Modeling a Progressive Disease Process Under Panel Observation

Amy Laird

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2013

Reading Committee:

Lurdes Inoue, Chair

Rebecca Hubbard

James Hughes

Volodymyr Minin

Program Authorized to Offer Degree:
Public Health – Biostatistics

University of Washington

Abstract

Modeling a Progressive Disease Process
Under Panel Observation

Amy Laird

Chair of the Supervisory Committee:
Associate Professor Lurdes Inoue
Biostatistics

Longitudinal studies are a useful tool for investigating the course of chronic diseases. Many chronic diseases are progressive and can be characterized by a set of health states. We can improve our understanding of the natural history of the disease by modeling the sequence of visited health states and the duration in each state. However, in most applications, subjects are observed intermittently. This observation scheme creates a major modeling challenge: the transition times are not known exactly, and in some cases the path through the health states is not known. Existing methods for modeling this type of data either impose strong parametric assumptions on the sojourn times in each state, or model time discretely and carry out inference nonparametrically, but both approaches have drawbacks.

We propose an alternative modeling approach that uses the principle of data augmentation. This method has several advantages: (1) it accommodates any parametric model for the sojourn times, including spline models; (2) it performs well under moderate sample sizes for suitable parametric choices for the sojourn time distributions; and (3) it does not require that subjects be observed in every health state. Using this approach it is possible to carry out inference about both the probability of taking a given path through the health states and the duration in each state. We evaluate the performance of our proposed approach via simulation study.

We extend our basic approach to accommodate the presence of left-censored entry into

the process. Further, we extend the approach to account for between-subject variability in the rate of progression through the process.

We apply a basic version of our proposed approach first to a study of HIV infection and progression to AIDS in a cohort of patients with hemophilia who were infected via contaminated blood transfusions. Our findings reflect those of the original study, and illustrate the need for flexible modeling of the duration in each health state. We also apply our proposed approach to a more detailed study of HIV/AIDS staging among untreated patients in Senegal who were infected with different strains of HIV. Our results indicate that patients with HIV-2 tend to progress more slowly than those infected with HIV-1 or both viruses, corroborating existing knowledge of the natural history of the disease process in each case.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The author wishes to express her sincere gratitude to her committee: Lurdes Inoue, Rebecca Hubbard, Jim Hughes, Barbra Richardson, Vladimir Minin, and Michael Emerman. In particular, I would like to thank Rebecca Hubbard for inspiring me to pursue this topic, and for her valuable feedback on my dissertation, which greatly improved it. Above all, I would like to thank my advisor, Lurdes Inoue, for her dedication and tireless efforts in helping me to develop as a researcher, and for the latitude she gave me to pursue a topic that I found intellectually interesting. I would also like to thank the Department of Biostatistics for their support throughout my graduate career.

DEDICATION

To my loving family, who never stopped believing in me.

Chapter 1

INTRODUCTION

Chronic diseases, such as cardiovascular disease, cancer, diabetes and HIV/AIDS, are the leading cause of death in the United States and worldwide. These diseases not only have an enormous economic impact through lost productivity and medical care spending, but they are also a major cause of disability and human suffering. It is important, therefore, to understand the natural history and etiology of these diseases. Further, since many chronic diseases are caused or made worse by modifiable factors such as diet and lifestyle, understanding factors affecting disease progression is critical.

For a number of chronic diseases, the evolution is characterized by visits to clinically relevant and ordered stages. Examination of the sequence of visited stages and duration in each stage can enhance our understanding of the natural progression of the disease and how demographic and clinical factors may have an impact on disease evolution. In this dissertation we develop statistical methodology to model progressive chronic disease.

Longitudinal follow-up of patients allows us to characterize the rate of disease progression. However, in general, we cannot assess patients' status continuously. Instead, in most applications, patients are assessed periodically, and disease status between two assessment points is unknown. This discrete observation scheme, known as panel observation, introduces challenges when analyzing longitudinal panel data.

