 5.13, we can conclude that our TVC-MRH models can estimate model parameters efficiently. For any given dataset, our TVC-MRH models can provide estimates either less variant than those getting from running piecewise exponential hazard models or at least with the same extent of variation. Overall, our TVC-MRH models can perform more stably and efficiently in comparison of piecewise exponential hazard models.

	H_0	Gender	Glucose	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
True:	1.00	0.48	0.70	0.13	0.07	0.08	0.08	0.14	0.21	0.08	0.20
TVC-NPM3											
2.5%	0.73	0.21	0.54	0.09	0.04	0.05	0.04	0.08	0.13	0.03	0.11
50%	0.97	0.52	0.71	0.13	0.07	0.08	0.08	0.13	0.19	0.08	0.20
mean	0.98	0.51	0.72	0.13	0.07	0.08	0.08	0.13	0.20	0.08	0.20
97.5%	1.26	0.82	0.91	0.19	0.11	0.12	0.13	0.20	0.31	0.15	0.31
TVC-EPEM3											
2.5%	0.76	0.23	0.56	0.10	0.04	0.05	0.05	0.08	0.14	0.03	0.12
50%	1.00	0.51	0.71	0.13	0.07	0.08	0.08	0.13	0.21	0.08	0.21
mean	1.00	0.50	0.71	0.14	0.07	0.08	0.08	0.14	0.21	0.08	0.21
97.5%	1.26	0.79	0.88	0.18	0.11	0.12	0.13	0.20	0.31	0.15	0.32

Table 5.11: Estimates and 95% probability intervals for all parameters of model TVC-NPM3 and model TVC-EPEM3 over 200 datasets with 200 subjects per set (time-varying covariates are generated from **cosine shape** functions, no missing data)

	H_0	Gender	Glucose	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
True:	1.00	0.48	0.70	0.13	0.07	0.08	0.08	0.14	0.21	0.08	0.20
TVC-NPM3											
2.5%	0.75	0.19	0.54	0.08	0.04	0.05	0.05	0.08	0.13	0.05	0.11
50%	0.96	0.51	0.73	0.13	0.07	0.08	0.08	0.13	0.19	0.08	0.19
mean	0.97	0.50	0.73	0.13	0.07	0.08	0.08	0.14	0.19	0.08	0.19
97.5%	1.26	0.80	0.87	0.19	0.11	0.13	0.13	0.20	0.29	0.13	0.29
TVC-EPEM3											
2.5%	0.76	0.18	0.53	0.08	0.04	0.04	0.05	0.08	0.13	0.04	0.11
50%	0.97	0.51	0.73	0.13	0.07	0.08	0.08	0.13	0.19	0.08	0.20
mean	0.98	0.50	0.73	0.13	0.07	0.08	0.08	0.14	0.20	0.08	0.20
97.5%	1.27	0.79	0.92	0.19	0.11	0.13	0.13	0.21	0.30	0.13	0.31

Table 5.12: Estimates and 95% probability intervals for all parameters of model TVC-NPM3 and model TVC-EPEM3 over 200 datasets with 200 subjects per set (time-varying covariates are generated from **five degree polynomials**, no missing data)

	H_0	Gender	Glucose	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
True:	1.00	0.48	0.70	0.13	0.07	0.08	0.08	0.14	0.21	0.08	0.20
TVC-NPM3 (200 datasets with 200 subjects per set)											
2.5%	0.79	0.19	0.51	0.08	0.04	0.05	0.05	0.08	0.13	0.04	0.10
50%	1.04	0.49	0.86	0.17	0.09	0.09	0.08	0.13	0.19	0.07	0.18
mean	1.03	0.50	0.87	0.18	0.09	0.09	0.09	0.13	0.19	0.08	0.18
97.5%	1.27	0.84	1.22	0.33	0.17	0.15	0.13	0.19	0.27	0.12	0.31
TVC-EPEM3 (200 datasets with 200 subjects per set)											
2.5%	0.78	0.18	-0.32	0.03	0.02	0.04	0.04	0.07	0.09	0.02	0.04
50%	1.10	0.48	0.73	0.14	0.07	0.08	0.08	0.13	0.20	0.08	0.19
mean	1.17	0.49	0.76	0.20	0.09	0.09	0.08	0.14	0.22	0.09	0.26
97.5%	2.03	0.83	1.74	0.80	0.25	0.19	0.14	0.21	0.47	0.26	0.97
TVC-NPM3 (200 datasets with 500 subjects per set)											
2.5%	0.88	0.29	0.39	0.09	0.04	0.06	0.06	0.10	0.14	0.05	0.10
50%	1.02	0.49	0.81	0.16	0.08	0.09	0.09	0.13	0.19	0.08	0.18
mean	1.03	0.49	0.80	0.17	0.08	0.09	0.09	0.13	0.20	0.08	0.20
97.5%	1.25	0.67	1.18	0.26	0.13	0.14	0.11	0.17	0.27	0.13	0.36
TVC-EPEM3 (200 datasets with 500 subjects per set)											
2.5%	0.88	0.28	-0.02	0.04	0.03	0.05	0.06	0.10	0.12	0.03	0.06
50%	1.03	0.49	0.70	0.13	0.07	0.08	0.08	0.13	0.20	0.08	0.19
mean	1.08	0.49	0.71	0.15	0.08	0.09	0.08	0.14	0.21	0.09	0.24
97.5%	1.53	0.66	1.42	0.36	0.14	0.15	0.12	0.18	0.35	0.20	0.66
TVC-NPM3 (200 datasets with 1000 subjects per set)											
2.5%	0.91	0.36	0.41	0.09	0.04	0.06	0.06	0.12	0.15	0.05	0.13
50%	1.02	0.48	0.73	0.14	0.08	0.09	0.08	0.14	0.20	0.08	0.20
mean	1.02	0.48	0.74	0.15	0.08	0.09	0.08	0.14	0.20	0.08	0.21
97.5%	1.16	0.62	1.05	0.23	0.12	0.12	0.10	0.17	0.26	0.12	0.32
TVC-EPEM3 (200 datasets with 1000 subjects per set)											
2.5%	0.89	0.36	0.20	0.06	0.03	0.05	0.06	0.11	0.15	0.05	0.11
50%	1.02	0.48	0.66	0.12	0.07	0.08	0.08	0.14	0.21	0.09	0.21
mean	1.04	0.48	0.67	0.13	0.07	0.08	0.08	0.14	0.21	0.09	0.23
97.5%	1.27	0.62	1.14	0.24	0.13	0.12	0.10	0.17	0.30	0.15	0.42

Table 5.13: Estimates and 95% probability intervals for all parameters of model TVC-NPM3 and model TVC-EPEM3 over 200 datasets with 200, 500, and 1000 subjects per set (time-varying covariates are generated from **linear** functions, no missing data)

Chapter 6

Hazard models with missing covariates and outcomes

In survival analysis, we always need to face the unavoidable scenario that some covariate values and event times are missing. In this chapter, we give an overview of models with missing covariates, and Frequentist and Bayesian approaches for hazard models with missing covariates. Also we talk about how different approaches of censoring may affect hazard model parameter estimation. Methods of missing time-varying covariate imputation are discussed in the end of this chapter.

6.1 Models with missing covariates

Missing data problems are commonly encountered in practice for researchers. There are many kinds of issues that can cause missing data for a study. Survey study will have missing data, when people accidently or intentionally skip some questions. In clinical trial, it is very likely that some people drop out the study before it ends, or can't make it to a few of the routinely scheduled tests. Errors happen in data entry can also causing missing data. In a long-time observation experiment, the failure of equipment would also make the data is unobservable while the equipment is being fixed. Since we only focus on missing covariates in this thesis, from on now, we will use missing data and missing covariates interchangeably.

Little and Rubin (1987) talk about three classes of missing covariates. The first classification is missing completely at random (MCAR). MCAR means that the missing data is completely independent of values of any covariates. When the data is MCAR, most analysis can still be

implemented to the subset with complete cases, but a loss of statistical power can occur. Little (1988) proposes a test for testing MCAR assumption and shows null distribution of this test statistic is asymptotically chi-squared-distributed.

Unfortunately, in clinical trail, most missing data doesn't fall into this class. Missing at random (MAR) is the second class. MAR occurs when the missing data is only dependent on the observed value of other covariates. By saying depending on the values of other covariates, we mean that for a covariate that we have more than one observed values along the time, if there are some missing values for it, the missing values are not dependent on the observed values this covariate, but only dependent on the observed values of the other covariates. For example, if females are more frequently to visit the doctor than the males, then the probability of missing a observation of blood pressure is higher for males than that of females, given an assumption that a patient will have blood pressure measured for each visit to hospital and we are carrying out a five-year study of the change of elder people's blood pressure. But if a patient decides to skip his scheduled follow-up since his last visit showed his blood pressure was in the range, then the missing is not MAR. Apart from the ignorable (i.e MCAR and MAR) missing data, the missing data is non-ignorable. When handling analysis with non-ignorable missing covariates, the missing data mechanism has to be incorporated too.

Methods of handling missing covariates can be roughly divided into the following four categories:

1. Complete data analysis
2. Imputation methods
3. Weighting methods
4. Model-based methods

Complete data analysis means that we only analyze units that are completely recorded and discard those with missing data. One advantage of this method is that it is easy to implement with any standard analysis methods. However, in order to get unbiased estimates, this method can

only be applied to data that is MCAR. In addition, subject removal will always sacrifice statistical power.

Imputation is any method that the missing data are filled with some estimates and then the whole data set is treated as complete for standard statistical methods. There are single imputation methods and multiple imputation methods. Single imputation means we only impute one value for a missing item. In statistical practice, mean imputation, hot deck imputation and regression imputation are some of the commonly used single imputation methods. Mean imputation substitutes the missing item with the mean of the other observed measures. In this way, the mean of the same variable will remain the same, but it will underestimate the true variance. The hot deck method uses values from similar responding units in the sample to fill the missing values. Unlike mean imputation, this method still retains the distribution of sampled values of a variable. When the missing value is replaced by the predicted value from a regression model over the observed data in the unit, the method is regression imputation. Multiple imputation, first proposed by Rubin (1978), is to fill a missing item with more than two imputed values from an appropriate distribution, which in statistical practice the distribution can be the missing data's posterior predictive distribution. The advantage of multiple imputation is that it incorporates the degree of uncertainty about which value to impute the way single imputation does not.

The weighting method is related to mean imputation method. By introducing design weights, the imputed value of a missing data is sampled from Horvitz-Thomposon estimator. In this method, the estimator of sample mean is unbiased.

In model-based methods, we always define a model for the missing data and make inference about parameters based on the likelihood function under that model, using procedures such as maximum likelihood and EM algorithm.

Regression analysis with missing values of the independent variables is excellently reviewed in Little (1992). Schafer (1997) explores both frequentist and Bayesian approaches for incomplete continuous and categorical multivariate data.

6.2 Frequentist and Bayesian approaches for hazard models with missing covariates

Missing data is an unavoidable problem in clinical trials, especially when covariates are repeatedly measured over time. Missing covariate values or survival times lead to incomplete data. In this section we will only talk about missing covariates. There is literature on both frequentist and Bayesian methods for survival models with MAR covariate data. Bayesian approaches to handle missing data problems have a few advantages and disadvantages with respect to frequentist approaches. Selecting a proper prior distribution can help overcome the problem of non-identifiable parameters in the likelihood function. Moreover, more information can be provided to parameters by adopting informative priors.

Since covariates can be either time-varying or fixed, first we review literature in handling scenario that all the covariates are fixed but some are missing. When the missing covariate values are missing at random (MAR), Expectation-maximization (EM) algorithm is commonly used in frequentist approaches. Ibrahim (1990) develops a general EM algorithm for obtaining maximum likelihood estimates for any generalized linear model (GLM) with data missing at random, with assumption that the unobserved covariates are random variables discretely distributed with finite range. In E step, the expected value of the log-likelihood function is calculated, with respect to the conditional distribution of missing data given observed data under the current estimate of the parameters. The M step maximizes quantity from the E step, which is the current estimate of the parameters. Repeat the E step and M step until it converges. Let $\mathbf{x} = (x_1, x_2, \ldots, x_p)'$ denote the covariates and assume they are random variables taking values from a finite set. Therefore, $\mathbf{x}$ has a multinomial distribution that can be parameterized by $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_r)'$. As in any GLM, we parameterize the conditional density of $y|\mathbf{x}$ from exponential family by $(\boldsymbol{\beta}, \phi)$, where y is the response. As from the exponential family, the density of y can be written as

$$f(y; \zeta, \phi) = \exp\left\{\frac{y\zeta - b(\zeta)}{\phi} + c(y, \phi)\right\} \tag{6.1}$$

where ϕ is the dispersion parameter and ζ is the canonical parameter, and the functions $b(\zeta)$ and

$c(y, \phi)$ are known. Therefore, by assuming $\phi_i = \phi/m_i$ for some known weights m_i, the complete data log-likelihood of the i-th individual

$$l(\boldsymbol{\theta}, \mathbf{x}_i, y_i) = \frac{y_i \mathbf{x}_i' \boldsymbol{\beta} - b(\mathbf{x}_i' \boldsymbol{\beta})}{\phi_i} + c(y_i, \phi_i) \tag{6.2}$$

Then the E step for all of the observations can be written as

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(s)}) = \sum_{i=1}^{n} \sum_{x_{\text{mis},i}} w_{i,(s)} l(\boldsymbol{\theta}, \mathbf{x}_i, y_i) = \sum_{i=1}^{n} \sum_{x_{\text{mis},i}} w_{i,(s)} \{l_{y_i|\mathbf{x}_i}(\boldsymbol{\beta}, \phi) + l_{\mathbf{x}_i}(\boldsymbol{\gamma})\} \tag{6.3}$$

In Equation (6.2) and Equation (6.3), n is the number of observations; $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi, \boldsymbol{\gamma})$; $\mathbf{x}_i = (\mathbf{x}_{\text{obs},i}, \mathbf{x}_{\text{mis},i})$, where $\mathbf{x}_{\text{obs},i}$ denotes the observed data and $\mathbf{x}_{\text{mis},i}$ denotes the missing data for the i-th individual; $\boldsymbol{\theta}^{(s)}$ is the current estimate of $\boldsymbol{\theta}$. Most importantly, the weight function

$$\begin{aligned} w_{i,(s)} &= p(\mathbf{x}_{\text{mis},i}|\mathbf{x}_{\text{obs},i}, y_i, \boldsymbol{\theta}^{(s)}) \\ &= \frac{p(y_i|\mathbf{x}_{\text{mis},i}, \mathbf{x}_{\text{obs},i}, \boldsymbol{\theta}^{(s)}) p(\mathbf{x}_{\text{mis},i}, \mathbf{x}_{\text{obs},i}|\boldsymbol{\theta}^{(s)})}{\sum_{\mathbf{x}_{\text{mis},i}} p(y_i|\mathbf{x}_{\text{mis},i}, \mathbf{x}_{\text{obs},i}, \boldsymbol{\theta}^{(s)}) p(\mathbf{x}_{\text{mis},i}, \mathbf{x}_{\text{obs},i}|\boldsymbol{\theta}^{(s)})} \\ &= \frac{p(y_i|\mathbf{x}_i, \boldsymbol{\theta}^{(s)}) p(\mathbf{x}_i|\boldsymbol{\theta}^{(s)})}{\sum_{\mathbf{x}_{\text{mis},i}} p(y_i|\mathbf{x}_i, \boldsymbol{\theta}^{(s)}) p(\mathbf{x}_i|\boldsymbol{\theta}^{(s)})} \end{aligned} \tag{6.4}$$

is defined as the conditional distribution of missing data given the current estimate of $\boldsymbol{\theta}$ and the observed data. Once we implement the weighted complete data log-likelihood in E step, the M step can be realized via Newton-Raphson algorithm.

