University of California
Los Angeles

# Bayesian Event History Analysis with Applications to Recurrent Episodes of Illicit Drug Use

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biostatistics

by

**Adam Jeffrey King**

2014

Abstract of the Dissertation

# Bayesian Event History Analysis with Applications to Recurrent Episodes of Illicit Drug Use

by

**Adam Jeffrey King**

Doctor of Philosophy in Biostatistics

University of California, Los Angeles, 2014

Professor Robert E. Weiss, Chair

Illicit drug use and concomitant problems such as high incarceration rates pose tremendous challenges to those directly involved and to society as a whole. In recent years, many medical, public health, and social science researchers have conducted studies to characterize the nature of these problems and look for potential solutions. These studies often record events such as subjects relapsing into drug use following periods of abstinence, and interest centers on finding or assessing the impact of risk factors for event occurrence. While analyses of these studies have often employed advanced statistical techniques borrowed from the social science literature, there have been few applications of modern Bayesian techniques or advanced event time models. In this dissertation, we develop a general Bayesian event history model with supporting computing tools, and we apply this work in a substance abuse context.

Because the field of survival and event history analysis is so broad, we first provide an extensive review of concepts and literature relevant to our subsequent methodological work. We then describe in detail our general Bayesian event history model. This model combines a number of features which are frequently omitted from analyses because available event time analysis software does not allow their simultaneous use. These features include simultaneous semiparametric incorporation of multiple continuous time-varying covariates, multiple event types or competing risks, and recurrent at-risk episodes. We provide novel

Markov chain Monte Carlo (MCMC) algorithms for obtaining posterior inferences from this model, evaluate their performance, and apply them to data on recurrent episodes of cocaine use in a population of illicit drug users.

Next, we extend our general Bayesian event history model to a full multistate model for histories in which subjects pass in between discrete states. In these histories, the state transitions are the events of interest, and the numbers and types of covariates and possible transition events depend on the subject's current state. We apply this multistate model to lifetime histories of cocaine use and incarceration following first use of cocaine. Finally, we conclude the dissertation with a description of our software implementation of our models and a discussion of future projects.

The dissertation of Adam Jeffrey King is approved.

Ronald S. Brookmeyer

Marc A. Suchard

Yih-Ing Hser

Robert E. Weiss, Committee Chair

University of California, Los Angeles

2014

*To my parents...*

*for nurturing my dreams.*

*And to my wife...*

*for making sure I pursue them.*

# Table of Contents

# List of Figures

# List of Tables

# Acknowledgments

This dissertation would not have been possible without the mentorship, support, and caring of many people. Chief among them is my advisor Rob Weiss. Through our weekly meetings, I not only learned about the practice and profession of statistics, but also learned what it means to be truly committed to helping students. I am also indebted to my committee members Ron Brookmeyer, Marc Suchard, and Yih-Ing Hser for providing their world class expertise in epidemiology, computing, and drug abuse research.

I want to thank Dr. Hser and my colleagues at the UCLA Integrated Substance Abuse Programs for providing financial support, as well as an enjoyable and stimulating work environment. I am also very grateful for receiving a generous fellowship from Amgen over the past two years. This support gave me the time to implement my dissertation work as a software package. I also want to thank UCLA Biostatistics faculty Tom Belin, Catherine Sugar, Ron Brookmeyer, and David Gjertson for giving me the chance to hone my teaching skills under their expert supervision.

Finally, I need to thank all the people who believed in me over the years. I couldn't have made it to the finish line without the steady encouragement from Rob Weiss and my wife Sherry. The hugs and general happiness I received from Sherry were also instrumental in my success. I never would have made it to UCLA in the first place without being spurred on by my undergraduate advisor Ted Shifrin. Finally, none of this would have been possible without the hard work, sacrifice, and love from my parents—this dissertation is their accomplishment as well.

# Vita

| | |
|---|---|
| 2005 | B.S., Mathematics, University of Georgia, Athens, GA |
| 2005–2007 | Teaching Assistant, Department of Mathematics, University of California, Los Angeles, CA |
| 2007 | M.A., Mathematics, University of California, Los Angeles, CA |
| 2007–2012 | Teaching Assistant, Department of Biostatistics, University of California, Los Angeles, CA |
| 2012–2014 | Amgen Graduate Fellowship |
| 2013–2014 | Graduate Student Researcher, Integrated Substance Abuse Programs, University of California, Los Angeles, CA |

# Publications

Konstantinos I. Papageorgiou, Ronald Mancini, Helene Chokron Garneau, Shu-Hong Chang, Imran Jarullazada, Adam King, Erin Forster-Perlini, Catherine Hwang, Raymond Douglas, Robert A. Goldberg (2012). A Three-Dimensional Construct of the Aging Eyebrow: The Illusion of Volume Loss. *Aesthetic Surgery Journal*, 32(1):46–57.

Konstantinos I. Papageorgiou, Catherine J. Hwang, Shu-Hong H. Chang, Imran Jarullazada, Helene Chokron Garneau, Michael J. Ang, Adam J. King, Ronald Mancini, Raymond S. Douglas, Robert A. Goldberg (2012). Thyroid-Associated Periorbitopathy: Eyebrow Fat and Soft Tissue Expansion in Patients With Thyroid-Associated Orbitopathy. *Archives of Ophthalmology*, 130(3):319–328.

Shu-Hong H. Chang, Konstantinos I. Papageorgiou, Michael J. Ang, Adam J. King, Ronald A. Goldberg (2013). High resolution ultrasound as an effective and practical tool to analyze eyebrow profile expansion in thyroid associated periorbitopathy. *Ophthalmic Plastic and Reconstructive Surgery*, 29(5):382–385.

# CHAPTER 1

# Introduction

We begin this chapter with an overview of the dissertation work, and give the rationale for the development of our novel models and algorithms. Next, we introduce the Treatment Utilization and Effectiveness Project (TUE) study, and review previous analyses of the TUE data. We conclude this chapter with a non-technical description of the methods developed in later chapters. Subsequent chapters contain an in-depth technical review of event history analysis, a detailed description of our general Bayesian event history model, an extension of this model to a full multistate model, and an overview of future work.

## 1.1 Overview of and Motivation for the Dissertation Work

Event histories consist of records of time points at which subjects experience certain events of interest. In the classical survival analysis setting, the timing of a single event is recorded for each subject. The Cox (1972) proportional hazards model has been developed extensively over the past four decades to analyze such event time outcomes, and we review many of these developments in the next chapter. In this dissertation, we develop hazard regression models for complex event histories consisting of observation of recurrent episodes of disease activity, behaviors, and circumstances. Here, the events of interest are occurrences which start or terminate episodes. This work is motivated by the Treatment Untilization and Effectiveness Project (TUE), a longitudinal study of illicit drug use in which all lifetime episodes of several types of drug use and related traits have been retrospectively recorded from each subject.

A common feature of event history datasets such as TUE is the discrete recording of the timings of events occurring in continuous time (Burridge, 1981; Allison, 1982). Specifically,

continuous time is partitioned into non-overlapping intervals (often corresponding to whole days, weeks, or months), and only the interval in which an event occurs is recorded. This special case of interval censoring, called *grouped time*, is responsible for the occurrence of tied observations in nominally-continuous survival data. To handle grouped time data containing large numbers of ties, Cox (1972) proposed modeling the *discrete hazard* $\lambda(t)$, which is the probability of event occurrence in the $t^{\text{th}}$ interval given the event did not occur in a previous interval. Specifically, he proposed the discrete time proportional hazards model,

$$\frac{\lambda(t)}{1-\lambda(t)} \equiv \frac{\lambda_0(t)}{1-\lambda_0(t)} \exp(\beta' x). \tag{1.1}$$

where $\lambda_0(t)$ is an arbitrary baseline discrete hazard and $x$ is a vector of regressors with unknown coefficients $\beta$. Because discrete hazards are true probabilities bounded between 0 and 1, this model replaces the continuous hazard rate in the continuous Cox model with the odds of the discrete hazard. Taking logs of both sides of (1.1) and writing $\beta_0(t) \equiv \text{logit}(\lambda_0(t))$, where $\text{logit}(\cdot)$ is the log-odds function, yields a logistic regression model

$$\text{logit}(\lambda(t)) \equiv \beta_0(t) + \beta' x \tag{1.2}$$

with one *person-time observation* for each discrete time point $t$ each subject was at risk for event occurrence. If the time grouping intervals are short relative to the typical duration of follow-up, this approach can yield a very large number of person-time observations, even with only moderate numbers of subjects (Fahrmeir and Lang, 2001; Browne et al., 2009).

Event history datasets also often contain a large number of distinct event types. In a *competing risks* event history, at each time point of follow up, a subject is at risk for multiple, mutually-exclusive events. For example, in a study following active cocaine users, subjects are simultaneously at risk for ceasing cocaine use due to arrest and voluntarily ceasing drug use. In a *multistate* event history, at each time point a subject is in one of several discrete states and is at risk for experiencing *transition events* wherein the subject passes into a different state. Here, each possible transition from one state to another is a distinct event type. Furthermore, each type of event may be observed multiple times in the same subject over the course of follow up. Accounting for the within-subject associations

between hazards of a single type of event at different follow up times and between the hazards of distinct types of events requires inclusion of multi-dimensional random effects in an event history regresson model. However, frequentist maximum likelihood (ML) inference procedures perform poorly when applied to discrete data with high-dimensional random effects, as ML inference requires approximating the marginal likelihood of parameters of interest with high-dimensional numerical integrations (Zeger and Karim, 1991).

A third common feature of event history data is that the hazard of event occurrence depends on several time-varing covariates, and these dependencies are nonlinear. These covariates often include time scales such as age, calendar time period, and time-at-risk (often called *duration* or *gap time*). Multistate event history models may have even more time scales, such as the time since a subject first entered a state or the cumulative amount of time spent in a state. The Cox (1975) partial likelihood estimation technique cannot be used when multiple arbitrary functions, representing the effects of multiple time scales, are simultaneously included (Keiding, 1990; Berzuini and Clayton, 1994). Even when only a single arbitrary baseline hazard function is required, estimation of the discrete Cox model via partial likelihood can be computationally intensive when the number of ties is large (Kalbfleisch and Prentice, 2002, p. 106).

The confluence of these three issues—large numbers of discrete person-time observations, large numbers of distinct event types requiring high-dimensional random effects, and the necessity of modeling multiple nonparametric functions—is commonly encountered in event history problems across many application areas. However, models, inference procedures, and statistical software are not currently available to simultaneously address all features of such event history data; the goal of this dissertation is to fulfill this need. We adopt a Bayesian approach to modeling and inference, with Gaussian Markov random field (GMRF) priors providing a convenient way to model the nonlinear effects of time scales and other continuous covariates. Obtaining posterior inferences via Markov chain Monte Carlo (MCMC) then circumvents the limitations of classical maximum partial and marginal likelihood estimation.

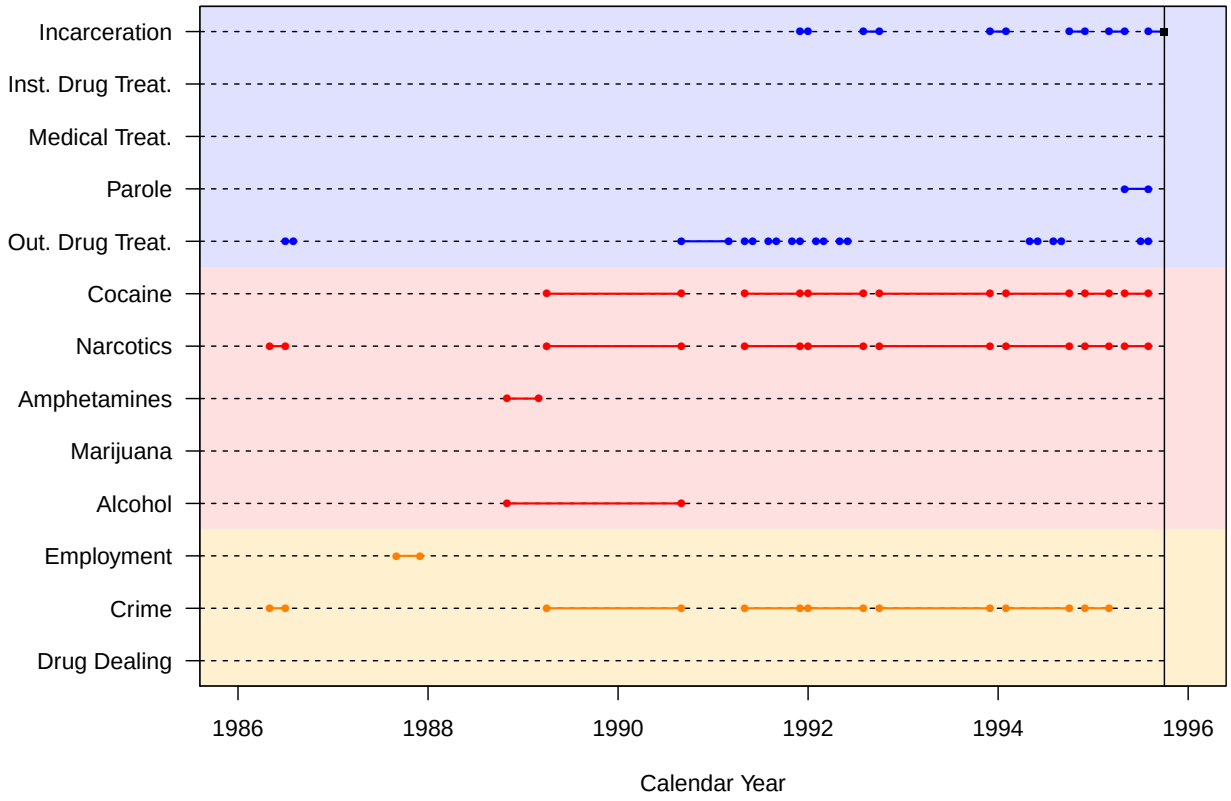
## 1.2 The Treatment Utilization and Effectiveness Project

The Treatment Utilization and Effectiveness Project (TUE) was an observational study of illicit drug users in Los Angeles County, California. It was conducted from 1992 to 1997 by the UCLA Drug Abuse Research Center, now part of the UCLA Integrated Substance Abuse Programs, and was funded by the National Institute on Drug Abuse. The principal investigators were Yih-Ing Hser, Douglas Anglin, and Douglas Longshore.

The TUE study was designed to achieve several objectives. First, many studies of drug users sample subjects from substance abuse treatment programs and facilities. However, these subjects may not be representative of the larger population of drug users. To examine the hidden populations of drug users not engaged in treatment, TUE recruited subjects from three sources thought to have a high prevalence of drug users: sexually transmitted disease (STD) clinics, hospital emergency rooms (ER's), and jails. Baseline interviews and urinalysis tests conducted with 5,168 subjects contacted at these sources confirmed high prevalence of illicit drug use (Hser et al., 1998). In particular, 8.5% of STD clinic, 18.1% of ER, and 52.8% of jail subjects tested positive for cocaine use.

A second objective of the TUE study was to obtain lifetime longitudinal records of drug use and related behaviors. From the 5,168 subjects given the baseline interview, 787 subjects were selected at two years follow-up to complete a *Natural History Interview* (NHI) component, described below, to elicit lifetime drug use histories. To be included in the follow-up sample, subjects 25 years of age or older at initial recruitment must have reported use of an illegal substance other than marijuana in the year prior to the initial recruitment interview, while subjects under 25 must have reported use of any drug including marijuana during that year. In addition, priority for selection was given to subjects reporting dependent use of an illicit substance and never having received treatment for addiction, as well as subjects with elevated risk of dependent drug use (due to criminal activity) or HIV infection (due to intravenous drug use). Among the 787 subjects selected, 14 were deceased, 18 refused to be interviewed, and 189 were available for interview but were never interviewed. The remaining 566 subjects were given the Natural History Interview, which we now describe in detail.

Figure 1.1: Complete NHI history from a single TUE subject. Solid colored lines represent episodes, and dashed black lines represent time spans during which the subject was not in an active episode of the corresponding trait. The vertical black line represents the time of interview, which right censors the final incarceration episode of this subject.




### 1.2.1 The Natural History Interview

The goal of the Natural History Interview is to provide a comprehensive picture of the lifetime drug use, drug treatment, and criminal history of interviewees. For each of a pre-specified collection of thirteen behaviors, traits, and circumstances, the NHI records all *episodes* of that characteristic, which are time spans of contiguous whole calendar months for which that characteristic was present or active for the subject taking the interview. In particular, for each episode, the interviewer records the starting and ending months of that episode, as well as answers to a collection of questions pertaining to that type of episode.

Figure 1.1 depicts the complete Natural History Interview data from a single TUE subject age 26 at the time of interview, which is depicted by the vertical black line at August, 1995. We see for example that this subject had six episodes of incarceration and seven episodes of cocaine use. Following the first episode of cocaine use, the subject began an outpatient drug treatment program, but each of the remaining cocaine use episodes ended with the subject's incarceration. Following each of these incarceration episodes, the subject immediately resumed cocaine use.

Since this retrospective data collection may involve asking subjects to recall episodes from many years earlier, a timeline of major life events is first constructed from available official records and the subject's memory to aid in recall of episodes around the times of those events. These events include arrests, incarcerations, parole, treatment, marriages, divorces, births of children, deaths of family members, employment changes, and geographical moves. Comparison with another measure, the Addiction Severity Index, suggests the NHI is reliable (Murphy et al., 2010), but the underreporting of cocaine use observed by Hser et al. (1999) may indicate limited validity, particularly among the STD and ER subjects.

### 1.2.2 Previous Analyses of the TUE NHI Data

A number of different statistical analysis methodologies have been used to analyze the TUE NHI data, and we briefly review them here. Hser et al. (2008) analyze NHI data from five studies, including TUE, to investigate the drug use and incarceration trajectories of subjects whose self-reported primary drug is heroin, cocaine, or methamphetamine (meth). For each subject, they examine all episodes of use of the subject's primary drug and episodes of incarceration during the first 10 years following first use of the primary drug. Episodes of drug use during this time are subclassified into *low-use* and *high-use* episodes depending on whether the subject used more frequently than 11 days per month on average during the episode. Furthermore, time spans during which the subject was neither using the primary drug nor incarcerated were considered *no-use* episodes. Thus, the entire 10-year time span each subject was examined was partitioned into episodes of four types: no-use, low-use,

high-use, and incarcerated. Separately for each primary drug type (heroin, cocaine, and meth), counts and average lengths of episodes of each type were summarized. Episodes were further subclassified based on the type of episode passed into at the end of the episode in question, and summaries were computed for each subtype. In addition to summarizing transitions between episodes, separate Cox models for each drug type were used to analyze time-to-quitting use of the primary drug, with incarceration and time-of-interview treated as independent censoring. Both fixed covariates (race, sex, and age of drug use onset) and time-varying covariates (number of usage episodes up to that point, treatment status, and legal supervision status) were used.

Prendergast et al. (2008) used NHI data from TUE and two other studies to predict post-first-treatment drug use, incarceration, and employment trajectories using pre-first-treatment drug use and incarceration trajectories. They first applied zero-inflated Poisson growth mixture models to the number of days of drug use and number of months of incarceration in each of the five years prior to first entry into drug abuse treatment. These mixture models identified four latent drug use trajectory classes (*low*, *high*, *decreasing*, and *increasing*) and three latent incarceration trajectories (*low*, *high*, and *increasing*). These categories were collapsed by including the *increasing* subjects in the *high* groups and including the *descreasing* subjects in the *low* groups, and subjects were cross-classified into four pre-treatment drug-incarceration trajectory classes: *low-low*, *low-high*, *high-low*, and *high-high*. This pre-treatment drug-incarceration classification was then used as the primary predictor variable in Poisson growth curve models of the number of days of drug use, number of months incarcerated, and number of months employed in the five years following the first drug treatment episode.

A primary goal of longitudinal studies of drug use and treatment is to assess the causal effect of past drug treatment on current drug use. In such studies, the time-dependent covariate of past drug use may be an intermediate variable for the causal effect of past treatment on current drug use, since past treatment may cause the subject to discontinue use in the past, and this abstinence may be maintained until the present. In addition, past drug use may also be a confounder of the relationship between past treatment and

current use, since past drug use may cause the subject to initiate treatment and this past use may be maintained until the present. Robins et al. (2000) call such a time-dependent covariate, which is both an intermediate and a confounder, a *time-dependent confounder*. They showed that the standard method of modeling current drug use as a function of both past drug treatment and past drug use (e.g. by including both in a multiple regression model) yields biased estimates of the treatment effect. The authors propose an unbiased alternative estimation method, called *marginal structural models* (MSM).

Motivated by this work, Li et al. (2010) applied MSM to TUE and two other studies to assess the effect of cumulative drug treatment on later drug use abstinence. *Total drug use abstinence*, defined as the total number of months subjects were neither using nor incarcerated during the $11^{\text{th}}$ through $15^{\text{th}}$ years following first use of their primary drug, was chosen as the outcome variable. Cumulative lifetime number of months of drug use and cumulative lifetime number of months of drug treatment, tallied at the end of each of the first 10 years following first use of the drug, were used as the time-dependent confounder and time-dependent treatment, respectively. The authors fit both MSM and standard regression models using these variables, and observed that the MSM showed that treatment resulted in a significant increase in drug use abstinence, while the standard regression showed no treatment effect.

Liang et al. (2010) jointly model time-to-death and percentage of follow-up time spent using alcohol and the subject's primary drug. The authors use a normal linear mixed model for the percentage of follow-up time (time from use initiation until death or censoring, excluding time spent incarcerated) that the subjects used alcohol, and a second linear mixed model for the percentage of follow-up spent using the subject's primary drug. These linear models for average usage level are linked to a Weibull proportional hazards model with frailty for the survival outcome of duration from drug use initiation to death, by including the usage variables as predictors in the survival model. They combine data from five studies, including TUE, using separate random effects to capture study-level heterogeneity in mortality, alcohol use, and drug use. Inferences are obtained from MCMC simlation using WinBUGS.

## 1.3 Non-technical Overview of the Dissertation Work

### 1.3.1 Event History and Multistate Models

As we saw in the previous section, complex datasets such as TUE may be analyzed with a number of different approaches. Consider an *event history* consisting of records of the time points at which subjects experience *transition events*, wherein they move from one state to another. A common practice is to first compute *aggregate* or *summary* measures of the event history, such as the percentage of time a subject spends in a certain state, and then treat these summaries as outcome variables. In contrast, an *event history analysis* directly models the probabilities of event occurrence over every time point of follow up. If we simultaneously model the probabilities of all possible transitions between states, the event history model is called a *multistate model*.

To illustrate, suppose for each subject we have recorded all episodes of cocaine use over a 10-year period, where an *episode* is a span of adjacent calendar months during which the subject was using the drug. A conventional longitudinal analysis might begin with computing the percentage of time spent using cocaine for each year of each subject's follow up period. We then have ten repeated measurements of this summary variable from each subject, which we can use as a continuous longitudinal outcome variable. An event history analysis of this dataset would instead model the risks of the subjects switching between the cocaine-using and non-using states. More specifically, for each month of follow up during which a subject was using cocaine, we would model the probability of them ceasing drug use the following month, and for each month during which a subject was not using cocaine, we would model the probability of them beginning or resuming drug use the next month.

To see why the event history analysis may be preferable to the conventional longitudinal analysis, suppose subjects are incarcerated for a substantial fraction of the time they are not using cocaine. In this case, a low value for the longitudinal outcome variable, percent of the year spent using cocaine, may be due to the subject spending most of the year in prison. Clearly, this is a much less desirable outcome than the subject spending the year in question in a state of voluntary abstinence from drug use, and whatever statistical model we

use should account for this difference.

The traditional longitudinal analysis could be adjusted by redefining the outcome as the percentage of time spent using cocaine among the part of each year during which the subject was not incarcerated. However, the outcome variable would then be undefined for years in which the subject was in prison the entire time, and treating such years as missing values would ignore the relationship between cocaine use and risk of arrest. A multivariate longitudinal model in which both the percentage of the year spent incarcerated and the percentage of non-incarcerated time spent using cocaine may be an improvement, but problems would still remain. For example, such a model would not be able to distinguish between a year in which the first half was spent using cocaine and the second half was spent in prison, and a year with the opposite pattern, with prison time followed by cocaine use. Observation of the former suggests cocaine use increased the subject's risk of arrest, while observation of the later suggests incarceration was not successful in rehabilitating the subject.

In contrast, the event history analysis can be naturally extended to accomodate the additional state of incarceration. We do this by creating a multistate model for the probabilities of all possible transitions between the three mutually-exclusive states *cocaine use*, *voluntary non-use*, and *incarceration*. Then if for example we observe a cocaine-using subject become incarcerated, our model will infer that the subject's cocaine use or other traits, such as past incarceration history, may have increased the subject's arrest risk.

The challenge, however, of building such an event history model is to specify all important risk factors for the occurrence of each type of transition event. For example, the risk of a subject voluntarily ceasing cocaine use in any given month may depend on the current age of the subject, how long the subject has been continuously using cocaine up to that point, and how many months the subject has been in drug abuse treatment during the previous year. Further complicating matters, predictors like age may not have a simple linear or quadratic relationship with the hazard of event occurrence, as is often assumed by standard regression models. We require a more flexible means of allowing the probabilities of transition events to depend on our predictor variables. Statisticians often refer to such flexible formulations as *nonparametric* components of a model.

Classical event time models and inference procedures encounter difficulties when incorporating multiple nonparametric components. Bayesian models on the other hand allow us to include arbitrary numbers of smooth, flexible functions to model the relationships between predictors and the risks of event occurrence. The algorithms for obtaining inferences from these models also perform better than algorithms for corresponding classical event history models. We now outline the Bayesian paradigm of statistical modeling and inference.

### 1.3.2 Bayesian Inference and Markov Chain Monte Carlo Computation

In *classical* or *frequentist statistics*, methods of estimation and hypothesis testing are devised so that if the experiment or study which furnished our data were replicated many times, the estimation and testing procedures would give us the correct answer a high proportion of the time. For example, suppose we follow a sample of male and female active cocaine users until they cease use of the drug, and then conduct a hypothesis test at the 5% significance level to determine whether the per-month rate of quiting differs between men and women. Then if there really is no difference in cocaine cessation rates between the sexes, we would correctly fail to find a significant difference 95% of the time. In addition, if we calculate a 95% confidence interval (CI) for the quitting rate in men, then we know that 95% of the time we take a sample and calculate such a CI, the interval will contain the true male cessation rate. This is *not*, however, the same thing as saying there is a 95% *probability* that the CI contains the true rate. To make such direct probability statements, we need a different paradigm of statistics.

In *Bayesian statistics*, uncertainty about unknown values, such as the cocaine cessation rates in men and women, is quantified using probability distributions. First, we specify a *prior probability distribution*, or *prior* for short, capturing our knowledge about the unknown parameters of our statistical model prior to examining the sample data. For instance, if we are 95% certain based upon our previous experience studying male cocaine users that their per-month quitting rate is between 4% and 16%, then we might express this prior belief with a probability distribution that has 95% of its mass, or concentration, between 0.04 and 0.16.

Once we have specified a prior distribution and a statistical model for how the unknown parameters relate to the observed data, we use a mathematical equation called *Bayes' theorem* to synthesize the prior and observed data into a *posterior probability distribution*. The posterior distribution captures both our prior knowledge about the unknown parameters and the information gained from the sample data. If after applying Bayes' theorem to our cocaine cessation example problem, we find the posterior distribution of the male quitting rate has 95% of its mass between .06 and .08, then we can say there is a 95% probability that the true rate at which men are ceasing cocaine use is between 6% and 8% per month.