The standard approach for multi-state progressive chronic disease under panel observation is to use a Markov model. While a Markov model permits statistical inference under panel observation, it does so at the cost of imposing a strong (Markov) assumption that may not be always reasonable. Our primary goal in this dissertation is to develop a model for a progressive multi-state process under a panel observation scheme that does not impose the Markov assumption. In particular, in a progressive multi-state model with panel observations, patients may either visit each of the disease stages or skip intermediate stages.

The framework we will develop allows for inference about the underlying disease trajectory.

The second goal of our work is to extend the methodology developed under our primary goal and accommodate features commonly encountered in applications. Specifically, there are several ways in which a patient's disease status information over time may be incomplete. The first, which we discussed previously, is that we observe patients at snapshots in time rather than continuously. In addition, we are generally not able to observe each patient from the time he enters the first disease stage. Thus, the time since a patient entered the first disease stage to the time of first observation is unknown (left censoring). Also, the delay may prevent some patients from being observed at all (left truncation). Finally, we may also stop observing each patient before he enters the final disease stage (right censoring). We develop methodology to accommodate the existence of left and right censoring, and we discuss ways in which our method could be modified to address left truncation of subjects.

Finally, a third goal of our work is to allow for between-subject variability in the rate of disease progression. This will allow us to investigate how demographic or clinical characteristics of patients may affect disease progression.

This dissertation is organized as follows. In Chapter 2 we provide the theoretical background for Markov and semi-Markov processes. We review the existing approaches for modeling multi-state data. Lastly, we introduce our motivating examples for this work.

In Chapter 3 we propose an approach for a progressive multi-state process under panel observation in which we assume that subjects visit each of the states in a sequence. We demonstrate the performance of our basic approach in a range of scenarios and for various levels of modeling flexibility, and compare the performance to that of existing methods. We apply our approach to our first motivating application.

Chapter 4 provides an extension of the basic method that allows for the skipping of intermediate disease stages. We illustrate this extension first in the special case of the illness-death model and then discuss the estimation approach in the general case that includes any number of disease stages. We illustrate the performance of the approach when the model includes three and four states.

Up to this point in our development, we have made the assumption that each patient is observed from the moment he enters the first disease stage. In Chapter 5 we develop

methodology to allow for left censoring. Performance of our proposed approach is assessed via simulation study.

Chapter 6 provides a framework that allows for between-subject variability in rate of disease progression. Differences in the rate of disease progression may be accounted for through either the probability of taking a given path through the states, or through the lengths of time spent, or sojourn times, in each state. We focus our development to address the latter and illustrate the performance of the approach via simulation study.

In Chapter 7 we apply our proposed approach to the main motivating example. Finally, in Chapter 8, we present concluding remarks and outline directions for future research.

Chapter 2

BACKGROUND

2.1 Introduction

The current standard approach to longitudinal multi-state data under panel observation is to use the Markov model (Meira-Machado et al., 2009; Foucher et al., 2007). Although it is applicable to a wide variety of situations, this model imposes heavy assumptions on the nature of the process. A related model, the semi-Markov model, is more flexible, but estimation is less tractable. In this chapter we examine and compare Markov and semi-Markov processes and review existing methods of estimation.

2.2 Markov and semi-Markov processes: background and notation

We are interested in modeling the disease process of interest as a semi-Markov process. In this section we define Markov processes, renewal processes, and semi-Markov processes, and examine their properties.

2.2.1 Markov processes.

We consider a stochastic process $\{Z(t), t \in [0, \infty)\}$ that takes on a finite set of states $\mathcal{S} = \{1, 2, \ldots, m\}$. In our applications, $Z(t)$ represents the disease state of a patient at time t, though we make no use of the numerical ordering of the states and treat them merely as labels. This process is *Markov* if for all $s, t \geq 0$ and for every $i, j \in \mathcal{S}$

$$P(Z(t+s) = j | Z(t) = i, Z(u) = z(u), 0 \leq u < s) = P(Z(t+s) = j | Z(t) = i) \doteq p_{ij}(t, t+s).$$

That is, we have a Markov process if, at each time point, the state of the process at a future time depends on the entire history only through the present state. We define the initial distribution of the process, $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)$, by $\phi_i \doteq P(Z(0) = i)$ for each state $i \in \mathcal{S}$, with