Lipsitz and Ibrahim (1996) extend the EM method in Ibrahim (1990) to survival data that may not from the class of generalized linear models, with missing covariates which are MAR and categorical. In this approach, failure time T_i for the i-th unit is assumed to have an arbitrary distribution, $p(T_i|\mathbf{x}_i, \boldsymbol{\beta})$, parameterized by $\boldsymbol{\beta}$ given covariate vector $\mathbf{x}_i$. Using the conventional notations for right-censored survival data, δ_i is the censoring indicator for the i-th unit and $y_i = \min(T_i, C_i)$, where C_i is the censoring time. The complete data log-likelihood function for the i-th unit can be formatted as

$$l(\boldsymbol{\beta}; y_i, \delta_i, \mathbf{x}_i) = \delta_i \log(p(y_i|\mathbf{x}_i, \boldsymbol{\beta})) + (1 - \delta_i) \log(S(y_i|\mathbf{x}_i, \boldsymbol{\beta})) \tag{6.5}$$

where $S(t|\mathbf{x}_i, \boldsymbol{\beta})$ is the survival function of failure time T_i. For example, when the failure time T_i has a Weibull distribution with scale parameter k,

$$l(\boldsymbol{\beta}; y_i, \delta_i, \mathbf{x}_i) = \delta_i \left\{ \log \left[\frac{\mathbf{x}_i'\boldsymbol{\beta}}{k} (\mathbf{x}_i'\boldsymbol{\beta} y_i)^{\frac{1}{k}-1} \right] + (-\mathbf{x}_i'\boldsymbol{\beta} y_i)^{\frac{1}{k}} \right\} + (1-\delta_i)(-\mathbf{x}_i'\boldsymbol{\beta} y_i^{\frac{1}{k}}) \tag{6.6}$$

Then the weighted EM method in Ibrahim (1990) can be used, given all the other assumptions are the same as in Ibrahim (1990).

Wei and Tanner (1990) propose a Monte Carlo version of the EM algorithm (MCEM), which relaxes the requirement of categorical covariates to continuous or mixed categorical and continuous covariates. In E step, the expected value of the log-likelihood function with respect to the conditional distribution of missing data z given observed data y under the current estimate of the parameters θ_0 can be formally written as:

$$Q(\theta, \theta_0) = \int_Z \log(p(\theta|z, y)) p(z|y, \theta_0) dz \tag{6.7}$$

where Z is the sample space for the latent data z. In M step, the goal still is to find the maximizer of function $Q(\theta, \theta_0)$ as an update of θ. The same as a general EM method, assume

$$Q_{i+1}(\theta, \theta^{(i)}) = \int_Z \log(p(\theta|z, y)) p(z|y, \theta^{(i)}) dz \tag{6.8}$$

is the Q function after the $(i+1)$-th iteration. Since the missing data is continuous or a mix of continuous and categorical, we can't evaluate the integral in (6.8) by setting a sum of finite terms. In addition, the specification of the integrand can make it impossible to evaluate it analytically. A Monte Carlo approach then is proposed to approximate $Q_{i+1}(\theta, \theta^{(i)})$ by $\frac{1}{m}\sum_{j=1}^{m} \log(p(\theta|z^{(j)}, y))$, where $z^{(1)}, z^{(2)}, \ldots, z^{(m)}$ are sampled from the conditional distribution $p(z|y, \theta^{(i)})$ and $\theta^{(i)}$ is the estimate of θ from the i-th iteration.

Ibrahim et al. (1999) develop a method for missing continuous or mixed categorical and continuous covariates for any parametric regression model using MCEM from Wei and Tanner (1990). Using the idea of composition method, the marginal distribution of covariates is modeled

as a product of one-dimensional conditional distributions,

$$\begin{aligned} p(x_{i1}, x_{i2}, \ldots, x_{ip}|\boldsymbol{\alpha}) &= p(x_{ip}|x_{i1}, \ldots, x_{i,p-1}, \alpha_p) \\ &\times p(x_{i,p-1}|x_{i1}, \ldots, x_{i,p-2}, \alpha_{p-1}) \\ &\times \cdots \times p(x_{i2}|x_{i1}, \alpha_2)p(x_{i1}|\alpha_1) \end{aligned} \tag{6.9}$$

where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$ is the p-dimensional covariate vector and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_p)$. α_j is a vector of the indexing parameters for the j-th conditional distribution and they are all different. Since our main interest is in the regression parameters, we treat the indexing parameters of marginal distribution of the covariates as nuisance parameters. Adopting the form in Equation (6.9), more flexibility is allowed for us to model the marginal distribution of the covariates and the number of nuisance parameters introduced in E step can be reduced.

Apart from EM-like methods, Lipsitz and Ibrahim (1998) propose a likelyhood-based approach to estimate the parameters of Cox's semiparametric proportional hazards model when some categorical covariate values are missing at random using a set of estimation equations and a feasible Monte Carlo method similar to MCEM algorithm proposed by Wei and Tanner (1990) to obtain parameter estimates. Using the same notations as in Equation (6.5), and T_i is assumed to follow the Cox proportional hazards regression model (Cox (1972)). Then the probability distribution for data $(y_i,\ \delta_i)$ given $\mathbf{x}_i$ of the i-th individual is proportional to

$$p(y_i, \delta_i|\mathbf{x}_i, \boldsymbol{\beta}) = \left[\lambda_0(y_i)e^{\mathbf{x}_i'\boldsymbol{\beta}}\right]^{\delta_i} \exp\{-e^{\mathbf{x}_i'\boldsymbol{\beta}}\Lambda_0(y_i)\} \tag{6.10}$$

where $\lambda_0(t)$ is an arbitrary baseline hazard function and $\Lambda_0(t)$ is the cumulative baseline hazard function which is defined as

$$\Lambda_0(t) = \int_0^t \lambda_0(u)\ du \tag{6.11}$$

For example if $\lambda_0(t)$ is defined as a two-piece function:

$$\lambda_0(t) = \begin{cases} \lambda_1(t), & \text{if } 0 < t \leqslant t_1 \\ \lambda_2(t), & \text{if } t_1 < x < \infty \end{cases} \tag{6.12}$$

where $\lambda_i(t)$ is parameterized by ϕ_i and $\boldsymbol{\phi} = (\phi_1, \phi_2)$, then we will have the corresponding log-likelihood function for the i-th subject as

$$l(\boldsymbol{\beta}, \boldsymbol{\phi}, y_i, \delta_i, \mathbf{x}_i) = \begin{cases} \delta_i[\log(\lambda_1(y_i)) + \mathbf{x}_i'\boldsymbol{\beta}] - \exp(\mathbf{x}_i'\boldsymbol{\beta}) \int_0^{y_i} \lambda_1(\tau)d\tau, & \text{if } 0 < y_i \leqslant t_1 \\ \delta_i[\log(\lambda_2(y_i)) + \mathbf{x}_i'\boldsymbol{\beta}] - \exp(\mathbf{x}_i'\boldsymbol{\beta})(\int_0^{t_1} \lambda_1(\tau)d\tau + \int_{t_1}^{y_i} \lambda_2(\tau)d\tau), & \text{if } t_1 < y_i < \infty \end{cases} \tag{6.13}$$

Adopting the counting process notation and considering complete data, the Cox partial likelihood score vector turns out as:

$$\mathbf{u}_\beta(\boldsymbol{\beta}) = \sum_{i=1}^{n} \int_0^{\infty} \{\mathbf{x}_i - \bar{\mathbf{x}}(s, \boldsymbol{\beta})\} dN_i(s) \tag{6.14}$$

Notations in Equation (6.14) are described in the following way: $N_i(t)$ is the failure indicator of the i-th subject at time t and $N_i(t) = 1$ if the observed failure time T_i is smaller than time t, otherwise $N_i(t) = 0$. $U_i(t)$ is the risk indicator of the i-th subject at time t, and $U_i(t) = 1$ if the i-th subject is at risk at time t, otherwise $U_i(t) = 0$. As a weighted average of $\mathbf{x}_i$'s,

$$\bar{\mathbf{x}}(s, \boldsymbol{\beta}) = \frac{\sum_{i=1}^{n} \mathbf{x}_i U_i(s) e^{\mathbf{x}_i'\boldsymbol{\beta}}}{\sum_{i=1}^{n} U_i(s) e^{\mathbf{x}_i'\boldsymbol{\beta}}} \tag{6.15}$$

and $dN_i(s) = N_i(s) - N_i(s^-)$ which is a binary random variable only taking value one if the failure time T_i equals s. The solution of $\mathbf{u}_\beta(\boldsymbol{\beta}) = 0$ is the maximum partial likelihood estimate $\hat{\boldsymbol{\beta}}$. Moreover, using the Breslow estimate (Breslow 1974), the baseline hazard function $\lambda_0(t)$ can be estimated as

$$\hat{\lambda}_0(t) = \frac{\sum_{i=1}^{n} dN_i(t)}{\sum_{i=1}^{n} U_i(t) e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}}} \tag{6.16}$$

Thus, an equation having $\hat{\lambda}_0(t)$ and $\hat{\boldsymbol{\beta}}$ as solution is constructed as below:

$$u_\lambda[\lambda_0(t), \boldsymbol{\beta}] = \sum_{i=1}^{n} \left[dN_i(t) - \lambda_0(t) U_i(t) e^{\mathbf{x}_i'\boldsymbol{\beta}} \right] \tag{6.17}$$

Eventually, Equation (6.10) can be estimated by

$$\hat{p}(y_i, \delta_i | \mathbf{x}_i, \boldsymbol{\beta}) = \left[\hat{\lambda}_0(y_i) e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}} \right]^{\delta_i} \exp\{-e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}} \hat{\Lambda}_0(y_i)\} \tag{6.18}$$

When we have data missing at random, models for covariate distribution are needed. Here we specify the marginal distribution of covariate $\mathbf{x}_i$ as in Equation (6.9). With the help of complete data maximum likelihood estimating equations

$$u_\alpha(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \frac{\partial \log(p(\mathbf{x}_i|\boldsymbol{\alpha}))}{\partial \boldsymbol{\alpha}} \tag{6.19}$$

the estimate of $\boldsymbol{\alpha}$, $\hat{\boldsymbol{\alpha}}$ can be solved from equation $u_\alpha(\boldsymbol{\alpha}) = 0$.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \lambda_0(t), \boldsymbol{\alpha})$, then all the estimates for complete data scenario turn out to be solutions satisfying the following estimating equations:

$$u(\boldsymbol{\theta}) = \begin{bmatrix} \mathbf{u}_\beta(\boldsymbol{\beta}) \\ u_\lambda[\lambda_0(t), \boldsymbol{\beta}] \\ u_\alpha(\boldsymbol{\alpha}) \end{bmatrix} = 0 \tag{6.20}$$

In order to deal with categorical data missing at random, we take the expectation of $u(\boldsymbol{\theta})$ with respect to the conditional distribution of the missing data given the observed data, and denote it as $u^*(\boldsymbol{\theta})$. Then solution to equation $u^*(\boldsymbol{\theta}) = 0$ is the estimate $\hat{\boldsymbol{\theta}}$. When the covariates are restricted to be categorical, $u^*(\boldsymbol{\theta})$ can be rewritten as:

$$u^*(\boldsymbol{\theta}) = \sum_{\mathbf{x}_{\text{mis},1}(j)}^{n_1} \cdots \sum_{\mathbf{x}_{\text{mis},n}(j)}^{n_n} p_{1j} \cdots p_{nj} \begin{bmatrix} \sum_{i=1}^{n} \int_0^\infty \{\mathbf{x}_i - \bar{\mathbf{x}}(s, \boldsymbol{\beta})\} dN_i(s) \\ \sum_{i=1}^{n} \left\{ dN_i(t) - \lambda_0(t) U_i(t) e^{\boldsymbol{\beta}' \mathbf{x}_i} \right\} \\ \sum_{i=1}^{n} \frac{\partial \log(p(\mathbf{x}_i|\boldsymbol{\alpha}))}{\partial \boldsymbol{\alpha}} \end{bmatrix} \tag{6.21}$$

In Equation (6.21), the conditional probability p_{ij} is defined as

$$\begin{aligned} p_{ij} &= \text{pr}[\mathbf{x}_{\text{mis},i} = \mathbf{x}_{\text{mis},i}(j) | \mathbf{x}_{\text{obs},i}, y_i, \delta_i, \boldsymbol{\theta}] \\ &= p[\mathbf{x}_{\text{mis},i}(j) | \mathbf{x}_{\text{obs},i}, y_i, \delta_i, \boldsymbol{\theta}] \\ &= \frac{p(y_i, \delta_i | \mathbf{x}_{\text{mis},i}(j), \mathbf{x}_{\text{obs},i}, \lambda, \boldsymbol{\beta}) p(\mathbf{x}_{\text{mis},i}(j), \mathbf{x}_{\text{obs},i} | \boldsymbol{\alpha})}{\sum_{\mathbf{x}_{\text{mis},i}} p(y_i, \delta_i | \mathbf{x}_i, \lambda, \boldsymbol{\beta}) p(\mathbf{x}_i | \boldsymbol{\alpha})} \end{aligned} \tag{6.22}$$

where $j = 1, \ldots, n_i$ are the indexes of the n_i distinct covariate patterns that $\mathbf{x}_{\text{mis},i}$ can take given

(y_i, δ_i) and $\sum_{j=1}^{n_i} p_{ij} = 1$. An EM-type algorithm is proposed to solve Equation (6.21) and details can be found in Lipsitz and Ibrahim (1998).

Papers that discuss Bayesian approaches to model survival data with MAR covariate values are rare. Ibrahim et al. (2001) detail Bayesian approaches to survival semiparametric models with MAR covariate data using informative priors. The approach is similar to Equation (6.12), but now the baseline hazard rate function is assumed to have a more general form. The time axis is partitioned into J intervals, $0 = t_0 < t_1 < t_2 < \ldots < t_{J-1} < t_J$, with t_J no less than the censoring time and a constant hazard rate λ_j within each interval $(t_{j-1}, t_j]$. Keeping all the other notations and assumptions the same as those for Equation (6.10), the complete data likelihood function for the i-th subject is

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda} | y_i, \delta_i, \mathbf{x}_i) = \left[\lambda_j e^{\mathbf{x}_i'\boldsymbol{\beta}}\right]^{\delta_i} \exp\{-e^{\mathbf{x}_i'\boldsymbol{\beta}}[\lambda_j(y_i - t_{j-1}) + \sum_{g=1}^{j-1} \lambda_g(t_g - t_{g-1})]\} \tag{6.23}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_J)$. Again the marginal distribution of covariates $\mathbf{x}_i$, $p(\mathbf{x}_i|\boldsymbol{\alpha})$, is specified as in Equation (6.9). Thus, the joint posterior density of $(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\alpha})$ condition on the observed data D_{obs} is

$$\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\alpha} | D_{\text{obs}}) \propto \prod_{i=1}^{n} \left\{ \int_{\mathbf{x}_{\text{mis},i}} L(\boldsymbol{\beta}, \boldsymbol{\lambda} | y_i, \delta_i, \mathbf{x}_i) p(\mathbf{x}_i|\boldsymbol{\alpha}) d\mathbf{x}_{\text{mis},i} \right\} \times \pi(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\alpha}) \tag{6.24}$$

where $\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\alpha})$ is the joint prior distribution of $(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\alpha})$ and it is constructed via the joint covariate distributions of observed historical data and parameter prior before historical data is observed.