Any inferences we draw from fitting a Bayesian model are completely determined by the posterior distribution, and, in turn, the posterior distribution is completely mathematically determined by Bayes' theorem. Unfortunately, this does not mean that drawing inferences from a Bayesian model is always straightforward. Indeed, if our statistical model has many unknown parameters, then the posterior distribution has a high number of dimensions. In such cases, calculating the posterior distribution directly using Bayes' theorem may be very difficult. However, because the posterior is a probability distribution, we may be able to learn about this distribution by drawing a sample from it and computing summary statistics of the sample values. This process is called inference via *Monte Carlo simulation*.

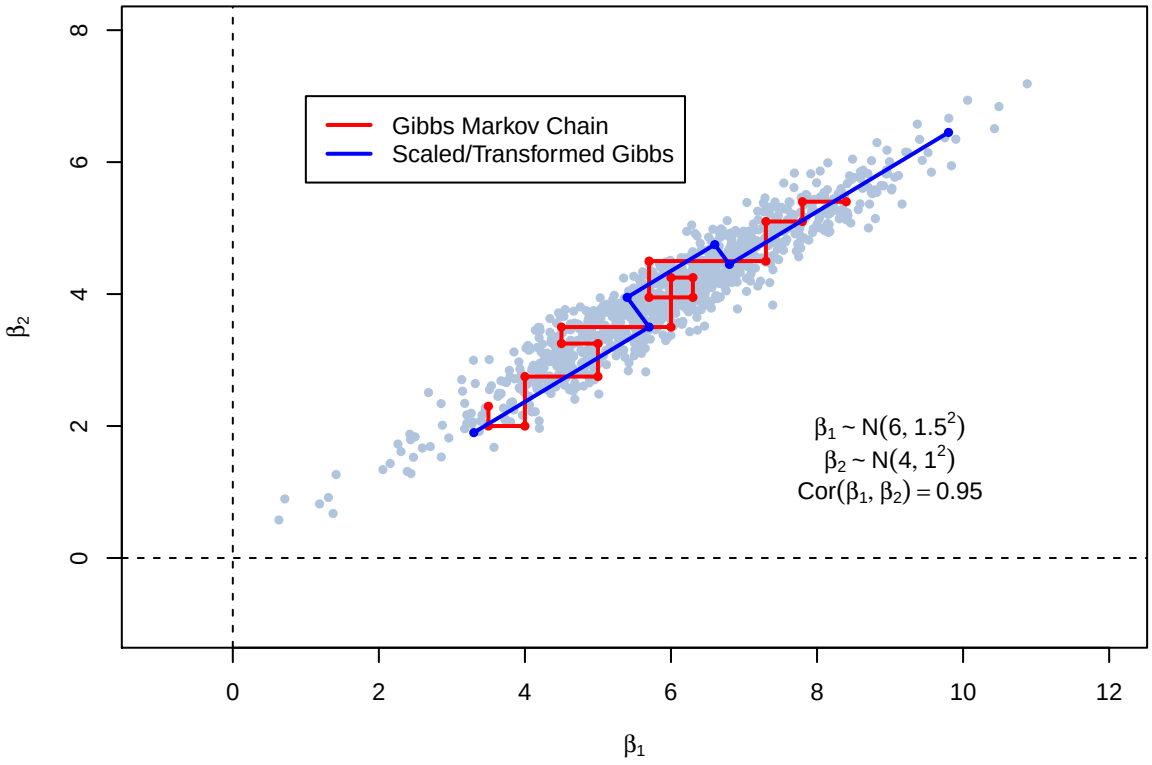
The most common Monte Carlo simulation technique used for Bayesian inference is called *Markov chain Monte Carlo* (MCMC). This technique randomly draws a sequence of sample points from the posterior distribution, called a *Markov chain*. The distribution of each succesive randomly drawn element of the chain is determined by the location of the previous element of the chain, the posterior distribution we are trying to sample from, and the particular MCMC algorithm chosen. Because each newly drawn point depends on the location of the previous point in the chain, the elements of the sample produced by the Markov chain are not independent. Instead, they are correlated, and the stronger this correlation is the less information the Markov chain provides about the posterior distribution. Thus it is important to design MCMC algorithms to minimize this correlation.

Figure 1.2 shows an example of a 2-dimensional posterior distribution, depicted by the ellipse-shaped cloud of gray-blue points. In red, we show a sequence of 19 points produced by

an MCMC algorithm known as the *Gibbs sampler*; this algorithm only allows one dimension (horizontal or vertical) to change at each successive move of the chain. As a result, it takes the chain many iterations to move from one side of the posterior distribution to the other. In blue, we show a modified Gibbs sampling MCMC algorithm which allows the chain to move along the major axis (that is, the longest direction) of the elliptic posterior distribution. By taking advantage of information about the shape of the posterior distribution, this chain is able to move further on average at each iteration than the unmodified Gibbs chain, and therefore produces a posterior sample with lower correlation.

In this dissertation, we develop novel MCMC algorithms for event history models with high-dimensional posterior distributions. Like in the simple 2-dimensional example above, our new algorithms will produce posterior samples with lower correlation by using information about the approximate shape of the posterior distribution of our event history models. We describe these algorithms in Chapter 3.

Figure 1.2: Two Gibbs sampler MCMC algorithms applied to a 2-dimensional multivariate normal posterior distribution.



# CHAPTER 2

# Event History Analysis Review

In this chapter, we provide an extensive review of event history analysis concepts and literature. We begin by reviewing basic survival analysis concepts. Subsequently, we cover Cox-type hazard regression models, alternative discrete and continuous survival models, multivariate survival analysis, and competing risks. Next, we briefly review the social science event history analysis literature. Finally, we discuss the issue of multiple time scales, and describe smoothing of time scale effects using Gaussian Markov random field priors and MCMC algorithms for obtaining inferences from models with such components.

## 2.1 Basic Survival Analysis Concepts and Notation

We begin with a brief review of basic survival analysis terminology and notation. More detailed introductions to survival methods may be found in Kalbfleisch and Prentice (2002) and Klein and Moeschberger (2003).

### 2.1.1 Event History Analysis and Survival Analysis

*Event history analysis* is the study of the timing of the occurrence of events (Tuma and Hannan, 1979; Aalen et al., 2008). More specifically, we are often interested in the duration of time a subject persists in a state in which they may possibly experience some predetermined type event. If, for example, the event of interest is *leaving a job*, then an event history analysis would examine the lengths of employment in a single job. Of course, the chances of a person leaving their current job depend on more than just how long the person has worked in that position, and an event history analysis may also consider how rates of leaving a job

depend on additional factors. These could include economic factors, such as current income or the local unemployment rate, or time variables other than how long the person has been employed, such as the person's age.

Methods for studying event times have been developed for and applied to a diverse collection of disciplines. For example, in medicine they are often used to study risk factors for death from a disease. Because the case of death being the event of interest is so common is biomedical applications, event history analysis is usually referred to as *survival analysis* by researchers in this field, even when the event is not death. Engineers, on the other hand, call the study of event times *reliability theory*, since they are often interested in the time until a mechanical part fails. This also explains how it became common to refer to event times as *failure times* even in cases where we would not usually refer to the event type as a "failure." Subsequently, we will use the terms *event history analysis* and *survival analysis* interchangeably. Moreover, we will refer to any kind of event as a *failure* and speak of the *risk* of a failure happening even when the event is a desirable outcome, such as voluntarily ceasing illegal drug use.

### 2.1.2 Continuous Time Survival Variables

In the following, we let $Y$ denote a continuous time survival variable. That is, $Y$ is a random variable denoting the elapsed time from some predefined *baseline* or *origin* time point, at which a subject is first at risk of experiencing an event, until that event occurs. Because this time is measured on a continuous scale, $Y$ may take on any positive real number value.

Let $f(t)$ and $F(t)$ denote, respectively, the probability density function (pdf) and cumulative distribution function (cdf) of $Y$; the *survival function* is defined

$$S(t) \equiv \Pr[Y > t] = \int_t^\infty f(t)dt = 1 - F(t). \tag{2.1}$$

The *hazard rate function*, alternatively known as the *failure rate* or *force of mortality*, is then defined

$$\lambda(t) \equiv \lim_{h \to 0^+} \frac{\Pr[t \leq Y \leq t + h | Y \geq t]}{h} = \frac{f(t)}{S(t)}. \tag{2.2}$$

This quantity measures the "instantaneous risk" a subject experiences in the next instant after having survived to time $t$, but it is *not* a probability, since it may take on any value between 0 and $\infty$. However, the probability of a failure occurring in the next $\Delta t$ units of time given that the subject has not already failed may be approximated by $\lambda(t) \cdot \Delta t$ when $\Delta t$ is sufficiently small.

Next, we define the *cumulative hazard* by

$$\Lambda(t) \equiv \int_0^t \lambda(u)du = \int_0^t \frac{f(u)}{S(u)}du = \int_0^t \frac{d}{du}[-\log S(u)]du = -\log S(t). \tag{2.3}$$

Because survival decreases with increasing cumulative hazard according to the equation

$$S(t) = \exp(-\Lambda(t)) = \exp\left(-\int_0^t \lambda(u)du\right), \tag{2.4}$$

the hazard rate $\lambda(t)$ completely determines the distribution of $Y$; thus we may specify a statistical model for $Y$ by giving a prescription for $\lambda(t)$. This approach will be discussed in detail in subsequent sections.

### 2.1.3 Discrete Time Survival Variables

Let the random variable $\tilde{Y}$ represent the time elapsed until some failure event occurs. If the collection of time points at which a failure may occur is discrete, then we say $\tilde{Y}$ is a discrete time survival variable. In this case, we may without loss of generality assume the collection of potential failure times is the positive integers $\{1, 2, 3, \ldots\}$.

Let $\tilde{f}(s)$ and $\tilde{F}(s)$ denote, respectively, the probability mass function (pmf) and cdf of $\tilde{Y}$. As in the continuous time case, we define the survival function by

$$\tilde{S}(s) \equiv \Pr[\tilde{Y} > s] = \sum_{k=s+1}^{\infty} \tilde{f}(k) = 1 - \tilde{F}(s). \tag{2.5}$$

The discrete hazard rate function is defined as the probability of experiencing the event at time point $s$ given that the subject has survived up to that point,

$$\tilde{\lambda}(s) \equiv \Pr[\tilde{Y} = s | \tilde{Y} \geq s] = \frac{\tilde{f}(s)}{\tilde{S}(s-1)}. \tag{2.6}$$

Unlike the continuous time hazard rate, the discrete hazard is a true probability, bounded between 0 and 1. But similar to the continuous case, the chances of surviving through the hazard at each successive time point decreases according to the equation

$$\tilde{S}(s) = \prod_{k=1}^{s} \Pr[\tilde{Y} > k | \tilde{Y} \geq k] = \prod_{k=1}^{s} (1 - \tilde{\lambda}(k)), \tag{2.7}$$

so that the hazard function $\tilde{\lambda}(s)$ completely determines the probability distribution of $\tilde{Y}$.

### 2.1.4 Censoring and Grouped Time

Frequently in studies involving time-to-event variables, we may not be able to observe the precise timing of the event of interest for all subjects. Instead, for some subjects we may only know that the event occurred in some specific range of times. For example, in a study where time to relapse of drug use in former users is measured, the study may end before some subjects experience a relapse. For these people, we know that their survival time is greater than the time elapsed between when they ceased drug use (at which point they became at risk for relapse) and the end of the study. When all we are able to observe about a subject's event time is that if the event occurred at all, it must have occurred after some time point $t$, that observation is said to be *right censored* at time $t$.

Alternatively, we may know that the event occurred between two time points $t_1$ and $t_2$ without knowing precisely where in this time interval the event occurred, a situation known as *interval censoring*. This could be the case, for instance, if we are measuring time to change in some clinical sign whose observation requires a laboratory procedure. In such a study all observations would be censored to the time intervals in between clinic visits.

Omitting censored observations from an analysis can lead to bias and loss of precision in estimates. In the drug relapse example, excluding the subjects who remained drug-free throughout the study period would cause a downward bias in survival time estimates. Moreover, in the example above where all observations were censored to intervals between lab measurments, removing the censored observations would eliminate the entire sample.

If every observation in a study with continuous time is censored to an interval from a

collection of possible intervals that is common to all observations, then we say that study has *grouped time*. More formally, we have grouped time data if there is a (possibly infinite) sequence of time points $\{0 = t_0 < t_1 < t_2 < \cdots\}$ such that all observations from all subjects are interval censored to intervals of the form $(t_i, t_j]$ or $(t_i, \infty)$. In such cases, we may either choose to model the original continuous survival time variable $Y$ using methods that allow interval-censored observations, or instead derive a discrete survival time variable

$$\tilde{Y} \equiv \min\{s | Y \leq t_s\}. \tag{2.8}$$

With this definition, $\tilde{Y} = 1$ when $Y \in (t_0, t_1]$, $\tilde{Y} = 2$ when $Y \in (t_1, t_2]$, and so on. Furthermore, the following relationships hold for all $s$,

$$Y \leq t_s \iff \tilde{Y} \leq s, \tag{2.9}$$

$$Y > t_s \iff \tilde{Y} > s, \tag{2.10}$$

$$F(t_s) = \tilde{F}(s), \tag{2.11}$$

$$S(t_s) = \tilde{S}(s). \tag{2.12}$$

This new discrete time variable $\tilde{Y}$ (which may itself be subject to interval and right censoring) is easy to compute from grouped time continuous data; in fact, grouped time data is often already encoded in such a discrete fashion. Moreover, $\tilde{Y}$ retains all the information in the original grouped-time dataset, since we may recover the original interval membership information about $Y$ from $\tilde{Y}$.

## 2.2 Cox Regression Models for Univariate Outcomes

The most commonly employed method of relating survival outcomes to covariates is regression modeling of the hazard rate function. As equations (2.4) and (2.7) imply, in the continuous and discrete cases respectively, a specification of the hazard rate is sufficient to characterize the distribution of a survival time variable. Moreover, regression models for hazard rates can easily be fit in the presence of right censoring that is independent of the survival outcome, as censored observations simply do not contribute information about the

hazard rate beyond their censoring times. In this section, we review both continuous and discrete versions of this approach.

### 2.2.1 Continuous Time Proportional Hazards Model

In his seminal 1972 paper "Regression Models and Life-Table," Sir David Cox introduced his now-famous proportional hazards model for continuous time outcomes. For an individual with a $p$-vector of covariate values $X(t)$ at time $t$, the *Cox proportional hazards* (CPH) model sets the hazard rate $\lambda(t; X(t))$ at time $t$ equal to the product of an unknown arbitrary nonnegative function of time $\lambda_0(t)$ and the exponential of a linear predictor $\beta' X(t)$,

$$\lambda(t; X(t)) \equiv \lambda_0(t) \exp(\beta' X(t)), \tag{2.13}$$

where $\beta$ is a vector of unknown regression coefficients. The *proportional hazards* name is due to the fact that the ratio of hazard rates for two individuals with respective time-invariant covariate vectors $X_1$ and $X_2$ does not depend on time, as

$$\frac{\lambda(t; X_1)}{\lambda(t; X_2)} = \frac{\lambda_0(t) \exp(\beta' X_1)}{\lambda_0(t) \exp(\beta' X_2)} = \exp(\beta'(X_1 - X_2)). \tag{2.14}$$

In particular, the hazard rate ratio for two individuals who differ only with respect to the $j^{\text{th}}$ covariate by the fixed amount $\Delta X_{\cdot j} \equiv X_{1j} - X_{2j}$ is $\exp(\beta_j \Delta X_{\cdot j})$. Conversely, non-proportionality of covariate effects may be introduced by including interactions between a covariate and some function of time as an additional time-dependent variable in $X$.

Because no functional form for the *baseline hazard rate* $\lambda_0(t)$ is assumed, the CPH model (2.13) is semiparametric. Often in applications, the regression coefficients $\beta$ are of primary interest, and $\lambda_0(t)$ is considered a nuisance parameter. Thus instead of maximizing a full likelihood, Cox (1972) bases inferences about $\beta$ on a "conditional likelihood" which does not contain $\lambda_0(t)$. Specifically, Cox conditions on the elapsed time between events in the dataset and the amount of censoring that takes place between observed events. The resulting likelihood of an observed event on subject $i$ is the (conditional) probability that a failure at time $t = Y_i$ belongs to subject $i$ as opposed to any other subject in the collection $R(t)$ of all

subjects still at risk at time $t$,

$$PL_i(\beta) \equiv \frac{\exp(\beta' X_i(t))}{\sum_{j \in R(t)} \exp(\beta' X_j(t))}. \tag{2.15}$$

However, Kalbfleisch and Prentice (1973) note that the product over all observed events $i$ of (2.15) is not strictly a marginal likelihood unless we make additional assumptions. Cox (1975) resolves this issue by defining a generalization of marginal and conditional likelihood called *partial likelihood*, for which the usual asymptotic results of ML estimation hold, and then noting that (2.15) is an instance of a partial likelihood.

### 2.2.2 Discrete Time Proportional Odds Model

Two distinct observations $Y_1 = y_1$ and $Y_2 = y_2$ are *tied* if $y_1 = y_2$. In theory, with continuous time tied observations occur with probability zero. However, in practice ties often occur as a result of rounding and implicit grouping of the continuous response times. To cover cases in which the number of ties is too large to address with simple modifications of the partial likelihood, Cox (1972) proposed a discrete time version of his continuous time proportional hazards model. The Cox *discrete proportional odds* (DPO) model

$$\frac{\tilde{\lambda}(s; \tilde{X}(s))}{1 - \tilde{\lambda}(s; \tilde{X}(s))} \equiv \frac{\tilde{\lambda}_0(s)}{1 - \tilde{\lambda}_0(s)} \exp(\tilde{\beta}' \tilde{X}(s)) \tag{2.16}$$

closely resembles its continuous analogue (2.13). To account for the fact that the discrete hazards $\tilde{\lambda}(s; \tilde{X}(s))$ at each time point $s$ are probabilities, and thus bounded between 0 and 1, proportionality is now in the odds of the hazard. Writing $\tilde{\beta}_0(s) \equiv \text{logit}(\tilde{\lambda}_0(s))$, we may bring the baseline hazard inside the linear predictor, yielding the reformulation

$$\text{logit}(\tilde{\lambda}(s; \tilde{X}(s))) \equiv \tilde{\beta}_0(s) + \tilde{\beta}' \tilde{X}(s). \tag{2.17}$$

Hence, the DPO model (2.16) relates the discrete hazard rate to covariates in precisely the same manner as a logistic regression model. This fact leads to a convenient method of fitting the DPO model using standard logistic regression software.

Suppose for subjects $i = 1, \ldots, N$ we observe discrete time survival data $(y_i, \delta_i, X_i)$, where $\delta_i$ is a censoring indicator encoded so that $\delta_i = 1$ means $\tilde{Y}_i = y_i$ (an observed event) and

$\delta_i = 0$ means $\tilde{Y}_i > y_i$ (a right censored observation). The covariate vector $X_i$ may be time varying, taking the value $X_i(s)$ at discrete time $s$. Assuming independent censoring and writing $\tilde{\lambda}_i(s) \equiv \tilde{\lambda}(s; X_i(s))$ for convenience, the likelihood is

$$L = \prod_{i=1}^{N} \left[ \tilde{\lambda}_i(y_i)^{\delta_i} (1 - \tilde{\lambda}_i(y_i))^{1-\delta_i} \prod_{s=1}^{y_i - 1} (1 - \tilde{\lambda}_i(s)) \right], \tag{2.18}$$

as the likelihood contribution of each observation $i$ is the probability of surviving through the first $y_i - 1$ time points times the probability of either failing or surviving at the last time point, depending on the censoring indicator $\delta_i$. All terms in the product (2.18) are probabilities related to covariates through a logit link function. This suggests that we can create a logistic regression dataset that has the same likelihood. For each subject $i$, we expand the data $(y_i, \delta_i, X_i)$ to $y_i$ logistic regression observations, one for each discrete time point subject $i$ was followed. These *person-time observations* have the form $(z, s, X)$, where $z$ is the binary response variable, $s$ is discrete time at risk, and $X$ is a covariate vector. The response variable $z$ is set equal to 0 for all observations corresponding to discrete time points at which the subject was observed to not fail. Conversely, we set $z = 1$ for time points that were observed failures; at most one such point is contributed by each survival time.

More formally, given the data $(y_i, \delta_i, X_i)$ from each subject $i$, we first create the $(y_i - 1)$ person-time observations

$$(0, 1, X_i(1)),\ (0, 2, X_i(2)),\ (0, 3, X_i(3)), \ldots,\ (0, y_i - 1, X_i(y_i - 1)). \tag{2.19}$$

Second, we create one additional observation for the last time point the subject was observed,

$$(\delta_i, y_i, X_i(y_i)). \tag{2.20}$$

We then have $\sum_{i=1}^{N} y_i$ logistic regression observations of the form $(z, s, X)$. Before fitting the model, the variable $s$ denoting discrete time needs to be recoded into $(\max y_i - 1)$ indicator variables to allow fitting separate intercepts for each discrete time point, which corresponds to having an arbitrary baseline hazard $\tilde{\beta}_0(s)$. Alternatively, other recodings of $s$ corresponding to other models for the baseline hazard are possible.

## 2.3 Alternate Discrete Models and Relationships Between Models

In this section, we describe several additional models for discrete time survival outcomes, and we examine relationships that exist between these models. We first compare the continuous and discrete Cox models (2.13) and (2.16), which leads to a discussion of a modification of the DPO model (2.16) called the *grouped proportional hazards* (GPH) model. Next we compare several ordinal logistic models to models already discussed, as well as to models for latent continuous failure times, such as the accelerated failure time model.

### 2.3.1 Grouping Continuous Survival Times

Suppose $Y$ is a continuous time survival variable with associated time-varying covariate vector $X(t)$ and hazard rate $\lambda(t; X(t))$. Assume $Y$ follows the CPH model (2.13), and let $\tilde{Y}$ be the corresponding grouped time variable, where the grouping intervals have endpoints chosen among the set of cut points $\{0 = t_0 < t_1 < t_2 < \cdots\}$. In other words, $\tilde{Y} = 1$ when $Y \in (t_0, t_1]$, $\tilde{Y} = 2$ when $Y \in (t_1, t_2]$, and so on. We assume the time-varying covariates $X(t)$ are constantly equal to $\tilde{X}(s)$ on the $s^{\text{th}}$ interval; that is, $X(t) = \tilde{X}(s)$ for all $t \in (t_{s-1}, t_s]$. Cox originally proposed the DPO model (2.16) for precisely this circumstance, when rounding or grouping of a continuous response leads to a large number of tied observations.

#### Comparison of the Continuous and Discrete Cox Models

Several authors have compared the coefficients $\beta$ from the continuous model (2.13) to the estimates of $\tilde{\beta}$ obtained from the discrete model (2.16). Thompson (1977) applies the DPO model to the leukemia remission data that Cox (1972) uses to illustrate the CPH model, where time is grouped into one-week intervals. He obtains similar estimates to the continuous time analysis, and then sketches a proof illustrating equivalence of Cox's continuous and discrete time models as the lengths of the grouping intervals approach zero. Abbott (1985) argues that the two models should give similar results when the probability of failure in each interval is small by approximating these probabilities with power series.

Indeed, we should not expect much descrepancy between odds ratio estimates from the DPO model (2.16) and hazard ratio estimates from the CPH model (2.13) when event probabilities in each interval are small and grouping intervals are short. This is because the odds ratio estimate should be approximately equal to the relative risk of event occurrence in that interval when the probability is small. In turn, the relative risk should be similar to the hazard ratio when the interval is short (in which case the continuous hazard rates are roughly constant) and probabilities are small. More succinctly, when intervals are short, $OR \approx RR \approx HR$, so that $\beta \approx \tilde{\beta}$.

Nevertheless, if the grouping intervals are not short and the discrete hazards are moderate or high, effect estimates can differ substantially. Kalbfleisch and Prentice (1973) provide a simple example of a CPH model with a single dichotomous covariate, where the baseline hazard is constantly equal to 0.2 and the coefficient $\beta = 0.5$. If these survival times are grouped into whole-integer intervals, the corresponding DPO coefficient for the comparison between the two groups is $\tilde{\beta} = 0.57$.

### Grouped Proportional Hazards Model

Kalbfleisch and Prentice (1973) and Prentice and Gloeckler (1978) propose addressing this discrepancy by replacing the logit link function in the DPO model with the complementary log-log link function $\text{cll}(p) \equiv \log(-\log(1-p))$. This yields the *grouped proportional hazards* (GPH) model

$$\text{cll}(\tilde{\lambda}(s; \tilde{X}(s))) = h(s) + \beta' \tilde{X}(s), \tag{2.21}$$

where $h(s)$ is the baseline discrete hazard on the complementary log-log scale. We use the same parameter $\beta$ in (2.21) as in the CPH model (2.13), since it turns out that if the continuous variable $Y$ follows a CPH model with parameter vector $\beta$, then the corresponding grouped outcome $\tilde{Y}$ follows the GPH model (2.21) *with the exact same parameter vector* $\beta$. Moreover, the discrete baseline hazard on the linear predictor scale may be expressed in terms of the continuous baseline hazard,

$$h(s) \equiv \log\left(\int_{t_{s-1}}^{t_s} \lambda_0(u)du\right). \tag{2.22}$$

To prove this, first note that it follows from (2.4) that

$$1 - \tilde{\lambda}(s; \tilde{X}(s)) = \Pr[Y > t_s | Y > t_{s-1}] = \exp\left( - \int_{t_{s-1}}^{t_s} \lambda(u; X(u)) du \right). \tag{2.23}$$

Taking negative logs of both sides, we get

$$\begin{aligned} -\log(1 - \tilde{\lambda}(s; \tilde{X}(s))) &= \int_{t_{s-1}}^{t_s} \lambda(u; X(u)) du \\ &= \int_{t_{s-1}}^{t_s} \lambda_0(u) \exp(\beta' X(u)) du \\ &= \int_{t_{s-1}}^{t_s} \lambda_0(u) \exp(\beta' \tilde{X}(s)) du \\ &= \int_{t_{s-1}}^{t_s} \lambda_0(u) du \cdot \exp(\beta' \tilde{X}(s)). \end{aligned} \tag{2.24}$$

Taking logs once more, we obtain

$$\log(-\log(1 - \tilde{\lambda}(s; \tilde{X}(s)))) = \log\left( \int_{t_{s-1}}^{t_s} \lambda_0(u) du \right) + \beta' \tilde{X}(s) = h(s) + \beta' \tilde{X}(s), \tag{2.25}$$

which is precisely the GPH model (2.21), as we wished to show.

Since the CPH and DPO models give similar results when the discrete hazards are small, the equivalence between CPH and GPH illustrated above implies that estimates from the GPH and DPO models will also not differ greatly. Indeed, Efron (1988) fits two-sample discrete time survival data with a parametric baseline hazard using both a logit link function as in the DPO model (2.16) and a complementary log-log link as in the GPH model (2.21). Estimates of the discrete hazard rate at each time point for each sample agreed between the two models to within 0.3%.

### 2.3.2 Ordinal Logistic Regression

Suppose once again that $\tilde{Y}$ is our discrete time survival outcome. Ignoring censoring for the moment, assume that the integer $J$ is chosen so large that no event time in our dataset exceeds $J$, i.e. $\tilde{Y} \leq J$ for all observations of $\tilde{Y}$. Then we may think of $\tilde{Y}$, which assumes values among the set of integers $\{1, 2, \ldots, J\}$, as an ordinal outcome, and model it using ordinal logistic regression techniques. Several different ordinal logistic regression models

exist, and they may be characterized by the probabilities they model and how they relate those probabilities to linear predictors.