$\phi_i \geq 0$ for each i and $\sum_{i \in \mathcal{S}} \phi_i = 1$. Conditional on $\boldsymbol{\phi}$, the process can be characterized by the matrix of transition probabilities $\mathbf{P}(t, t+s) = [p_{ij}(t, t+s)]$ for $s, t \geq 0$. Alternatively, the process can be characterized by the matrix $\mathbf{Q}(t) = [q_{ij}(t)]$ of transition intensities, defined for $t \geq 0$ as

$$q_{ij}(t) \doteq \lim_{s \to 0} \frac{p_{ij}(t, t+s)}{s} \quad \text{for } i, j \in \mathcal{S}, j \neq i,$$

and

$$q_{ii}(t) \doteq -\sum_{j \neq i} q_{ij}(t) \quad \text{for each } i \in \mathcal{S}.$$

The $q_{ij}(\cdot)$ are also known as *cause-specific hazard functions* (Prentice et al., 1978). It follows from the definition of the process that $p_{ij}(t, t) = \delta_{ij}$ for all $i, j \in \mathcal{S}$ and all t, where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise (Kronecker's delta). In matrix notation, this becomes $\mathbf{P}(t, t) = \mathbf{I}$.

If the transition probabilities depend only on the elapsed time s and not on the chronological time t, then the Markov process is *homogeneous* and we write $p_{ij}(t, t+s) \equiv p_{ij}(s)$ and $q_{ij}(t) \equiv q_{ij}$. If $\mathbf{P}(s)$ is the transition probability matrix for a homogeneous Markov process, then as a consequence of the Markov property and homogeneity, $\mathbf{P}(s)$ satisfies the *Chapman-Kolmogorov equation*:

$$p_{ij}(s) = \sum_{k \in \mathcal{S}} p_{ik}(u) p_{kj}(s-u), \;\; 0 < u < s.$$

In matrix form this set of equations becomes

$$\mathbf{P}(s) = \mathbf{P}(u)\mathbf{P}(s-u), \;\; 0 < u < s.$$

From the Chapman-Kolmogorov equations we can derive the forward and backward equations:

$$\frac{d}{ds}\mathbf{P}(s) = \mathbf{Q}\mathbf{P}(s) = \mathbf{P}(s)\mathbf{Q},$$

which can be solved to yield

$$\mathbf{P}(s) = \exp(\mathbf{Q}s) \doteq \sum_{n=0}^{\infty} \frac{\mathbf{Q}^n s^n}{n!},$$

where $\mathbf{Q}^0$ is defined to be the identity matrix $\mathbf{I}$. This last equation makes clear that for a homogeneous Markov process, the matrix of transition intensities and the matrix of transition probabilities give equivalent characterizations of the process.

If the domain of the Markov process is the set of nonnegative integers $\mathbb{Z}^* = \{0, 1, 2, \dots\}$ rather than a real interval, then the process $\{Z_t, t \in \mathbb{Z}^*\}$ is called a *Markov chain* and the Markov assumption reduces to

$$P(Z_{t+1} = j | Z_t = i, Z_{t-1} = z_{t-1}, \dots, Z_0 = z_0) = P(Z_{t+1} = j | Z_t = i) \doteq p_{ij}(t), \quad t \in \mathbb{Z}^*.$$

A Markov chain is uniquely characterized by its initial distribution $\boldsymbol{\phi}$ and transition probability matrix $\mathbf{P}(\cdot)$. Similar to a Markov process, a Markov chain is *homogeneous* if the transition probabilities do not depend on chronological time so that $p_{ij}(t) \equiv p_{ij}$. For a homogeneous Markov chain p_{ij} gives the probability of making a transition from state i to state j in one step, but we can also consider the probability of being in state j several steps after being in state i. The matrix of n-step transition probabilities, denoted $\mathbf{P}^{(n)}$, is given by $\mathbf{P}^{(n)} = \mathbf{P}^n$, the matrix of one-step transition probabilities raised to the n^{th} power. This follows from the discrete-time version of the Chapman-Kolmogorov equations.