There is also literature on handling non-ignorable missing data in survival models using frequentist methods. In this scenario, in addition to the marginal distribution of covariates aforementioned with MAR data, the missing data mechanism is now must also be considered. Leong et al. (2001) propose a model to handle categorical but non-ignorable missing data through a sequence of one-dimensional conditional distributions. The approach follows the same fashion in Lipsitz and Ibrahim (1998), and keeps all the notations the same as in Equation (6.21), but adds another estimating equation about missing data mechanism to Equation (6.21). $\mathbf{R}_i = (R_{i,1}, \ldots, R_{i,K})'$ is defined as the missing covariate indicator for the covariates of the i-th subject. If $x_{i,k}$ is observed,

then $R_{i,k}$ equals one, otherwise zero. They define the missing data mechanism as the conditional distribution of $\mathbf{R}_i$ given $(y_i, \delta_i, \mathbf{x}_i)$ parameterized by vector $\boldsymbol{\phi}$, with notation $p(\mathbf{r}_i|y_i, \delta_i, \mathbf{x}_i, \boldsymbol{\phi})$. Defined in this way, $p(\mathbf{r}_i|y_i, \delta_i, \mathbf{x}_i, \boldsymbol{\phi})$ follows the multinomial distribution with 2^K possible outcomes. Therefore, the maximum likelihood estimate of $\boldsymbol{\phi}$ over complete data is the solution of equation $\mathbf{u}_\phi(\boldsymbol{\phi}) = 0$, where

$$u_\phi(\boldsymbol{\phi}) = \sum_{i=1}^{n} \frac{\partial \log(p(\mathbf{r}_i|y_i, \delta_i, \mathbf{x}_i, \boldsymbol{\phi}))}{\partial \boldsymbol{\phi}} \tag{6.25}$$

In order to reduce the number of nuisance parameters used in $p(\mathbf{r}_i|y_i, \delta_i, \mathbf{x}_i, \boldsymbol{\phi})$, instead of using a multinomial logistic regression, $p(\mathbf{r}_i|y_i, \delta_i, \mathbf{x}_i, \boldsymbol{\phi})$ is represented as a product of a sequence of one dimensional conditional distributions:

$$\begin{aligned} p(r_{i1}, r_{i2}, \ldots, r_{iK}|\boldsymbol{\phi}) = {} & p(r_{iK}|r_{i1}, \ldots, r_{i,K-1}, \delta_i, y_i, \mathbf{x}_i, \phi_K) \\ & \times p(r_{i,K-1}|r_{i1}, \ldots, r_{i,K-2}, \delta_i, y_i, \mathbf{x}_i, \phi_{K-1}) \\ & \times \cdots \times p(r_{i2}|r_{i1}, \delta_i, y_i, \mathbf{x}_i, \phi_2) p(r_{i1}|\delta_i, y_i, \mathbf{x}_i, \phi_1) \end{aligned} \tag{6.26}$$

where each of the distribution can be modeled by logistic regression, as r_{ik} is a binary outcome. Incorporating the missing data mechanism, the conditional probability p_{ij} becomes

$$\begin{aligned} p_{ij} &= \mathrm{pr}[\mathbf{x}_{\mathrm{mis},i} = \mathbf{x}_{\mathrm{mis},i}(j)|\mathbf{x}_{\mathrm{obs},i}, y_i, \delta_i, \mathbf{r}_i, \boldsymbol{\theta}] \\ &= p[\mathbf{x}_{\mathrm{mis},i}(j)|\mathbf{x}_{\mathrm{obs},i}, y_i, \delta_i, \mathbf{r}_i, \boldsymbol{\theta}] \\ &= \frac{p(y_i, \delta_i|\mathbf{x}_{\mathrm{mis},i}(j), \mathbf{x}_{\mathrm{obs},i}, \lambda, \boldsymbol{\beta}) p(\mathbf{x}_{\mathrm{mis},i}(j), \mathbf{x}_{\mathrm{obs},i}|\boldsymbol{\alpha}) p(\mathbf{r}_i|\mathbf{x}_{\mathrm{mis},i}(j), \mathbf{x}_{\mathrm{obs},i}, y_i, \delta_i, \boldsymbol{\phi})}{\sum_{\mathbf{x}_{\mathrm{mis},i}} p(y_i, \delta_i|\mathbf{x}_i, \lambda, \boldsymbol{\beta}) p(\mathbf{x}_i|\boldsymbol{\alpha}) p(\mathbf{r}_i|\mathbf{x}_i, y_i, \delta_i, \boldsymbol{\phi})} \end{aligned} \tag{6.27}$$

Parameter estimates can be obtained via the method described in Lipsitz and Ibrahim (1998).

Herring et al. (2004) propose a model for proportional hazards with non-ignorably missing covariates that are categorical or continuous or mixed. The approach is very similar to the one discussed in Leong et al. (2001), except the conditional density Equation (6.27) used in the E step now consists of an integral rather than a sum.

$$p[\mathbf{x}_{\mathrm{mis},i}|\mathbf{x}_{\mathrm{obs},i}, y_i, \delta_i, \mathbf{r}_i, \boldsymbol{\theta}] = \frac{p(y_i, \delta_i|\mathbf{x}_{\mathrm{mis},i}, \mathbf{x}_{\mathrm{obs},i}, \lambda, \boldsymbol{\beta}) p(\mathbf{x}_{\mathrm{mis},i}, \mathbf{x}_{\mathrm{obs},i}|\boldsymbol{\alpha}) p(\mathbf{r}_i|\mathbf{x}_{\mathrm{mis},i}, \mathbf{x}_{\mathrm{obs},i}, y_i, \delta_i, \boldsymbol{\phi})}{\int_{\mathbf{x}_{\mathrm{mis},i}} p(y_i, \delta_i|\mathbf{x}_i, \lambda, \boldsymbol{\beta}) p(\mathbf{x}_i|\boldsymbol{\alpha}) p(\mathbf{r}_i|\mathbf{x}_i, y_i, \delta_i, \boldsymbol{\phi}) d\mathbf{x}_{\mathrm{mis},i}} \tag{6.28}$$

Monte Carlo Expectation Maximization(MCEM) algorithm is used to obtain parameter estimates.

In a model, we may have either fixed covariates or time-varying covariates, or both. Besides missing fixed covariates, timing-varying covariates can also be missing. There is literature in Bayesian approaches about non-ignorably missing covariates in survival models with time-varying covariates. Bradshaw et al. (2010) propose a Fully Bayesian(FB) approach to model proportional hazards with non-ignorably missing time-varying covariates. The joint likelihood consists a series of conditional distributions: the marginal distribution of covariates; the Cox proportional hazards regression model the distribution of the event; the distribution of missing data mechanism. With non-informative priors specified for model parameters, this approach will yield similar posterior mean and standard deviations of parameter estimates to those from maximum likelihood method. Moreover, the FB framework is not only less computationally intensive than MCEM framework (Herring et al. 2004) for this model , but also can yield variance estimates more easily.

In this approach, failure time T_i for the i-th unit is assumed to have a Cox piecewise exponential hazard distribution given the covariates, $p(T_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\beta}, \boldsymbol{\lambda})$, parameterized by $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ given covariate vector $\mathbf{x}_i$ and $\mathbf{z}_i$. For the i-th subject, $\mathbf{x}_{ik} = (x_{ik1}, \ldots, x_{ikp})$ denotes the k-th measurement of p completely observed variables. and $\mathbf{z}_{ik} = (z_{ik1}, \ldots, z_{ikq})$ denotes the k-th measurement of q variables with possible missing values, where k takes value from 1 to K_i. Using the conventional notations for right-censored survival data, δ_i is the censoring indicator for the i-th unit and $y_i = \min(T_i, U_i)$, where U_i is the censoring time. The whole time axis is partitioned into J intervals, $0 = t_0 < t_1 < t_2 < \ldots < t_{J-1} < t_J$, with t_J no less than the maximum of y_i and a constant hazard rate λ_j within each interval $(t_{j-1}, t_j]$. Noticing that the total number of measurement K_i may smaller than J, we therefore introduce notations for covariates over each interval. For interval $(t_{j-1}, t_j]$, the covariates for subject i in it is defined as $\mathbf{x}^*_{ij} = (x^*_{ij1}, \ldots, x^*_{ijp})'$ and $\mathbf{z}^*_{ij} = (z^*_{ij1}, \ldots, z^*_{ijq})'$, where x^*_{ijl} and z^*_{ijl} can be imputed from observations of previous interval and following interval. In order to model the missing data mechanism, missing data indictor $\mathbf{r}_i$ is introduced. $\mathbf{r}_{ik} = (r_{ik1}, \ldots, r_{ikq})$ is the missingness indicator for $\mathbf{z}_{ik}$. If z_{ikl} is missing, r_{ikl} will be 1, otherwise 0. If y_i fall into interval $(t_{j-1}, t_j]$, the complete data likelihood function for the i-th unit can be formatted as

$$L(\boldsymbol{\beta}, \boldsymbol{\lambda}|y_i, \delta_i, \mathbf{x}_i, \mathbf{z}_i) = \Big[\lambda_j \exp(\mathbf{x}_{ij}^{*'}\boldsymbol{\beta}_1 + \mathbf{z}_{ij}^{*'}\boldsymbol{\beta}_2)\Big]^{\delta_i} \times \exp\{-\exp(\mathbf{x}_{ij}^{*'}\boldsymbol{\beta}_1 + \mathbf{z}_{ij}^{*'}\boldsymbol{\beta}_2)\lambda_j(y_i - t_{j-1}) - \sum_{g=1}^{j-1}\exp(\mathbf{x}_{ig}^{*'}\boldsymbol{\beta}_1 + \mathbf{z}_{ig}^{*'}\boldsymbol{\beta}_2)\lambda_g(t_g - t_{g-1})\} \tag{6.29}$$

where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_J)$ denotes the baseline hazard rate vector and the covariate effects vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ consists two subvectors $\boldsymbol{\beta}_1$ with dimension $p \times 1$ and $\boldsymbol{\beta}_2$ with dimension $q \times 1$, which are associated with covariates $\mathbf{x}_{ik}^*$ and $\mathbf{z}_{ik}^*$ respectively.

The joint distribution of the missing covariates $\mathbf{z}_i$ is modeled via a series of one-dimensional conditional distributions as following:

$$\begin{aligned} p_{\mathbf{z}}(\mathbf{z}_i|\mathbf{x}_i, \boldsymbol{\alpha}) &= p(z_{iK_iq}|z_{iK_i1}, \ldots, z_{iK_i(q-1)}, \mathbf{z}_{i(K-1)}, \ldots, \mathbf{z}_{i1}, \mathbf{x}_{iK_i}, \alpha_{K_iq}) \\ &\times \cdots \times p(z_{iK_i1}|\mathbf{z}_{i(K_i-1)}, \ldots, \mathbf{z}_{i1}, \mathbf{x}_{iK_i}, \alpha_{K_i1}) \\ &\times \cdots \times p(z_{i(K_i-1)q}|z_{i(K_i-1)1}, \ldots, z_{i(K_i-1)(q-1)}, \mathbf{z}_{i(K_i-2)}, \ldots, \mathbf{z}_{i1}, \mathbf{x}_{i(K_i-1)}, \alpha_{(K_i-1)q}) \\ &\times \cdots \times p(z_{i(K_i-1)1}|\mathbf{z}_{i(K_i-2)}, \ldots, \mathbf{z}_{i1}, \mathbf{x}_{i(K_i-2)}, \alpha_{(K_i-1)1}) \\ &\times \cdots \times p(z_{i1q}|z_{i11}, \ldots, z_{i1(q-1)}, \mathbf{x}_{i1}, \alpha_{1q}) \\ &\times \cdots \times p(z_{i11}|\mathbf{x}_{i1}, \alpha_{11}) \end{aligned} \tag{6.30}$$

Similarly, the joint distribution of the missing data mechanism $\mathbf{r}_i$ is modeled via a series of one-dimensional conditional distributions as following:

$$\begin{aligned} p_{\mathbf{r}}(\mathbf{r}_i|y_i, \mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\phi}) &= p(r_{iK_iq}|r_{iK_i1}, \ldots, r_{iK_i(q-1)}, \mathbf{r}_{i(K-1)}, \ldots, \mathbf{r}_{i1}, \mathbf{x}_{iK_i}, \mathbf{z}_{iK_i}, y_i, \phi_{K_iq}) \\ &\times \cdots \times p(r_{iK_i1}|\mathbf{r}_{i(K_i-1)}, \ldots, \mathbf{r}_{i1}, \mathbf{x}_{iK_i}, \mathbf{z}_{iK_i}, y_i, \phi_{K_i1}) \\ &\times \cdots \times p(r_{i(K_i-1)q}|r_{i(K_i-1)1}, \ldots, r_{i(K_i-1)(q-1)}, \mathbf{r}_{i(K_i-2)}, \ldots, \mathbf{r}_{i1}, \mathbf{x}_{i(K_i-1)}, \mathbf{z}_{i(K_i-1)}, y_i, \phi_{(K_i-1)q}) \\ &\times \cdots \times p(r_{i(K_i-1)1}|\mathbf{r}_{i(K_i-2)}, \ldots, \mathbf{r}_{i1}, \mathbf{x}_{i(K_i-2)}, \mathbf{z}_{i(K_i-2)}, y_i, \phi_{(K_i-1)1}) \\ &\times \cdots \times p(r_{i1q}|r_{i11}, \ldots, r_{i1(q-1)}, \mathbf{x}_{i1}, \mathbf{z}_{i1}, y_i, \phi_{1q}) \\ &\times \cdots \times p(r_{i11}|\mathbf{x}_{i1}, y_i, \phi_{11}) \end{aligned} \tag{6.31}$$

In all, substituting Equation (6.29), (6.30) and (6.31) into the following equation, we can have the complete data likelihood.

$$\begin{aligned} l(\boldsymbol{\beta}, \lambda, \alpha, \phi) &= \prod_{i=1}^{n} p(\mathbf{r}_i, y_i, \mathbf{z}_i | \mathbf{x}_i, \boldsymbol{\beta}, \lambda, \alpha, \phi) \\ &= \prod_{i=1}^{n} p_{\mathbf{r}}(\mathbf{r}_i | y_i, \mathbf{z}_i, \mathbf{x}_i, \phi) p_y(y_i | \mathbf{z}_i, \mathbf{x}_i, \boldsymbol{\beta}, \lambda) p_{\mathbf{z}}(\mathbf{z}_i | \mathbf{x}_i, \alpha) \end{aligned} \tag{6.32}$$

Consequently, given observed data, the joint posterior distribution of the parameters is proportional to:

$$l(\boldsymbol{\beta}, \lambda, \alpha, \phi | y, \mathbf{r}, \mathbf{x}, \mathbf{z}) \propto \left(\prod_{i=1}^{n} \int_{\mathbf{z}_i} p(y_i, \mathbf{r}_i, \mathbf{x}_i, \mathbf{z}_i | \boldsymbol{\beta}, \lambda, \alpha, \phi) \, d\mathbf{z}_i \right) \times p(\boldsymbol{\beta}, \lambda, \alpha, \phi) \tag{6.33}$$

6.3 Summary of approaches

Of all the literature on missing covariates in survival models, both Bayesian and frequentist methods have been proposed. The ideas of these methods are similar in that they all consider models for event times and models for covariates. When the missing data is non-ignorable, then models for missing data mechanism are included.

In frequentist approaches, the failure time of subject i can either be modeled via a parametric distribution, for example, a Weibull distribution (Lipsitz and Ibrahim 1996); or via the non-parametric Breslow estimate (Breslow 1974) as in estimation equation approaches (e.g., Lipsitz and Ibrahim (1998) and Leong et al. (2001)).

In Bayesian approaches, using proportional hazard regression model, a general form of complete data time-to-event likelihood function for the i-th unit can be formatted as

$$L(\boldsymbol{\beta} | y_i, \delta_i, \mathbf{x}_i) = \left[\lambda_0(y_i) e^{\mathbf{x}_i'(y_i)\boldsymbol{\beta}} \right]^{\delta_i} \exp\{ - \int_0^{y_i} e^{\mathbf{x}_i'(\tau)\boldsymbol{\beta}} \lambda_0(\tau) d\tau \} \tag{6.34}$$

where the failure time of the i-th subject is T_i; censoring time is C_i; $y_i = \min(T_i, C_i)$; δ_i is the conventional censoring indicator; $\lambda_0(t)$ is the baseline hazard rate function, $\mathbf{x}_i(t)$ is the vector of the observed covariates for the i-th subject at time t and vector $\boldsymbol{\beta}$ denotes the regression coefficients.

There are different ways to evaluate Equation (6.34) or its logarithmic value equivalently. Here we assume the regression coefficients and covariates $\mathbf{x}_i(t)$ are constants. Frequently, the

integral in Equation (6.34) doesn't have a closed form, therefore approximation is needed. One typical way is assuming the baseline hazard rate $\lambda_0(t)$ is a piecewise constant function based on a partition of the time axis. In this way, the integral turns out to be a finite sum that is easy to calculate, such as in Equation (6.23). When covariates $\mathbf{x}_i(t)$ are time-dependent, the integral can be approximated similar to Equation (6.29). The hazard rate over each interval can be assumed to have the same non-informative independent prior by sharing the same hyperparameters. For instance, in Bradshaw et al. (2010), given λ_j is the hazard rate of the j-th time interval, the prior density of λ_j is specified as gamma distributions with shape and inverse scale parameters of 0.01. In our MRH model, we also assume our baseline hazard rate is a piecewise constant function of time, but the priors of hazard rate λ_j's are not independent and generally they are not the same either. Only with certain assumptions, hazard rate prior over each interval will have the same gamma density, details can be found in Section 1.3.2.

When it comes to model the covariates and missing data mechanism, the same technique is adopted. The whole joint marginal distribution of covariates is modeled by a series of one-dimensional conditional distributions. The missing data mechanism is modeled in the same fashion if needed. Details of how they are modeled can be found in Ibrahim et al. (1999), Ibrahim et al. (2001), Leong et al. (2001), Herring et al. (2004) and Bradshaw et al. (2010). This approach to model the joint marginal density will reduce the number of nuisance parameters that have to be specified effectively. Equation (6.30) represents the model of missing time-varying covariates. For subject i, we can see each missing covariate z_{ikl} is sequentially conditioning on other $\mathbf{z}$ values at the k-th measurement, all $\mathbf{z}$ values prior to the k-th measurement and the fully observed covariate data $\mathbf{x}_{ik}$.