### Continuation Ratio Logit Model

The first class of models we consider specifies the distribution of $\tilde{Y}$ by modeling the conditional probabilities $\Pr[\tilde{Y} = s|\tilde{Y} \geq s]$. Of course, we recognize that these probabilities are precisely the discrete hazard rates $\tilde{\lambda}(s)$, which have appeared in the DPO model (2.16) and GPH model (2.21). When the logit link function is used as in (2.16), Agresti (2002) refers to this ordinal model as the *continuation ratio logit* (CRL) model. The more general class of models where the link function is unspecified is referred to as *sequential models* in Tutz (1991). Since these models are equivalent to the discrete hazard approach previously discussed, we do not consider them further here.

### Ordinal Proportional Odds and Proportional Hazards Models

The *cumulative link* models, introduced in McCullagh (1980), relate the *cumulative probabilities* $\Pr[\tilde{Y} \leq s]$ to covariates through a link function. In particular, McCullagh introduced the *ordinal proportional odds* (OPO) and *ordinal proportional hazards* (OPH) models, which relate the cumulative probabilities to linear predictors with the logit and complementary log-log links respectively,

$$\text{logit}(\Pr[\tilde{Y} \leq s]) \equiv \theta_s + \bar{\beta}'X \tag{2.26}$$

$$\text{cll}(\Pr[\tilde{Y} \leq s]) \equiv c_s + \beta'X. \tag{2.27}$$

The models use different intercept terms for each cumulative probability $\Pr[\tilde{Y} \leq s]$, and since these probabilities are increasing in $s$, the models require that the intercept terms are increasing in $s$ as well. The same covariate coefficient vector is used for each value of $s$, and this feature gives the models the properties they are named after. For the ordinal proportional odds model (2.26), the odds ratio of a given cumulative probability for a $\Delta$-unit increase in $X_i$ will be $\exp(\Delta\bar{\beta}_i)$ regardless of what $s$ is.

On the other hand, for the ordinal proportional hazards model (2.27), it is not obvious what "hazard" the proportionality might be in. It turns out that this model is equivalent to the grouped proportional hazards model (2.21) for discrete survival outcomes, with the same covariate coefficient vector $\beta$. Since as we saw in the last section the GPH model (2.21) is implied by a CPH model for the underlying continuous response $Y$, the OPH model (2.27) is also implied when a continuous time proportional hazards model underlies the ordinal outcome.

**Equivalence of the OPH and GPH Models**

The equivalence of the two discrete proportional hazards models was first explicitly noted in Laara and Matthews (1985). In this section, we prove this equivalence and provide expressions for the parameters of each model in terms of those of the other.

To see that the two models are equivalent, first assume the GPH model (2.21) holds,

$$\text{cll}(\tilde{\lambda}(s)) \equiv \log(-\log(1-\tilde{\lambda}(s))) \equiv h_s + \beta' X. \tag{2.28}$$

Taking exponentials gives an equation we'll need momentarily,

$$-\log(1-\tilde{\lambda}(s)) = \exp(h_s + \beta' X). \tag{2.29}$$

Next, by equation (2.7) we have

$$1 - \Pr[\tilde{Y} \leq s] = \Pr[\tilde{Y} > s] = \prod_{i=1}^{s}(1-\tilde{\lambda}(i)). \tag{2.30}$$

Taking negative logs and substituting using equation (2.29) gives

$$-\log(1-\Pr[\tilde{Y} \leq s]) = \sum_{i=1}^{s} -\log(1-\tilde{\lambda}(i)) = \sum_{i=1}^{s} \exp(h_i + \beta' X) = \exp(\beta' X)\sum_{i=1}^{s}\exp(h_i). \tag{2.31}$$

Taking logs once more, we have

$$\log(-\log(1-\Pr[\tilde{Y} \leq s])) = \beta' X + \log\left(\sum_{i=1}^{s}\exp(h_i)\right), \tag{2.32}$$

which is precisely the OPH model (2.27), with $c_s \equiv \log(\sum_{i=1}^{s}\exp(h_i))$.

Conversely, assume the OPH model holds,

$$\text{cll}(\Pr[\tilde{Y} \leq s]) \equiv \log(-\log(1 - \Pr[\tilde{Y} \leq s])) \equiv c_s + \beta' X. \tag{2.33}$$

We first take exponentials to obtain a more algebraically useful form of the model,

$$-\log(1 - \Pr[\tilde{Y} \leq s]) = \exp(c_s + \beta' X). \tag{2.34}$$

Now then, we have

$$1 - \tilde{\lambda}(s) = 1 - \Pr[\tilde{Y} = s | \tilde{Y} \geq s] = \Pr[\tilde{Y} > s | \tilde{Y} > s - 1] = \frac{1 - \Pr[\tilde{Y} \leq s]}{1 - \Pr[\tilde{Y} \leq s - 1]}. \tag{2.35}$$

Taking negative logs and using equation (2.34) yields

$$\begin{aligned} -\log(1 - \tilde{\lambda}(s)) &= -\log(1 - \Pr[\tilde{Y} \leq s]) - (-\log(1 - \Pr[\tilde{Y} \leq s - 1]) \\ &= \exp(c_s + \beta' X) - \exp(c_{s-1} + \beta' X) \\ &= (\exp(c_s) - \exp(c_{s-1})) \exp(\beta' X). \end{aligned} \tag{2.36}$$

Finally, taking logs one more time gives us back the GPH model

$$\log(-\log(1 - \tilde{\lambda}(s))) = \log(\exp(c_s) - \exp(c_{s-1})) + \beta' X, \tag{2.37}$$

where $h_s \equiv \log(\exp(c_s) - \exp(c_{s-1}))$.

### Cumulative Link and Accelerated Failure Time Models

In the preceding sections we defined the OPO model (2.26) and OPH model (2.27) as the ordinal models which relate the cumulative probabilities $\Pr[\tilde{Y} \leq s]$ to linear predictors with the logit and complementary log-log links respectively. We may motivate the choice of other link functions with the following latent variable derivation. Suppose an underlying continuous random variable $Y$ follows the linear regression model

$$Y \equiv \beta' X + \epsilon. \tag{2.38}$$

As usual, we let $\tilde{Y}$ denote the discretized version of $Y$, though since $Y$ may now in general take on negative values, we no longer assume the cut points $\{t_0 < t_1 < t_2 < \cdots < t_J\}$ defining the discretization are all positive numbers. If the error term $\epsilon$ has cdf $G$, then

$$\Pr[\tilde{Y} \leq s] = \Pr[Y \leq t_s] = \Pr[\beta' X + \epsilon \leq t_s] = \Pr[\epsilon \leq t_s - \beta' X] = G(t_s - \beta' X), \tag{2.39}$$

so $\tilde{Y}$ follows the *cumulative link* model with link function $G^{-1}$,

$$G^{-1}(\Pr[\tilde{Y} \leq s]) = t_s - \beta' X. \tag{2.40}$$

In particular, $\tilde{Y}$ follows a *cumulative probit* model whenever it is the discretization of some latent continuous quantity for which a linear regression model with normal errors holds.

Returning to the survival analysis context, suppose now that $Y$ follows the *accelerated failure time* (AFT) model

$$Y \equiv \exp(\beta' X + \epsilon), \tag{2.41}$$

where as before $\epsilon$ is an error term with cdf $G$. Then

$$\begin{aligned} \Pr[\tilde{Y} \leq s] &= \Pr[Y \leq t_s] \\ &= \Pr[\exp(\beta' X + \epsilon) \leq t_s] \\ &= \Pr[\beta' X + \epsilon \leq \log(t_s)] \\ &= \Pr[\epsilon \leq \log(t_s) - \beta' X] \\ &= G(\log(t_s) - \beta' X), \end{aligned} \tag{2.42}$$

and so

$$G^{-1}(\Pr[\tilde{Y} \leq s]) = \log(t_s) - \beta' X. \tag{2.43}$$

Hence, the AFT model (2.41) for the continuous survival time $Y$ implies that the corresponding grouped outcome $\tilde{Y}$ will follow a cumulative link model, where the link function is the inverse of the cdf of the error term from the AFT model.

### 2.3.3 Review of Model Naming

As we have just seen, models for discrete survival outcomes often have multiple equivalent characterizations, and different ways of viewing the models have sometimes lead to redundant naming. In particular, the discrete time Cox model (2.16), which we have labeled the *discrete proportional odds* (DPO) model, has been called several different names in the literature, and we briefly review these here.

First, as previously mentioned, the DPO model is equivalent to the *continuation ratio logit* model in the ordinal logistic regression context, as described in Agresti (2002). Cupples et al. (1988) have characterized the DPO model in the case of a constant baseline hazard and time-varying covariates as a *generalized person-years approach*. Because estimation in this context involves *pooling* all observations into a single sample before running a logistic regression routine, they also call this method the *pooling of repeated observations* (PRO) approach. For the same reason, D'Agostino et al. (1990) use the phrase *pooled logistic regression*. Since the second example in Cox (1975) consists of a grouped life table, Efron (1988) calls the DPO model *partial logistic regression*. Finally, Ingram and Kleinman (1989) call the Cox discrete survival model *person-time logistic regression*, to distinguish it from ordinary logistic regression applied to dichotomized survival outcomes.

## 2.4 Multivariate Survival Analysis

When more than one outcome variable is observed on the same study subject, these observations will often not be independent. Ignoring this dependence can lead to bias and underestimation of standard errors in effect estimates. Two broad approaches exist for obtaining estimates without these problems. First, the dependence between multiple variables observed on the same subject is explicitly accounted for by including latent variables known as *random effects* in the model. The second approach is to fit models which specify the marginal distributions of the multiple observations without making strong assumptions about the exact nature of the association between the variables, and then to utilize formulas for the covariance of the parameter estimates which are asymptotically correct regardless of the true nature of the variables' association. In this section, we briefly review these two approaches in a general survival analysis context, and then review adaptations of these methods to the discrete time case. More in-depth introductions to these topics may be found in Hougaard (2000) and Cook and Lawless (2007).

### 2.4.1 Random Effects and Frailty

The random effects approach models the association between multiple variables with the inclusion of latent variables called random effects. These models make a *conditional independence assumption*: conditional on these unobserved random effects, the multiple observations are presumed independent. Random effects models may in general also be referred to as *mixed effects*, *hierarchical*, *multilevel*, *variance components*, or *subject-specific* models. In survival analysis specifically, these models are usually called *frailty models*, since the random effect may be thought of as capturing how "frail," or prone to experiencing failure, the subject is.

#### Frailty in Univariate Models

The concept of frailty was initially proposed in Vaupel et al. (1979) in the univariate survival analysis setting to model unobserved differences between individuals in their "endowment for longevity." In particular, these authors define frailty to be a latent random variable $z$, distinct for each subject, multiplying the hazard rate function for that subject. Using our notation, the continuous time Cox proportional hazards model with frailty states that the *subject-specific* hazard rate for a subject with frailty term $z$ is

$$\lambda_{SS}(t; X(t), z) \equiv z \cdot \lambda_0(t) \exp(\beta' X(t)). \tag{2.44}$$

The purpose of including the frailty term in univariate settings is for that term to capture the aggregate effect of unobserved variables and factors (such as genetic or lifestyle traits) on the hazard. This effectively partitions variability in survival times into two sources: unobserved risk factors captured by frailty and "simple randomness" (Hougaard, 1995). The hazard rate for a population of individuals may then be derived from a specification such as (2.44) for the subject-specific hazard (the hazard *conditional* on the frailty $z$) by averaging over the distribution of the frailty $z$ in the population still at risk for event occurrence. More formally, if $g(z; t, X(0,t))$ is the density function for the frailty among individuals with covariate history $X(0,t)$ still alive at time $t$, then the *population-average* hazard rate

is defined as

$$\lambda_{PA}(t; X(t)) \equiv \int_0^\infty \lambda_{SS}(t; X(t), z) g(z; t, X(0,t)) dz. \tag{2.45}$$

Unfortunately, given a CPH model (2.44) for the subject-specific hazard, the population-average hazard (2.45) may no longer follow a CPH model, even if the covariates are constant. The reason for this is that when a covariate value $X$ changes, not only does the term $\exp(\beta' X)$ change, but the conditional distribution $g(z; t, X)$ also changes. If one were to fit a CPH model not accounting for frailty to data following a frailty model (2.44), the shape of the fitted baseline hazard function may be quite different from that of the frailty model, and the parameter estimates from the fitted population-average model will be attenuated compared to their true underlying subject-specific values. Because experiencing greater hazard in the past shifts the location of the distribution of the frailty term downward more quickly as time passes (intuitively, the extra risk weeds out high-frailty individuals more quickly, leaving a lower average frailty population remaining), the effect of a risk factor on a population hazard becomes increasingly small as time passes compared to its effect on a particular subject's hazard.

### Frailty in Multivariate Models

Frailty terms represent unmeasured factors which affect the risk of the occurrence of the event of interest. When the association among multiple survival observations may be attributed to the different observations having related frailties, then inclusion of frailty random effects may be sufficient to account for the association. In the simplest case, different observations thought to be associated may share the exact same single frailty term, which is fixed over time. For example, the frailty term may be used to account for genetic similarities of siblings whose survival times are being observed. This model, which is analogous to the random intercept linear model, is known as the *shared frailty model* (Hougaard, 2000, Chap. 7). The gamma distribution is the most common choice for the distribution of the shared frailty at time zero, owing to the simple mathematical properties that result from this assumption. For example, the conditional distribution of the frailty given survival to a certain time point

is also gamma distributed. The log-normal distribution, on the other hand, is appealing since it is equivalent to including a normal random effect inside the linear predictor if a CPH model such as (2.44) is used.

One limitation of the standard shared frailty model is that because the frailty is fixed in time, it can only model *long-term dependence*, that is, situations in which increased risk for one variable implies that the hazard for a different associated variable is *always* increased throughout time. In contrast, some collections of variables may exhibit *short-term dependence*, where the occurrence of an event for one variable implies only a temporary increase in the hazard for another variable (Hougaard, 2000, Chap. 11). This may be the case when subjects share a common changing environment. In this case occurrence of an event in one subject implies that the shared environment may be especially risky at that time. We can account for such dependence by replacing the shared random variable $z$ with a stochastic process $z(t)$ that is shared among associated survival outcomes.

Another extension of the basic shared frailty model is to incorporate different functions of the random frailty term $z$ into the hazard rates for the different survival outcomes (Hougaard, 2000, Chap. 10). For example, we can model negative association between pairs of variables by multiplying the hazard for one variable by $z$ as before, and multiplying the hazard of the other variable by $1/z$. More generally, we can model differing effects of the latent covariates represented by the frailty by multiplying the hazards of each outcome by different powers of the random variable $z$, which we estimate from the data. For identifiability purposes, we must fix one of these powers. Alternatively, we can model different but related frailties with a *bona fide* multivariate distribution for the frailty. Common choices for this distribution include an additive gamma model and a multivariate log-normal distribution.

### 2.4.2 Marginal Models

The second general approach to accounting for association among survival outcomes is to specify models for the marginal distributions of the survival times and then utilize standard error and covariance estimates for the parameters of these models which are asymptoti-

cally correct regardless of the true association stucture of the survival outcomes (Hougaard, 2000, Chap. 13). The marginal distributions are usually specified with CPH models for the population-average hazard (2.45), which is the average hazard experienced among individuals with the given covariate values who are still alive at the given time.

There are two approaches to fitting marginal models. Wei et al. (1989) fit the marginal CPH models for the different outcomes separately. They show that the joint distribution of the covariate coefficient estimates is asymptotically multivariate normal, and provide estimates for the covariances of these coefficients. Alternatively, one may specify a "working model" for the covariances between the survival times, and utilize a version of the estimating equations approach introduced in Liang and Zeger (1986) and Zeger and Liang (1986).

### 2.4.3 Review of Multivariate Discrete Time Survival Literature

Ten Have and Uttal (1994) is the first work addressing multivariate discrete survival outcomes. They present both random effects and estimating equations approaches for this data, and the subject-specific and population-average discrete hazards, respectively, are both related to linear predictors with a logit link function as in the DPO model (2.16). They apply their models to a dataset consisting of 10 repeated measurements per subject of survival outcomes where failure may occur at one of three discrete time points, or may be right censored after the third time point. Due to the small number of discrete time points relative to the size of the sample, which contained 89 subjects all with complete data, they employ an arbitrary baseline hazard function for the effect of discrete time.

Their subject-specific model uses a single time-constant 3-dimensional multivariate normal random effect for each subject. The three components of the random effect are a random intercept, random measurement occasion number (whose coefficient takes values 1 through 10 for the 10 repeated measurement occasions of the survival outcome for each subject), and random occasion number squared. Hence, a subject's frailty is fixed in discrete time within each survival observation, but varies over each subject's 10 measurement occasions according to a quadratic function whose coefficients are the latent 3-dimensional random effect.

They fit the random effects model using a Bayesian Gibbs sampling approach with non-informative priors, and compare the results to those from estimating equations approaches with both independence and exchangeability working assumptions for the correlation. The random effects and exchangeability approaches performed similarly, both being superior to the independence approach.

Guo and Lin (1994) also develop estimating equations methods for discrete survival data, but utilize the GPH model (2.21) for the marginal survival times instead of the DPO model (2.16) which Ten Have and Uttal use. They provide simulation results which indicate that their robust sandwich covariance estimator derived under an independence assumption produces good estimates, whereas the naive Fisher information covariance estimate produces results with substantial bias. Since they illustrate their methods with the same dataset as Ten Have and Uttal use, they do not have to address the problem of estimating a large number of parameters for the baseline hazard.

Hedeker et al. (2000) treat multivariate discrete survival outcomes by including multivariate normal or rectangular random effects in the linear predictor. They model the survival outcomes by using both the logit and complementary log-log link functions with both discrete hazard probabilities and cumulative probabilities. In other words, they use all four models (2.16), (2.21), (2.26), and (2.27) introduced earlier for relating discrete outcomes to covariates. Their example data only uses three discrete timepoints, so as in Ten Have and Uttal (1994) and Guo and Lin (1994), they use separate baseline hazard parameters for each discrete time.

## 2.5 Competing Risks

Often in time-to-event studies, followup for some subjects ends before they experience the event of interest. When this end to followup results from an inability to observe a subject who may still be at risk, it is called *censoring*. However, followup may also end because the subject experiences a *competing risk*, an event other than the event of interest which makes them no longer at risk of experiencing the event of interest. For example, in a study of

the time to voluntary cessation of cocaine use, a subject may be arrested before voluntarily quitting the drug. If occurrence of a competing risk is not of substantive interest in the study *and* this occurrence may be assumed to be *independent* of the occurrence of the primary event of interest, then observations ending with these other events may simply be considered censored. When this is not the case, the competing risk must be accounted for in the statistical analysis and interpretation. In this section, we review methods for incorporating competing risks into survival models.

### 2.5.1 Competing Risks in Continuous Time

Suppose $Y$ is a continuous survival outcome, which may be terminated by any of $R$ competing risks, including any primary event of interest. Each observation of $Y$ consists of a pair $(y, \delta)$, where $y$ is the study time and $\delta$ is an event-type indicator. For right censored observations $(Y > y)$, we do not know the type of event the subject experienced, and this is indicated by $\delta = 0$. For observed events $(Y = y)$, we know the event type $r$ the subject experienced (where $r \in \{1, \ldots, R\}$), and this is indicated by $\delta = r$. In the special case $R = 1$, where there is only one type of event, $\delta$ is simply a censoring indicator, coded as $\delta = 1$ for an observed event and $\delta = 0$ for a censored observation.

One approach to the competing risks problem is to presume the existence of separate event times $Y_r$ for each risk $r$, at least $R - 1$ of which are *latent*, since we may only possibly observe the timing of the first risk to occur, $Y = \min(Y_1, \ldots, Y_R)$. This approach however has two problems. First, the joint distribution of $(Y_1, \ldots, Y_R)$ is not identifiable from the observable data $(y, \delta)$ without specifying additional untestable assumptions (Prentice et al., 1978). Second, it is often not clear what the physical meaning of the latent times $Y_r > Y$ is when an earlier competing risk, such as death, prevents them from ever happening. Giving them meaning would require specifying a counterfactual circumstance in which the competing risk that did occur has been removed, or rendered inoperable, in some physical manner.

An alternative approach advocated by Prentice et al. (1978) is to express the problem in terms of the *cause-specific hazards* of each risk, which are the instantaneous rates at which

events of each type are occurring given that no event of *any* type has occurred up to that time point. Letting $Y$ be the time of the occurrence of the event and $D \in \{1, \ldots, R\}$ denote which type of event occurred, we formally define the cause-specific hazard $\lambda^{(r)}(t)$ for risk $r$ at time $t$ by

$$\lambda^{(r)}(t) \equiv \lim_{h \to 0^+} \frac{\Pr[t \leq Y \leq t + h, D = r | Y \geq t]}{h}. \tag{2.46}$$

These hazards are identifiable from competing risks survival data $(y, \delta)$. Conversely, Prentice et al. show that the likelihood of a sample of observations $(y, \delta)$ may be written entirely in terms of the $\lambda^{(r)}$. Thus, they conclude that these hazards are "the basic estimable quantities in the competing risks framework."

The likelihood for continuous time competing risks data factors into separate pieces each depending on only one of the cause-specific hazards. Each of these factors is precisely the likelihood that would result from treating all other competing risks as censoring (Prentice et al., 1978). This implies that the cause-specific hazards may be estimated separately, using methods designed for single-risk data. In particular, if each of the $\lambda^{(r)}(t)$ follow a CPH model (2.13) with separate covariate effect parameters and baseline hazards, then Cox's partial likelihood (2.15) may be used to estimate the parameters for each cause-specific hazard.

### 2.5.2 Competing Risks in Discrete Time

The *discrete time cause-specific hazard* $\tilde{\lambda}^{(r)}(s)$ is defined to be the probability that risk $r$ occurs at time $s$ given that none of the risks has occurred at any previous time point. If $\tilde{Y} \in \{1, 2, \ldots, \}$ is the discrete survival time and $D \in \{1, \ldots, R\}$ is the risk that occurs, we have

$$\tilde{\lambda}^{(r)}(s) \equiv \Pr[\tilde{Y} = s, D = r | \tilde{Y} \geq s]. \tag{2.47}$$

Since for each discrete time $s$, the hazards $\{\tilde{\lambda}^{(1)}(s), \ldots, \tilde{\lambda}^{(R)}(s)\}$ are probabilities of mutually exclusive events, their sum can be at most 1. This suggests that unlike in the continuous time case, these hazards cannot be estimated separately. Indeed, the likelihood of an observation $(\tilde{y}, \delta)$ does not factor into separate components for each competing risk, so estimation must

be performed for all competing risks simultaneously.

For the single-risk DPO model (2.16), we related the single discrete hazard to covariates with a logit link function. Analogously, we may relate the multiple discrete cause-specific hazards to covariates in the same manner as a multinomial logistic model, yielding the *multinomial proportional odds* (MPO) model

$$\tilde{\lambda}^{(r)}(s; X(s)) \equiv \frac{\exp\left(\tilde{\beta}_0^{(r)}(s) + \tilde{\beta}^{(r)} X(s)\right)}{1 + \sum_{m=1}^{R} \exp\left(\tilde{\beta}_0^{(m)}(s) + \tilde{\beta}^{(m)} X(s)\right)}. \tag{2.48}$$

This model provides the same benefit as the single-risk DPO model of being able to perform estimation by first expanding the data to person-time format, with one observation for each discrete time point each subject was at risk, and then using standard logistic regression software. This approach was first described in detail in Allison (1982), and has been applied in competing risks analyses of waiting times for liver transplant and hospital length of stay (Gibbons et al., 2003; Barnett et al., 2009).

### 2.5.3 Multistate Models

In a *multistate* event history, subjects move between different discrete states over the course of follow up, and the events of interest are those marking transitions between states (Hougaard, 1999; Andersen and Keiding, 2002; Putter et al., 2007). The competing risks are then the collection of states the subject may directly transition into from the subject's current state, which will vary depending on the current state. Consequently, a *multistate model* for a multistate history requires a different competing risks survival model for each state a subject may occupy.

More formally, a multistate model requires three specifications. First, we must specify the collection of all discrete states the subjects may occupy during the study follow up period. Usually this collection is finite, in which case we label the states with the positive integers $s = 1, \ldots, S$. Second, for each state $s$, we specify the collection of all states $r$ for which it is possible to directly transition from state $s$ to state $r$ without passing into any other state in between. We say such states $r$ are *directly reachable* from $s$, and write $s \to r$ as a shorthand

representation of a direct transition. For example, suppose the state of never having used cocaine is state 1, active cocaine use is state 2, treatment for cocaine abuse is state 3, and non-treatment abstinence from cocaine use following first use is state 4. Then a transition from state $s = 1$ to state $r = 3$ or $r = 4$ is not possible, since in either case a subject must pass through state 2 in between. Third, for each state $s$ we must specify a model for the durations of episodes subjects spend in $s$, where each state $r$ directly reachable from $s$ represents a competing risk for termination of an episode in state $s$.

The primary challenge in constructing a multistate model lies in specifying the cause-specific hazards of each competing risks model. These hazards often depend simultaneously on multiple time scales and functions of the subject's past state history, such as the time since the subject last changed states, the number of previous episodes spent in the current state, the cumulative amount of time spent in the current state, and the time since the the subject first entered the state. Even after accounting for these and other observable factors, the hazards of different transition events may be associated within subjects, requiring introduction of latent random effects. We address these challenges with a Bayesian multistate model in Chapter 4.

## 2.6 Review of Event History Analysis Literature

As noted in Section 2.1.1, researchers in many different fields analyze event time data. Much of the work we have reviewed up to this point has come from the statistics and biostatistics literature. However, a parallel and somewhat separate literature on event time data has been developed by statisticians working in the social sciences, where survival analysis is usually called *event history analysis*. This event history analysis literature tends to address problems more similar to the TUE data than the general statistics survival analysis literature; in particular, they often address discrete time data. Hence we now provide a brief review of event history analysis publications.

Allison (1982) provides a comprehensive overview of event history analysis. This paper presents both the DPO model (2.16) and GPH model (2.21), and describes fitting the DPO

model. The cause-specific hazards approach to discrete time competing risks data using the MPO model (2.48) is outlined. He also discusses choice of time scale, noting that in principle one may incorporate more than one time scale into the linear predictor. Finally, he covers repeated events and explains how to allow the hazard of future events to depend on history of past events.

Yamaguchi (1990) uses discrete time survival methods to model times until transitions between two states. He proposes simultaneous modeling of two distinct two-state processes observed on the same subject using a multinomial logit model for the combined state of the two processes assumed at each discrete time point. He applies his model to a premarital cohabitation process (cohabiting or not cohabiting) and a marijuana process (using or not using), using data obtained from a timeline follow-back interview. The methods proposed do not account for between-subject heterogeneity or dependence of the transition probabilities on the history of the processes beyond using the two previous states as covariates.

Singer and Willett (1993) and Willett and Singer (1993) provide introductions to discrete time survival analysis targeted towards education and psychology researchers, respectively. These tutorials are similar to Allison (1982), though not as technical or comprehensive. In addition, Willett and Singer (1995) illustrates the application of discrete survival methods to repeated *spells* or *episodes* of a trait for which subjects pass between the states of the trait being present and absent. Like Yamaguchi (1990), repeated observation of the time until entry into the active state and the time until exit from the active state are modeled together, though heterogeneity is not accounted for.

Steele et al. (2004) address a scenario similar to that in Willett and Singer (1995), in which transition hazards for entry into and exit from a state are jointly modeled. In Steele et al.'s work, within-individual association between the hazard of one transition type across repeated episodes is modeled by sharing the same frailty term across repeated observations. In addition, frailties for different transition types (e.g., entrance into or exit from a state) are allowed to be correlated within individuals; the vector of frailty terms on the linear predictor scale for each individual is assumed to follow a multivariate normal distribution. The authors also introduce the concept of a *transition within the same state*, where for one

or more states, an episode of being in that state may end followed immediately by the start of another episode in that same state. They apply their model to a contraception use status process with two states—using contraception and not using contraception. A change of contraception method is considered a transition within the contraception using state, and this event is treating as a competing risk for the end of contraception use episodes (the other risk being termination of contraception use). They handle discrete time competing risks using the MPO model (2.48).