Examination of the transition probability matrix $\mathbf{P}(\cdot)$ can yield insight into the behavior of the Markov chain. Considering a homogeneous Markov chain, if $p_{ii} = 1$, then state i, called an *absorbing state*, cannot be left once it is entered (Chiang, 1980, p. 114; Limnios and Oprişan, 2001, p. 86). A state j is said to be *accessible* from state i if $p_{ij}^n > 0$ for some $n \geq 0$ (Ross, 1996, p. 168).

States of a Markov process may be classified in an analogous way by examining the Markov kernel $\mathbf{Q}(\cdot)$.

2.2.2 Renewal processes.

As we will see in Chapter 3, the Markov assumption is strong, so we will consider a stochastic process with a less stringent assumption, known as a semi-Markov process. It is useful first to consider some concepts from the related field of renewal theory.

Consider a sequence of independent, identically distributed, nonnegative random variables $\{X_n,\ n \geq 1\}$. We define an associated sequence of random variables $\{T_n,\ n \geq 0\}$ by

$$T_n = X_1 + \cdots + X_n,\ \ n \geq 1 \quad \text{and} \quad T_0 = 0;$$

this sequence is called a *renewal process* and T_n is called the n^{th} renewal time. We additionally define $N(t) = \sup\{n \geq 0 :\ T_n \leq t\}$, the number of renewals that have occurred by time t.

Renewal processes often arise in reliability theory, which considers a system in which a part fails at a random time and is replaced by an identical part. In this context, T_n is the time at which the n^{th} replacement part is installed, and $N(t)$ is the number of times that the part has been replaced by time t.

The n^{th} renewal time T_n is so named because the process is probabilistically restarted at each renewal time as a consequence of the $\{X_n\}$ being independent and identically distributed.

2.2.3 Markov renewal processes and semi-Markov processes.

To discuss a process for which the Markov assumption is relaxed, we turn to a framework that separates the evolution of the process into its sequence of states and sequence of sojourn times, where a *sojourn time* is the length of time between two consecutive transitions.

We consider a discrete two-dimensional stochastic process, called a *J-X process*, $(J - X) = \{(J_n, X_n),\ n \geq 0\}$, where the J-process represents the states visited and the X-process represents the sojourn times in each of those states. Hence $X_n \geq 0$ and $J_n \in \mathcal{S}$, $\mathcal{S} = \{1, 2, \ldots, m\}$, for each $n \geq 0$, and by convention $X_0 = 0$ almost surely. The process begins in state J_0, where it remains for time X_1 before making a transition to state J_1 and,

in general, remains in state J_n for time X_{n+1} before making a transition to a state J_{n+1} (see Figure 2.1).

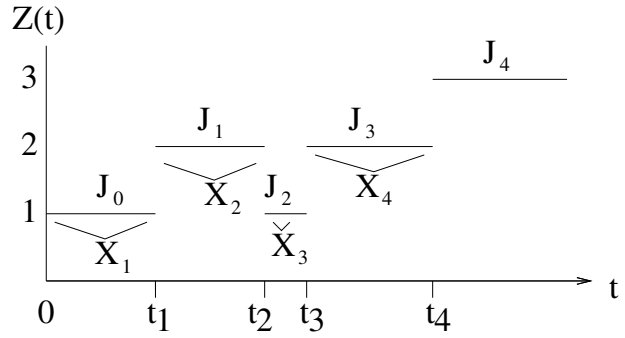


Figure 2.1: Example of a J-X process, showing relationships between $\{(J_n, X_n),\ n \geq 0\}$ and associated process $Z(\cdot)$. Under the semi-Markov assumption, the J-X process is a Markov renewal process, and $Z(\cdot)$ is the associated semi-Markov process.

The time at which the n^{th} transition occurs is given by

$$T_n \doteq \sum_{r=1}^{n} X_r,\ n \geq 1 \quad \text{and} \quad T_0 = 0.$$

Hence $X_n = T_n - T_{n-1}$ for $n \geq 1$.