Frequentist approaches are all EM-type algorithms, no matter the covariates are missing at random(MAR) or non-ignorably missing. First, by setting up either the likelihood functions or a set of estimation equations, then the parameter estimation steps are implemented by either regular Expectation-maximization(EM) algorithm (Dempster et al. 1977) or an extended EM algorithm, such as Monte Carlo EM (Wei and Tanner 1990). When the missing covariates all are discrete

random variables with finite range, the E-step can be set up as a sum of weighted complete data log-likelihood and regular EM method is used (e.g., Lipsitz and Ibrahim (1996), Lipsitz and Ibrahim (1998) and Leong et al. (2001)). When the missing covariates are continuous or mixed categorical and continuous covariates, the E-step consists of an integral rather than a sum and MCEM algorithm is used (e.g., Herring et al. (2004)). In M-step, to find a maximizer of the quantity from E-step, Newton-Raphson is always used.

There is limited literature on Bayesian approaches to survival analysis with missing covariate data. Ibrahim et al. (2001) describe Bayesian approaches to survival semiparametric models with MAR covariate data using informative priors in details and Bradshaw et al. (2010) propose methods about non-ignorably missing time-vary covariates in survival models. Gibbs sampler is used to obtain estimates of parameters. Within each iteration of Gibbs sampler, the full conditionals could be sampled from either directly, or using adaptive rejection sampling (ARS) algorithm (Gilks and Wild 1992) or other sampling methods, depending on the property of conditional posterior distributions.

6.4 Approaches for censored data

Besides missing covariate values, missing response is also a common scenario in survival analysis. In this section we are going to discuss some common statistical methods to analyse censored data. There are four basic classes for handling censored data.

First, the same as our previous discussion, we can also use complete data analysis only over the uncensored complete observations. With a usual censoring percentage of 50% or more in clinical study, this approach will significantly decrease our sample size and undoubtedly lead to biased inferences. Only when the censored data is MCAR can we get unbiased estimates via this method. Secondly, we can impute the missing survival time by a left-point imputation, which assumes that all censored data fails right after the censoring time. Or we can impute the censored data with right-point, given assumption that these cases never fail. The survival probability will be either underestimated or overestimated with these two methods, so neither of them are ideal for censored

data. Third, analysis can be carried out based on dichotomized data. Instead of considering right-censoring and interval-censoring, we only study if the event happens or not within a time-window. Subjects is indicated as 1 if a failure is observed, otherwise as 0. Then standard methods such as logistic regression and contingency table can be used for analysis. This method has some disadvantages, including: that we can't tell if a subject is drop-out the study before it ends or it is censored at the end of the study, we can't model the variability in the timing of the event, and for models like this, we can't incorporate any time-varying covariates. The last class is likelihood-based approach. A common characteristic of likelihood-based approaches is that adjustment is made based on whether or not an individual observation is censored. The non-parametric Kaplan-Meier estimator of the survival function and the Cox model are both likelihood-based approaches. These four methods are described further in Leung et al. (1997).

6.5 Practical implications

Now we consider MRH model with missing time-varying covariates. In a MRH model, when the time resolution is picked, we have the time axis partitioned into 2^M even length time intervals. Regardless of whether the time-varying covariates are measured routinely or randomly for a subjects, it is highly possible that for a covariate, we may have no measurement within one or more time intervals. This scenario can be considered as missing data for this predesigned resolution. In order to approximate integrals in calculating cumulative hazard function, our MRH model assumes that within each time interval, the baseline hazard rate is constant, and the measurement for each covariate is also a constant for a subject. With this assumption, we need only one measurement for each covariate in a time interval for each subject. The following are the good ways of handling missing data in a MRH model:

- Use the previous measurement to substitute the missing one, or use the average of the previous and next measurement to substitute the missing value. This is a very rough estimate of the missing covariate, especially when we have multiple missing observation in

successive time intervals.

- Treat the missing measurement as parameters, and model the distribution of missing covariates, sampling the missing value iteratively in the Gibbs sampler procedure.

Apart from missing data, we may also have multiple measurements for a covariate in the same bin for a subject. Theoretically, we need to integrate each covariate as a function of time in calculating MLEs. But in general, we don't know the exactly form of covariates as function of time. As shown before, for a subject, an average of all the measurements for the same covariate within a time interval can be used as a good and easy approximation of covariate value we need for MRH model, since we always only have discrete measurements along the time. If we want to be more accurate, we can fit these covariate data with a regression model, then average the integral of the model over each time interval, and these average values will be used in our MRH model for this covariate. And this can also be used to impute missing covariates.

Chapter 7

Evaluating MRH models with missing time-varying covariates with simulated data

In this chapter, we evaluate MRH models with missing time-varying covariates using simulated datasets. We first talk about how to generate missing time-varying covariates from a full dataset. Then we propose methods to impute missing time-varying covariates. In addition, we show results of applying MRH models to simulated datasets with missing time-varying covariates using their corresponding imputed datasets. In the end, we discuss how the shape of function for a time-varying covariate will affect missing data imputation.

7.1 Generating missing time-varying covariates

First, we generate datasets with missing time-varying covariates. We still use the datasets generated in Section 5.2 to build datasets with missing covariate values. For each dataset, we assume no gender information is missing and we want to have some of the time-varying covariate glucose values missing completely at random with a given percentage. For example, if we want 5% of glucose values missing complete at random for a dataset, then for each glucose value in that dataset, we run a Bernoulli distribution with success probability 95% for it. If the outcome is 1, we keep this glucose value, otherwise we discard it. Then our missing covariate dataset is generated. Using this procedure, we generate datasets having 5% and 15% of glucose values missing completely at random. The full datasets we using are the 200 datasets of size 200 whose time-varying covariates are generated from cosine shape functions and the 200 datasets of size 200 whose time-varying

covariates are generated from five degree polynomials from Section 5.2.

7.2 Missing data imputation

When some of the covariate values in a dataset are missing, we can't implement our TVC-MRH model discussed in Chapter 5 directly to the dataset. We need to impute the missing covariates first. We impute the missing covariates through the following steps.

(1) For a patient with one or more missing glucose values, we fit a curve based on the other available glucose values and associate generate time t of this patient. This assumption is reasonable. In real life, if we have a measurement on the chart, unusually the measure time will also be on file. Here we use polynomials to fit the data, polynomials can be degree one or more. And we select the relatively best model by comparing their AIC values.

(2) Once we figure out the model of a patient's glucose value, we start to generate predictive value of glucose $G^*(t)$ at each given time t (the t's we use to generate glucose value $G(t)$ for this patient, discarding the ones associated with missing data). The predictive value is generated from predictive intervals of regression models. Statistic

$$\frac{G^*(t) - \hat{G}(t)}{S\sqrt{1 + \frac{1}{n} + \frac{n(t-\bar{t})^2}{n\sum t_i^2 - (\sum t_i)^2}}} \tag{7.1}$$

has a student's t distribution with degree of freedom $n - p$, where p is the number of parameters in the selected model. Here $G^*(t)$ is the predictive value at time t; $\hat{G}(t)$ is the fitted value at time t; n is the number of data points used in fitting the model; S is the residual standard error; t_i's are the measure times associated to the given glucose values. For instance, given a subject with glucose value $G(t_1)$, $G(t_2)$, $G(t_3)$, $G(t_5)$, $G(t_6)$, $G(t_8)$ and their corresponding measure times, t_1, t_2, t_3, t_5, t_6 and t_8. We sample a set of $G^*(t_i)$, $i = 1, 2, 3, 5, 6, 8$ using Equation (7.1). Then we fit the data we get in this step and model is selected by comparing AIC still.

(3) For each missing glucose value of this subject, we impute it by integrating the fitted curve from step (2) over its associated time interval and dividing by the interval length. Still use the example from step 7.1, we impute $G(t_4)$ and $G(t_7)$ in this step.

(4) We repeat step (2) to step (3) enough times, such as 1000, then we get a sample of $G(t_4)$ estimates and a sample of $G(t_7)$ estimates. Then we calculate the mean and standard error of those estimates, for $G(t_4)$ and $G(t_7)$ separately.

For a dataset, we run the above steps, for each subject. In the end, we will have a estimated mean and standard error for each missing glucose values. Now for each missing data, we generate a value using a normal distribution of mean equal to its estimated mean and standard deviation equal to its estimated standard error. And these values together with the initial available values will consist a full dataset. We repeated this five times, so in the end we get 5 imputed full dataset per one datasets with missing time-varying covariates.

In step (1), we mentioned that fit a curve based on the other available glucose values and associate generate time t of this patient. There are two ways of understanding other available glucose values. If the event of our interest is failure/death, then there is no way for us to have the glucose values after the observed failure time for a subject. In this case, we can at most have all the glucose values before the failure time available. Thus, only those values can be used to fit the glucose value associated curve in step (1). But if the event of our interest is recurrence, then it is possible that we can also have glucose values available after the event happens until the termination of the study. In other words, in step (1) we can use all available values in the record to fit a model regardless of the event time. In this chapter, we impute missing values for the same dataset using both of these two ways. In order to make the reference clear, we name the dataset imputed from the first way as conditional imputed missing time-varying covariates(C-MTVC), since the availability of time-varying covariates is also depending on the event time. And we name the dataset imputed from the second way as unconditional imputed missing time-varying covariates(MTVC).

And we also have another way of imputing missing covariates. We refer it as deterministic imputation. The former one in this section is referred as non-deterministic imputation. This deterministic imputation procedure is very similar to the one discussed aforementioned. Step (1) is the same. In step (2), instead of sampling a set of predictive values $G^*(t)$, we directly use the fitted model from step (1) as the curved needed in step (3) and impute the missing value. All 3 steps just need to be run once and no step (4) any more. Combining with the two different types of step (1), we give a name for dataset imputed from this procedure as conditional deterministic imputed missing time-varying covariates(EC-MTVC) and unconditional deterministic imputed missing time-varying covariates(E-MTVC). In all, we may have four different types of imputed datasets, together with conditional imputed missing time-varying covariates(C-MTVC), unconditional imputed missing time-varying covariates(MTVC).

Later on when we compare MRH models over datasets imputed from different methods, we will first include the type of the imputed datasets and the followed by the MRH model we use. For example, notation EC-MTVC-NPM3 means the missing data are imputed from unconditional and exact imputation procedure described early in this section and we use a 3-level MRH model without any pruning over this imputed dataset.

Eventually, for each dataset with missing data generated in Section 7.1, we impute five MTVC datasets, five C-MTVC datasets, an E-MTVC dataset and an EC-MTVC dataset for it.

7.3 Analysis of parameter estimates

First, for each set of data from Section 7.2, we implemented MRH strategy NPM3 (3-level model with time-varying covariates without any pruning) and EPEM3 (piecewise exponential hazard model with $2^3 = 8$ equal width intervals). The details of applying MRH models to datasets with time-varying covariates can be found in Section 5.1. To fresh our memory, the notation are described as following:

- **E-MTVC-NPM3:** dataset is imputed from unconditional and deterministic imputation procedure and a 3-level MRH model without any pruning is used
- **MTVC-NPM3:** dataset is imputed from unconditional and non-deterministic imputation procedure and a 3-level MRH model without any pruning is used
- **EC-MTVC-NPM3:** dataset is imputed from conditional and deterministic imputation procedure and a 3-level MRH model without any pruning is used
- **C-MTVC-NPM3:** dataset is imputed from conditional and non-deterministic imputation procedure and a 3-level MRH model without any pruning is used
- **E-MTVC-EPEM3:** dataset is imputed from unconditional and deterministic imputation procedure and a piecewise exponential hazard model with $2^3 = 8$ equal width intervals is used
- **MTVC-EPEM3:** dataset is imputed from unconditional and non-deterministic imputation procedure and a piecewise exponential hazard model with $2^3 = 8$ equal width intervals is used
- **EC-MTVC-EPEM3:** dataset is imputed from conditional and deterministic imputation procedure and a piecewise exponential hazard model with $2^3 = 8$ equal width intervals is used
- **C-MTVC-EPEM3:** dataset is imputed from conditional and non-deterministic imputation procedure and a piecewise exponential hazard model with $2^3 = 8$ equal width intervals is used

For MRH models, MCMC chains with 200000 iterations for each of the 200 datasets were run separately. The first 50000 iterations of each MCMC chain was discarded as the burn-in, and every 10th sample from the chain was kept to reduce autocorrelation. In the end, for the group of E-MTVC-NPM3 and EC-MTVC-NPM3, 15000 posterior samples per dataset were used to derive

posterior PMRH estimates (posterior means), resulting in 200 sets of estimates. For the group of MTVC-NPM3 and C-MTVC-NPM3, since each of them had five chains for the same dataset, we first merged the 15000 posterior samples from five chains in the same group together. We then had 75000 posterior samples per dataset per group, and similarly they were used to derive posterior PMRH estimates (posterior means), resulting in 200 sets of estimates. And we use these means to calculate the corresponding 95% probability intervals for each parameter of our interest.

All the simulations were coded in R and run on a supercomputer with 1368 nodes, each containing two hex-core 2.8Ghz Intel Westmere processors with 12 cores per node and 2GB of RAM per core. Since all the datasets are of size 200 and the model applied to them is the same, for each chain the computation time is about the same. It takes about 2.1 hours to complete 200000 iterations no matter how the dataset was imputed.

For the piecewise exponential hazard models, we examine the MLEs for model parameters. For the group of E-MTVC-EPEM3 and EC-MTVC-EPEM3, we have 200 sets of MLEs separately. For the group of MTVC-NPM3 and C-MTVC-NPM3, since each of them had five sets of MLEs for the same dataset, we take the average of MLEs for each parameter first, still resulting in a set of MLEs per dataset. Again we have 200 sets of MLEs separately in the end. And we use these MLEs to calculate the corresponding 95% probability intervals for each parameter of our interest.

In Table 7.1 and Table 7.2, we give the 95% probability intervals for parameters of model TVC-NPM3, E-MTVC-NPM3, MTVC-NPM3, EC-MTVC-NPM3 and C-MTVC-NPM3 over 200 datasets with 200 subjects per set. The time-varying covariates in the original datasets are generated from five degree polynomial Equation (5.11) and 5% of the glucose values are missing completely at random for the associated dataset with missing covariates.

From the result of TVC-NPM3, we can see when there is no missing data. The true parameter values can be estimated efficiently. Among all other four models involving missing data imputation, model E-MTVC-NPM3 performs the best, the followed by model EC-MTVC-NPM3, then model MTVC-NPM3, and model C-MTVC-NPM3. It is not surprising because when we implemented model E-MTVC-NPM3, for each subject, the missing data was imputed by making full use of

all available covariate values regardless of the corresponding event time. And instead of using multiple predictive curve, the missing value was imputed only using the fitted model average over its associated interval, which definitely introduce less uncertainty to the imputed value. More information and less uncertainty surely will lead to a better model performance. Comparing the result from model MTVC-NPM3 and model C-MTVC-NPM3, and result from model E-MTVC-NPM3 and model EC-MTVC-NPM3 we can find that with less information used in data imputation process and other imputation steps be the same , conditional imputed datasets will always result in the worse performance for a same model. Comparing the result from model E-MTVC-NPM3 and model MTVC-NPM3, and result from model EC-MTVC-NPM3 and model C-MTVC-NPM3 we can find that with the same amount of information used in data imputation process, exact imputed datasets will always lead to better performance for a same model, since less uncertainty are brought into the imputation process. We can also tell that no matter how the datasets are imputed, the estimates for constant covariate effect always can be approximated to a certain accuracy. But when the other parameter estimates are being affected too much, the constant covariate effect estimates will still be influenced, since parameter estimates are all dependent on the other parameter values in MRH models. Result of model C-MTVC-NPM3 is an example of showing this trend.

In Table 7.3, we show the 95% probability intervals for parameters of model TVC-EPEM3, E-MTVC-EPEM3, MTVC-EPEM3, EC-MTVC-EPEM3 and C-MTVC-EPEM3 over 200 datasets with 200 subjects per set. The time-varying covariates in the original datasets are generated from five degree polynomial Equation (5.11) and 5% of the glucose values are missing completely at random for the associated dataset with missing covariates. When comparing the result in Table 7.3 with the result in Table 7.1, we can still find that our MRH models with time-varying covariates can yield equal or less variant parameter estimates than those from piecewise exponential hazard models for the same dataset.