Steele (2011) generalizes this work to simultaneous modeling of multiple binary state processes, a situation similar to that treated by Yamaguchi. The models for two (or more) binary state processes are linked by allowing the random effects for different processes to be correlated and allowing the transition hazards for one process to depend on the current state and history of a different process.

## 2.7 Time Scales

In survival studies, the risk of a subject experiencing the event of interest is often associated with one or more notions of time. For example, if the event of interest is recurrence of cancer, then a subject's risk may be related to the time elapsed between the last remission of cancer and the current time. In this case, the notion of time is referred to as *time-at-risk*, since it measures how long the subject has been at risk of experiencing the event of interest. In this section, we review issues related to the incorporation of this and other notions of time into models for survival outcomes.

### 2.7.1 Age-Period-Cohort Effects

In addition to time-at-risk, *age* (or *time since birth*) is also often related to the hazard of event occurrence. In fact, in studies of mortality from all causes, these two notions of time coincide, since one becomes at risk for death the moment they are born. Other distinct notions of time are however also associated with event occurrence rates. The absolute calendar time the subject is observed, often referred to as the *period* of observation, may also be related

to the risk the subject experiences at that moment. For example, the risk of death from heart disease for a 60-year-old man in the period January, 1950 is most likely higher than the risk for a 60-year-old man observed in the period January, 2000, due to improvements in medical treatment for heart disease which occurred during the 50-year span separating the two periods. Furthermore, the time of birth, called the birth *cohort*, can have an association with the risk a subject experiences throughout the rest of their lifetime. This may be due to differences between birth cohorts in nutrition during infancy and early childhood. As another example, the 1990 birth cohort will likely have lower cumulative lifetime exposure to cigarette smoke than the 1940 cohort, even if we compare the groups at common ages or periods.

In principle, the three time variables *age*, *period*, and *cohort* may represent distinct underlying effects on a subject's hazard of event occurrence. Unfortunately, these effects are in general not identifiable from observable data. The reason for this is that any two of the variables completely determine the other; in particular it is always true that

$$\text{cohort} + \text{age} = \text{period}. \tag{2.49}$$

Hence, we can only observe data on the two-dimensional slice of the three-dimensional age-period-cohort space defined by equation (2.7.1). Much published work on age-period-cohort effects has focused on this identifiability problem. Usually, the suggested solution involves imposition of specific functional forms or restrictions on the effects of these variables (Fienberg and Mason, 1979). However, the estimates obtained are sensitive to which particular restriction is imposed, and choice of the identifiability constraints must be based upon information external to the sample data.

### 2.7.2 Incorporation of Multiple Time Scales

The two most widely used survival analysis techniques—namely Kaplan-Meier estimation (Kaplan and Meier, 1958) and the CPH model (2.13)—only allow nonparametric incorporation of a single time variable, as their estimation procedures cannot be extended to simultaneously include multiple time scales (Keiding, 1990). Thus any additional time variables

must be incorporated in a less satisfactory fashion, either by stratification on a discretized version of that variable or by including a parametric function of the variable inside the linear predictor of the Cox model. For this reason, research on multiple time scales has focused on choice of a single time scale (Kom et al., 1997; Thiebaut and Benichou, 2004; Imbens, 1994). Alternatively, some authors have considered conditions under which multiple time variables may be collapsed into a single variable (Oakes, 1995).

### 2.7.3 Nonparametric Bayesian Treatment of Time-Varying Effects

There are several Bayesian approaches to allowing survival outcomes to depend on time. With a single continuous time scale, one may use a model which leaves the functional form of this dependence unspecified, and supply prior information about how time relates to survival. For example, the Dirichlet process prior may be placed on the distribution of survival times (Susarla and Van Ryzin, 1976; Ferguson and Phadia, 1979), or a gamma process prior may be placed on the cumulative hazard (Kalbfleisch, 1978). However, these formulations have several drawbacks, such as discrete realizations of the survival time distributions, contradicting a basic assumption of continuous time. Of greater concern is the fact that the Dirichlet process prior induces *negative* association between the risk of event occurrence at nearby time points. This means, for example, that if an event is observed at time $t = 50.0$ years of age, then the posterior probability of event occurrence at times $t = 49.9$ years and $t = 50.1$ years of age are *lower* than if the event at $t = 50.0$ had not been observed. This directly contradicts our prior knowledge that the relationship between risk and time should generally be smooth.

Alternatively, one may assume that the hazard rates are piecewise-constant with respect to the time variable. Let $\{t_0 < t_2 < \cdots t_M\}$ be a discretization of the single time axis under consideration such that there is no person-time observation from any subject before $t_0$ or after $t_M$. We assume that the baseline hazard $\lambda_0(t)$ for the time variable is constantly equal to $\lambda_m$ for $t \in (t_{m-1}, t_m]$, and we let $\beta_m \equiv \log \lambda_m$ denote these rates on the log scale. We may then place independent priors on these hazards, such as $\lambda_m \sim \text{Gamma}(a, b)$ or $\beta_m \sim N(\mu, \sigma^2)$

(Ibrahim et al., 2001).

However, if some intervals do not contain many observed events, posterior estimates of the $\lambda_m$ may lack precision and vary erratically over nearby intervals. To address this problem, several authors have suggested introducing prior correlation for the $\lambda_m$, which allows neighboring $\lambda_m$ to borrow strength from one another and prevents implausibly large variation of the $\lambda_m$ over short time spans. We now describe such an approach using multivariate normal priors specified via sets of conditional distributions, and then review work on MCMC inference for models incorporating these components.

### 2.7.4 Dynamic, CAR, and GMRF Models

*State space* or *dynamic generalized linear models* extend GLM's by allowing the regression parameter vector $\theta$ to vary with a single discrete time scale $t$ (Harrison and Stevens, 1976; West et al., 1985). The states $\theta_t$ of the model are then presumed to follow a stochastic *transition* or *evolution* equation,

$$\theta_t = G_t\theta_{t-1} + w_t, \tag{2.50}$$

where $G_t$ is a known transition matrix and $w_t$ is normal with mean zero and known covariance. Essentially, this approach places a multivariate autoregressive prior process on the $\theta_t$ to enable a Bayesian version of a varying-coefficient model (Hastie and Tibshirani, 1993). This approach has seen extensive application in Bayesian survival analysis, where it is used in piecewise-constant formulations to approximate both smooth baseline hazards and smoothly varying coefficients (Gamerman, 1991; Fahrmeir, 1994; Fahrmeir and Wagenpfeil, 1996; Fahrmeir and Knorr-Held, 1997).

A more general class of prior distributions was introduced by Besag (1974) and subsequent co-authors to model spatial correlation in geostatistical and imaging applications (see, e.g., Besag et al., 1991). Instead of specifying the distribution of $\theta = (\theta_1, \ldots, \theta_n)$ forward in time $t$ as in (2.50), the distribution of $\theta$ is specified implicitly via its *full conditionals* $p(\theta_t|\theta_{-t})$, where $\theta_{-t} = \{\theta_s \mid s \neq t\}$. Such specifications are called *conditional autoregressions (CAR)* or *Markov random fields (MRF)* (Besag and Kooperberg, 1995). The Markov name

refers to the feature that the distributions $p(\theta_t|\theta_{-t})$ often only depend on a small number of elements in $\theta_{-t}$, which correspond to nodes $s$ which are spatially or temporally close to node $t$. Typically, the full conditionals $p(\theta_t|\theta_{-t})$ are presumed Gaussian, in which case the resulting joint distribution is multivariate normal and called a *Gaussian Markov random field (GMRF)*. The distribution can then be specified using the inverse covariance matrix, or *precision matrix*, $Q$ of the multivariate normal,

$$p(\theta|Q) \propto \exp(-\tfrac{1}{2}\theta' Q\theta). \tag{2.51}$$

The Markov property of the full conditionals ensures that the matrix $Q$ is sparse, which leads to fast algorithms for sampling $\theta$ (Rue, 2001; Rue and Held, 2005).

### 2.7.5 MCMC for Models with GMRF Components

The specifications (2.50) and (2.51) induce prior correlation among the $\theta_t$. This leads to high posterior correlations when the likelihood is weak, making block MCMC updating of the $\theta_t$ necessary to avoid slow MCMC mixing. Carter and Kohn (1994) observed slow mixing when using single-site updates for the states in a dynamic model with a normal response variable, and found that including all state parameters in a single block Gibbs step speeded mixing dramatically. In the case of a GLM with a non-normal response variable, the full conditional of a GMRF $\theta$ whose elements appear in the linear predictors of the model will generally not be possible to sample from directly. However, in the special case where each linear predictor $\eta_i$ also contains its own independent normal term $\epsilon_i$, a reparameterization replacing $\{\epsilon_i\}$ with $\{\eta_i\}$ in a Gibbs sampling scheme will result in a multivariate normal full conditional for $\theta$ (Besag et al., 1995). The parameters $\epsilon_i$ are not always desired in the model though, and such reparameterizations may slow mixing in cases where the data variance is high (Gelfand et al., 1995).

Several authors have attempted to devise full conditional approximating block Metropolis-Hastings (M-H) proposals for $\theta$ or its subvectors. Gamerman (1998) proposed sampling from the normal posterior (calculated with the Kalman filter) of a weighted least squares reformulation of a dynamic GLM. However, he reported very low acceptance rates, suggesting

inadequacies of his full conditional approximation. Shephard and Pitt (1997) used an iterative procedure at each step to calculate a second-order Taylor series expansion of the log full conditional density, which is a multivariate normal density and so may be easily sampled from as a Metropolis-Hastings proposal. They also encountered low acceptance rates with large block sizes.

Knorr-Held (1999) suggested a much simpler updating strategy, which forgos the goal of accurate approximation of the full conditional. Proposals for a parameter block $\theta_{t_1}, \ldots, \theta_{t_2}$ are drawn from the prior distribution conditional on the other parameters in the model including the other $\theta_t$ ($t < t_1$ or $t > t_2$). These *conditional prior proposals* are very easy to sample, being multivariate normal with easy to calculate moments, and result in Metropolis-Hastings acceptance ratios which simplify to the likelihood ratio. Unfortunately, acceptance rates are low whenever the likelihood is strong compared to the prior, as the proposal does not take the observed data into account.

Rue (2001) and Knorr-Held and Rue (2002) constructed full conditional approximating block M-H proposals of $\theta$ for Poisson disease mapping problems. They combined quadratic approximations of the exponential terms of the Poisson log likelihood with the GMRF prior densities to produce a GMRF approximate full conditional density that may be sampled from as an M-H proposal. They reported high acceptance rates of 20–80%, even when updating blocks of hundreds of parameters. Our M-H algorithms in Chapter 3 employ a similar strategy of log likelihood approximation, but for a different observation model with a much larger number of observations. Moreover, whereas the linear predictor specifications of Knorr-Held and Rue follow strict hierarchical structures with no covariates, our models allow much more complicated linear predictor specifications with multiple covariates and arbitrarily crossed factors.

# CHAPTER 3

# A General Bayesian Event History Model

In this chapter, we present a general Bayesian event history model, featuring competing risks, recurrent events, and semiparametric incorporation of time-varying covariates including multiple time scales. Smoothing of the arbitrary functions representing the effects of each continuous covariate is accomplished using Gaussian Markov random field (GMRF) priors. To overcome computational difficulties due to the large person-time datasets involved, we provide several novel MCMC algorithms and efficient data structures for their software implementation. As we show in Section 3.2, the fact that these large collections of person-time observations produce only moderate numbers of observed events simplifies the derivation and calculation of efficient block Metropolis-Hastings proposals.

The rest of this chapter is organized as follows. In Section 3.1, we detail the semiparametric, competing risks, recurrent event history model we use throughout the remainder of the chapter. Section 3.2 introduces novel computing tools for efficiently obtaining inferences from this model. In Section 3.3, we apply our model and algorithms to data on recurrent episodes of cocaine use from Treatment Utilization and Effectiveness Project subjects. Extensions of our model are discussed in the next two chapters.

## 3.1 Bayesian Event History Model

In this section, we outline our discrete time competing risks observation model, which is based upon the *multinomial proportional odds* model (2.48) from Section 2.5.2. We then discuss the components of the linear predictors and their prior specification.

### 3.1.1 Discrete Time Competing Risks Observation Model

Let $i = 1, \ldots, I$ index subjects, $j = 1, \ldots, J_i$ index repeated at-risk episodes, or *recurrent events*, within subject $i$, and $t = 1, \ldots, T_{ij}$ index discrete time points at which subject $i$ is at risk for his $j^{\text{th}}$ recurrent event. The value $t$, often called *time-at-risk*, *duration*, or *gap time*, always resets to $t = 1$ at the beginning of each successive episode, so that $T_{ij}$ is the discrete duration of episode $j$ from subject $i$. We also let $r = 1, \ldots, R$ index the distinct competing risks under consideration, and let $\Delta_{ij} \in \{1, \ldots, R\}$ denote the competing risk which terminates the $j^{\text{th}}$ at-risk period from subject $i$ at time $t = T_{ij}$.

The *discrete time cause-specific hazard* $\lambda_{ij}^{(r)}(t)$ is the probability that the $j^{\text{th}}$ recurrent event from subject $i$ occurs due to risk $r$ at time $t$ elapsed since the subject became at risk for the $j^{\text{th}}$ recurrence, given that the event had not yet occurred by time $t-1$,

$$\lambda_{ij}^{(r)}(t) \equiv \Pr[T_{ij} = t, \Delta_{ij} = r \mid T_{ij} \geq t]. \tag{3.1}$$

For convenience, we also let $\lambda_{ij}^{(0)}(t)$ denote the probability that none of the competing risks occurs at time $t$ given the subject was still at risk at that point, so that $\lambda_{ij}^{(0)}(t) = 1 - \sum_{r=1}^{R} \lambda_{ij}^{(r)}(t)$. For all $i$, $j$, and $t$, the collection $\{\lambda_{ij}^{(0)}(t), \ldots, \lambda_{ij}^{(R)}(t)\}$ are probabilities of mutually exclusive events, representing the $R+1$ possible outcomes at person-time observation $(i, j, t)$. We relate these probabilities to $R$ linear predictors $\{\eta_{ij}^{(1)}(t), \ldots, \eta_{ij}^{(R)}(t)\}$ using the multinomial logit link function,

$$\eta_{ij}^{(r)}(t) \equiv \log\left(\frac{\lambda_{ij}^{(r)}(t)}{\lambda_{ij}^{(0)}(t)}\right), \quad r = 1, \ldots, R, \tag{3.2}$$

which allows the $\lambda_{ij}^{(r)}(t)$ to assume any values subject to $0 < \lambda_{ij}^{(r)}(t) < 1$ and $\sum_{r=0}^{R} \lambda_{ij}^{(r)}(t) = 1$.

### 3.1.2 Linear Predictor and Prior Specification

The linear predictors $\eta_{ij}^{(r)}(t)$ can include covariates that vary between subjects, between at-risk episodes within subjects, and between time points within episodes. Let $X_{m,ij}(t)$ denote the value of covariate $m$ for subject $i$ at time $t$ during the $j^{\text{th}}$ at-risk episode. The effects of each covariate $X_m$ are modeled with separate functions $\beta_m^{(r)}(x)$ for each competing risk $r$;

we also include intercept terms $\beta_0^{(r)}$, giving

$$\eta_{ij}^{(r)}(t) \equiv \beta_0^{(r)} + \sum_{m=1}^{M} \beta_m^{(r)}(X_{m,ij}(t)). \tag{3.3}$$

The functions $\beta_m^{(r)}(x)$ are of three types, depending on the type of covariate. In all cases, however, the functions assume only a finite number of values, labeled $\beta_m^{(r)}[1], \ldots, \beta_m^{(r)}[K_m]$.

### Continuous Covariates

For continuous $X_m$, we use piecewise-constant functions $\beta_m^{(r)}(x)$ with GMRF priors to approximate arbitrary smooth functions. We may then, for example, incorporate duration dependence by setting $X_{1,ij}(t) = t$ for all $i$, $j$, and $t$, so that $\beta_1^{(r)}(t)$ corresponds to a cause-specific baseline hazard for $r = 1, \ldots, R$.

Let $c_{m,0} < c_{m,1} < \cdots < c_{m,K_m}$ be a known collection of change points for the effect of covariate $m$. We assume $\beta_m^{(r)}(x)$ is constant and equal to $\beta_m^{(r)}[k]$ for all $x$ in the $k^{\text{th}}$ interval defined by these points,

$$\beta_m^{(r)}(x) \equiv \beta_m^{(r)}[k], \quad \text{for all } x \in (c_{m,k-1}, c_{m,k}]. \tag{3.4}$$

We then place a *first-order random walk* GMRF prior on the parameters $\beta_m^{(r)}[1], \ldots, \beta_m^{(r)}[K_m]$,

$$\beta_m^{(r)}[k] \mid \beta_m^{(r)}[1], \ldots, \beta_m^{(r)}[k-1] \sim N(\beta_m^{(r)}[k-1], \sigma_m^{(r)}), \quad k = 2, \ldots, K_m, \tag{3.5}$$

with a flat prior for the first parameter $\beta_m^{(r)}[1]$. This is an improper, or *intrinsic*, GMRF, as it is invariant to level shifts added to all the $\beta_m^{(r)}[k]$ (Besag and Kooperberg, 1995). Denoting by $IG(a, b)$ the inverse gamma distribution with shape $a$ and scale $b$ and writing $\sigma = \sigma_m^{(r)}$, we complete the specification by assigning priors $\sigma \sim IG(a, b)$, which gives

$$p(\sigma|a, b) = \frac{b^a}{\Gamma(a)} \sigma^{-a-1} \exp\left(\frac{-b}{\sigma}\right). \tag{3.6}$$

### Categorical Covariates

The second type of covariate is an unordered categorical variable with a small or moderate number of categories. Suppose $X_m$ may only take on values in the set $\{x_{m,1}, \ldots, x_{m,K_m}\}$. We

allow the functions $\beta_m^{(r)}(x)$, $r = 1, \ldots, R$, to have different values for each of these covariate values, and we write

$$\beta_m^{(r)}(x_{m,k}) \equiv \beta_m^{(r)}[k], \quad \text{for } k = 1, \ldots, K_m. \tag{3.7}$$

We require a prior for the column vector of parameters $\theta = (\beta_m^{(r)}[1], \ldots, \beta_m^{(r)}[K_m])'$, with three properties: exchangeability of the components $\theta_k = \beta_m^{(r)}[k]$, borrowing of strength among the $\theta_k$, and invariance to level shifts. Writing $K = K_m$ for convenience, and letting $I_K$ denote the $K \times K$ identity matrix and $J_K$ the $K \times K$ matrix of 1's, define the singular prior precision matrix

$$Q \equiv \tau \tfrac{K}{K-1} \left( I_K - \tfrac{1}{K} J_K \right), \tag{3.8}$$

which then defines an improper prior for $\theta$ via equation (2.51). The *a priori* specified conditional prior precision parameter $\tau$ controls the degree of borrowing of strength in the prior, which is easy to see from the full conditionals

$$\theta_k \mid \theta_{-k} \sim N\left( \frac{1}{K-1} \sum_{l \neq k} \theta_l,\ \tau^{-1} \right). \tag{3.9}$$

**Random Effects**

Finally, we consider categorical variables with large numbers of categories, in which case it is desirable to allow the data to inform the between-category variability. For example, we may use the variable $X_m$ to model associations among recurrent events within subjects by setting $X_{m,ij}(t) = i$ for all $i$, $j$, and $t$. The random effects $\beta_m^{(r)}(i)$ are then *shared frailties*, as described in Section 2.4.1.

As in the fixed effects case, we assume $X_m$ takes values in $\{x_{m,k}\}_{k=1}^{K_m}$ and write $\beta_m^{(r)}(x_{m,k}) \equiv \beta_m^{(r)}[k]$. Let $\beta_m[k] = (\beta_m^{(1)}[k], \ldots, \beta_m^{(R)}[k])'$ denote the vector of the $k^{\text{th}}$ category's random effects for all competing risks $r$; we allow these effects to be correlated within categories by assigning a multivariate normal distribution to $\beta_m[k]$,

$$\beta_m[k] \overset{iid}{\sim} N_R(0, \Sigma_m), \quad k = 1, \ldots, K_m. \tag{3.10}$$

Letting $IW(\nu, S)$ denote the inverse Wishart distribution with $\nu$ degrees of freedom and

scale $S$, we assign hyperpriors $\Sigma_m \sim IW(\nu, S)$, so that

$$p(\Sigma_m|\nu, S) \propto |\Sigma_m|^{-(\nu+R+1)/2} \exp\left(-\tfrac{1}{2}\text{trace}\left(S\Sigma_m^{-1}\right)\right) \tag{3.11}$$

and $E[\Sigma_m] = S/(\nu - R - 1)$.

## 3.2 Computing

Event history analyses often involve very large person-time datasets with hazards of event occurrence which continually vary within and between subjects. Obtaining inferences from such models using MCMC requires repeated evaluation of the log likelihood of these person-time observations, which is computationally intensive given the dataset sizes. Hence, it is necessary to ensure that each such traversal of the likelihood is as efficient as possible, both in terms of computer time required and mixing speed of the resulting MCMC routine. While much work has been done over the past two decades on MCMC for hierarchical generalized linear models, the particular problems posed by event history datasets have up to this point not been addressed.

We now describe three novel MCMC techniques for efficiently obtaining inferences from the model in the previous section. The first involves creation of several auxiliary data structures to speed updating and evaluation of the likelihood. The other techniques use approximations of the log likelihood to construct block Metropolis-Hastings updates.

### 3.2.1 Data Structures and Likelihood Evaluation

#### Likelihood

First we establish some notation. Let the pair $(t_{ij}, \delta_{ij})$ be the data observed for the $j^{\text{th}}$ recurrent event from subject $i$, where $\delta_{ij} = r \geq 1$ means risk $r$ was observed at time $t = t_{ij}$ and $\delta_{ij} = 0$ denotes right censoring at $t_{ij}$. We assume that any right censoring occurs at the end of the corresponding time interval. For each $r = 0, \ldots, R$, set $\delta_{ij}^{(r)} = 1$ if $\delta_{ij} = r$ and $\delta_{ij}^{(r)} = 0$ otherwise, and let $y_{ij}^{(r)}(t)$ be person-time level event indicators, where $y_{ij}^{(r)}(t) = 1$ if

$t = t_{ij}$ and $r = \delta_{ij}$, with $y_{ij}^{(r)}(t) = 0$ otherwise.

Letting $\theta$ denote all parameters of our model and $D = \{(t_{ij}, \delta_{ij})\}$ denote the observed data, the likelihood is

$$\begin{aligned} L(\theta|D) &= \prod_{i=1}^{I}\prod_{j=1}^{J_i}\left(\prod_{t=1}^{t_{ij}-1}\lambda_{ij}^{(0)}(t)\right)\left(\prod_{r=0}^{R}\left(\lambda_{ij}^{(r)}(t_{ij})\right)^{\delta_{ij}^{(r)}}\right) \\ &= \prod_{i=1}^{I}\prod_{j=1}^{J_i}\left(\prod_{t=1}^{t_{ij}}\lambda_{ij}^{(0)}(t)\right)\left(\prod_{r=1}^{R}\left(\exp\left(\eta_{ij}^{(r)}(t_{ij})\right)\right)^{\delta_{ij}^{(r)}}\right) \\ &= \prod_{i=1}^{I}\prod_{j=1}^{J_i}\prod_{t=1}^{t_{ij}}\left[\left(1+\sum_{r=1}^{R}\exp\left(\eta_{ij}^{(r)}(t)\right)\right)^{-1}\prod_{r=1}^{R}\exp\left(y_{ij}^{(r)}(t)\cdot\eta_{ij}^{(r)}(t)\right)\right], \end{aligned} \tag{3.12}$$

where in the second line we used equation (3.2) to replace $\lambda_{ij}^{(r)}(t_{ij})$ with $\lambda_{ij}^{(0)}(t_{ij})\exp(\eta_{ij}^{(r)}(t_{ij}))$ and moved the term $\lambda_{ij}^{(0)}(t_{ij})$ into the first product. In the third line, we have replaced the indicators $\delta_{ij}^{(r)}$ with $y_{ij}^{(r)}(t_{ij})$ and included the exponential terms for $t < t_{ij}$, which is possible since for those $t$ we have $y_{ij}^{(r)}(t) = 0$. In addition, in the third line we have used the substitution $\lambda_{ij}^{(0)}(t) = (1 + \sum_{r=1}^{R}\exp(\eta_{ij}^{(r)}(t)))^{-1}$ implied by the multinomial logit link function (3.2).

Unlike the continuous time cause-specific hazards model, the discrete time model likelihood (3.12) does not factor into separate components for each cause-specific hazard. Thus separate hazard estimation is not possible with the discrete model, which makes sense as otherwise this could result in incompatible estimates of the $\lambda_{ij}^{(r)}(t)$, which sum over $r$ to greater than 1.

### Person-time Indexing

The likelihood (3.12) is a product over all person-time observations of the expression in large brackets. Currently, we are using three indices $i$, $j$, and $t$ to specify person-time observations. For convenience, we now pass to a single index $n = 1, \ldots, N$ for all person-time observations, where $N = \sum_{i=1}^{I}\sum_{j=1}^{J_i} t_{ij}$ is the total number of person-time observations. The order in which $n$ indexes these observations can be arbitrary, but for optimal computer memory access performance, we recommend *row major order*, where $n = 1$ corresponds to

$(i, j, t) = (1, 1, 1)$, $n = 2$ corresponds to $(i, j, t) = (1, 1, 2)$,..., $n = J_1 + 1$ corresponds to $(i, j, k) = (1, 2, 1)$, and so on. We now replace the $ij$ subscript and function argument $t$ with the single subscript $n$ in all our notation; e.g., $\lambda_{ij}^{(r)}(t)$ and $X_{m,ij}(t)$ become simply $\lambda_n^{(r)}$ and $X_{m,n}$. Thus the likelihood (3.12) may be re-written

$$L(\theta|D) = \prod_{n=1}^{N} \left[ \left( 1 + \sum_{r=1}^{R} \exp\left(\eta_n^{(r)}\right) \right)^{-1} \prod_{r=1}^{R} \exp\left(y_n^{(r)} \cdot \eta_n^{(r)}\right) \right]. \tag{3.13}$$

**Discretized Predictors and Multinomial Aggregation**

For all three types of covariates we consider, the effects are represented by functions $\beta_m^{(r)}(x)$ for each risk $r$ additively included in the appropriate linear predictors. Each of these functions, in turn, is represented by the finite collection of values $\{\beta_m^{(r)}[1], \ldots, \beta_m^{(r)}[K_m]\}$ the function assumes. For each entry $x = X_{m,n}$ in our set of covariate values, let $k_{m,n}$ denote the integer $k$ such that $\beta_m^{(r)}(x) = \beta_m^{(r)}[k]$. Using these *discretized predictors* $k_{m,n}$ and the person-time index $n$, the linear predictor specification (3.3) may be rewritten

$$\eta_n^{(r)} \equiv \beta_0^{(r)} + \sum_{m=1}^{M} \beta_m^{(r)}[k_{m,n}]. \tag{3.14}$$

Depending on the number and types of covariates, as well as the discretizations chosen for any continuous covariates, for certain subsets $\{n_1, \ldots, n_b\}$ of person-time observations we will have $k_{m,n_1} = \cdots = k_{m,n_b}$ for all $m$. These observations will then always have the same values for the corresponding linear predictors (3.14). In this case, we can store only one copy of the covariate vector $(k_{1,n_1}, \ldots, k_{M,n_1})$ along with the size $b$ of the subset and the total number of observed events $y^{(r)} \equiv y_{n_1}^{(r)} + \cdots + y_{n_b}^{(r)}$ for each competing risk $r$. The likelihood contribution of these $b$ person-time observations can then be replaced by the single multinomial likelihood term

$$\left( 1 + \sum_{r=1}^{R} \exp\left(\eta_{n_1}^{(r)}\right) \right)^{-b} \prod_{r=1}^{R} \exp\left(y^{(r)} \cdot \eta_{n_1}^{(r)}\right), \tag{3.15}$$

which may speed evaluation of the likelihood (3.13) substantially; we call this replacement *multinomial aggregation*. For example, the model described in Section 3.3.1 has $M = 7$ covariates, including subject-specific random effects and three continuous time scales. Despite using discretizations of the continuous time scales with around 50 change points each,

multinomial aggregation in this case reduced the number of terms in the likelihood by 17%, yielding a substantial reduction in computation time per MCMC iteration. We recommend performing this aggregation in practice, especially for models without random effects and for models with fewer time-varying continuous predictors, but for the sake of simplicity we ignore its implementation in subsequent algorithm descriptions.