We assume that $P(X_0 = 0) = 1$ and that

$$P(J_0 = i) = \phi_i \ \text{ for each } \ i \in \mathcal{S}, \ \text{ with } \ \phi_i \geq 0 \ \text{ for all } \ i \ \text{ and } \ \sum_{i \in \mathcal{S}} \phi_i = 1;$$

$\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)$ is the initial distribution of the process. As in the previous subsection we define $N(t) = \sup\{n \geq 0 :\ T_n \leq t\}$, the number of transitions made during $[0, t]$. The semi-Markov assumption is that for all $s \geq 0$,

$$\begin{aligned} P(J_n = j, X_n \leq s | (J_k, X_k),\ k = 0, 1, \ldots, n-1) &= P(J_n = j, X_n \leq s | J_{n-1}, T_{n-1}, n-1) \\ &\doteq {}^{(n-1)}K_{J_{n-1}j}(T_{n-1}, s) \end{aligned}$$

for $n \geq 1$ and $j \in \mathcal{S}$, where for each i and j, ${}^{(n-1)}K_{ij}(\cdot,\cdot)$ is a real-valued function satisfying

$$ {}^{(n-1)}K_{ij}(t, t+s) = 0 \quad \text{for} \;\; s \leq 0 \;\; \text{or} \;\; t \leq 0 $$

and

$$ \lim_{s\to\infty} \sum_{j\in\mathcal{S}} {}^{(n-1)}K_{ij}(t, t+s) = 1 \quad \text{for each} \;\; i \in \mathcal{S}, \; t \geq 0. $$

If this assumption holds, then $\{(J_n, X_n)\}$ is a *Markov renewal process* and the associated process $Z(t) \doteq J_{N(t)}$ is a *completely nonhomogeneous semi-Markov process* (refer to Figure 2.1). We can interpret this assumption as the statement that the future of the process depends on the entire history only through the current state, J_{n-1}, the elapsed chronological time, T_{n-1}, and the number of transitions between states, $n-1$, that the process has made.

This very general formulation of a semi-Markov process includes several important special cases. If the kernel $[{}^{(n-1)}K_{ij}(\cdot,\cdot)]$ does not depend on the number of transitions, then the process is nonhomogeneous semi-Markov (Iosifescu-Manu, 1972), and if additionally the kernel does not depend on the chronological time t, then the process is homogeneous semi-Markov (Lévy, 1954a,b; Smith, 1955). We discuss the latter case in more detail.

If for all $s \geq 0$,

$$ \begin{aligned} P(J_n = j, X_n \leq s | (J_k, X_k), \; k = 0, 1, \ldots, n-1) &= P(J_n = j, X_n \leq s | J_{n-1}) \\ &\doteq K_{J_{n-1}j}(s), \end{aligned} $$

where each $K_{ij}(\cdot)$ is a real-valued function satisfying

$$ K_{ij}(s) = 0 \quad \text{for} \;\; s \leq 0 $$

and

$$ \lim_{s\to\infty} \sum_{j\in\mathcal{S}} K_{ij}(s) = 1 \quad \text{for each} \;\; i \in \mathcal{S}, $$

then the associated process $Z(\cdot)$ is a *homogeneous semi-Markov process*. The assumption for such a process is that the future evolution depends on the history only through the current state of the process and the elapsed time in this state. This assumption is much weaker than the homogeneous Markov assumption. From this point forward we consider only homogeneous semi-Markov processes, referred to in many sources as simply semi-Markov processes, and we assume that each element of the *semi-Markov kernel* $K_{ij}(\cdot)$ is absolutely continuous.

To form a link with renewal theory, we note that the two-dimensional process $\{(J_n, T_n), n \geq 0\}$ is known as a *Markov renewal process*. As we will see, Markov renewal processes represent a marriage of Markov chain theory and renewal theory: in particular, the special case of a Markov renewal process with one state is a renewal process. We have seen that a renewal process probabilistically restarts itself at each renewal time. Analogously, a homogeneous semi-Markov process is probabilistically restarted at each transition time, but the future evolution of the process at each transition time depends on the current state.

A semi-Markov process can be uniquely characterized by its initial distribution $\boldsymbol{\phi}$ and the *kernel* $\mathbf{K}$. We can examine the marginal process $\{J_n, n \geq 0\}$ and the process $\{X_n, n \geq 0\}$ conditional on $\{J_n, n \geq 0\}$, which we call the J- and X-processes, respectively. Using the semi-Markov assumption and the Lebesgue Monotone Convergence Theorem (Pyke, 1961a), we can show that the J-process is a homogeneous Markov chain, called the *embedded Markov chain* of the semi-Markov process, and that it is governed by the transition probability matrix defined by $p_{ij} \doteq \lim_{s\to\infty} K_{ij}(s)$ for all $i, j \in \mathcal{S}$.