In Table 7.4 we show the result of the 95% probability intervals for parameters of model E-MTVC-NPM3 and E-MTVC-EPEM3 over 200 datasets with 200 subjects per set. The time-varying covariates in the original datasets are generated from five degree polynomial Equation

	H_0	Gender	Glucose	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
True:	1.00	0.48	0.70	0.13	0.07	0.08	0.08	0.14	0.21	0.08	0.20
TVC-NPM3											
2.5%	0.75	0.19	0.54	0.08	0.04	0.05	0.05	0.08	0.13	0.05	0.11
50%	0.96	0.51	0.73	0.13	0.07	0.08	0.08	0.13	0.19	0.08	0.19
mean	0.97	0.50	0.73	0.13	0.07	0.08	0.08	0.14	0.19	0.08	0.19
97.5%	1.26	0.80	0.87	0.19	0.11	0.13	0.13	0.20	0.29	0.13	0.29
E-MTVC-NPM3 (5% of glucose values are missing completely at random)											
2.5%	0.75	0.20	0.56	0.08	0.04	0.05	0.05	0.08	0.12	0.05	0.12
50%	0.96	0.51	0.73	0.13	0.07	0.08	0.08	0.13	0.19	0.08	0.19
mean	0.97	0.51	0.73	0.13	0.07	0.08	0.08	0.14	0.19	0.08	0.19
97.5%	1.25	0.79	0.87	0.19	0.11	0.13	0.13	0.21	0.29	0.13	0.30
MTVC-NPM3 (5% of glucose values are missing completely at random)											
2.5%	0.80	0.16	0.06	0.07	0.04	0.04	0.05	0.08	0.13	0.05	0.14
50%	1.04	0.50	0.51	0.12	0.07	0.07	0.08	0.13	0.19	0.09	0.27
mean	1.05	0.49	0.47	0.12	0.07	0.08	0.08	0.13	0.20	0.10	0.27
97.5%	1.32	0.76	0.83	0.19	0.10	0.12	0.12	0.20	0.29	0.16	0.46
EC-MTVC-NPM3 (5% of glucose values are missing completely at random)											
2.5%	0.78	0.19	0.48	0.08	0.04	0.05	0.05	0.08	0.13	0.05	0.13
50%	0.98	0.51	0.67	0.13	0.07	0.08	0.08	0.13	0.19	0.09	0.21
mean	1.00	0.50	0.67	0.13	0.07	0.08	0.08	0.14	0.20	0.09	0.21
97.5%	1.29	0.78	0.82	0.19	0.11	0.13	0.13	0.20	0.29	0.14	0.33
C-MTVC-NPM3 (5% of glucose values are missing completely at random)											
2.5%	0.95	0.14	0.00	0.07	0.03	0.04	0.04	0.08	0.13	0.07	0.26
50%	1.15	0.46	0.02	0.11	0.06	0.07	0.07	0.12	0.20	0.12	0.38
mean	1.16	0.45	0.02	0.11	0.06	0.07	0.07	0.13	0.21	0.12	0.39
97.5%	1.45	0.72	0.08	0.15	0.09	0.11	0.11	0.19	0.32	0.19	0.54

Table 7.1: Estimates and 95% probability intervals for parameters of model TVC-NPM3, E-MTVC-NPM3, MTVC-NPM3, EC-MTVC-NPM3 and C-MTVC-NPM3 over 200 datasets with size 200 (time-varying covariates are generated from **five degree polynomial** and **5%** of glucose values are missing completely at random in the original datasets)

	$R_{1,0}$	$R_{2,0}$	$R_{2,1}$	$R_{3,0}$	$R_{3,1}$	$R_{3,2}$	$R_{3,3}$
True:	0.37	0.55	0.55	0.65	0.50	0.40	0.29
TVC-NPM3							
2.5%	0.30	0.43	0.44	0.51	0.32	0.29	0.18
50%	0.38	0.55	0.55	0.65	0.50	0.41	0.30
mean	0.38	0.55	0.55	0.65	0.50	0.41	0.30
97.5%	0.47	0.66	0.66	0.78	0.66	0.56	0.43
E-MTVC-NPM3 (5% of glucose values are MCAR)							
2.5%	0.30	0.43	0.44	0.51	0.31	0.29	0.18
50%	0.38	0.55	0.55	0.65	0.50	0.41	0.30
mean	0.38	0.55	0.55	0.65	0.50	0.41	0.30
97.5%	0.47	0.66	0.66	0.78	0.65	0.56	0.44
MTVC-NPM3 (5% of glucose values are MCAR)							
2.5%	0.24	0.44	0.34	0.51	0.31	0.27	0.15
50%	0.34	0.55	0.47	0.65	0.49	0.40	0.27
mean	0.33	0.55	0.48	0.64	0.50	0.40	0.27
97.5%	0.43	0.66	0.65	0.77	0.66	0.54	0.39
EC-MTVC-NPM3 (5% of glucose values are MCAR)							
2.5%	0.29	0.44	0.43	0.52	0.31	0.29	0.18
50%	0.37	0.56	0.53	0.65	0.50	0.41	0.29
mean	0.37	0.55	0.53	0.65	0.50	0.41	0.29
97.5%	0.45	0.66	0.64	0.78	0.65	0.56	0.42
C-MTVC-NPM3 (5% of glucose values are MCAR)							
2.5%	0.21	0.43	0.30	0.51	0.31	0.26	0.14
50%	0.27	0.55	0.40	0.65	0.49	0.38	0.24
mean	0.27	0.55	0.40	0.65	0.49	0.38	0.25
97.5%	0.34	0.66	0.49	0.78	0.65	0.53	0.36

Table 7.2: Estimates and 95% probability intervals for split parameters of model TVC-NPM3, E-MTVC-NPM3, MTVC-NPM3, EC-MTVC-NPM3 and C-MTVC-NPM3 over 200 datasets with size 200 (time-varying covariates are generated from **five degree polynomial** and **5%** of glucose values are missing completely at random(MCAR) in the original datasets)

	H_0	Gender	Glucose	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
True:	1.00	0.48	0.70	0.13	0.07	0.08	0.08	0.14	0.21	0.08	0.20
TVC-EPEM3											
2.5%	0.76	0.18	0.53	0.08	0.04	0.04	0.05	0.08	0.13	0.04	0.11
50%	0.97	0.51	0.73	0.13	0.07	0.08	0.08	0.13	0.19	0.08	0.20
mean	0.98	0.50	0.73	0.13	0.07	0.08	0.08	0.14	0.20	0.08	0.20
97.5%	1.27	0.79	0.92	0.19	0.11	0.13	0.13	0.21	0.30	0.13	0.31
E-MTVC-EPEM3 (5% of glucose values are MCAR)											
2.5%	0.76	0.19	0.54	0.08	0.04	0.04	0.04	0.08	0.13	0.04	0.11
50%	0.97	0.51	0.73	0.13	0.07	0.08	0.08	0.14	0.19	0.08	0.19
mean	0.98	0.50	0.73	0.13	0.07	0.08	0.08	0.14	0.20	0.08	0.20
97.5%	1.27	0.79	0.92	0.19	0.11	0.13	0.13	0.21	0.30	0.13	0.32
MTVC-EPEM3 (5% of glucose values are MCAR)											
2.5%	0.80	0.16	0.07	0.07	0.03	0.04	0.04	0.08	0.13	0.05	0.13
50%	1.02	0.49	0.58	0.12	0.07	0.07	0.08	0.13	0.20	0.09	0.25
mean	1.04	0.49	0.54	0.12	0.07	0.08	0.08	0.13	0.20	0.09	0.26
97.5%	1.34	0.78	0.86	0.19	0.11	0.13	0.12	0.20	0.30	0.15	0.45
EC-MTVC-EPEM3 (5% of glucose values are MCAR)											
2.5%	0.79	0.18	0.47	0.08	0.03	0.04	0.04	0.08	0.13	0.05	0.13
50%	0.99	0.50	0.66	0.13	0.07	0.08	0.08	0.13	0.20	0.08	0.22
mean	1.01	0.49	0.66	0.13	0.07	0.08	0.08	0.14	0.20	0.09	0.22
97.5%	1.32	0.78	0.84	0.19	0.11	0.13	0.13	0.21	0.30	0.14	0.35
C-MTVC-EPEM3 (5% of glucose values are MCAR)											
2.5%	0.87	0.13	0.04	0.07	0.03	0.04	0.04	0.08	0.13	0.05	0.18
50%	1.10	0.47	0.32	0.12	0.06	0.07	0.07	0.13	0.20	0.11	0.33
mean	1.11	0.47	0.33	0.12	0.06	0.07	0.07	0.13	0.21	0.11	0.33
97.5%	1.41	0.74	0.66	0.17	0.10	0.12	0.12	0.20	0.31	0.17	0.50

Table 7.3: Estimates and 95% probability intervals for all parameters of model TVC-EPEM3, E-MTVC-EPEM3, MTVC-EPEM3, EC-MTVC-EPEM3 and C-MTVC-EPEM3 over 200 datasets with size 200 (time-varying covariates are generated from **five degree polynomial** and **5%** of glucose values are missing completely at random(MCAR) in the original datasets)

(5.11) and 15% of the glucose values are missing completely at random for the associated dataset with missing covariates. As expected, our MRH models with time-varying covariates still yield equal or less variant parameter estimates than those from piecewise exponential hazard models. As we discussed above, among all the strategies we used to impute missing covariates in a dataset, the unconditional exact imputation approach will perform relatively the best, so when we increase the percentage of missing value we can still get very attractive estimates for parameters. And if we compare the result in Table 7.4 with the result of model MTVC-NPM3 in Table 7.1, we can find that even with a higher percentage of missing data, if we use exact imputation in our miss data imputation process, we may possibly still get better estimates for parameters. Based on the conclusions we draw from Table 7.1, we don't even need to bother the other three imputation strategies, since the more data is missing, surely they will perform even worse. And in both Table 7.1 and Table 7.4 we can tell as more data is missing, or the less available data can be used in imputation process, or use non-exact imputation step, all will lead to the estimates of time-varying covariates become smaller and the estimates of the baseline cumulative hazard function become larger correspondingly. Since any of these three conditions happens, more uncertainty is brought to the imputed dataset.

Exactly as what we did to the imputed dataset associated to five degree polynomial group, now we run all the models over 200 datasets with 200 subjects per set. The time-varying covariates in the original datasets are generated from cosine shape Equation (5.10) and 5% of the glucose values are missing completely at random for the associated dataset with missing covariates. In Table 7.5 and Table 7.6, we give the 95% probability intervals for all parameters of running MRH models. In Table 7.7, are the 95% probability intervals for all parameters of running piecewise exponential hazard models. From the result of TVC-NPM3 and TVC-EPEM3, again, we can see when there is no missing data. The true parameter values can be estimated efficiently. But for all the other models, the estimates for baseline cumulative hazard function and the glucose effects seem to be approximated very badly. In comparison with result from Table 7.1, we may guess that the reason causing this difference is due to the different data structures of the original time-varying

	H_0	Gender	Glucose	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
E-MTVC-NPM3 (15% of glucose values are MCAR)											
2.5%	0.76	0.18	0.53	0.08	0.04	0.05	0.05	0.08	0.12	0.05	0.13
true	1.00	0.48	0.70	0.13	0.07	0.08	0.08	0.14	0.21	0.08	0.20
50%	0.96	0.52	0.72	0.13	0.07	0.08	0.08	0.13	0.19	0.08	0.19
mean	0.98	0.50	0.71	0.13	0.07	0.08	0.08	0.14	0.19	0.08	0.20
97.5%	1.24	0.77	0.87	0.19	0.11	0.13	0.13	0.20	0.29	0.13	0.30
E-MTVC-EPEM3 (15% of glucose values are MCAR)											
2.5%	0.76	0.18	0.52	0.08	0.04	0.04	0.04	0.08	0.12	0.04	0.12
true	1.00	0.48	0.70	0.13	0.07	0.08	0.08	0.14	0.21	0.08	0.20
50%	0.97	0.51	0.71	0.13	0.07	0.08	0.08	0.13	0.19	0.08	0.20
mean	0.99	0.50	0.72	0.13	0.07	0.08	0.08	0.14	0.20	0.08	0.20
97.5%	1.26	0.77	0.93	0.19	0.11	0.13	0.13	0.21	0.30	0.13	0.32

Table 7.4: Estimates and 95% probability intervals for all parameters of model E-MTVC-NPM3 and E-MTVC-EPEM3 over 200 datasets with size 200 (time-varying covariates are generated from **five degree polynomial** and **15%** of glucose values are missing completely at random in the original datasets)

covariates, since except the datasets, both the imputation methods and MRH models used in this two case are completely the same. And MRH models with time-varying covariates provide equal or less variant parameter estimates as well.

For completeness, in Table 7.8 we still show the result of the 95% probability intervals for all parameters of model E-MTVC-NPM3 and E-MTVC-EPEM3 over 200 datasets with 200 subjects per set. The time-varying covariates in the original datasets are generated from cosine shape function Equation (5.10) and 15% of the glucose values are missing completely at random for the associated dataset with missing covariates. As expected, the glucose estimator exhibits poor performance, analogous to the MRH models.

	H_0	Gender	Glucose	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
True:	1.00	0.48	0.70	0.13	0.07	0.08	0.08	0.14	0.21	0.08	0.20
TVC-NPM3											
2.5%	0.73	0.21	0.54	0.09	0.04	0.05	0.04	0.08	0.13	0.03	0.11
50%	0.97	0.52	0.71	0.13	0.07	0.08	0.08	0.13	0.19	0.08	0.20
mean	0.98	0.51	0.72	0.13	0.07	0.08	0.08	0.13	0.20	0.08	0.20
97.5%	1.26	0.82	0.91	0.19	0.11	0.12	0.13	0.20	0.31	0.15	0.31
E-MTVC-NPM3 (5% of glucose values are MCAR)											
2.5%	0.90	0.18	0.00	0.09	0.06	0.07	0.06	0.09	0.15	0.04	0.13
50%	1.16	0.50	0.04	0.13	0.10	0.11	0.10	0.14	0.23	0.10	0.23
mean	1.17	0.50	0.07	0.13	0.10	0.12	0.10	0.14	0.23	0.11	0.24
97.5%	1.53	0.83	0.18	0.19	0.15	0.17	0.17	0.23	0.36	0.19	0.35
MTVC-NPM3 (5% of glucose values are MCAR)											
2.5%	0.92	0.19	0.0	0.09	0.06	0.08	0.06	0.09	0.15	0.04	0.13
50%	1.16	0.50	0.0	0.13	0.10	0.11	0.10	0.14	0.23	0.10	0.24
mean	1.18	0.50	0.0	0.13	0.10	0.12	0.10	0.14	0.23	0.11	0.24
97.5%	1.57	0.81	0.0	0.19	0.15	0.18	0.17	0.23	0.36	0.19	0.36
EC-MTVC-NPM3 (5% of glucose values are MCAR)											
2.5%	0.92	0.18	0.01	0.09	0.06	0.08	0.06	0.09	0.15	0.04	0.14
50%	1.17	0.50	0.03	0.13	0.10	0.11	0.10	0.14	0.23	0.10	0.24
mean	1.18	0.50	0.03	0.13	0.10	0.12	0.10	0.14	0.23	0.11	0.24
97.5%	1.56	0.81	0.03	0.19	0.15	0.18	0.17	0.23	0.36	0.19	0.35
C-MTVC-NPM3 (5% of glucose values are MCAR)											
2.5%	0.92	0.18	0.0	0.09	0.06	0.08	0.06	0.09	0.15	0.04	0.13
50%	1.17	0.50	0.0	0.13	0.10	0.11	0.10	0.14	0.23	0.10	0.24
mean	1.18	0.49	0.0	0.13	0.10	0.12	0.10	0.14	0.23	0.11	0.24
97.5%	1.57	0.81	0.0	0.19	0.15	0.18	0.17	0.23	0.36	0.19	0.36

Table 7.5: Estimates and 95% probability intervals for parameters of model TVC-NPM3, E-MTVC-NPM3, MTVC-NPM3, EC-MTVC-NPM3 and C-MTVC-NPM3 over 200 datasets with size 200 (time-varying covariates are generated from **cosine shape** functions and **5%** of glucose values are missing completely at random(MCAR) in the original datasets)

	$R_{1,0}$	$R_{2,0}$	$R_{2,1}$	$R_{3,0}$	$R_{3,1}$	$R_{3,2}$	$R_{3,3}$
True:	0.37	0.55	0.55	0.65	0.50	0.40	0.29
TVC-NPM3							
2.5%	0.30	0.46	0.43	0.54	0.35	0.28	0.15
50%	0.38	0.55	0.54	0.64	0.50	0.39	0.30
mean	0.38	0.55	0.54	0.65	0.51	0.39	0.30
97.5%	0.47	0.66	0.66	0.77	0.66	0.52	0.47
E-MTVC-NPM3 (5% of glucose values are MCAR)							
2.5%	0.31	0.41	0.41	0.47	0.38	0.27	0.16
50%	0.39	0.51	0.53	0.57	0.53	0.38	0.32
mean	0.39	0.51	0.53	0.58	0.53	0.38	0.31
97.5%	0.47	0.62	0.65	0.70	0.69	0.52	0.52
MTVC-NPM3 (5% of glucose values are MCAR)							
2.5%	0.31	0.41	0.41	0.46	0.38	0.27	0.15
50%	0.39	0.51	0.52	0.57	0.53	0.38	0.31
mean	0.39	0.51	0.52	0.57	0.53	0.38	0.31
97.5%	0.47	0.62	0.64	0.70	0.70	0.52	0.51
EC-MTVC-NPM3 (5% of glucose values are MCAR)							
2.5%	0.31	0.41	0.41	0.47	0.38	0.27	0.15
50%	0.39	0.51	0.52	0.57	0.53	0.38	0.31
mean	0.39	0.51	0.52	0.58	0.53	0.38	0.31
97.5%	0.47	0.62	0.64	0.71	0.70	0.52	0.51
C-MTVC-NPM3 (5% of glucose values are MCAR)							
2.5%	0.32	0.41	0.41	0.46	0.38	0.27	0.15
50%	0.39	0.51	0.52	0.57	0.53	0.38	0.31
mean	0.39	0.51	0.52	0.58	0.53	0.38	0.31
97.5%	0.47	0.62	0.63	0.70	0.70	0.52	0.51