**Data Structures for Efficient Computation**

We obtain posterior inferences from our model via Metropolis-Hasting (M-H) algorithms, and we derive efficient proposals for these block updates in the following subsection. Here, we detail the construction of auxiliary data structures designed to reduce the computational burden of the log likelihood ratio evaluation step of these M-H updates. In particular, we avoid the explicit construction and manipulation of very large, sparse design matrices. The logarithm of the likelihood (3.13) is

$$l(\theta|D) = \sum_{n=1}^{N} \left[ -\log\left(1 + \sum_{r=1}^{R} \exp\left(\eta_n^{(r)}\right)\right) + \sum_{r=1}^{R} y_n^{(r)} \cdot \eta_n^{(r)} \right]. \tag{3.16}$$

Due to the large number $N$ of person-time observations in our event history datasets, evaluation of the log likelihood ratio $l(\theta^*|D) - l(\theta|D)$, where $\theta$ denotes the current values of the model parameters and $\theta^*$ is the proposed update, consumes the majority of the computer time of each MCMC scan.

For each covariate $m$, we define blocks of linear predictor parameters by specifying contiguous subsets $k_1 : k_2 \equiv \{k_1, k_1 + 1, \ldots, k_2\}$ of the discretized predictor values $1, \ldots, K_m$ and setting $B_{k_1:k_2} \equiv \{\beta_m^{(r)}[k] \mid k_1 \leq k \leq k_2,\ 1 \leq r \leq R\}$. To enable faster calculation of the log likelihood ratio for updating the parameters in the block $B_{k_1:k_2}$ all at once, we store three auxiliary data structures:

1. We pre-compute the set $n_{m,k_1:k_2} \equiv \{n \mid k_1 \leq k_{m,n} \leq k_2\}$, which is the subset of person-time observations whose hazards depend on any of the parameters in $B_{k_1:k_2}$; we write $n_{m,k} \equiv n_{m,k:k}$ for short when $k_1 = k_2$. The expression in brackets in (3.16) then only needs to be evaluated for $n \in n_{m,k_1:k_2}$, as the remaining terms will cancel in

$l(\theta^*|D) - l(\theta|D)$. We call the structure $\{n_{m,k_1:k_2} \mid m = 1, \ldots, M,$ category subsets $k_1 : k_2\}$ a *doubly-ragged array*, because the number of subsets of discretized predictor values $k_1 : k_2$ used in the chosen blocking scheme varies with the covariate $m$, and in turn the size of the set $n_{m,k_1:k_2}$ varies with both $m$ and $k_1 : k_2$.

2. We store the values of all linear predictors $\eta_n^{(r)}$ evaluated at the current model parameters $\theta$, which we denote by $\eta_n^{(r)}(\theta)$. Then these values do not need to be re-calculated from scratch at each step of the MCMC algorithm, and $\eta_n^{(r)}(\theta^*) = \eta_n^{(r)}(\theta) + \Delta\beta_m^{(r)}[k_{m,n}]$ is also simple to evaluate, where $\Delta\beta_m^{(r)}[k_{m,n}] \equiv \beta_m^{(r)}[k_{m,n}]^* - \beta_m^{(r)}[k_{m,n}]$ denotes the proposed change in the parameter $\beta_m^{(r)}[k_{m,n}]$.

3. We pre-compute $y_{m,k}^{(r)} \equiv \sum_{n \in n_{m,k}} y_n^{(r)}$ for each $k \in k_1 : k_2$, which is the number of observed events for cause $r$ whose hazards depends on the parameter $\beta_m^{(r)}[k]$. Then we make the replacement

$$\sum_{n \in n_{m,k_1:k_2}} \sum_{r=1}^{R} y_n^{(r)} \eta_n^{(r)}(\theta^*) - y_n^{(r)} \eta_n^{(r)}(\theta) = \sum_{k \in k_1:k_2} \sum_{r=1}^{R} y_{m,k}^{(r)} \cdot \Delta\beta_m^{(r)}[k] \tag{3.17}$$

in the calculation of the log likelihood ratio. Note that a sum over person-time observations $n \in n_{m,k_1:k_2}$ on the left-hand side of (3.17) has been replaced with a sum over categories $k \in k_1 : k_2$ on the right, which has a much smaller number of terms. In addition, we show in the next subsection that $y_{m,k}^{(r)}$ is a good approximation to the Fisher information for the parameter $\beta_m^{(r)}[k]$, a fact which helps us compute efficient proposals.

Our novel combination of linear predictor formulation and supporting data structures for efficient MCMC updates enables fitting of rich event history models to very large person-time datasets.

### 3.2.2 Block Random Walk Metropolis Step

We propose two Metropolis-Hasting algorithms for sampling from the posterior. The first is a block *random walk Metropolis* (RWM) algorithm which uses multivariate normal proposals

centered at the current parameter values. We describe this method here, and then provide a block *full conditional approximating* (FCA) Metropolis-Hastings algorithm in the following subsection.

Given a contiguous subset $k_1 : k_2 = \{k_1, k_1+1, \ldots, k_2\}$ of the discretized values $1, \ldots, K_m$ of the predictor $X_m$ and its associated block of parameters $B \equiv \{\beta_m^{(r)}[k] \mid k_1 \leq k \leq k_2,\ 1 \leq r \leq R\}$, we need to find a covariance matrix $\Sigma_B$ which ensures that the RWM proposal

$$B^* \sim N(B, \Sigma_B) \tag{3.18}$$

provides efficient MCMC mixing. Gelman et al. (1996) show that we should choose $\Sigma_B$ proportional to the covariance $\Sigma_C$ of the full conditional distribution $p(B \mid \theta \backslash B, D)$; specifically, we should set $\Sigma_B \equiv (2.4/\sqrt{|B|})^2 \Sigma_C$, where $|B| = R \cdot (k_2 - k_1 + 1)$ is the dimension of the block $B$. Because $\Sigma_C$ is difficult to calculate directly, we approximate the corresponding precision matrix $Q_C \equiv \Sigma_C^{-1}$ using the Hession matrix of the log full conditional density with respect to the parameters in $B$. The second derivatives of the log conditional prior density $p(B \mid \theta \backslash B)$ are available immediately, since for all three covariate classes described in Section 3.1.2, the conditional prior is normal with precision $Q_P$ which is a simple function of conditioned upon parameters.

We approximate the second derivatives of the log likelihood (3.16) as follows. Letting $\beta_1 \equiv \beta_m^{(r_1)}[\tilde{k}_1]$ and $\beta_2 \equiv \beta_m^{(r_2)}[\tilde{k}_2]$ denote two parameters in $B$, we first have

$$\frac{\partial^2 l(\theta|D)}{\partial \beta_1^2} = \sum_{n \in n_{m,\tilde{k}_1}} -\lambda_n^{(r_1)}(1 - \lambda_n^{(r_1)}) \approx \sum_{n \in n_{m,\tilde{k}_1}} -\lambda_n^{(r_1)} \approx -y_{m,\tilde{k}_1}^{(r_1)}. \tag{3.19}$$

The first approximate equality follows from the fact that the hazards $\lambda_n^{(r)}$ are generally small probabilities, and the second is due to the fact that the aggregate hazard $\lambda_{m,k}^{(r)} \equiv \sum_{n \in n_{m,k}} \lambda_n^{(r)}$ for a collection $n_{m,k}$ of person-time observations should be roughly equal to the total number of events $y_{m,k}^{(r)} \equiv \sum_{n \in n_{m,k}} y_n^{(r)}$ observed among that collection. For the mixed partial derivatives, when $\tilde{k}_1 = \tilde{k}_2$ and $r_1 \neq r_2$ we have

$$\frac{\partial^2 l(\theta|D)}{\partial \beta_1 \partial \beta_2} = \sum_{n \in n_{m,\tilde{k}_1}} \lambda_n^{(r_1)} \lambda_n^{(r_2)} \approx 0, \tag{3.20}$$

because the product of the hazards $\lambda_n^{(r_1)}\lambda_n^{(r_2)}$ is small relative to the sizes of the terms in (3.19). Finally, the mixed partials are identically 0 when $\tilde{k}_1 \neq \tilde{k}_2$, since $\beta_m^{(r_1)}[\tilde{k}_1]$ and $\beta_m^{(r_2)}[\tilde{k}_2]$ have no person-time observations in common.

We combine these log likelihood approximations with the conditional prior precision matrix $Q_P$ of the block $B$ by adding the numbers of observed events $y_{m,k}^{(r)}$ along the diagonal of $Q_P$ to produce our approximation $\hat{Q}_C$ to the full conditional precision $Q_C$. Setting $\Sigma_B \equiv (2.4/\sqrt{|B|})^2\hat{Q}_C^{-1}$, we can then sample from the proposal (3.18) using an efficient method which takes advantage of the sparsity of $\hat{Q}_C$ (Rue and Held, 2005, Chap. 2). Importantly, computing $\hat{Q}_C$ and sampling from (3.18) does not require any computation which scales with $N$, so computation time in each MCMC scan is still dominated by log likelihood evaluation. As we show in Section 3.3.2, these proposals result in nearly-optimal acceptance rates with no tuning or adaptation required and with little increase in computation time per iteration compared to using fixed proposals.

### 3.2.3 Block Full Conditional Approximating Metropolis-Hastings Step

The *effective sample size* for a parameter $\theta$ for an MCMC run of a certain length is the size of an independent sample from the posterior having the same standard error for estimating the posterior mean of $\theta$ as the standard error of the mean from the MCMC run. The *efficiency* of the MCMC algorithm for $\theta$ is then defined as the effective sample size divided by the length of the MCMC chain. For example, if an MCMC run of 10,000 iterations provides the same information about the posterior mean of $\theta$ as an independent sample from the posterior of size 180, then the effective sample size of this run for $\theta$ is 180, and the MCMC efficiency is $180/10000 = 1.8\%$.

It is critical for maximizing MCMC efficiency to update groups of highly-correlated parameters within the same block. When using GMRF priors with fine discretizations of continuous covariates, these groups of highly-correlated parameters can be large, necessitating large blocks $B$. However, the maximum efficiency of RWM algorithms is roughly $0.3/|B|$, and thus declines with increasing block size $|B|$ (Gelman et al., 1996). We have found in

practice for our models that as $|B|$ increases, the advantages of block updating of highly-correlated parameters are often negated by the RWM block size efficiency penalty, leading to poor performance with MCMC efficiencies of 1% or less. For these circumstances, we have developed block Metropolis-Hastings proposals which seek to approximate the entire shape, including location information, of the full conditional distributions, not just their dispersion as with RWM. Unlike with RWM, these *full conditional approximating* (FCA) M-H steps should be nearly as efficient as exact draws from the full conditionals.

Our strategy for creating such proposals is to combine the Gaussian conditional prior with a multivariate second-order Taylor series approximation of the log likelihood. In particular, we expand $l(B|\theta\backslash B, D)$ as a function of the block parameters $B \equiv \{\beta_m^{(r)}[k] \mid k_1 \leq k \leq k_2,\ 1 \leq r \leq R\}$ about their current values in the chain $\tilde{B} \equiv \{\tilde{\beta}_m^{(r)}[k] \mid k_1 \leq k \leq k_2,\ 1 \leq r \leq R\}$. Letting $=_B$ and $\approx_B$ denote equality and approximate equality as a function of $B$ up to an *additive* constant, we have

$$\begin{aligned} l(B|\theta\backslash B, D) &\approx_B \sum_{r=1}^{R}\sum_{k=k_1}^{k_2}\left[\frac{\partial l(\tilde{B})}{\partial\beta_m^{(r)}[k]}\left(\beta_m^{(r)}[k]-\tilde{\beta}_m^{(r)}[k]\right)\right. \\ &\qquad \left. +\frac{1}{2}\frac{\partial^2 l(\tilde{B})}{\partial\beta_m^{(r)}[k]^2}\left(\beta_m^{(r)}[k]-\tilde{\beta}_m^{(r)}[k]\right)^2\right] \\ &=_B \sum_{r=1}^{R}\sum_{k=k_1}^{k_2}\left[\left(\frac{\partial l(\tilde{B})}{\partial\beta_m^{(r)}[k]}-\frac{\partial^2 l(\tilde{B})}{\partial\beta_m^{(r)}[k]^2}\cdot\tilde{\beta}_m^{(r)}[k]\right)\beta_m^{(r)}[k]\right. \\ &\qquad \left. +\frac{1}{2}\left(\frac{\partial^2 l(\tilde{B})}{\partial\beta_m^{(r)}[k]^2}\right)\beta_m^{(r)}[k]^2\right]. \end{aligned} \tag{3.21}$$

We omit the terms of the Taylor expansion with mixed partial derivatives, since by (3.20) and the associated discussion these derivatives are all either close to zero or identically zero. For the two fixed effects covariate cases in Section 3.1.2, this allows sampling of the parameters $B^{(r)} \equiv \{\beta_m^{(r)}[k] \mid k_1 \leq k \leq k_2\}$ separately for each risk $r$, since the $B^{(r)}$ are then conditionally independent.

The first derivatives of the log likelihood evaluated at the current parameter values $\tilde{B}$ are given by

$$\frac{\partial l(\tilde{B})}{\partial\beta_m^{(r)}[k]} = \sum_{n\in n_{m,k}} y_n^{(r)} - \tilde{\lambda}_n^{(r)} = y_{m,k}^{(r)} - \sum_{n\in n_{m,k}} \tilde{\lambda}_n^{(r)}, \tag{3.22}$$

which requires evaluating the hazards $\tilde{\lambda}_n^{(r)}$ under the current values of the model parameters; the second derivatives may then either be approximated as in (3.19), or calculated exactly from the $\tilde{\lambda}_n^{(r)}$. Once the necessary derivatives are calculated, the coefficients of the linear terms $\beta_m^{(r)}[k]$ and quadratic terms $\beta_m^{(r)}[k]^2$ in (3.21) are combined with the linear and quadratic terms from the log prior density to produce a quadratic approximation to the log full conditional density. Because quadratic log density functions correspond to normal distributions, we can then sample from this density as our M-H proposal for the block $B$. The main difficulty with this approach is that it requires computing the hazards $\tilde{\lambda}_n^{(r)}$ under the current parameter values. Furthermore, to calculate the proposal density ratio for the accept/reject step of the M-H algorithm, we must also compute these hazards under the proposed values of the block parameters. However, we demonstrate in Section 3.3.2 that the dramatic improvement in mixing more than makes up for the additional computation time per M-H update.

## 3.3 TUE Application

We apply the models and algorithms developed in the previous sections to analyze the durations of repeated episodes of cocaine use from the Treatment Utilization and Effectiveness Project. The data we utilize comes from a subsample of 408 TUE subjects who completed the Natural History Interview described in Section 1.2.1 and reported at least one cocaine use episode. Episodes of cocaine use are recorded as contiguous segments of whole calendar months. For example, if a subject began using cocaine in January of 1988 and ceased using in June of the following year, that would be recorded as one episode of cocaine use covering the 18 discrete time points corresponding to all 12 months of 1988 and the first 6 months of 1989.

### 3.3.1 Data and Model

In total, the 408 subjects contributed 1,527 episodes and 29,645 person-month observations of cocaine use. Hence, on average each subject had $1527/408 = 3.74$ cocaine use episodes

with mean duration 29645/1527 = 19.4 months. We consider $R = 2$ competing causes for cocaine episode termination. First, subjects may cease using cocaine due to arrest, and we let competing risk $r = 1$ be *incarceration*. Voluntary use cessation, or *stop-use* for short, is then risk $r = 2$. Of the 1527 cocaine use episodes, 689 ended with incarceration, 738 ended with stop-use, and 100 were right censored due to ongoing cocaine use at the time of the interview.

We consider $M = 7$ covariates in total. Sex and race are included as categorical fixed effects and given the exchangeable priors described in Section 3.1.2, with small conditional prior precisions $\tau$, making the priors relatively noninformative. Number of episodes of cocaine use up to and including the current episode is also treated as categorical, with the effect remaining the same for sixth and subsequent episodes for a total of six categories. Current duration of the episode is included using the GMRF smoothing prior given in Section 3.1.2. We allow the duration effect to vary at the ends of each of the first 60 months of cocaine use and at the ends of each of the sixth through tenth years of use; this discretization of the duration time scale results in 66 categories. Similarly, we include current age and current absolute calendar year. Finally, we include subject-specific random effects. Table 3.1 summarizes these covariates.

### 3.3.2 MCMC Performance Evaluation

We evaluate the performance of the algorithms described in the previous section by comparing three MCMC schemes applied to our example dataset. For each predictor $m$ and category $k \in \{1, \ldots, K_m\}$, the *single-category RWM* (SC-RWM) scheme updates the parameters $\{\beta_m^{(1)}[k], \beta_m^{(2)}[k]\}$ together in a block of size $R = 2$; we do not consider any scheme which updates the parameters $\beta_m^{(1)}[k]$ and $\beta_m^{(2)}[k]$ separately, because this would require evaluating the log likelihood twice as often as schemes which update these parameters together. The *multi-category RWM* (MC-RWM) scheme retains the single-category updating for the parameters of the categorical covariates and random effects, but uses RWM updates of blocks of the form $B \equiv \{\beta_m^{(r)}[k] \mid k_1 \leq k \leq k_2,\ r = 1, 2\}$ for the three time scales (duration, age,

Table 3.1: Covariate specification used for modeling the durations of repeated episodes of cocaine use.

| $m$ | Name | Covariate Type | Categories $K_m$ | Values/Discretization |
|---|---|---|---|---|
| 1 | Subject ID | Random Effect | 408 | 1,. . . ,408 |
| 2 | Sex | Categorical | 2 | Male, Female |
| 3 | Race | Categorical | 3 | Black, Hispanic, White |
| 4 | Episode Number | Categorical | 6 | 1,. . . ,5, $\geq 6$ |
| 5 | Episode Duration | Continuous | 66 | months 1,. . . ,60, |
| | | | | years 6,. . . ,10, $\geq 10$ years |
| 6 | Age | Continuous | 46 | 10,. . . ,55 years of age |
| 7 | Calendar Time | Continuous | 35 | calendar years 1963,. . . ,1997 |

and calendar time). We use blocks of 6 adjacent categories $k$ (for 12 total parameters per block), with the exception of the last blocks for age and calendar time, which had 4 and 5 categories respectively. Finally, the *multi-category FCA* (MC-FCA) scheme replaces the block RWM updates for the time scale parameters in the MC-RWM scheme with the block FCA M-H algorithm of Section 3.2.3.

We ran each of the three schemes for 20,000 iterations, following 2,000 iterations discarded as burn-in. We first examine the adequacy of the approximations used to derive the M-H proposals by comparing the observed acceptance rates in the three runs to their theoretically optimal values. For the first scheme, we updated all fixed-effects parameters in 158 blocks each of size 2; we ignore the random effects in this comparison, as for these parameters we pre-computed fixed proposal distributions to speed computation. Gelman et al. (1996) show that the optimal RWM acceptance rate for blocks of size 2 is .352, and our observed rates summarized in the first line of Table 3.2 are quite close to this value for all 158 blocks.

The rates from the 25 multi-category RWM updates in the second scheme are summarized in the second line of Table 3.2; all of these blocks are of size 12, except for one block of size 10 and one of size 8. The optimal rate lies in between the $\approx .26$ value Gelman et al. give for

Table 3.2: Summaries of Metropolis-Hastings proposal acceptance rates for fixed-effects parameters. The first line summaries single-category (block size 2) RWM acceptance rates of all fixed-effects parameters. The second and third lines summarize acceptance rates for updates of 25 blocks each containing roughly 12 parameters representing the time scale effects.

| | Number | Optimal | Observed Rates | | | | |
|---|---|---|---|---|---|---|---|
| Scheme | of Blocks | Rate | Min. | $25^{\text{th}}$ | Median | $75^{\text{th}}$ | Max. |
| SC-RWM | 158 | .352 | .333 | .351 | .355 | .361 | .385 |
| MC-RWM | 25 | $\approx .25$ | .245 | .254 | .257 | .260 | .276 |
| MC-FCA | 25 | 1.000 | .754 | .825 | .861 | .952 | .996 |

blocks of size 8–10 and the asymptotically optimal rate of .234 as the block size increases indefinitely. As in the single category case, the observed multi-category RWM rates are very close to optimal. Finally, for the full conditional approximating proposals used in the MC-FCA scheme, the goal is to accept as many proposals as possible. Typically almost 90% of these proposals are accepted, indicating that our Gaussian approximation to the full conditionals developed in Section 3.2.3 is quite accurate.

Next, we compare MCMC efficiencies of the three schemes for estimating the posterior means of the parameters representing the time scale effects. Here *efficiency* means the factor by which we must multiply the number of MCMC iterations to obtain the size of an independent sample from the posterior having the same Monte Carlo standard error for estimating the posterior mean; we estimate efficiency using the `effectiveSize()` function from the `coda` R package (Plummer et al., 2012). These results are summarized in Table 3.3.

The SC-RWM and MC-RWM algorithms performed very similarly in almost all cases; it appears the advantages of updating correlated parameters in the same blocks were canceled by the $\approx$ 6-fold lower theoretical maximum efficiency possible when performing RWM updates of blocks 6-times larger. The only case where the performance of the SC-RWM and MC-RWM schemes diverged was for the parameters representing the duration effect on

Table 3.3: Summary statistics of observed MCMC efficiencies of the parameters representing each time scale effect for each MCMC scheme.

| Predictor | Number of Parameters | Scheme | Observed Efficiency | | | | |
|---|---|---|---|---|---|---|---|
| | | | Min. | $25^{\text{th}}$ | Median | $75^{\text{th}}$ | Max. |
| Duration | 132 | SC-RWM | .005 | .007 | .018 | .057 | .096 |
| | | MC-RWM | .003 | .007 | .014 | .024 | .032 |
| | | MC-FCA | .039 | .093 | .128 | .340 | .445 |
| Age | 92 | SC-RWM | .002 | .004 | .005 | .008 | .013 |
| | | MC-RWM | .002 | .004 | .006 | .007 | .014 |
| | | MC-FCA | .023 | .034 | .044 | .074 | .106 |
| Calendar Time | 70 | SC-RWM | .002 | .005 | .008 | .013 | .022 |
| | | MC-RWM | .003 | .006 | .009 | .012 | .018 |
| | | MC-FCA | .026 | .072 | .095 | .121 | .214 |

the stop-use competing risk. Here, the SC-RWM scheme performs better because too little smoothing occurs for this effect, leading to reduced posterior correlation of the associated parameters. We discuss this problem further in the following subsection.

The MC-FCA scheme consistently outperforms the two RWM schemes by a factor of roughly 10. This Metropolis-Hastings scheme provides the advantages of blocking without the drawbacks of lower acceptance rates and smaller proposed steps of large-block-size RWM. However, the dramatically improved MCMC mixing comes at a cost of increased computation time per MCMC iteration, as this scheme requires evaluation of the first and second derivatives of the log likelihood under both the current and proposed values of the parameter block being updated. FCA updates appear to take about three times as long as RWM updates for any blocking scheme; because the MC-FCA scheme used these updates for three of the seven covariates, we expect it to overall run about twice as slow as the other two schemes. Indeed, the SC-RWM and MC-RWM schemes ran at 10.75 and 10.85 iterations per second, respectively, in R 3.0 on a 1.6GHz PC, while the MC-FCA scheme only produced

5.48 iterations per second. Hence, overall the FCA algorithm is 3-5 times as efficient as RWM.

### 3.3.3 Model Selection and Inferences

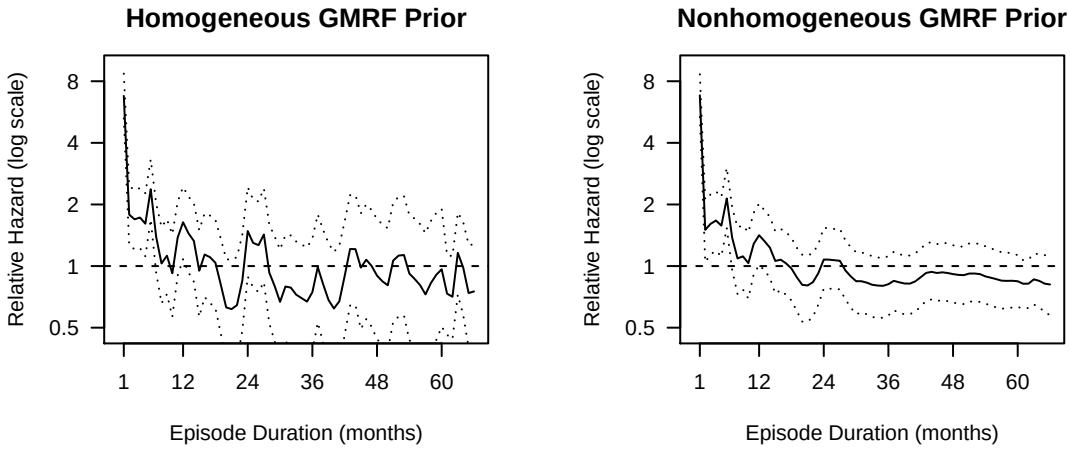
Our model allows the hazards of each of the two competing risks, incarceration and voluntary drug use cessation, to depend on seven covariates, as given in Table 3.1. We now consider whether and how these covariates are related to these risks. We begin by addressing the lack of smoothing of the duration effect noted in the previous section, and then proceed to interpret posterior summaries of the covariate effects.

#### Nonhomogeneous GMRF Duration Effect Prior

The SC-RWM scheme performed better than the MC-RWM scheme for updating the duration effect parameters, while the two schemes performed similarly for the other time scale effects. Because we expect MC-RWM, which uses larger blocks than SC-RWM, to perform better with increasing posterior correlation, this suggests that our GMRF prior induced less posterior correlation among the duration parameters than was induced for the other time scale effects. Indeed, the first panel of Figure 3.1 depicts posterior medians and 95% credible intervals for the duration effect on the stop-use competing risk. We observe a roughly 4-fold decrease in risk of voluntary use cessation following the first month of the cocaine use episode. This large, well-estimated first-month change results in a high inferred value for the increment variance $\sigma_5^{(2)}$ of the first-order random walk GMRF prior (3.5), which in turn leads to the inadequate smoothing evident in the left panel of Figure 3.1.