To discuss the X-process, we define the functions for $s \geq 0$

$$F_{ij}(s) \doteq \begin{cases} \frac{K_{ij}(s)}{p_{ij}}, & p_{ij} > 0; \\ 1_{(s\geq 1)}, & p_{ij} = 0 \end{cases}$$

for each i and j and

$$H_i(s) \doteq \sum_{j\in\mathcal{S}} K_{ij}(s)$$

for each i. We can show the following: for all $s \geq 0$,

$$\begin{aligned} F_{ij}(s) &= P(X_n \leq s | J_{n-1} = i, J_n = j) \\ H_i(s) &= P(X_n \leq s | J_{n-1} = i); \end{aligned}$$

these are known respectively as the conditional and unconditional distributions of the sojourn time in state i. We note that the above definition of $F_{ij}(\cdot)$ in the case that $p_{ij} = 0$ is arbitrary; if it is impossible to make a transition from i to j, then there is no need to consider the distribution of the sojourn time in i before transitioning to j. Let $f_{ij}(\cdot)$ be the density corresponding to $F_{ij}(\cdot)$; as we will soon see, the former exists because of our assumption that the semi-Markov kernel is absolutely continuous.

We have seen that the marginal J-process of a semi-Markov process is a homogeneous Markov chain, so a natural question is whether the behavior of the marginal X-process has a similar independence property. Since the n^{th} sojourn time X_n depends on the current state J_{n-1}, the situation is clearly different for the X-process. The X-process does have the property of conditional independence from the J-process. Specifically, for any natural numbers $n_1, \ldots, n_k$ with $n_1 < \cdots < n_k$ and any nonnegative real numbers $x_{n_1}, \ldots, x_{n_k}$, we have (see Janssen and Manca, 2006)

$$P(X_{n_1} \leq x_{n_1}, \ldots, X_{n_k} \leq x_{n_k} | J_{n_1-1}, J_{n_1}, \ldots, J_{n_k-1}, J_{n_k}) = F_{J_{n_1-1}J_{n_1}}(x_{n_1}) \cdots F_{J_{n_k-1}J_{n_k}}(x_{n_k});$$

that is, the sojourn times $X_{n_1}, \ldots, X_{n_k}$ are conditionally independent given their respective current and next states, $J_{n_1-1}, J_{n_1}, \ldots, J_{n_k-1}, J_{n_k}$. Considering the Markov renewal process $\{(J_n, S_n),\ n \geq 0\}$, this proposition follows from the fact that transitions between states are renewal events, so the times at which they occur are renewal times.

We can express each element of the kernel in a natural way as the product of the respective transition probability and the conditional sojourn time distribution:

$$K_{ij}(s) = F_{ij}(s) \cdot p_{ij} \quad \text{for} \quad s \geq 0.$$

We noted previously that the semi-Markov process can be uniquely characterized by $(\boldsymbol{\phi}, \mathbf{K})$,

and the argument here makes clear (Janssen and Manca, 2006) that it can also be characterized by $(\boldsymbol{\phi}, \mathbf{P}, \mathbf{F})$. From standard survival analysis we know that under some regularity conditions, the time to failure can be characterized by the cumulative distribution function $F(\cdot)$ or the hazard function $h(\cdot)$. By analogy, if X_n is the sojourn time in state $J_{n-1} = i$ before going to state $J_n = j$, then we can characterize the distribution of X_n by F_{ij} or equivalently by the conditional hazard function, for $s \geq 0$:

$$h_{ij}(s) \doteq \lim_{\Delta s \downarrow 0} \frac{1}{\Delta s} P(s \leq X_n < s + \Delta s | J_{n-1} = i, J_n = j,\ X_n \geq s).$$

Hence the distribution of the conditional sojourn time in state i before proceeding to state j can be uniquely characterized by the hazard function $h_{ij}(\cdot)$, for each $i \neq j \in \mathcal{S}$, so we can characterize the semi-Markov process alternatively by $(\boldsymbol{\phi}, \mathbf{P}, \mathbf{h})$.