Table 7.6: Estimates and 95% probability intervals for split parameters of model TVC-NPM3, E-MTVC-NPM3, MTVC-NPM3, EC-MTVC-NPM3 and C-MTVC-NPM3 over 200 datasets with size 200 (time-varying covariates are generated from **cosine shape** functions and **5%** of glucose values are missing completely at random(MCAR) in the original datasets)

	H_0	Gender	Glucose	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
True:	1.00	0.48	0.70	0.13	0.07	0.08	0.08	0.14	0.21	0.08	0.20
TVC-EPEM3											
2.5%	0.76	0.23	0.56	0.10	0.04	0.05	0.05	0.08	0.14	0.03	0.12
50%	1.00	0.51	0.71	0.13	0.07	0.08	0.08	0.13	0.21	0.08	0.21
mean	1.00	0.50	0.71	0.14	0.07	0.08	0.08	0.14	0.21	0.08	0.21
97.5%	1.26	0.79	0.88	0.18	0.11	0.12	0.13	0.20	0.31	0.15	0.32
E-MTVC-EPEM3 (5% of glucose values are MCAR)											
2.5%	0.91	0.16	0.00	0.09	0.06	0.07	0.05	0.09	0.15	0.03	0.13
50%	1.16	0.49	0.07	0.13	0.09	0.11	0.10	0.14	0.24	0.10	0.24
mean	1.18	0.49	0.11	0.13	0.10	0.11	0.10	0.14	0.24	0.10	0.24
97.5%	1.56	0.82	0.45	0.20	0.15	0.17	0.17	0.24	0.37	0.18	0.37
MTVC-EPEM3 (5% of glucose values are MCAR)											
2.5%	0.93	0.18	0.0	0.09	0.06	0.08	0.05	0.09	0.15	0.04	0.13
50%	1.19	0.49	0.0	0.13	0.10	0.11	0.10	0.14	0.24	0.10	0.25
mean	1.20	0.48	0.0	0.13	0.10	0.12	0.10	0.14	0.24	0.11	0.25
97.5%	1.60	0.81	0.0	0.19	0.15	0.18	0.17	0.23	0.37	0.20	0.38
EC-MTVC-EPEM3 (5% of glucose values are MCAR)											
2.5%	0.86	0.16	0.02	0.09	0.06	0.06	0.05	0.09	0.15	0.03	0.10
50%	1.13	0.50	0.16	0.13	0.09	0.11	0.10	0.14	0.23	0.09	0.23
mean	1.14	0.50	0.21	0.13	0.09	0.11	0.10	0.14	0.24	0.10	0.23
97.5%	1.50	0.82	0.65	0.19	0.15	0.16	0.17	0.22	0.35	0.18	0.35
C-MTVC-EPEM3 (5% of glucose values are MCAR)											
2.5%	0.93	0.17	0.0	0.09	0.06	0.08	0.05	0.09	0.15	0.04	0.13
50%	1.19	0.49	0.0	0.13	0.10	0.11	0.10	0.14	0.24	0.10	0.25
mean	1.20	0.48	0.0	0.13	0.10	0.12	0.10	0.14	0.24	0.11	0.25
97.5%	1.60	0.80	0.0	0.19	0.15	0.18	0.17	0.23	0.37	0.19	0.38

Table 7.7: Estimates and 95% probability intervals for all parameters of model TVC-EPEM3, E-MTVC-EPEM3, MTVC-EPEM3, EC-MTVC-EPEM3 and C-MTVC-EPEM3 over 200 datasets with size 200 (time-varying covariates are generated from **cosine shape** functions and **5%** of glucose values are missing completely at random(MCAR) in the original datasets)

	H_0	Gender	Glucose	d_1	d_2	d_3	d_4	d_5	d_6	d_7	d_8
E-MTVC-NPM3 (15% of glucose values are MCAR)											
2.5%	0.92	0.18	-0.01	0.09	0.06	0.08	0.06	0.09	0.15	0.04	0.13
true	1.00	0.48	0.70	0.13	0.07	0.08	0.08	0.14	0.21	0.08	0.20
50%	1.16	0.50	0.02	0.13	0.10	0.11	0.10	0.14	0.23	0.10	0.23
mean	1.17	0.50	0.02	0.13	0.10	0.12	0.10	0.14	0.23	0.11	0.24
97.5%	1.53	0.81	0.09	0.19	0.15	0.18	0.17	0.22	0.36	0.19	0.35
E-MTVC-EPEM3 (15% of glucose values are MCAR)											
2.5%	0.93	0.17	-0.01	0.09	0.06	0.08	0.05	0.09	0.15	0.04	0.13
true	1.00	0.48	0.70	0.13	0.07	0.08	0.08	0.14	0.21	0.08	0.20
50%	1.18	0.49	0.03	0.13	0.10	0.11	0.10	0.14	0.24	0.10	0.24
mean	1.19	0.49	0.03	0.13	0.10	0.12	0.10	0.14	0.24	0.11	0.25
97.5%	1.55	0.80	0.10	0.19	0.15	0.18	0.17	0.23	0.37	0.19	0.37

Table 7.8: Estimates and 95% probability intervals for all parameters of model E-MTVC-NPM3 and E-MTVC-EPEM3 over 200 datasets with size 200 (time-varying covariates are generated from **cosine shape** functions and **15%** of glucose values are missing completely at random(MCAR) in the original datasets)

7.4 Theoretical analysis

Recall that, in Section 4.2, we discussed about cumulative hazard function estimation. Consider the i-th subject with failure time y_i, over each time interval $[t_{s-1}, t_s]$, if we use the average value of $G(t)$ of this interval to replace $G(t)$ within this interval, then the general form of overall error bound of approximating the his/her cumulative function $H(y_i)$ is showed in Equation (4.19).

In our simulated data study, we generate the glucose value per each subject in a dataset use a function

$$G(t) = a_0 + a_1 \cos(a_2 t) \tag{7.2}$$

Therefore, with all the assumption the same as in Section 4.2, the overall error for approximating the i-th unit's cumulative hazard function becomes

$$
\begin{aligned}
|\text{error}| &\leqslant \frac{1}{2!}\beta^2 \max(\lambda_0(\tau))e^{\max(G(\tau))\beta}\left\{\int_0^{y_i} G^2(\tau)d\tau - \sum_{s=1}^{j-1}\frac{\left(\int_{t_{s-1}}^{t_s} G(\tau)d\tau\right)^2}{t_s - t_{s-1}} - \frac{\left(\int_{t_{j-1}}^{y_i} G(\tau)d\tau\right)^2}{y_i - t_{j-1}}\right\} \\
&= \frac{1}{2!}\beta^2 \max(\lambda_0(\tau))e^{\max(G(\tau))\beta} \\
&\quad \times\left\{\sum_{s=1}^{j-1}\frac{a_1^2}{4a_2}(\sin(a_2t_s) - \sin(a_2t_{s-1})) + \frac{a_1^2}{2}(t_s - t_{s-1}) - \frac{a_1^2(\sin(a_2t_s) - \sin(a_2t_{s-1}))^2}{a_2^2(t_s - t_{s-1})}\right. \\
&\quad \left. + \frac{a_1^2}{4a_2}(\sin(a_2y_i) - \sin(a_2t_{j-1})) + \frac{a_1^2}{2}(y_i - t_{j-1}) - \frac{a_1^2(\sin(a_2y_i) - \sin(a_2t_{j-1}))^2}{a_2^2(y_i - t_{j-1})}\right\} \\
&= \frac{1}{2!}\beta^2 \max(\lambda_0(\tau))e^{\max(G(\tau))\beta} \\
&\quad \times\left\{\frac{a_1^2}{4a_2}\sin(a_2y_i) + \frac{a_1^2}{2}y_i - \sum_{s=1}^{j-1}\frac{a_1^2(\sin(a_2t_s) - \sin(a_2t_{s-1}))^2}{a_2^2(t_s - t_{s-1})} - \frac{a_1^2(\sin(a_2y_i) - \sin(a_2t_{j-1}))^2}{a_2^2(y_i - t_{j-1})}\right\}
\end{aligned}
\tag{7.3}
$$

Further, in our MRH model, although we have time-varying covariates, but within each interval, even for these time-varying covariates we still assume it is a constant. In other words, in each interval, we just choose a $G(t)$ value to represent the whole $G(t)$ values, when we calculate our true cumulative hazard function and this representing value is selected completely at random. With this assumption, when trying to approximate the value of a time-varying covariate using its average value over that interval, the error can even bigger. In this scenario, Equation (4.13) becomes

$$
\begin{aligned}
|\text{error}| &\leqslant \frac{1}{2!}\beta^2 \max(\lambda_s)e^{\max(G(\tau))\beta}\left\{\sum_{s=1}^{j-1}\int_{t_{s-1}}^{t_s}(a_1\cos(a_2t_s^*) - \frac{a_1(\sin(a_2t_s) - \sin(a_2t_{s-1}))}{a_2(t_s - t_{s-1})})^2 d\tau\right. \\
&\quad \left. + \int_{t_{j-1}}^{y_i}(a_1\cos(a_2t_j^*) - \frac{a_1(\sin(a_2y_i) - \sin(a_2t_{j-1}))}{a_2(y_i - t_{j-1})}))^2 d\tau\right\} \\
&\leqslant \frac{1}{2!}\beta^2 \max(\lambda_s)e^{\max(G(\tau))\beta}\left\{\sum_{s=1}^{j-1}\int_{t_{s-1}}^{t_s} a_1^2 d\tau + \int_{t_{j-1}}^{y_i} a_1^2 d\tau\right\} \\
&= \frac{1}{2!}\beta^2 \max(\lambda_s)e^{\max(G(\tau))\beta}a_1^2 y_i
\end{aligned}
\tag{7.4}
$$

In Equation (7.4), we assume the failure time y_i falls into the j-th time interval and t_s^* denotes a randomly picked time in the s-th time interval $[t_{s-1}, t_s]$. So we can see when use the average value of $G(t)$ over each interval to replace the covariate value in the corresponding interval, and use it to

approximate the cumulative hazard function up to failure time, sometimes the error can be very big as showed in the last line of Equation (7.4).

Similarly, when we deal with missing data, if we use the average value of the fitted curve to be the imputed value of missing data in a interval, the error for calculating hazard increment of the interval can be as large as

$$|\text{error}| \leqslant \frac{1}{2!}\beta^2\lambda_s e^{M_s\beta}a_1^2(t_s - t_{s-1}) \tag{7.5}$$

Here we assume the missing data is in the s-th interval, β is positive and M_s is the maximum value of $G(t)$ over interval $[t_{s-1}, t_s]$. When β is less than 0, corresponding error bound can be figured out.

Moreover, the bounds from Equation (7.5) are calculated based on that we have the true expression of $G(t)$. In real life, we will never have the true expression of covariate curve, so we will always fit a curve for a covariate using its available values. In many case this fitted curve may not reflect the fact, therefore, the error in Equation (7.5) will be even larger. In Figure 7.1 , Figure 7.2 and Figure 7.3 we showed exmaples of how fitted curves can vary from the real values, for the case that one, two, or three out of eight covariate values are missing in a subject's record. When it comes to take average of the fitted curve as the missing value over corresponding intervals, the difference of the true missing value and the average will be even large.

In all, from Figure 7.1 , Figure 7.2 and Figure 7.3, we can conclude that, if we want to use the average value of the covariate over an interval to represent the missing value, intuitively,the covariate need to be very "flat", within each interval from our initial partition of the time axis. Mathematically, for each time interval in the study, the covariate function should have a very small L^2 norm of the difference of itself and its mean over that interval.

From all the simulation tests result and the above discussion, we now have better understanding about how MRH models can perform with missing time-varying covariates. First, it seems no matter what method we use, when the covariate has a pattern of going up and down quite frequently over the study time with a large range, it is very hard to use average value of the covariate within

that interval to substitute the missing value. This coincides with our conclusion drawn aforementioned, since for a function behaves in this way, it is too various and the L^2 norm of itself minus its mean within that interval won't be small. For instance, if over the interval the covariate happen to have a full cycle of cosine function, then the average will be zero. But when we take covariate values in our survival model, for one time interval, when we have a covariate value available, we just make the covariate have this value in this interval. Since this value can be any number in the range of cosine function. It is improper to use its average to replace the missing value.

Secondly, when it comes to fit models with available data, we may also not get good fitted curves. When we use polynomials to fit cosine function base curve, as the cases we analyse here, depending on how the given data spreading sometimes even the fitted model itself will not be good, therefore the average of integral values can't be good. I randomly pick a unit's full data of eight glucose values, including all the glucose values and their corresponding measure times. And the original glucose values are from the dataset we generated in Section 5.3 with cosine shape functions. I randomly discard one or two or three values from these 8 values. Then I use polynomials to fit the rest data, best model is selected by AIC values. I then calculate the predictive value of each missing data using their real measure times. In real life, we may not have the times, so it will be even worse. In Figure 7.1 we have only one data missing at a time, in Figure 7.2 we have two data missing at a time,and in Figure 7.3 we have three data missing at a time. In these Figures, observed data is in blue, the true value of missing data is in red and the predictive values are in green. When there is only one data missing, when the data is missing will affect its prediction value a lot. As in Figure 7.1, we can see when it is missing at the beginning or the end of a study, the prediction value will not be closed to the real one. When a value is missing in the middle of a study, the prediction of it would be much closer to the real one. But when more than one data is missing, no matter when the value is missing, the predictions will become worse and worse. Also even with the same number of missing data, when the data are missing will affect the fitted model and prediction value a lot.