The source of the problem is that our GMRF prior (3.5) is *homogeneous*; that is, the increment variance $\sigma_m^{(r)}$ is the same for all $k$. A more sensible prior would reflect our prior knowledge that the hazards should change more rapidly earlier in the cocaine use episodes. For example, the chances of a subject voluntarily stopping use should change more between months 1 and 2 of cocaine use than between months 50 and 51. We model this by introducing constant scaling factors $s_m^{(r)}[k]$, $k = 1, \ldots, K_m - 1$, in front of $\sigma_m^{(r)}$, yielding the

Figure 3.1: Posterior medians (solid lines) and point-wise 95% credible intervals (dotted lines) of the duration effects on the stop-use competing risk, using the homogeneous GMRF prior (left panel) and nonhomogeneous prior (right panel).




*nonhomogeneous* random walk GMRF prior

$$\beta_m^{(r)}[k] \mid \beta_m^{(r)}[1], \ldots, \beta_m^{(r)}[k-1] \sim N(\beta_m^{(r)}[k-1], s_m^{(r)}[k-1]\sigma_m^{(r)}), \quad k = 2, \ldots, K_m. \tag{3.23}$$

For the duration effect, we choose the constants $s_m^{(r)}[k]$ so that $s_m^{(r)}[k] \propto 1/k$ and $\sum_k s_m^{(r)}[k] = K_m - 1$. With this choice, for a given value of $\sigma_m^{(r)}$, the aggregate variance $\sum_k s_m^{(r)}[k]\sigma_m^{(r)}$ of the random walk prior across the entire time scale is the same as for the homogeneous prior (3.5), making it sensible to compare inferred values of $\sigma_m^{(r)}$ between models using the two priors.

The right panel of Figure 3.1 shows posterior summaries of the duration effect on the stop-use risk using the nonhomogeneous prior. At earlier durations of cocaine use, the estimates are similar to those obtained with the homogeneous prior. At later durations, more borrowing of strength occurs, leading to a smoother inferred relationship with narrower credible intervals. The posterior mean of $\sigma_5^{(2)}$ decreases 4-fold from 0.163 to 0.042 when passing from the homogeneous to nonhomogeneous prior, which also shows the nonhomogeneous prior induces much more smoothing across the time scale. Meanwhile, for the incarceration competing risk, we observe no dramatic change in hazard during the first month, and the

two priors give similar results, with the posterior mean of $\sigma_5^{(1)}$ virtually unchanged when passing between the two formulations. We adopt the nonhomogeneous prior for the duration effects on both competing risks, and retain the homogeneous formulations for the age and calendar time effects.

**Inferences**

We now examine in detail the inferred relationships between each of the covariates and each of the competing risks. Due to the lack of identifiability of level shifts in the additive linear predictor formulation (3.14), we first center each set of parameters $\beta_m^{(r)}[1], \ldots, \beta_m^{(r)}[K_m]$ by subtracting from each parameter the set mean. We then plot posterior quantiles (medians and 95% credible interval endpoints) of the centered $\beta_m^{(r)}[k]$ versus $k$ for each risk $r$ and predictor $m$. In these plots, the horizontal axis has been labeled with the covariate values corresponding to the category values $k$, and the vertical axis has been labeled with the exponentials of the centered parameter values, so that these labels correspond to approximate hazard ratios comparing subjects with category value $k$ to the mean hazard of all categories of that covariate.

Figure 3.2 presents these inference plots for the three categorical covariates. We see from the plot in the first panel that the hazard of incarceration among active cocaine users does not depend strongly on the subject's sex. Because the 95% credible interval endpoints for the male category are 0.906 and 1.319, the credible interval for the hazard ratio comparing men to women is $(0.906^2, 1.319^2) = (0.82, 1.74)$. Thus men have between an 18% lower and 74% higher risk of incarceration during any given month of cocaine use than women with the same values for the other covariates. On the other hand, women are between 3% and 84% more likely to voluntarily stop cocaine use in any given month than comparable men. In the second row of Figure 3.2, we see that Hispanic active cocaine users are incarcerated at a rate roughly 40% higher than for black or white users, which are incarcerated at similar rates, though 95% credible intervals for the three groups all overlap substantially. White subjects stopped using cocaine voluntarily at rate around 75% higher than black or Hispanic users,

who stopped use at similar rates.

Number of episodes of cocaine use up to and including the current episode appears to be strongly related to both competing risks, though in opposite ways. The hazard of becoming incarcerated during an episode of cocaine use for subjects having had three or more previous episodes of use is roughly 2.5 times higher than for subjects during their first episode of use, all else equal. On the other hand, the hazard of a subject voluntarily stopping use drops by almost 80% between the first and sixth episodes of cocaine use. Because we have included subject-specific random effects, these are *subject-specific interpretations*. This means for example that the increased incarceration hazard at later episodes is not due to observing later episodes more frequently from subjects with higher propensities for arrest; individual subjects' risks of incarceration are actually increasing with their number of cocaine use episodes. These findings underscore the importance of intervening earlier in subjects' drug use careers.

The time scale effects are summarized in Figure 3.3. Incarceration risk appears elevated during the first year of a cocaine use episode, and then steadily drops off thereafter, decreasing by almost half between month 12 and month 60 of continuous cocaine use. The hazard of a subject voluntarily ceasing cocaine use are 4.5 times lower in the second month of use than in the first month. Over the subsequent 30 months of use, the stop-use hazard drops by another factor of 2, leveling off thereafter. Spikes in the hazard at months 6, 12, and 24 may represent recall errors due to subjects rounding the durations of episodes they report. These spikes do not show up in the incarceration hazard plot, because official arrest records were used to aid recall.

Our model does not provide clear evidence of associations between the current age of the subjects and the risks of either incarceration or voluntary use cessation. Twenty-year-old subjects may be roughly one third more likely than forty-year-olds to become incarcerated while using cocaine, and subjects in their teenage years may be a third more likely to voluntarily quit than older subjects. In contrast, absolute calendar time has a very strong relationship with both competing risks. Incarceration rates for active cocaine users increased 4-fold between 1980 and 1995, coinciding with the crack cocaine epidemic in the Los Angeles

Table 3.4: Mean deviance ($\bar{D}$), deviance at the posterior mean ($D(\bar{\theta})$), effective number of parameters ($p_D$), and deviance information criterion (DIC) for models with all predictors included or one predictor removed.

| Model | $\bar{D}$ | $D(\bar{\theta})$ | $p_D$ | DIC |
|---|---|---|---|---|
| All Predictors | 11120.7 | 10655.8 | 465.0 | 11585.7 |
| Sex Removed | 11123.9 | 10659.8 | 464.1 | 11588.0 |
| Race Removed | 11137.0 | 10677.9 | 459.0 | 11596.0 |
| Episode Removed | 11244.8 | 10834.4 | 410.5 | 11655.3 |
| Duration Removed | 11128.6 | 10606.6 | 522.0 | 11650.6 |
| Age Removed | 11120.4 | 10660.9 | 459.5 | 11579.9 |
| Calendar Time Removed | 11353.4 | 10947.7 | 405.8 | 11759.2 |

region. The chances of a subject voluntarily stopping use appear to have increased even more during the same time frame. However, this could be partially attributable to recall bias, with subjects having a more fine-grained recollection of cocaine use episodes (and thus a greater frequency of reported voluntary termination events) in the years just before the interviews were conducted in the mid 1990's.

We conclude based upon these posterior summaries that among our predictors only age may not be related to the risk of cocaine use termination. To confirm these findings, we calculated the deviance information criterion (DIC; Spiegelhalter et al., 2002) for the model with all predictors included and the six models having one of the fixed effects predictors removed. From Table 3.4, we see that the model excluding age attains the lowest DIC, beating the model including all predictors by 6. The model excluding sex appears only slightly worse than the model including all predictors, which makes sense given that the credible intervals for the sex effect on the stop-use risk just barely miss the null value of 1, while the intervals for the incarceration risk intersect. Removal of any other predictors results in substantial increases in DIC.

Figure 3.2: Posterior medians (solid lines) and point-wise 95% credible intervals (dotted lines) of the categorical covariate effects on the risk of incarceration (left column) and stop-use (right column).

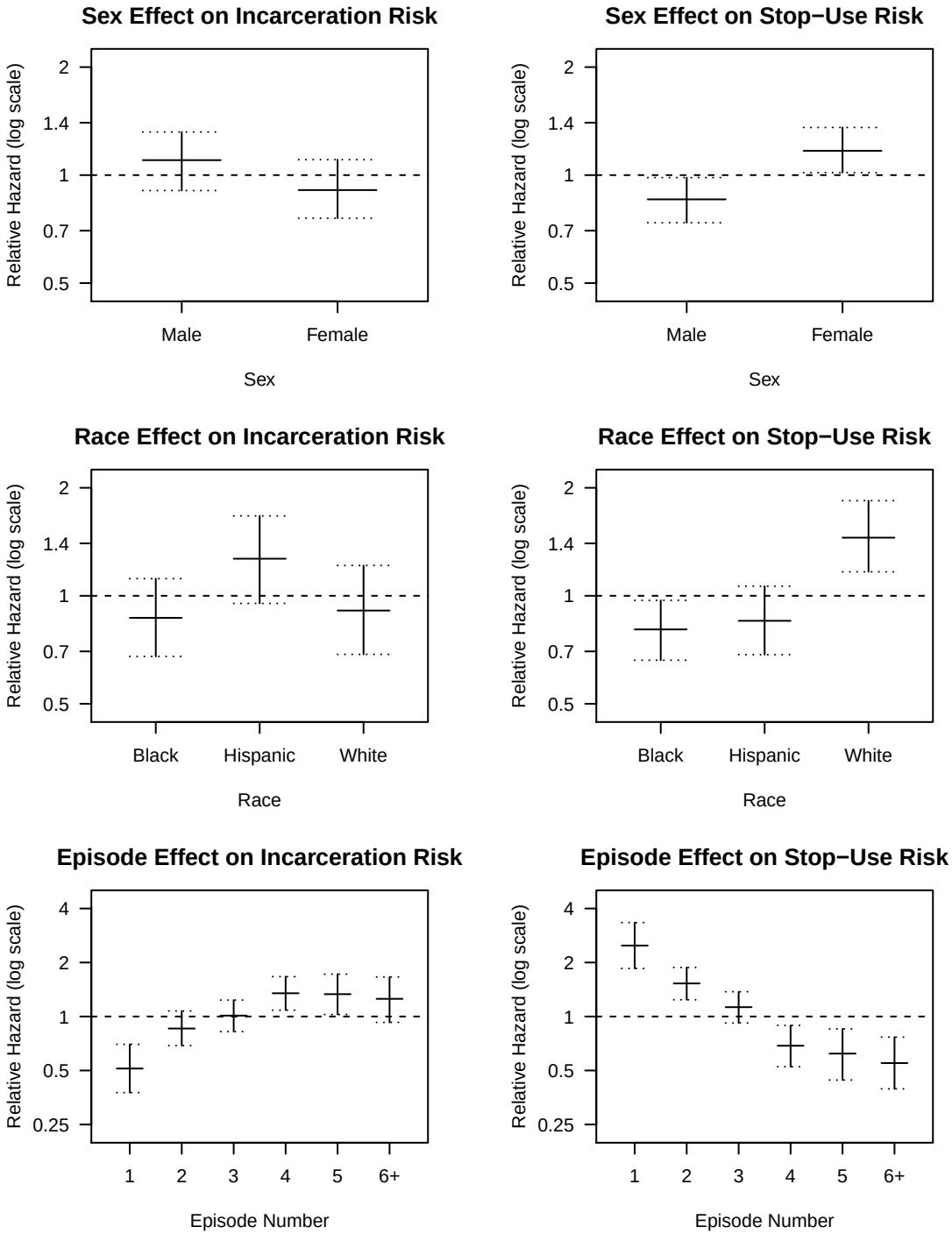
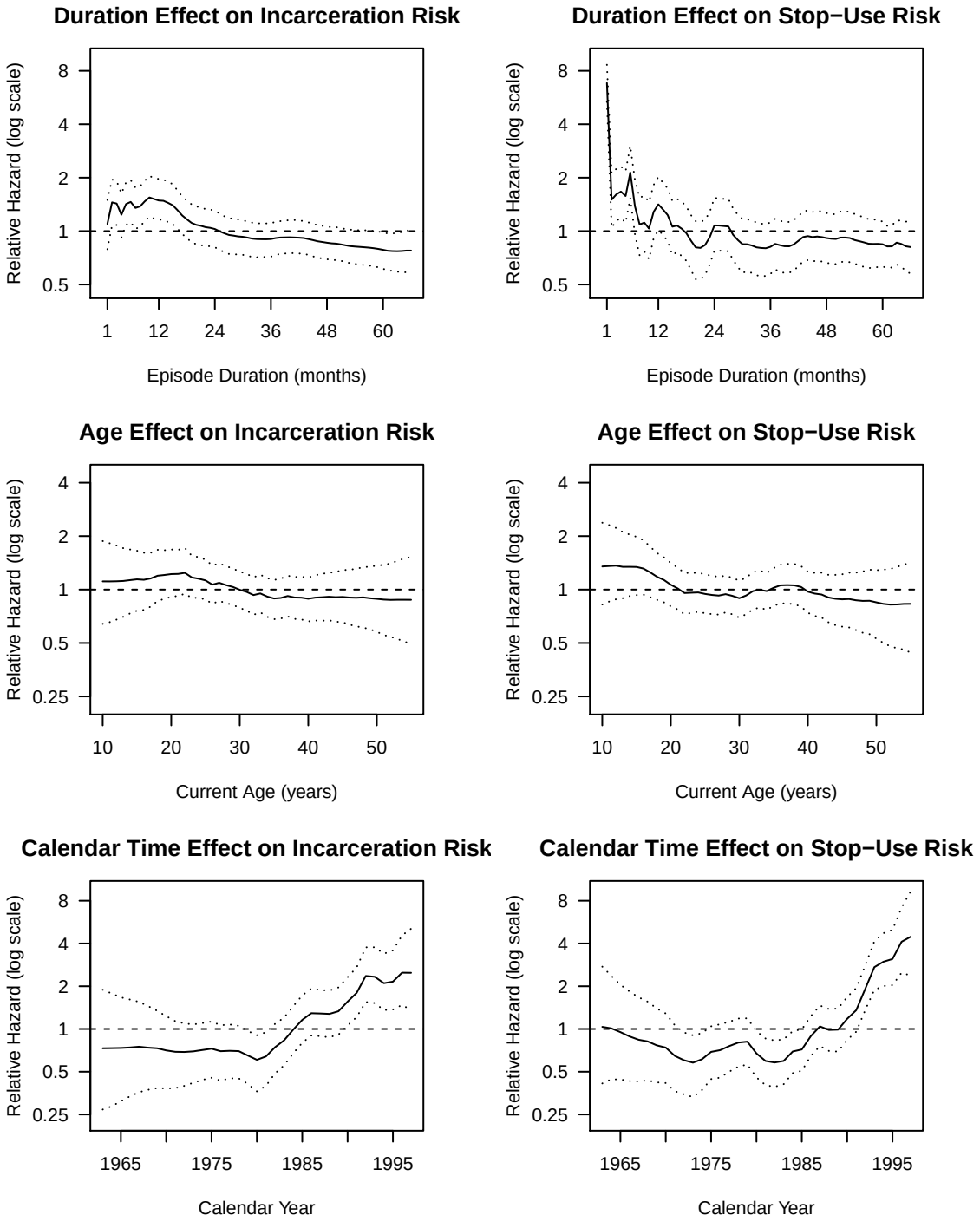


Figure 3.3: Posterior medians (solid lines) and point-wise 95% credible intervals (dotted lines) of the time scale effects on the risk of incarceration (left column) and stop-use (right column).



# CHAPTER 4

# A General Bayesian Multistate Model

In this chapter, we extend the competing risks event history model of the previous chapter to a *multistate model*, as defined in Section 2.5.3, for histories in which subjects pass in between discrete states. We begin by extending the notation and definitions of the previous chapter to accommodate dependence on state $s$, and then show how multiple competing risks event history models may be linked to form a full multistate model. In Section 4.2, we use this model to analyze lifetime histories of episodes of cocaine use, episodes of abstinence from use, and episodes of incarceration from Treatment Utilization and Effectiveness Project subjects.

## 4.1 Multistate Model Specification

To specify a complete model for the state each subject assumes at each discrete time point of observation, we must, for each state $s = 1, \ldots, S$, give a model for the probabilities of subjects passing from state $s$ to each state $r$ directly reachable from $s$. Each of these $S$ sub-models is a discrete time competing risks event history model of the type considered in Chapter 3. Thus, we extend the model given by the cause-specific hazards definition (3.1), the link relationship (3.2), and linear predictor specification (3.3) to a general multistate model context by allowing the data and parameters to depend on the state $s$.

Let $i = 1, \ldots, I$ index the subjects in our dataset, $j = 1, \ldots, J_i^{(s)}$ index episodes in which subject $i$ is continually in state $s$, and $t = 1, \ldots, T_{ij}^{(s)}$ index the discrete time points that subject $i$ spends in his $j^{\text{th}}$ episode in state $s$. The variable $t$ resets to the value 1 each time a subject enters a new state, so that $t$ represents the discrete time since the subject last changed states; this variable is often called *gap time* or *duration*. Finally, we let $R^{(s)} \subset \{1, \ldots, S\}$

denote the set of all states $r$ directly reachable from $s$, and $\Delta_{ij}^{(s)} \in R^{(s)}$ denote the state entered upon termination of subject $i$'s $j^{\text{th}}$ episode in state $s$.

The *transition hazard rate* $\lambda_{ij}^{(s,r)}(t)$ is the probability that subject $i$ moves from state $s$ to state $r$ at the $t^{\text{th}}$ discrete time point of $i$'s $j^{\text{th}}$ episode in state $s$ given that $i$ had not exited $s$ at an earlier time in that episode,

$$\lambda_{ij}^{(s,r)}(t) \equiv \Pr[T_{ij}^{(s)} = t, \Delta_{ij}^{(s)} = r \mid T_{ij}^{(s)} \geq t]. \tag{4.1}$$

As in the competing risks event history model of Chapter 3, for each $s$, $i$, $j$, and $t$, we relate the transition hazards $\{\lambda_{ij}^{(s,r)}(t) \mid r \in R^{(s)}\}$ to linear predictors $\eta_{ij}^{(s,r)}(t)$ with the multinomial logit link function,

$$\eta_{ij}^{(s,r)}(t) \equiv \log\left(\frac{\lambda_{ij}^{(s,r)}(t)}{\lambda_{ij}^{(s,s)}(t)}\right), \quad r \in R^{(s)}, \tag{4.2}$$

where $\lambda_{ij}^{(s,s)}(t) \equiv 1 - \sum_{r \in R^{(s)}} \lambda_{ij}^{(s,r)}(t)$ is the probability of remaining in state $s$ at time $t$.

For each state $s$, we let the hazard rates for transitions out of $s$ depend on covariates $X_m^{(s)}$, $m = 1, \ldots, M^{(s)}$. The values of these covariates may vary with subject, episode, and time. Let $X_{m,ij}^{(s)}(t)$ denote the value of $X_m^{(s)}$ for subject $i$ at time $t$ during his $j^{\text{th}}$ episode in state $s$. We model the covariate effects with functions $\beta_m^{(s,r)}(x)$ included in the linear predictors,

$$\eta_{ij}^{(s,r)}(t) \equiv \beta_0^{(s,r)} + \sum_{m=1}^{M^{(s)}} \beta_m^{(s,r)}(X_{m,ij}^{(s)}(t)). \tag{4.3}$$

The covariates may be of three types, as described in Section 3.1.2. First, continuous covariate effects are modeled using the Gaussian Markov Random Fields smoothing approach outlined in Section 3.1.2. Second, for categorical covariate effects we allow the function $\beta_m^{(s,r)}(x)$ to take different values for each $x$, and place an exchangeable prior on these function values.

The third case is covariates $X_m^{(s)}$ that correspond to random effects. For example, if $X_{1,ij}^{(s)}(t) = i$ for all $s$, $i$, $j$, and $t$, then the first covariate for each state is a subject identifier, and its effects on the transition hazards are subject-specific random effects. Suppose more generally that covariate $m$ is the same random effects identifier for all states $s$, and let $k$ be a value assumed by the $X_{m,ij}^{(s)}(t)$. Then for each state $s$ and state $r \in R^{(s)}$, $\beta_m^{(s,r)}(k)$ is the

random effect for category $k$ on the $s \to r$ transition hazard. We allow these random effects to be correlated across all transition types. Let $\beta_m(k)$ be the vector of all category $k$ random effects $\beta_m^{(s,r)}(k)$ listed in some fixed order. We place a multivariate normal prior on $\beta_m(k)$,

$$\beta_m(k) \sim N_d(0, \Sigma_m), \tag{4.4}$$

where $d = \sum_{s=1}^{S} |R^{(s)}|$ is the number of possible transition types in the multistate model. We complete the specification by placing the inverse Wishart hyperprior (3.11) on $\Sigma_m$, with $d$ replacing $R$ as the dimension value.

Obtaining posterior inferences from this model proceeds as in the previous chapter, using any of the MCMC algorithms from Section 3.2.

## 4.2 TUE Cocaine and Incarceration History Application

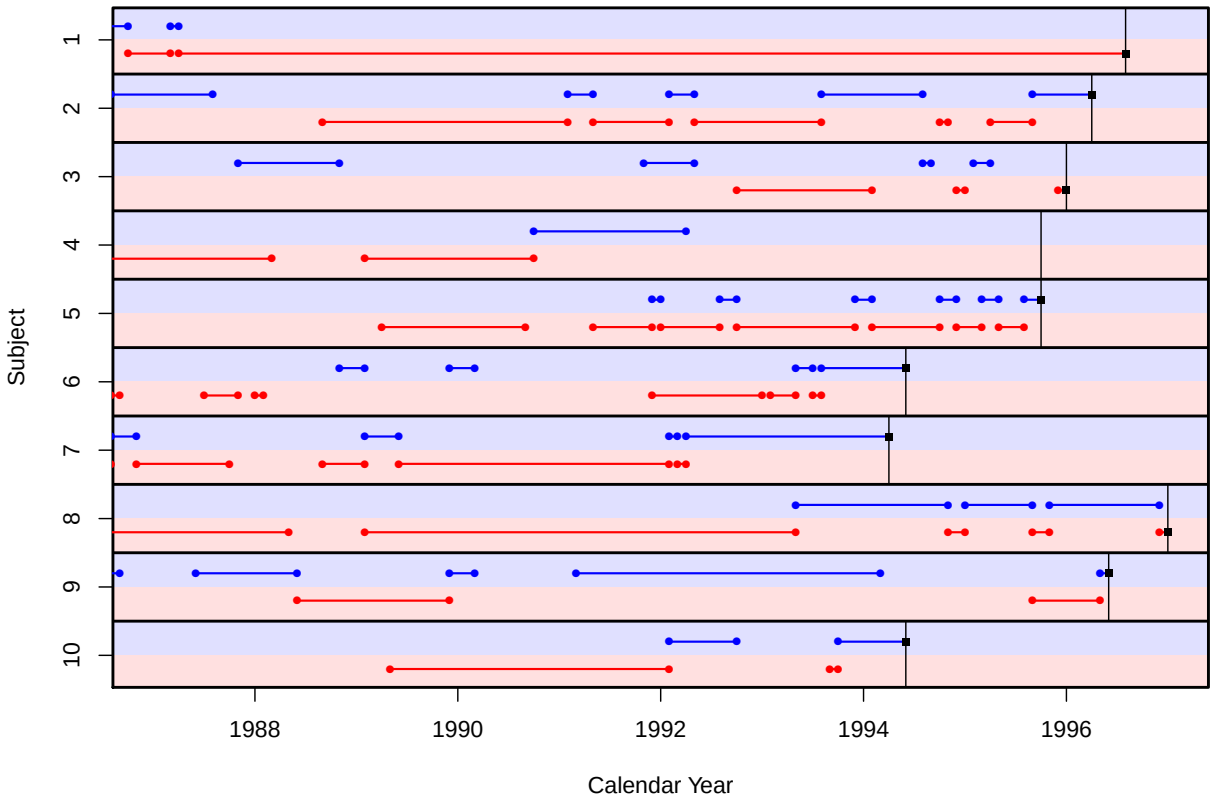
We analyze complete cocaine and incarceration history from the time of first cocaine use until the time of interview in the sample of 408 TUE subjects used with the event history model in Section 3.3.1. These histories may be broken into three mutually exclusive states: *active cocaine use* (C), *incarceration* (I), and *non-using* while not incarcerated (N). Table 4.1 summarizes the number of episodes spent in each state across all subjects. In total, the subjects contributed 63,492 person-months of follow up, for 13.0 years of observation per subject on average. Among this time, 47% was spent using cocaine, 13% was spent incarcerated, and 40% was spent not using cocaine while not incarcerated.

Figure 4.1 shows all episodes of cocaine use (red) and incarceration (blue) occurring between 1987 and 1997 for ten TUE subjects; spans with neither red nor blue line segments are episodes of non-use. Patterns of use and imprisonment vary greatly between subjects. The first subject uses cocaine continually for the entire decade preceding his interview. The fourth subject never resumes cocaine use following a 2-year incarceration episode, while the fifth subject frequently passes back and fourth between jail and cocaine use in the five years prior to the time of interview.

Table 4.1: Summaries of all episodes of cocaine use (C), incarceration (I), and non-use (N) following first use of cocaine in 408 TUE subjects. All subject histories began at the time of first cocaine use and were right censored at the time of interview.

| State | Total Num. of Epis. | Num. Epis. Ending With: C | I | N | Censoring | Total Person-Month Obs. | Avg. Epis. Dur. (months) |
|---|---|---|---|---|---|---|---|
| C | 1527 | | 697 | 729 | 101 | 29645 | 19.4 |
| I | 1102 | 507 | | 512 | 83 | 8567 | 7.8 |
| N | 1241 | 612 | 405 | | 224 | 25280 | 20.4 |
| all | 3870 | 1119 | 1102 | 1241 | 408 | 63492 | 16.4 |

Figure 4.1: Cocaine and incarceration history spanning 1987 to 1997 from ten TUE subjects. Blue and red line segments represent episodes of incarceration and cocaine use, respectively. Vertical black lines signify the time of interview, with squares indicating right censoring.



### 4.2.1 Covariates

For each of the three states, we model the hazards of transitions into each of the two other states using all 7 covariates used in the example in Section 3.3.1 plus one additional covariate, so that we have $M^{(s)} = 8$ total covariates for each state $s$. The additional covariate we include is the previous state just prior to entering the current state, which is categorical with the categories being the two states other than the current state. Sex and race are included as categorical fixed effects, with race having three categories (black, Hispanic, and white). Number of episodes spent in that state up to and including the current episode is also included as a categorical covariate, with episode numbers of 6 and higher grouped into the same category. A subject identifier is included, so that our model has subject-specific random effects for each of the six possible transition types.

Episode *duration*, also called *gap time* because it is the gap between transition events, is included as a continuous covariate with the piecewise-constant nonhomogeneous GMRF model of Section 3.3.3. For cocaine use and non-use episodes, we use the same discretization of the duration time scale as we used for the model in Section 3.3, with the duration effect varying at the ends of each of the first 60 months and at the ends of years 6 through 10. Because as presented in Table 4.1, incarceration episodes are typically much shorter than cocaine use and non-use episodes, for this state we allow the duration effect to vary at the ends of each of the first 36 months and at the ends of years 4 and 5. For all states, age and calendar time are included with effects varying at the boundaries of whole years and homogeneous GMRF priors.