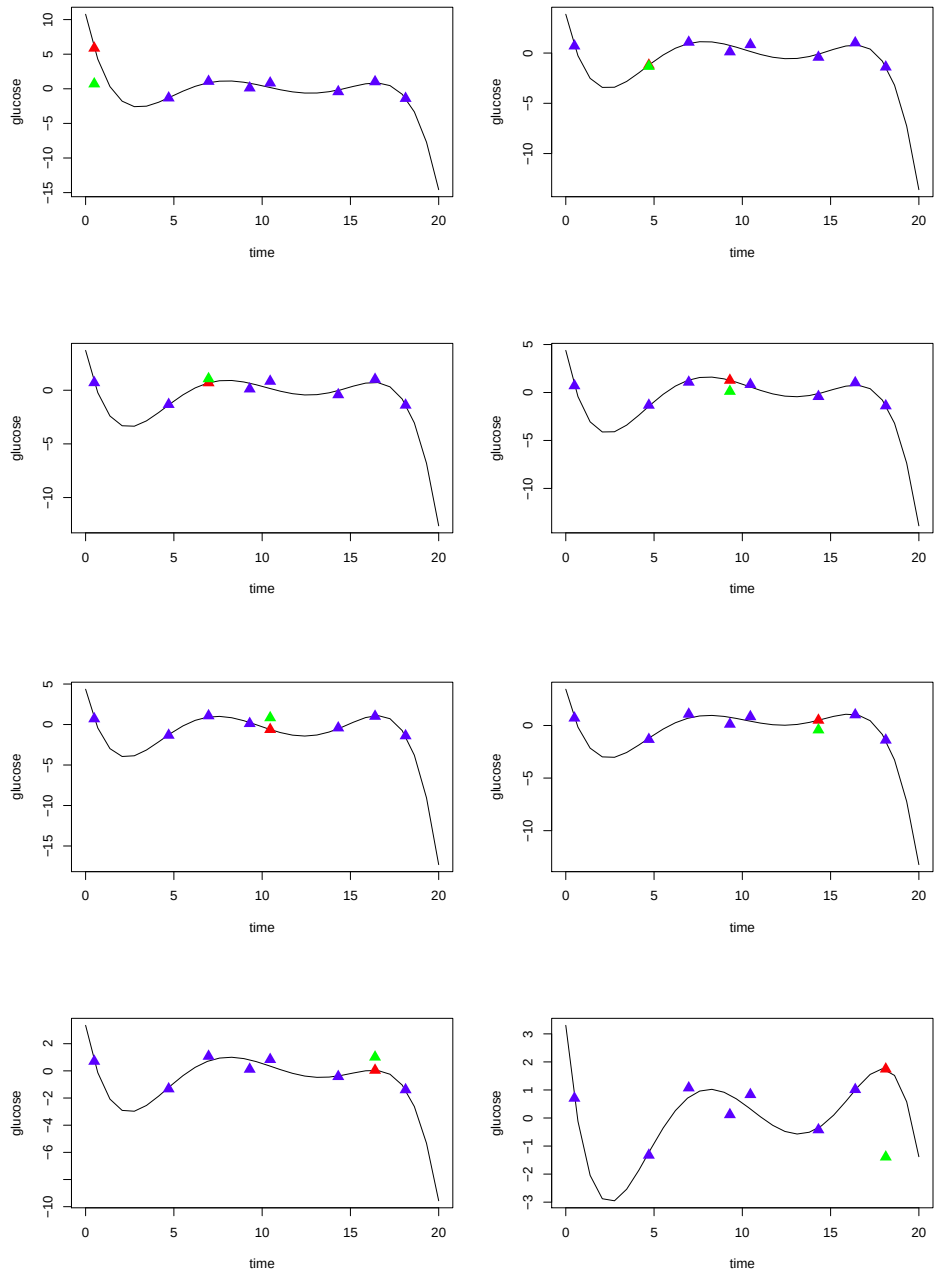

Figure 7.1: Expectation of imputed missing value (in red), missing values are in green, available values are in blue. (original data is generated using cosine like function)

In the same fashion, I plot Figure 7.4, Figure 7.5 and Figure 7.6, where the original full covariate values of a subject is just from the dataset we generated in Section 5.3 with five degree polynomials. Again, we can find that the percentage of covariate values is missing and when the

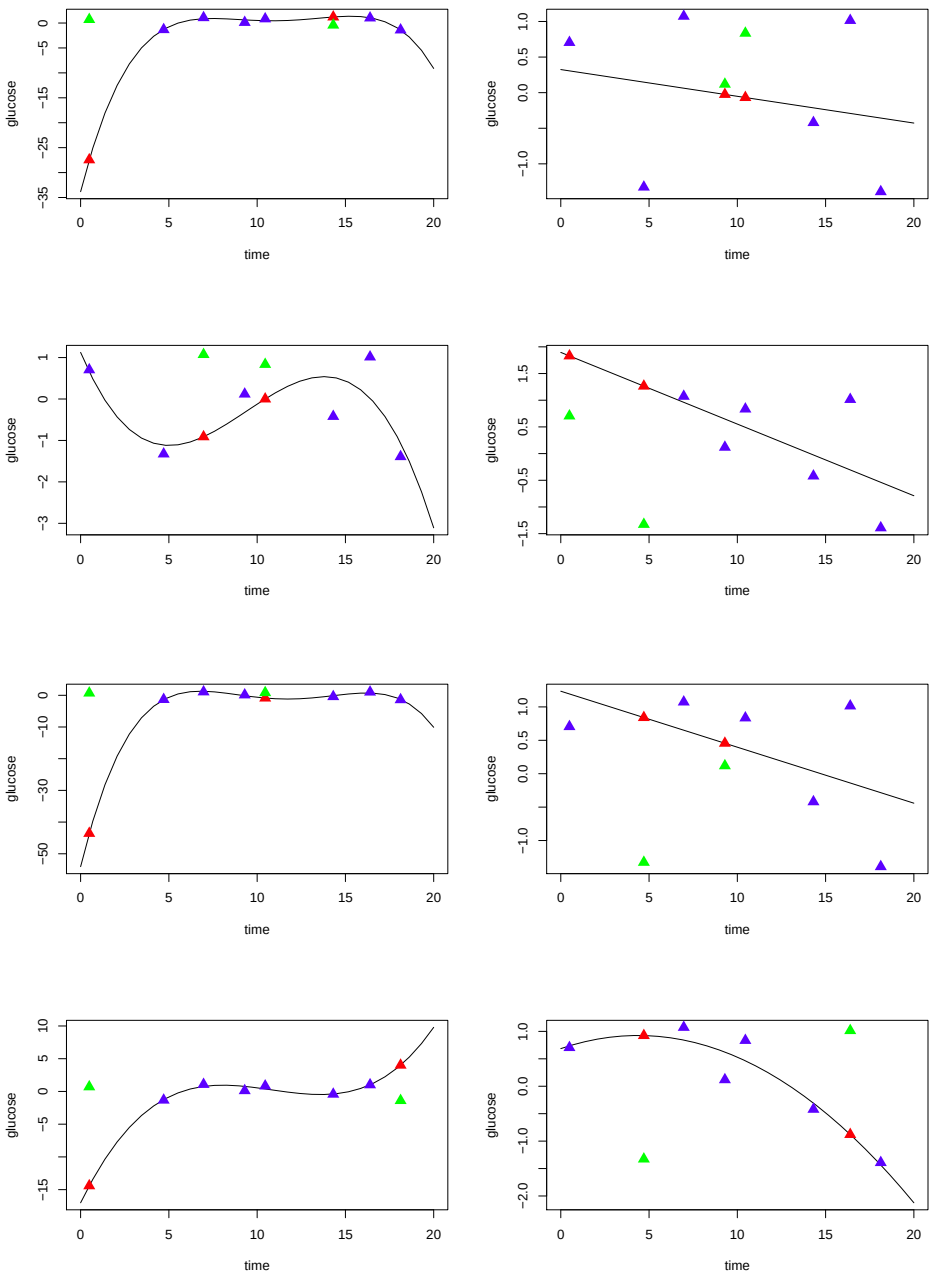

Figure 7.2: Expectation of imputed missing value (in red), missing values are in green, available values are in blue. (original data is generated using cosine like function)

data is missing both will affect the fitted model and therefore may severely affect the missing data imputation that we discussed in Section 7.2.

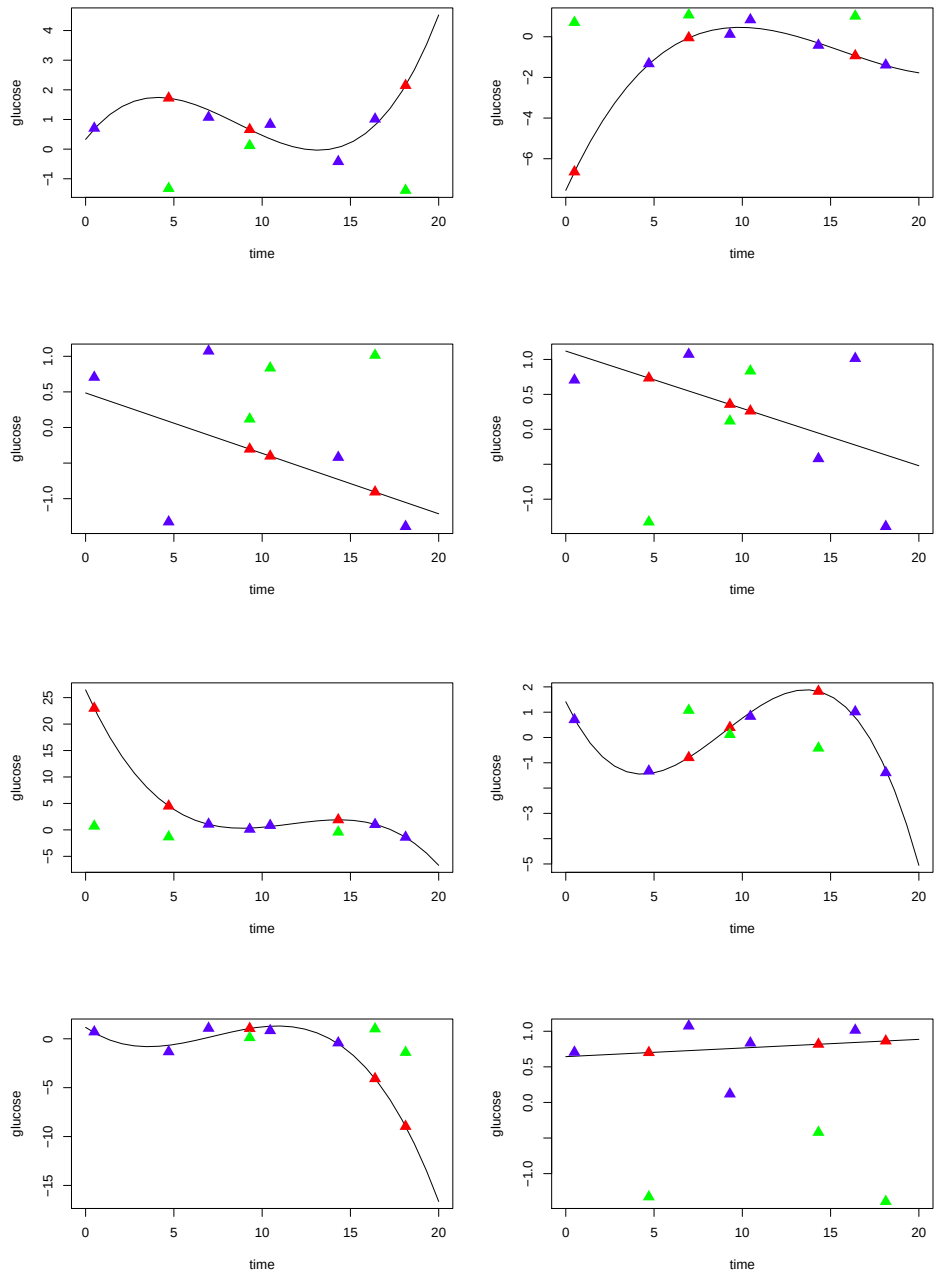

Figure 7.3: Expectation of imputed missing value (in red), missing values are in green, available values are in blue. (original data is generated using cosine like function)

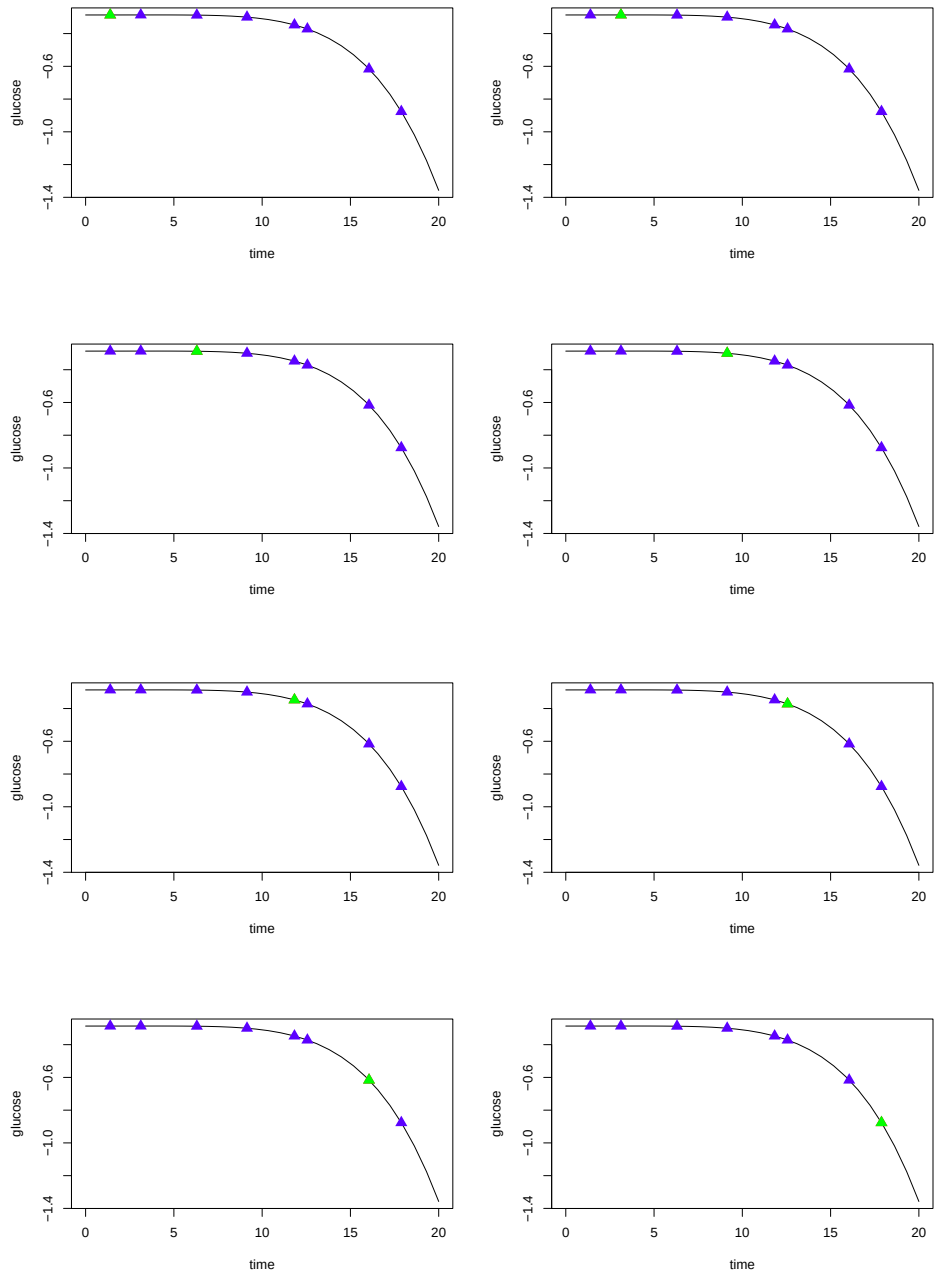

Figure 7.4: Expectation of imputed missing value (in red), missing values are in green, available values are in blue. (original data is generated using five degree polynomial no random noise added)

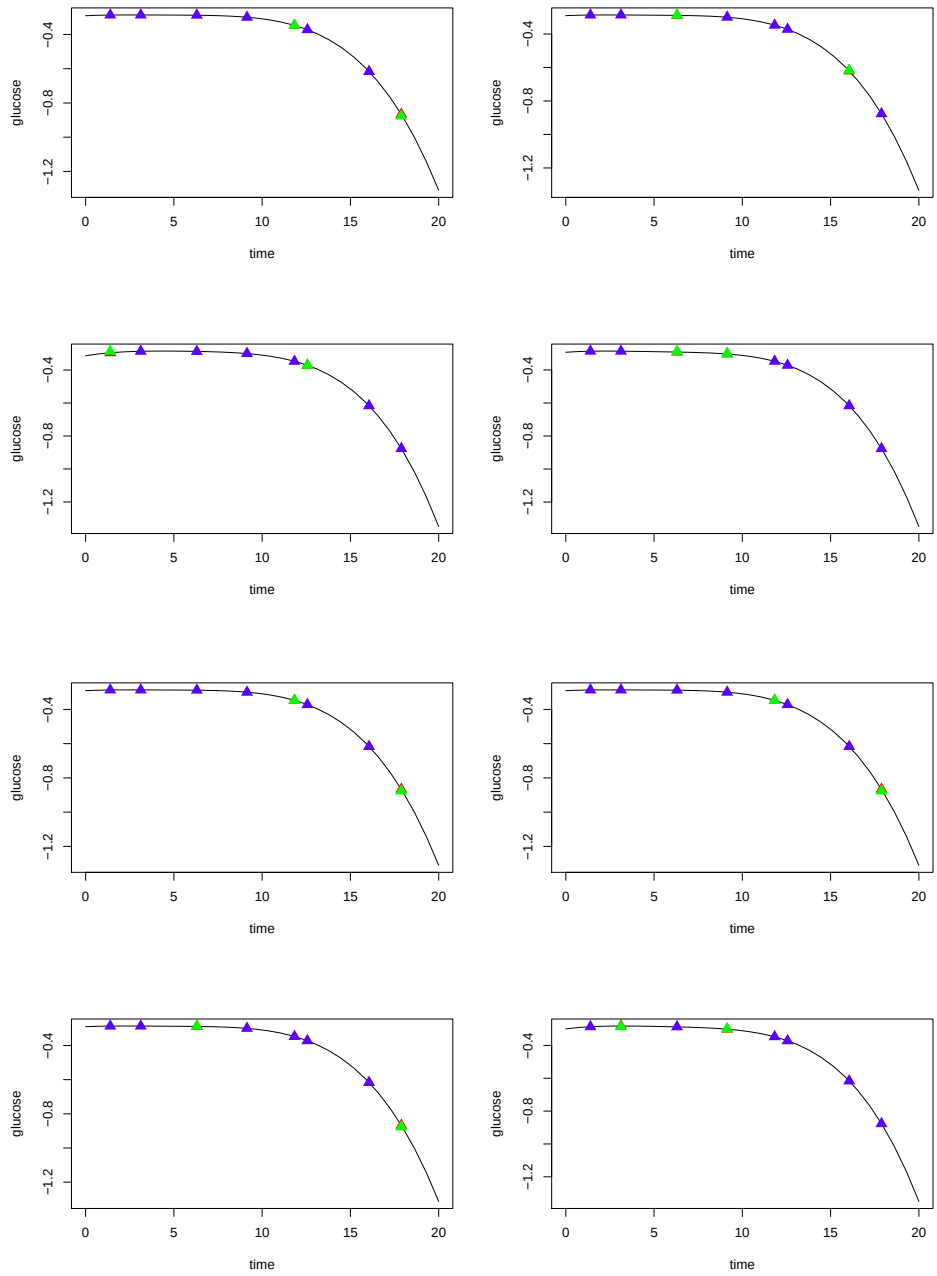

Figure 7.5: Expectation of imputed missing value (in red), missing values are in green, available values are in blue. (original data is generated using five degree polynomial no random noise added)