### 4.2.2 Inferences

We obtained inferences from our multistate model applied to the dataset described in Table 4.1 using the single-category random walk Metropolis (SC-RWM) algorithm from Sections 3.2.2 and 3.3.2, with all six random effects from each subject updated in the same block. We ran this MCMC routine for 220,000 iterations in total, discarding the first 20,000 for burn-in and storing only every tenth iteration following the burn-in period. Table 4.2 summarizes

Table 4.2: Summaries of effective sample sizes for all model parameters from 20,000 posterior samples obtained by thinning 200,000 post-burn-in MCMC iterations.

| Parameter Type | Number of Parameters | Effective Sample Size | | | | |
|---|---|---|---|---|---|---|
| | | Min. | $25^{\text{th}}$ | Median | $75^{\text{th}}$ | Max. |
| Categorical Fixed Effects | 78 | 451 | 1364 | 1846 | 3302 | 11850 |
| Continuous Fixed Effects | 806 | 190 | 450 | 720 | 1161 | 5223 |
| Random Effects | 2448 | 322 | 3433 | 5101 | 6753 | 12220 |
| GMRF Hyperparameters | 18 | 712 | 997 | 1412 | 1638 | 3242 |
| Random Effects Covariance | 21 | 193 | 449 | 617 | 687 | 924 |

effective sample sizes from the resulting 20,000 thinned, post-burn-in samples. Generally, effective sample sizes range from several hundred to several thousand, which should be sufficient for accurate inferences.

Posterior summaries of the effects of each of the seven fixed effects covariates on each of the six transition hazard rates are displayed in Figures 4.2–4.8. In each of these plots, the vertical axis represents the hazard for a particular covariate value relative to the average hazard over all values of that covariate. To illustrate, the last panel of Figure 4.2 depicts posterior summaries of the relationship between sex and the risk of incarceration for subjects not currently using cocaine. The posterior median for the male category is slightly over 1.4, which means that all else equal, males have a 40% higher hazard of incarceration than the average hazard of males and females. Alternatively, male subjects are about $1.4/0.7 = 2$ times more likely to become incarcerated than female subjects while not actively using cocaine.

We now highlight some findings from the inference plots. From Figure 4.2, we see that women voluntarily cease cocaine use at a higher rate than men, and also exit periods of incarceration into both cocaine-use and non-use states more quickly than men. Women are more likely to resume cocaine use during periods of non-use, but are incarcerated at much lower rates than men while not using cocaine. In Figure 4.3, we observe that Hispanic

cocaine users are at higher risk for incarceration than black or white cocaine users, while white cocaine users are more likely to quit using than blacks or Hispanics. Black subjects transition from incarceration to non-use at a 40% lower rate than Hispanics or whites, and transition back to incarceration at a roughly 30% lower rate than the other races. Black subjects also relapse into cocaine use at twice the rate of the other races.

Figure 4.4 shows the effect of the number of episodes spent in the current state on the hazard of transitions out of that state. In the first row of plots, we see that the chances of incarceration increase with the number of episodes of cocaine use, while the chances of voluntarily stopping use decrease with additional cocaine use episodes. Episodes of incarceration last longer for subjects having had more previous periods of incarceration. In Figure 4.5, we observe that for all transitions $s \to r$ other than non-use to cocaine-use, the hazard of the transition is higher if the subject's state just prior to entering $s$ was $r$.

Turning to the duration effects presented in Figure 4.6, we see that the hazard of incarceration among subjects actively using cocaine gradually drops 2-fold from the second through fifth years of continual use. The hazard of voluntary use cessation drops 4-fold following the first month of cocaine use, and then is reduced by an additional factor of 2 over the subsequent two year period. This chances of exiting incarceration into both active-use and non-use states increase after the first month of incarceration, and then do not change much thereafter. Finally, the hazards of relapse into cocaine use and incarceration fall with increasing durations spent in the non-use state. In particular, the relapse risk drops 4-fold from the second to $26^{\text{th}}$ month of non-use, while incarceration risk drops 2-fold during the same period.

In Figure 4.7, we observe no significant age effects on the transition hazards, with the exception of the incarceration to cocaine use transition. Subjects in their 40's are 3 times more likely to exit incarceration into a cocaine using state than subjects in their teenage years. The first and last panels of Figure 4.8 show that from the early 1980's to early 1990's, rates of incarceration quadrupled for active cocaine users and doubled for non-active users. The rate of voluntarily ceasing use also appears to have quadrupled duing this same decade-long period, and the rate of relapse into cocaine use for subjects not currently using doubled

Table 4.3: Posterior summaries of the subject random effects covariance matrix $\Sigma$. We give summaries of the random effect standard deviations for each transition between the cocaine use (C), incarceration (I), and non-use (N) states. We also give summaries of the random effects correlations for the five pairs of random effects with the largest magnitude inferred correlations.

| | | Posterior Quantiles | | |
|---|---|---|---|---|
| Parameter | Description | $2.5^{\text{th}}$ | Median | $97.5^{\text{th}}$ |
| $\sqrt{\Sigma_{11}}$ | C $\rightarrow$ I random effect std. dev. | 1.05 | 1.26 | 1.51 |
| $\sqrt{\Sigma_{22}}$ | C $\rightarrow$ N random effect std. dev. | 0.67 | 0.87 | 1.07 |
| $\sqrt{\Sigma_{33}}$ | I $\rightarrow$ C random effect std. dev. | 0.94 | 1.20 | 1.49 |
| $\sqrt{\Sigma_{44}}$ | I $\rightarrow$ N random effect std. dev. | 0.48 | 0.66 | 0.86 |
| $\sqrt{\Sigma_{55}}$ | N $\rightarrow$ C random effect std. dev. | 0.61 | 0.78 | 0.97 |
| $\sqrt{\Sigma_{66}}$ | N $\rightarrow$ I random effect std. dev. | 1.04 | 1.33 | 1.66 |
| $\Sigma_{16}/\sqrt{\Sigma_{11}\Sigma_{66}}$ | C $\rightarrow$ I and N $\rightarrow$ I effect correl. | 0.67 | 0.82 | 0.92 |
| $\Sigma_{24}/\sqrt{\Sigma_{22}\Sigma_{44}}$ | C $\rightarrow$ N and I $\rightarrow$ N effect correl. | -0.05 | 0.30 | 0.59 |
| $\Sigma_{35}/\sqrt{\Sigma_{33}\Sigma_{55}}$ | I $\rightarrow$ C and N $\rightarrow$ C effect correl. | -0.05 | 0.28 | 0.55 |
| $\Sigma_{45}/\sqrt{\Sigma_{44}\Sigma_{55}}$ | I $\rightarrow$ N and N $\rightarrow$ C effect correl. | -0.73 | -0.45 | -0.10 |
| $\Sigma_{56}/\sqrt{\Sigma_{55}\Sigma_{66}}$ | N $\rightarrow$ C and N $\rightarrow$ I effect correl. | -0.54 | -0.27 | 0.03 |

from the 1970's to 1990's. However, these last two results may partially reflect recall bias, with subject better able to recall events nearer to the times of interview in the mid 1990's.

Finally, posterior summaries of the covariance matrix of the subject random effects appear in Table 4.3. There is a very large amount of heterogeneity among study subjects with respect to their hazards of transitioning between states. For example, the posterior median estimate of the standard deviation of the cocaine to incarceration transition random effect is 1.26. This means a subject with a propensity for incarceration while using cocaine that is 2 standard deviations above average has a hazard of incarceration that is roughly $\exp(2 \cdot 1.26) \approx 12$ times higher than a typical subject with a random effect of 0. Several of the transition

random effects exhibit moderate or strong correlations with one another. For all three states, a subject's propensities to transition into that state from each of the two other states are positively correlated. The C $\rightarrow$ I and N $\rightarrow$ I random effects exhibited an especially strong correlation, with a posterior median estimate of 0.82, which says that subjects with a high arrest risk while using cocaine also tend to have a high arrest risk while not using, and vice versa. The negative correlation between the I $\rightarrow$ N and N $\rightarrow$ C effects also makes sense, as subjects with a higher likelihood of not resuming cocaine use immediately following incarceration have a lower risk of resuming cocaine use while not incarcerated.

Figure 4.2: Posterior medians (solid lines) and point-wise 95% credible intervals (dotted lines) of the effects of sex on the hazards of each transition type. The vertical axis is the transition hazard rate on a log scale relative to the average hazard over all covariate values (dashed line).

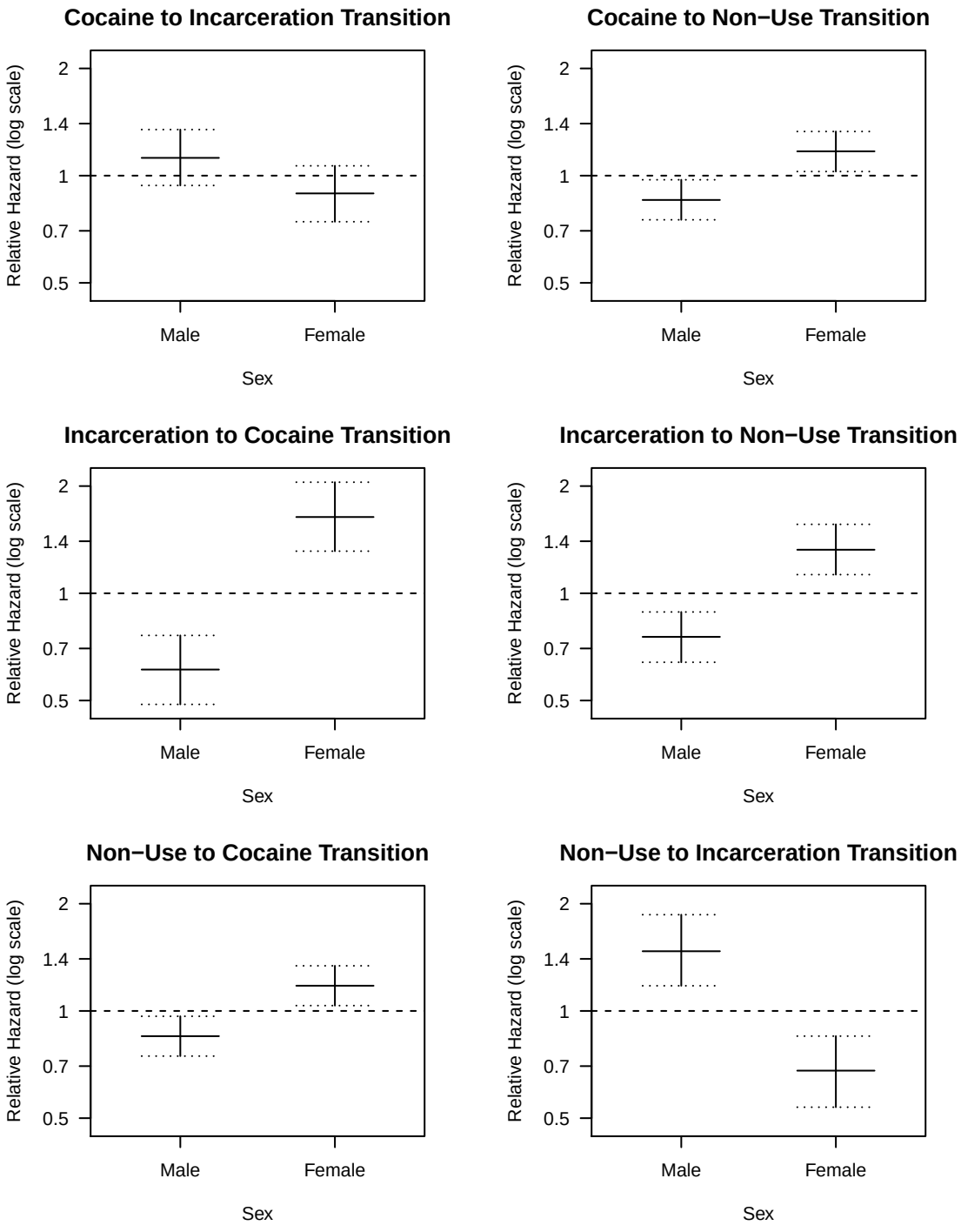


Figure 4.3: Posterior medians (solid lines) and point-wise 95% credible intervals (dotted lines) of the effects of race on the hazards of each transition type. The vertical axis is the transition hazard rate on a log scale relative to the average hazard over all covariate values (dashed line).

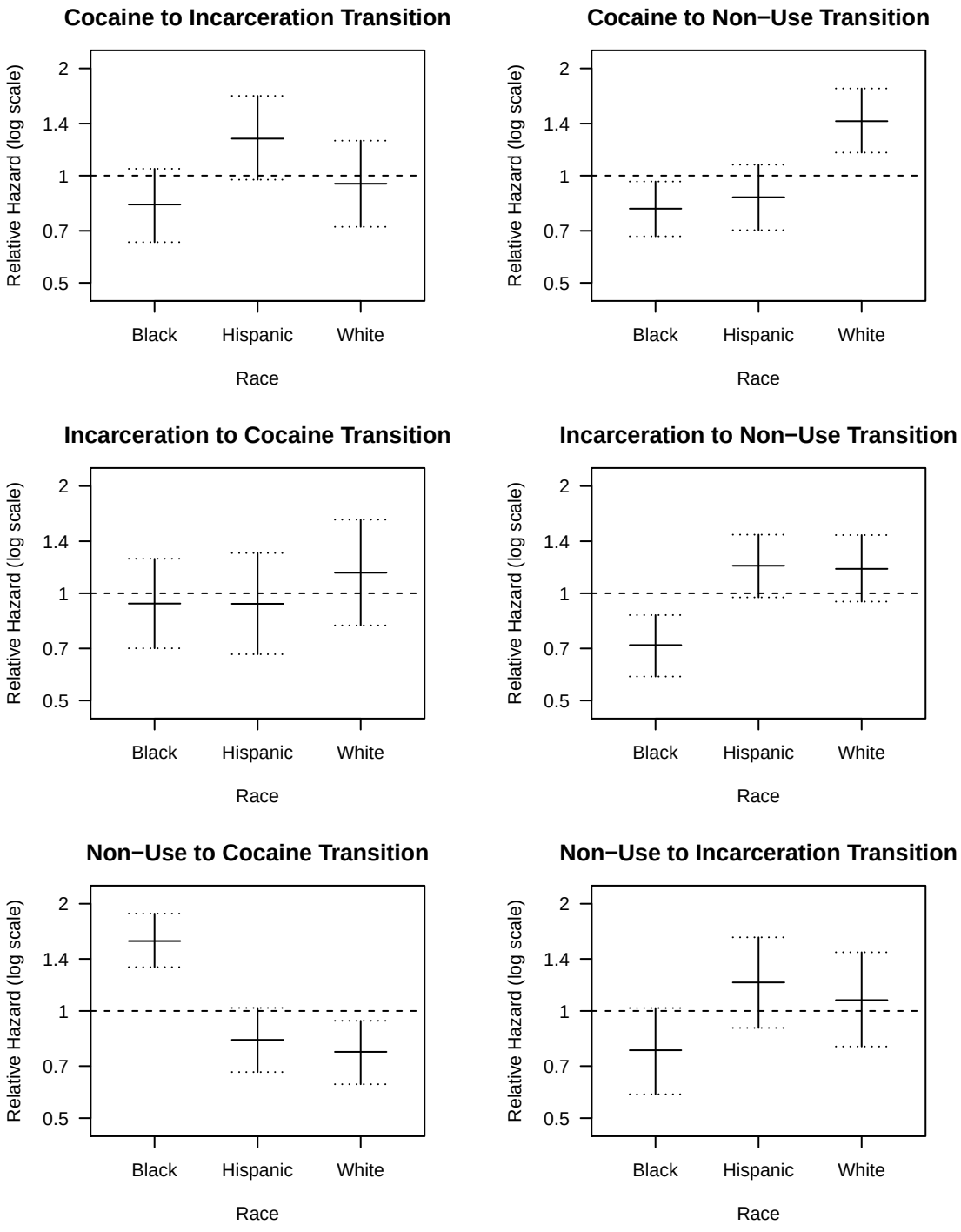


Figure 4.4: Posterior medians (solid lines) and point-wise 95% credible intervals (dotted lines) of the effects of the number of episodes spent in the current state on the hazards of each transition type. The vertical axis is the transition hazard rate on a log scale relative to the average hazard over all covariate values (dashed line).

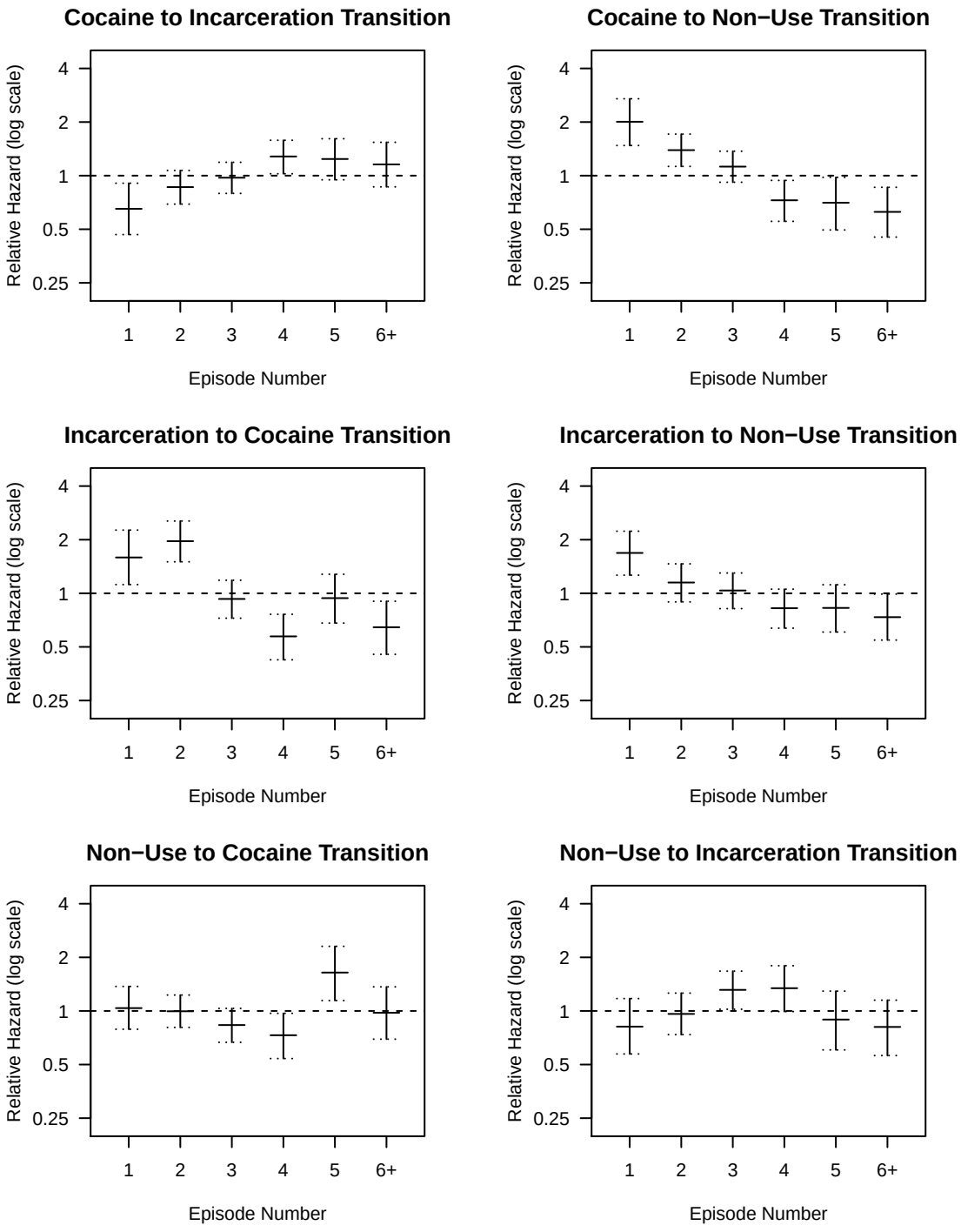


Figure 4.5: Posterior medians (solid lines) and point-wise 95% credible intervals (dotted lines) of the effects of the last state just prior to entering the current state on the hazards of each transition type. The vertical axis is the transition hazard rate on a log scale relative to the average hazard over all covariate values (dashed line).

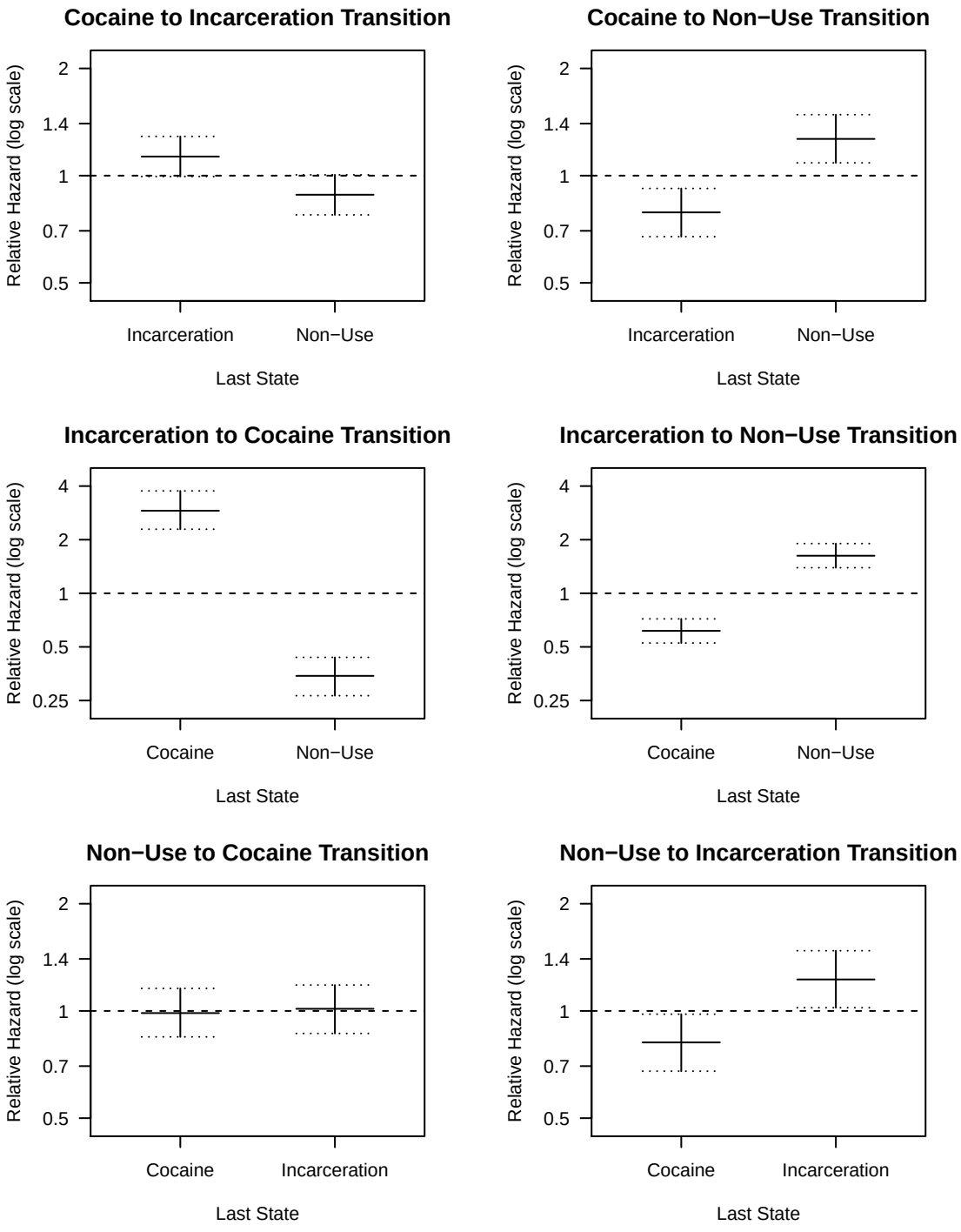


Figure 4.6: Posterior medians (solid lines) and point-wise 95% credible intervals (dotted lines) of the effects of the current duration of the current episode on the hazards of each transition type. The vertical axis is the transition hazard rate on a log scale relative to the average hazard over all covariate values (dashed line).

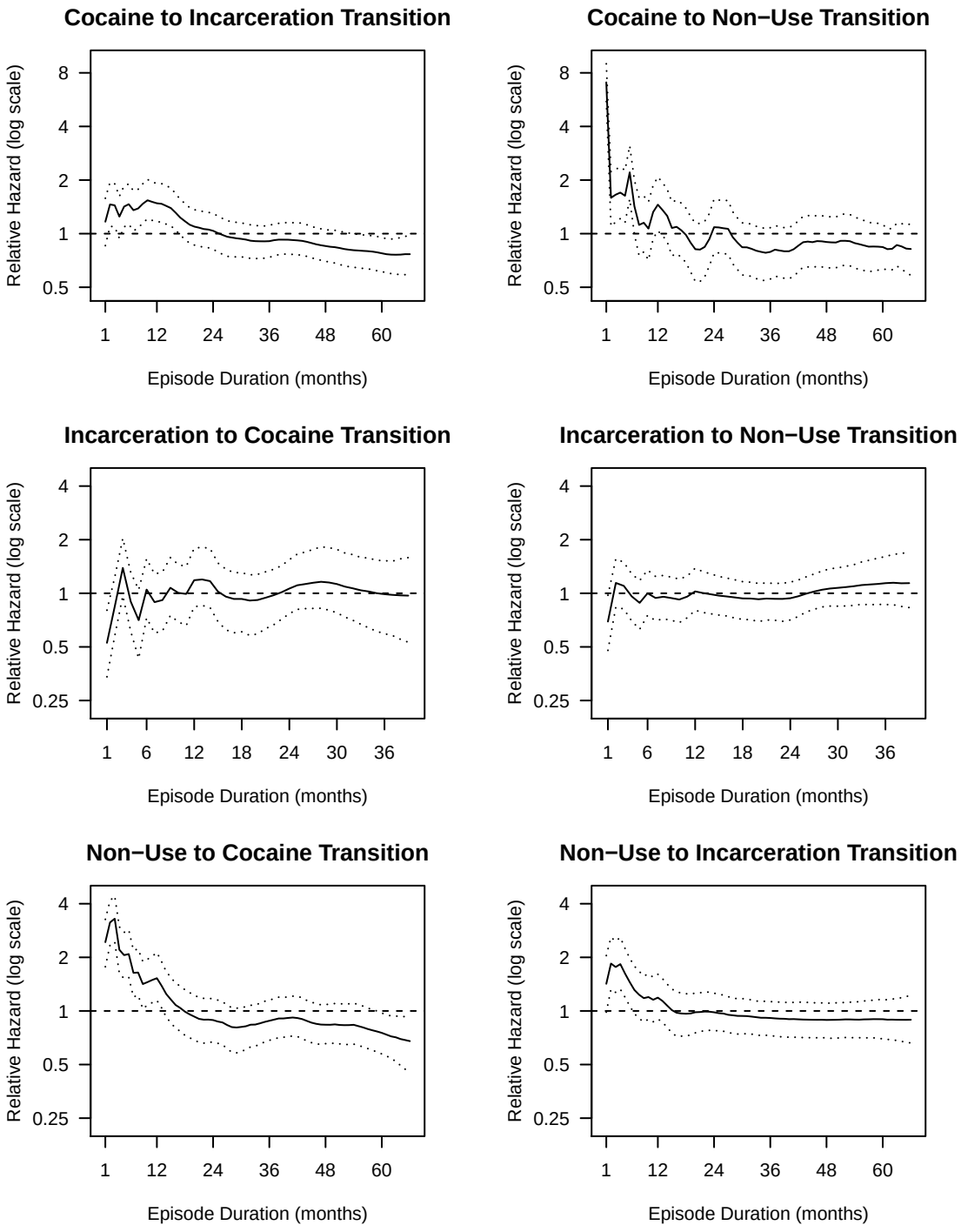


Figure 4.7: Posterior medians (solid lines) and point-wise 95% credible intervals (dotted lines) of the effects of current age on the hazards of each transition type. The vertical axis is the transition hazard rate on a log scale relative to the average hazard over all covariate values (dashed line).