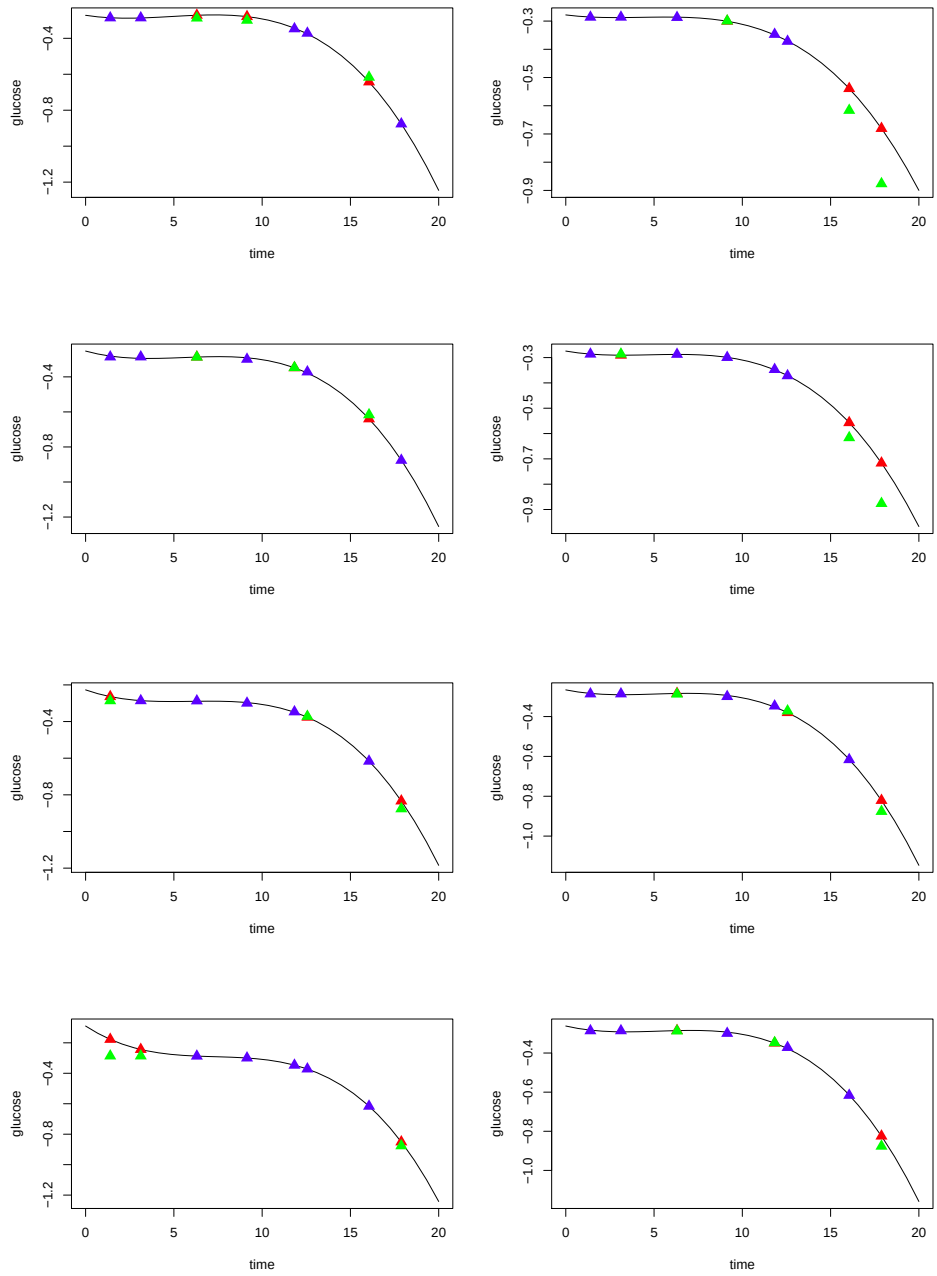

Figure 7.6: Expectation of imputed missing value (in red), missing values are in green, available values are in blue. (original data is generated using five degree polynomial no random noise added)

Chapter 8

Conclusions and future work

8.1 Conclusions

The first part of this thesis we propose an extension to the multiresolution model (MRH) prior that can accommodate data-driven varying smoothness of the hazard rate function over time. The method still relies on the tree-like prior structure for the hazard rate, but makes data-driven choices about using identical hazard rates in adjoint time intervals. The method, which we call pruned MRH (PMRH) method, is studied in simulated data, revealing computational savings, stable estimation, and inferential procedures for the hazard rate, with little impact on covariate effect and the overall hazard estimates. The model is applied to a prostate cancer study, where it was used to jointly estimate the baseline hazard function and the impact of treatment, Gleason score, and age on hazard over time.

The second part investigates how PMRH models can be implemented in data with time-varying covariates, and verify that PMRH models also work well in these data. Apart from full datasets, we also study the missing data imputation strategies and how MRH models can perform with datasets having missing time-varying covariates, using their imputed full datasets from different approaches. Among the four imputation approaches we propose, we show that unconditional exact imputation will lead us to the best MRH model performance when all the other conditions are the same. We find that no matter how much data is missing and no matter which imputation approach is used, the estimates for constant covariate effects can mostly be approximated to certain accuracy. But the estimates for baseline cumulative hazard function have a trend to be larger

and larger, and time-varying covariate effects are getting more and more close to zero, when more covariate values are missing , and the conditional imputation or non-exact imputation or mixed are used. Further more, we investigate how the shape of the original covariate function can affect missing data imputation and therefore affect the performance of MRH models. In order to impute missing data effectively, the L^2 norm of the original covariate function minus the mean of it within each interval from the time axis partition in MRH models need to be small. Moreover, beside the shape of the original covariate function,we show that the percentage of missing covariate values and the time they are missing also would affect the fitted model for covariates. The higher percentage of missing covariates and more missing data at the beginning or close to the failure/censoring time, will make the model fitting more difficult, and result in worse missing data imputation.

8.2 Future work

In this thesis we have shown that our extended MRH models and approaches can solve some survival analysis problems. But there are still more extensions can be made. Currently, all the models we study having the assumption that the covariate effects are time independent. On the other hand, in survival analysis, we also have encountered time-varying effects in applications. Therefore, we are going to incorporate time-varying effects to our PMRH models and MRH models with time-varying covariates in our future work.

Bibliography

Aalen, O. and Gjessing, H. "Understanding the shape of the hazard rate: A process point of view." Statistical Science, 16:1–22 (2001).

Andersen, P., Borgan, O., Gill, R., and Keiding, N. Statistical Methods Based on Counting Processes. Berlin: Springer–Verlag (1993).

Anderson, J. and Senthilselvan, A. "A two-step regression model for hazard functions." Applied Statistics, 31:44–51 (1982).

Antoniadis, A., Gregoire, G., and Nason, G. "Density and hazard rate estimation for right-censored data by using wavelet methods." Journal of the Royal Statistal Society, Series B, 61:63–84 (1999).

Arjas, E. and Gasbarra, D. "Nonparametric Bayesian inference from right censored survival data, using the Gibbs sampler." Statistica Sinica, 4:505–524 (1994).

Austin, P. C. "Generating survival times to simulate Cox proportional hazards models with time-varying covariates." Statistics in Medicine, 31:3946–3958 (2012).

Bouman, P., Dignam, J., Dukic, V., and Meng, X. "A multiresolution hazard model for multi-center survival studies: Application to tamoxifen treatment in early stage breast cancer." Journal of the American Statistical Association, 102:1145–1157 (2007).

Bouman, P., Dukic, V., and Meng, X. "Bayesian multiresolution hazard model with application to an AIDS reporting delay study." Statistica Sinica, 15:325–357 (2005).

Bradshaw, P., Ibrahim, J., and Gammon, M. "A Bayesian proportional hazards regression model with non-ignorably missing time-varying covariates." Statistics in Medicine, 29:3017–3029 (2010).

Breslow, N. "Covariance analysis of censored survival data." Biometrics, 30:89–99 (1974).

Burridge, J. "Empirical Bayes Analysis of Survival Time Data." Journal of the Royal Statistal Society, Series B, 43:65–75 (1981).

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

Dempster, A. P., , Laird, N. M., and Rubin, D. B. "Maximum Likelihood from Incomplete Data via the EM Algorithm." Journal of the Royal Statistical Society - Series B, 39:1–38 (1977).

Dignam, J., Dukic, V., Anderson, S., Mamounas, E., Wickerham, D., and Wolmark, N. "Hazard of recurrence and adjuvant treatment effects over time in lymph node-negative breast cancer." Breast Cancer Research and Treatment, 116:595–602 (2009).

Dukic, V. and Dignam, J. "Bayesian Hierarchical Multiresolution Hazard Model for the Study of Time-Dependent Failure Patterns in Early Stage Breast Cancer." Bayesian Analysis, 2:591–610 (2007).

Fisher, B., Costantino, J., Redmond, C., Poisson, R., Bowman, D., Couture, J., Dimitrov, N., Wolmark, N., Wickerham, D., and Fisher, E. "A randomized clinical trial evaluating tamoxifen in the treatment of patients with node-negative breast cancer who have estrogen-receptor-positive tumors." The New England Journal of Medicine, 320:479–484 (1989).

Fisher, B., Dignam, J., Bryant, J., DeCillis, A., Wickerham, D., Wolmark, N., Costantino, J., Redmond, C., Fisher, E., Bowman, D., Deschenes, D., Dimitrov, N., Margolese, R., Robidoux, A., Shibata, H., Terz, J., Paterson, A., Feldman, M., Farrar, W., Evans, J., and Lickley, H. "Five versus more than five years of tamoxifen therapy for breast cancer patients with negative lymph nodes and estrogen receptor positive tumors." Journal of the National Cancer Institute, 88:1529–1542 (1996).

Fisher, L. and Lin, D. "Time-dependent covariates in the Cox proportional-hazards regression model." Annual Reviews, 20:143–157 (1999).

Geman, S. and Geman, D. "Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images." IEEE Transactions on Pattern Analysis and Machine Intelligence, 6:721–741 (1984).

Gilks, W., Best, N., and Tan, K. "Adaptive rejection Metropolis sampling." Applied Statistics, 44:455–472 (1995).

Gilks, W. and Wild, P. "Adaptive rejection sampling for Gibbs sampling." Applied Statistics, 41:337–348 (1992).

Gore, S., Pocock, S., and Kerry, G. "Regression models and non-proportional hazards in the analysis of breast cancer survival." Applied Statistics, 33:176–195 (1984).

Gray, R. "Some diagnostic methods for Cox regression models through hazard smoothing." Biometrics, 46:93–102 (1990).

—. "Flexible methods for analyzing survival data using splines, with application to breast cancer prognosis." Journal of the American Statistical Association, 87:942–951 (1992).

—. "Hazard rate regression using ordinary nonparametric regression smoothers." Journal of Computational and Graphical Statistics, 5:190–207 (1996).

Herring, A., Ibrahim, J., and Lipsitz, S. "Non-ignorable missing covariate data in survival analysis: a case-study of an International Breast Cancer Study Group trial." Applied Statistics, 53:293–310 (2004).

Hjort, N. "Nonparametric Bayes Estimators Based on Beta Processes in Models for Life History Data." The Annals of Statistics, 18:1259–1294 (1990).

Horwitz, E., Bae, K., Hanks, G., Porter, A., Grignon, D., Brereton, H., Venkatesan, V., Lawton, C., Rosenthal, S., Sandler, H., and Shipley, W. "Ten-year follow-up of Radiation Therapy Oncology Group protocol 92-02: a phase III trial of the duration of elective androgen deprivation in locally advanced prostate cancer." J Clin Oncol, 2497–504 (2008).

Ibrahim, J. "Incomplete Data in Generalized Linear Models." Journal of the American Statistical Association, 85:765–769 (1990).

Ibrahim, J., Chen, M.-H., and Lipsitz, S. "Monte Carlo EM for Missing Covariates in Parametric Regression Models." Biometrics, 55:591–596 (1999).

Ibrahim, J., Chen, M.-H., and Sinha, D. Bayesian Survival Analysis. New York: Springer (2001).

Kalbfleisch, J. "Bayesian analysis of survival distribution." Journal of the Royal Statistal Society, Series B, 40:214–221 (1978).

Kolaczyk, E. "Bayesian multiscale models for Poisson processes." Journal of the American Statistical Association, 94:920–933 (1999).

Lee, J. and Kim, Y. "A new algorithm to generate beta processes." Technical report. Department of Statistics, Pennsylvania State University (2002).

Leong, T., Lipsitz, S., and Ibrahim, J. "Incomplete covariates in the Cox model with applications to biological marker data." Applied Statistics, 50:467–484 (2001).

Leung, K., Elashoff, R. M., and Afifi, A. A. "Censoring Issues in Survival Analysis." Annu. Rev. Public Health, 18:83–104 (1997).

Lipsitz, S. and Ibrahim, J. "Using the EM-Algorithm for Survival Data with Incomplete Categorical Covariates." Lifetime Data Analysis, 2:5–14 (1996).

—. "Estimating equations with incomplete categorical covariates in the Cox model." Biometrics, 54:1002–1013 (1998).

Little, R. "Regression With Missing X's: A Review." Journal of the American Statistical Association, 87:1227–1237 (1992).

Little, R. J. A. "A Test of Missing Completely at Random for Multivariate Data with Missing Values." Journal of the American Statistical Association, 83:1198–1202 (1988).

Little, R. J. A. and Rubin, D. B. Statistical Analysis with Missing Data. New York: John Wiley & Sons (1987).

Nieto-Barajas, L. and Walker, S. "Markov beta and gamma processes for modeling hazard rates." Scandinavian Journal of Statistics, 29:413–424 (2002).

Nowak, R. and Kolaczyk, E. "A statistical multiscale framework for Poisson inverse problems." IEEE Transactions on Information Theory, 46:1811–1825 (2000).

Pilepich, M., Winter, K., John, M., Mesic, J., Sause, W., Rubin, P., Lawton, C., Machtay, M., and Grignon, D. "Phase III Radiation Therapy Oncology Group (RTOG) trial 86-10 of androgen deprivation adjuvant to definitive radiotherapy in locally advanced carcinoma of the prostate." International Journal of Radiation Oncology Biology Physics, 50:1243–1252 (2001).

Pilepich, M., Winter, K., Lawton, C., Krisch, R., Wolkov, H., Movsas, B., Asbell, E. H. S., and Grignon, D. "Androgen suppression adjuvant to definitive radio- therapy in prostate carcinoma: Long-term results of phase III RTOG 85-31." International Journal of Radiation Oncology Biology Physics, 61:1285–1290 (2005).

Prentice, R., Kalbfleisch, J., Peterson, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. "The Analysis of Failure Times in the Presence of Competing Risks." Biometrics, 34:541–554 (1978).

Rubin, D. B. "Multiple imputations in sample surveys-A phenomenological Bayesian approach to nonresponse." Imputation and Editing of Faulty or Missing Survey Data. U.S. Department of Commerce, 1–23 (1978).

Schafer, J. Analysis of Incomplete Multivariate Data. London: Chapman & Hall (1997).

Sinha, D. and Dey, D. "Semiparameteric Bayesian analysis of survival data." Journal of the American Statistical Association, 92:1195–1212 (1997).

Walker, S. and Mallick, B. "Hierarchical Generalized Linear Models and Frailty Models with Bayesian Nonparameteric Mixing." Journal of the Royal Statistical Society - Series B, 59:845–860 (1997).

Wei, G. and Tanner, M. "A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms." Journal of the American Statistical Association, 85:699–704 (1990).