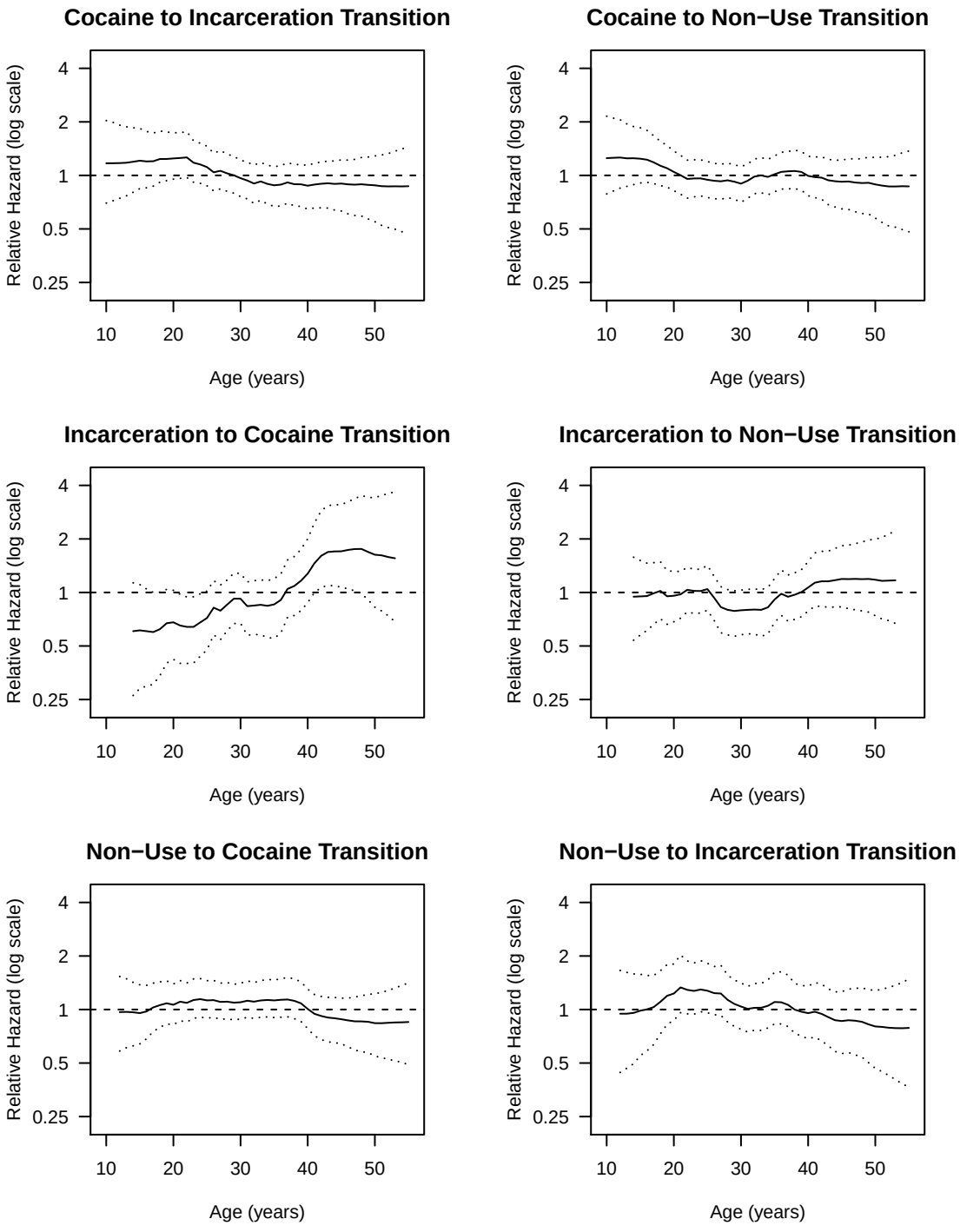
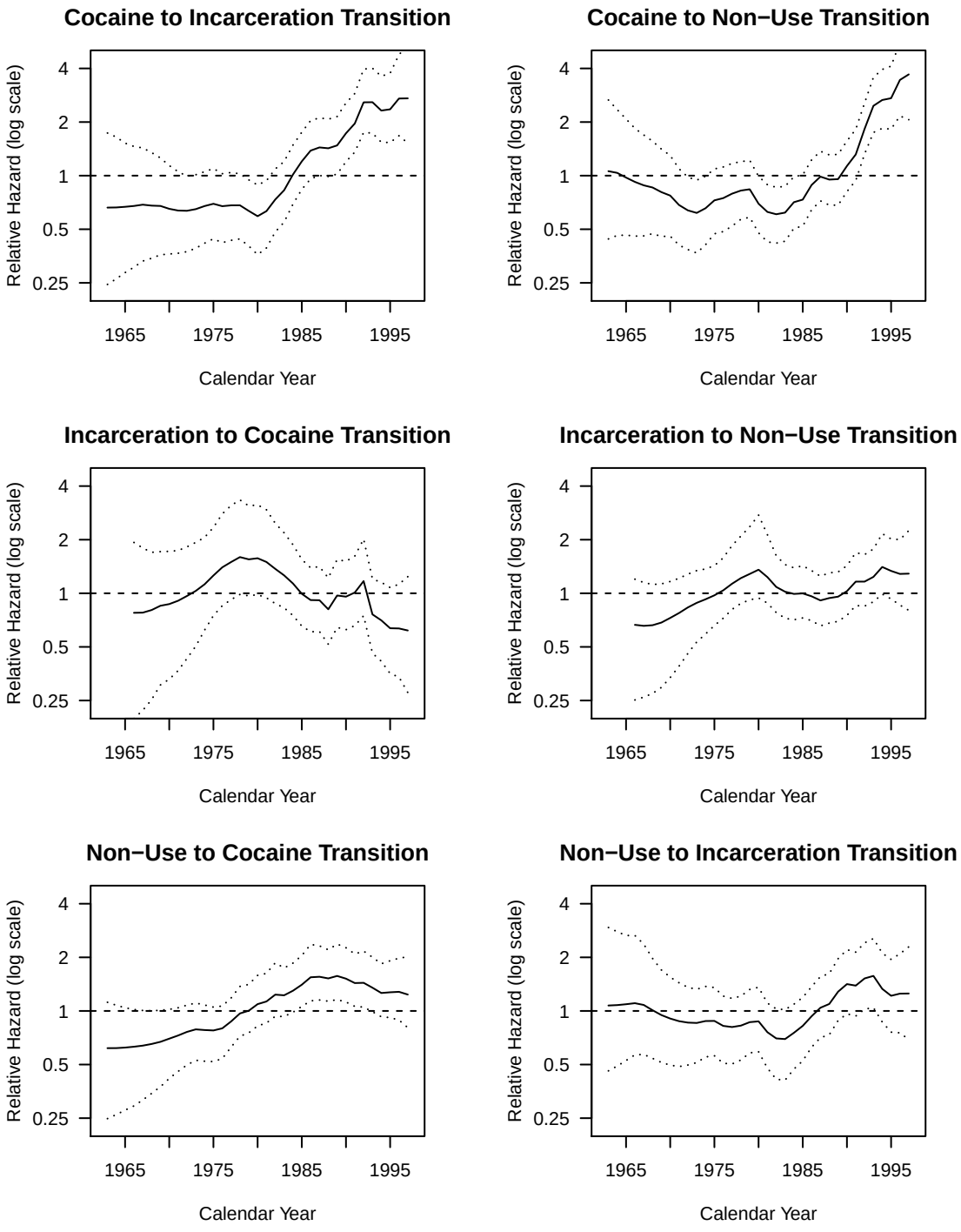


Figure 4.8: Posterior medians (solid lines) and point-wise 95% credible intervals (dotted lines) of the effects of current calendar time on the hazards of each transition type. The vertical axis is the transition hazard rate on a log scale relative to the average hazard over all covariate values (dashed line).



# CHAPTER 5

# Ongoing and Future Work

In this chapter, we discuss work in progress as well as plans for future work. First, we describe the R-language routines used to fit the models from Chapters 3 and 4, and discuss features to be added to this code prior to its publication as an R package. Next, we describe future methodological extensions of our event history models.

## 5.1 R Software Package

### 5.1.1 Current Functionality

We have implemented the Bayesian competing risks event history model of Section 3.1 and the Bayesian multistate model of Section 4.1 in the full generality allowed by the notation in these sections. These implementations consist of R-language routines for performing the MCMC algorithms of Section 3.2. More specifically, the code allows specifications with:

1. Arbitrary numbers of states, with the competing risks model corresponding to the case of a single state
2. For each state, arbitrary numbers of competing risks or transition types
3. For each state, arbitrary numbers and types of covariates (categorical or continuous)
4. For each covariate of each state, separate prior hyperparameters for each event type
5. Arbitrary numbers of random effects, which are correlated across all event types
6. For each covariate of each state, choice of MCMC algorithm (random walk Metropolis or full-conditional-approximating Metropolis-Hastings, with user-selected blocking)

Although Section 3.3.2 shows the FCA algorithm to be much more efficient than RWM, we provide the option of using RWM because the FCA algorithm can fail at burn-in, due to the log likelihood approximations it uses being poor when the parameter values are very far from regions of high posterior mass.

The generality of our implementation is achieved via extensive use of R list objects and multidimensional arrays. Compared to an R implementation of a competing risks model with fixed, hard-coded predictor and outcome variables, our general competing risks code runs approximately 10% slower on the same model using the same inference algorithms. In addition, the general multistate model R code applied to a single-state competing risks model runs about 15% slower than the general competing risks R code; this is due to the necessity of placing almost all variables from the competing risks model inside another level of R list objects to enable dependence upon state. In aggregate, there is an approximate 25% slowdown for our fully-general multistate model code compared to an ad-hoc model implementation using separate variable names for each combination of risk, predictor, and state, as well separately named auxiliary variables and MCMC parameter storage variables. We consider this slowdown to be a very modest cost compared to the cost of having to re-implement the inference algorithms for each new dataset we wish to analyze.

### 5.1.2 Additional Covariate Options

Interactions between predictor variables is a common and useful regression model feature. With our current software implemenation, it is possible to include interactions between categorical covariates by replacing these variables with a single categorical variable taking on the values of all possible pairs of values of the two predictors. In addition, it is currently possible to include an interaction between a categorical and continuous variable by creating one copy of the continuous covariate for each value of the categorical covariate; each continuous covariate copy is then set to be constant for observations with values of the categorical covariate other than the value that corresponds to that continuous covariate copy. However, this interaction coding is somewhat cumbersome, and a more direct approach of having a

separate arbitrary function representing the continuous covariate effect for each categorical value would be a valuable future addition to our software. General interactions between two or more continuous covariates are not possible using the current software. However, we could model arbitrary smooth functions of two continuous covariates using the GMRF approach described in Section 2.7.4 and extend our MCMC algorithms to this case using the sparse matrix algorithms of Rue and Held (2005). We could use such an approach to include general spatial and spatio-temporal covariates as well.

Certain predictor variables may plausibly only affect the hazard of a subset of the competing risks or transition types for a given state. For example, the chances of a non-using subject relapsing into cocaine use may depend on the current street price of cocaine, but a non-using subject's incarceration risk may not be related to cocaine prices. With our current model, any predictor variable included for one competing risk for a given state must be included for all competing risks for that state. A simple modification of our MCMC routines could fix the parameters representing the covariate effects on a subset of the cause-specific hazards to zero. This would not leave these cause-specific hazards completely unrelated to the covariate values though, as with the multinomial logit link function, a change in any of the linear predictors adjusts all the event probabilites. However, with relatively rare events, changes to these cause-specific hazards would be small. Finally, we could extend our model to allow parameters to be shared across competing risks or transition types. For example, a covariate such as law enforcement presence in a neighborhood may be presumed to have the same effect on the incarceration hazard of both cocaine-using and non-using subjects.

### 5.1.3 Alternate Competing Risks Model

With the multistate model of Chapter 4, we modeled the *transition hazard rates* (4.1), which are cause-specific hazards for the competing causes of episode termination. Such a modeling approach is sensible when the different types of transitions represent *different causes or reasons* for the current episode to terminate. For example, the two possible transitions out of the cocaine use state in the model in Section 4.2 are moves to the non-use and incarceration

states, each of which represents a distinct cause of the subject ceasing cocaine use; in the former case the subject chooses to quite, while in the later case the subject is forced to quit.

However, for other states, the different states a subject may pass into do not represent different *causes* of the subject leaving the current state. A subject leaving the incarceration state may pass into either the cocaine-use or non-use states, but cocaine-use and non-use are not competing causes for leaving prison. Instead, subjects are allowed to leave prison, and then decide whether to resume cocaine use. This issue made interpretation of the transition hazard rates out of the cocaine use state in Section 4.2.2 unclear. For such cases, we could use an alternative two-stage competing risks model, where we first model the aggregate hazard of leaving the current state, and then conditional on the time of leaving the current state, we model the new state the subject passes into with a multinomial logistic model. Because our current software implementation is capable of handling both stages of this model, this extension will only require recoding of the person-time dataset generation.

### 5.1.4 Interface Development

The current R implementation requires the user to specify, in the notation of Section 3.2.1, the covariate values $X_{m,n}$ and competing risks event indicators $y_n^{(r)}$ for each person-time observation $n = 1, \ldots, N$. However, this *person-time level* specification is not necessary if there are no covariates that vary within episodes other than time scales which may be computed for each person-time observation based upon the starting and ending times of each episode. For example, if we know the calendar time of each subject's birth, then we can compute their current ages at each time point of each episode so long as we know the calendar time each episode began and ended. In such cases, we could allow the user to specify the data at the *episode level*, with one record for each episode. Expansion of this dataset to person-time format could then be automated and hidden from the user.

Another user friendly feature, particularly for non-statisticians, would be the inclusion of default priors. Often in Bayesian statistics, it is not possible to specify a single prior which is sensible for a wide class of applications. However, in event history problems, linear predictor

parameters are approximate log hazard ratios; rarely are these hazard ratios greater than 50 or less than 1/50, so linear predictor parameters are usually between -4 and 4. If for example we have a time scale that has been discretized into 40 values, and across the entire scale we do not expect the function representing the effect to change by more than 4, then a sensible value for the first-order random walk GMRF prior increment standard deviation might be 4/40=0.1, giving an increment variance of 0.01. A scaled inverse chi-squared prior for this variance could then be supplied with scale hyperparameter 0.01 and a small degrees of freedom hyperparameter.

## 5.2 Model Extensions

### 5.2.1 Inferred Predictor Discretization and Clustering

While many continuous covariates have uniformly smooth relationships with the hazards of event occurrence, others may have irregular relationships. In Section 3.3.3, we observed a dramatic decrease in the hazard of voluntary cocaine use cessation following the first month of cocaine use, with much more gradual changes in the hazard in subsequent months. We addressed this irregularity by allowing the increment variance of our random walk GMRF smoothing prior to vary, creating a nonhomogeneous random walk. An alternative approach to our nonhomogenous formulation would be to allow the change points in the time scale discretization to be random, as in Haneuse et al. (2008). In this case, a change point for the duration effect on voluntary cessation would then be inferred to be between the first and second months of drug use, with change points inferred to be more dispersed in later months. In addition to allowing random discretizations of the continuous covariates, we could also extend our model by allowing random clustering for the categorical fixed or random effects. For example, clustering of similar subjects could replace the multivariate normal subject-specific random effects.

### 5.2.2 New Inference Targets via Simulation

Our event history and multistate models allow us to make inferences about the hazards of subjects transitioning between states. However, transition hazard rates are not necessarily the most relevant target of inference. Instead, we may for example wish to predict the marginal probability that a subject will be using cocaine at a certain future time point, or the probability that the subject will remain indefinitely abstinent, given the history of the subject up to the present. We can make inferences about such probabilities and other quantities of interest by simulating posterior predictive histories of subjects given a hypothetical partial history of the subject and our observed data.

# Bibliography


Aalen, O. O., Borgan, O., and Gjessing, H. K. (2008). *Survival and Event History Analysis: A Process Point of View*. Springer.

Abbott, R. D. (1985). Logistic regression in survival analysis. *American Journal of Epidemiology*, 121(3):465–471.

Agresti, A. (2002). *Categorical Data Analysis*. Wiley, 2nd edition.

Allison, P. D. (1982). Discrete-time methods for the analysis of event histories. *Sociological Methodology*, 13(1):61–98.

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

Barnett, A. G., Batra, R., Graves, N., Edgeworth, J., Robotham, J., and Cooper, B. (2009). Using a longitudinal model to estimate the effect of methicillin-resistant Staphylococcus aureus infection on length of stay in an intensive care unit. *American Journal of Epidemiology*, 170(9):1186–1194.

Berzuini, C. and Clayton, D. (1994). Bayesian analysis of survival on multiple time scales. *Statistics in Medicine*, 13(8):823–838.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36(2):192–236.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science*, 10(1):3–41.

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika*, 82(4):733–746.

Besag, J., York, J., and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43(1):1–59.

Browne, W. J., Steele, F., Golalizadeh, M., and Green, M. J. (2009). The use of simple reparameterizations to improve the efficiency of Markov chain Monte Carlo estimation for multilevel models with applications to discrete time survival models. *Journal of the Royal Statistical Society, Series A*, 172(3):579–598.

Burridge, J. (1981). Empirical Bayes analysis of survival time data. *Journal of the Royal Statistical Society, Series B*, 43(1):65–75.

Carter, C. K. and Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika*, 81(3):541–553.

Cook, R. J. and Lawless, J. F. (2007). *The Statistical Analysis of Recurrent Events*. Springer.

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

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

Cupples, L. A., D'Agostino, R. B., Anderson, K., and Kannel, W. B. (1988). Comparison of baseline and repeated measure covariate techniques in the Framingham Heart Study. *Statistics in Medicine*, 7(1–2):205–218.

D'Agostino, R. B., Lee, M.-L., Belanger, A. J., Cupples, L. A., Anderson, K., and Kannel, W. B. (1990). Relation of pooled logistic regression to time dependent Cox regression analysis: The Framingham Heart Study. *Statistics in Medicine*, 9(12):1501–1515.

Efron, B. (1988). Logistic regression, survival analysis, and the Kaplan-Meier curve. *Journal of the American Statistical Association*, 83(402):414–425.

Fahrmeir, L. (1994). Dynamic modelling and penalized likelihood estimation for discrete time survival data. *Biometrika*, 81(2):317–330.

Fahrmeir, L. and Knorr-Held, L. (1997). Dynamic discrete-time duration models: Estimation via Markov chain Monte Carlo. *Sociological Methodology*, 27(1):417–452.

Fahrmeir, L. and Lang, S. (2001). Bayesian inference for generalized additive mixed models based on Markov random field priors. *Journal of the Royal Statistical Society, Series C*, 50(2):201–220.

Fahrmeir, L. and Wagenpfeil, S. (1996). Smoothing hazard functions and time-varying effects in discrete duration and competing risks models. *Journal of the American Statistical Association*, 91(436):1584–1594.

Ferguson, T. S. and Phadia, E. G. (1979). Bayesian nonparametric estimation based on censored data. *The Annals of Statistics*, 7(1):163–186.

Fienberg, S. E. and Mason, W. M. (1979). Identification and estimation of age-period-cohort models in the analysis of discrete archival data. *Sociological Methodology*, 10:1–67.

Gamerman, D. (1991). Dynamic Bayesian models for survival data. *Journal of the Royal Statistical Society, Series C*, 40(1):63–79.

Gamerman, D. (1998). Markov chain Monte Carlo for dynamic generalized linear models. *Biometrika*, 85(1):215–227.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995). Efficient parametrisations for normal linear mixed models. *Biometrika*, 82(3):479–488.

Gelman, A., Roberts, G. O., and Gilks, W. R. (1996). Efficient Metropolis jumping rules. In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., editors, *Bayesian Statistics 5*, pages 599–607.

Gibbons, R. D., Duan, N., Meltzer, D., Pope, A., Penhoet, E. D., Dubler, N. N., Francis, C., Gill, B., Guinan, E., Henderson, M., Ildstad, S. T., King, P. A., Martinez-Maldonado, M., Mclain, G. E., Murray, J., Nelkin, D., Spellman, M. W., and Pitluck, S. (2003). Waiting for organ transplantation: results of an analysis by an Institute of Medicine committee. *Biostatistics*, 4(2):207–222.

Guo, S. W. and Lin, D. Y. (1994). Regression analysis of multivariate grouped survival data. *Biometrics*, 50(3):632–639.

Haneuse, S. J.-P. A., Rudser, K. D., and Gillen, D. L. (2008). The separation of timescales in Bayesian survival modeling of the time-varying effect of a time-dependent exposure. *Biostatistics*, 9(3):400–410.

Harrison, P. J. and Stevens, C. F. (1976). Bayesian forecasting. *Journal of the Royal Statistical Society, Series B*, 38(3):205–247.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B*, 55(4):757–796.

Hedeker, D., Siddiqui, O., and Hu, F. B. (2000). Random-effects regression analysis of correlated grouped-time survival data. *Statistical Methods in Medical Research*, 9(2):161–179.

Hougaard, P. (1995). Frailty models for survival data. *Lifetime Data Analysis*, 1(3):255–273.

Hougaard, P. (1999). Multi-state models: A review. *Lifetime Data Analysis*, 5(3):239–264.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer.

Hser, Y.-I., Boyle, K., and Anglin, M. D. (1998). Drug use and correlates among sexually transmitted disease patients, emergency room patients, and arrestees. *Journal of Drug Issues*, 28(2):437–454.

Hser, Y.-I., Evans, E., Huang, D., Brecht, M.-L., and Li, L. (2008). Comparing the dynamic course of heroin, cocaine, and methamphetamine use over 10 years. *Addictive Behaviors*, 33(12):1581–1589.

Hser, Y.-I., Maglione, M., and Boyle, K. (1999). Validity of self-report of drug use among STD patients, ER patients, and arrestees. *American Journal of Drug and Alcohol Abuse*, 25(1):81–91.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001). *Bayesian Survival Analysis*. Springer.

Imbens, G. W. (1994). Transition models in a non-stationary environment. *The Review of Economics and Statistics*, 76(4):703–720.

Ingram, D. D. and Kleinman, J. C. (1989). Empirical comparisons of proportional hazards and logistic regression models. *Statistics in Medicine*, 8(5):525–538.

Kalbfleisch, J. D. (1978). Non-parametric Bayesian analysis of survival time data. *Journal of the Royal Statistical Society, Series B*, 40(2):214–221.

Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika*, 60(2):267–278.

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

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Keiding, N. (1990). Statistical inference in the Lexis diagram. *Philosophical Transactions of the Royal Society A*, 332(1627):487–509.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, 2nd edition.

Knorr-Held, L. (1999). Conditional prior proposals in dynamic models. *Scandinavian Journal of Statistics*, 26(1):129–144.

Knorr-Held, L. and Rue, H. (2002). On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics*, 29(4):597–614.

Kom, E. L., Graubard, B. I., and Midthune, D. (1997). Time-to-event analysis of longitudinal follow-up of a survey: Choice of the time-scale. *American Journal of Epidemiology*, 145(1):72–80.

Laara, E. and Matthews, J. N. S. (1985). The equivalence of two models for ordinal data. *Biometrika*, 72(1):206–207.

Li, L., Evans, E., and Hser, Y.-I. (2010). A marginal structural modeling approach to assess the cumulative effect of drug treatment on later drug use abstinence. *Journal of Drug Issues*, 40(1):221–240.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Liang, L.-J., Huang, D., Brecht, M.-L., and Hser, Y.-I. (2010). Differences in mortality among heroin, cocaine, and methamphetamine users: A hierarchical Bayesian approach. *Journal of Drug Issues*, 40(1):121–140.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B*, 42(2):109–142.

Murphy, D. A., Hser, Y.-I., Huang, D., Brecht, M.-L., and Herbeck, D. M. (2010). Self-report of longitudinal substance use: A comparison of the UCLA natural history interview and the addiction severity index. *Journal of Drug Issues*, 40(2):495–515.

Oakes, D. (1995). Multiple time scales in survival analysis. *Lifetime Data Analysis*, 1(1):7–18.

Plummer, M., Best, N., Cowles, K., Vines, K., Sarkar, D., and Almond, R. (2012). *coda: Output analysis and diagnostics for MCMC*. Version 0.16-1, URL http://cran.r-project.org/web/packages/coda/.

Prendergast, M., Huang, D., and Hser, Y.-I. (2008). Patterns of crime and drug use trajectories in relation to treatment initiation and 5-year outcomes: An application of growth mixture modeling across three data sets. *Evaluation Review*, 32(1):59–82.

Prentice, R. L. and Gloeckler, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics*, 34(1):57–67.

Prentice, R. L., Kalbfleisch, J. D., Peterson, A. V., Flournoy, N., Farewell, V. T., and Breslow, N. E. (1978). The analysis of failure times in the presence of competing risks. *Biometrics*, 34(4):541–554.

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

Robins, J. M., Hernan, M. A., and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560.

Rue, H. (2001). Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society, Series B*, 63(2):325–338.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman & Hall/CRC.

Shephard, N. and Pitt, M. K. (1997). Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, 84(3):653–667.

Singer, J. D. and Willett, J. B. (1993). It's about time: Using discrete-time survival analysis to study duration and the timing of events. *Journal of Educational Statistics*, 18(2):155–195.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B*, 64(4):583–639.

Steele, F. (2011). Multilevel discrete-time event history models with applications to the analysis of recurrent employment transitions. *Australian and New Zealand Journal of Statistics*, 53(1):1–26.

Steele, F., Goldstein, H., and Browne, W. (2004). A general multilevel multistate competing risks model for event history data, with an application to a study of contraceptive use dynamics. *Statistical Modelling*, 4(2):145–159.

Susarla, V. and Van Ryzin, J. (1976). Nonparametric Bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association*, 71(356):897–902.

Ten Have, T. R. and Uttal, D. H. (1994). Subject-specific and population-averaged continuation ratio logit models for multiple discrete time survival profiles. *Applied Statistics*, 43(2):371–384.

Thiebaut, A. C. M. and Benichou, J. (2004). Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. *Statistics in Medicine*, 23(24):3803–3820.

Thompson, W. A. (1977). On the treatment of grouped observations in life studies. *Biometrics*, 33(3):463–470.

Tuma, N. B. and Hannan, M. T. (1979). Approaches to the censoring problem in analysis of event histories. *Sociological Methodology*, 10:209–240.

Tutz, G. (1991). Sequential models in categorical regression. *Computational Statistics & Data Analysis*, 11(3):275–295.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, 16(3):439–454.

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

West, M., Harrison, P. J., and Migon, H. S. (1985). Dynamic generalized linear models and Bayesian forecasting. *Journal of the American Statistical Association*, 80(389):73–83.

Willett, J. B. and Singer, J. D. (1993). Investigating onset, cessation, relapse, and recovery: Why you should, and how you can, use discrete-time survival analysis to examine event occurrence. *Journal of Consulting and Clinical Psychology*, 61(6):952–965.

Willett, J. B. and Singer, J. D. (1995). It's deja vu all over again: Using multiple-spell discrete-time survival analysis. *Journal of Educational and Behavioral Statistics*, 20(1):41–67.

Yamaguchi, K. (1990). Logit and multinomial logit models for discrete-time event-history analysis: a causal analysis of interdependent discrete state processes. *Quality and Quantity*, 24(3):323–341.

Zeger, S. L. and Karim, M. R. (1991). Generalized linear models with random effects; a gibbs sampling approach. *Journal of the American Statistical Association*, 86(413):79–86.

Zeger, S. L. and Liang, K.-Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, 42(1):121–130.