2.2.4 Comparison of Markov and semi-Markov processes.

We can now consider the implications of the Markov and semi-Markov assumptions for the underlying process. We have seen that for a semi-Markov process, we can think of the sequence of states and the sojourn times conditional on the states separately, with the distribution of the conditional sojourn times left unspecified. On the other hand, for a nonhomogeneous Markov process, we can show that the memoryless property implies that the sojourn times in any state i are distributed exponentially with rate $q_i(t) \doteq -\sum_{j \neq i} q_{ij}(t)$ at chronological time t, which reduces to $q_i = -\sum_{j \neq i} q_{ij}$ for a homogeneous Markov process (Ross, 1996, p. 232). Hence, for a Markov process, if the process is in state i, the transition intensities to other states do not depend on the elapsed time s in state i, whereas in a semi-Markov process, the transition intensities are allowed to depend on the elapsed time in i.

More precisely, a homogeneous Markov process is a homogeneous semi-Markov process with kernel $\mathbf{K}(\cdot)$ defined for $s \geq 0$ by

$$K_{ij}(s) = p_{ij} \cdot [1 - \exp(-\lambda_i s)],$$

with $\lambda_i > 0$ and $p_{ii} = 0$ for all $i \in \mathcal{S}$. That is, the process is Markov if the distribution of the sojourn time in state i conditional on j being the next state to be visited, is exponential with a rate parameter that depends only on i; and if the process cannot make a transition from state i back to itself, which is a consequence of how a Markov process is defined. The above Markov process is governed by the kernel $\mathbf{Q} = [q_{ij}]$ with $q_{ij} = p_{ij} \cdot \lambda_i$ for $i \neq j$ and $q_{ii} = \sum_{j' \neq i} q_{ij'}$.

It is common in the literature to impose the assumption that each transition must be to a different state in defining a semi-Markov process so that parameters are identifiable (Ouhbi and Limnios, 1999). If a semi-Markov process has no absorbing states and does not satisfy this assumption, it may be transformed to a related semi-Markov process for which the assumption is satisfied (Pyke, 1961a; Limnios and Oprişan, 2001, p. 86). Specifically, we may transform the *J-X* process such that the semi-Markov kernel of the transformed process is given by

$$K'_{ij}(s) = \begin{cases} \frac{K_{ij}(s)}{1-p_{ii}}, & i \neq j; \\ 0, & i = j, \end{cases}$$

where $K_{ij}(\cdot)$ is an element of the semi-Markov kernel of the original process. Although it is not in the literature to our knowledge, it stands to reason that a similar transformation may be applied to a process with absorbing states:

$$K'_{ij}(s) = \begin{cases} \frac{K_{ij}(s)}{1-p_{ii}}, & p_{ii} \neq 1,\ i \neq j; \\ 0, & p_{ii} \neq 1,\ i = j; \\ K_{ij}(s), & p_{ii} = 1. \end{cases}$$

Since transitions from a state to itself are not relevant in the current context, we assume from now on that transitions from a non-absorbing state must be to a different state for semi-Markov processes.

Often it is natural to discuss distributions in terms of their hazard functions. For each $(i, j) \in \mathcal{S}$ with $i \neq j$ we distinguish two associated hazard functions of a semi-Markov

process: the conditional hazard function $h_{ij}(\cdot)$, which corresponds to $F_{ij}(\cdot)$, was introduced previously, and the cause-specific hazard function $\eta_{ij}(\cdot)$, which corresponds to $K_{ij}(\cdot)$, is defined for $s \geq 0$ as

$$\eta_{ij}(s) \doteq \lim_{\Delta s \downarrow 0} \frac{1}{\Delta s} P(s \leq X_n < s + \Delta s,\ J_n = j | J_{n-1} = i, X_n \geq s).$$

We can show that, for $i \neq j$ and $s \geq 0$,

$$\eta_{ij}(s) = \frac{p_{ij} \cdot f_{ij}(s)}{S_{i\cdot}(s)}$$

where $S_{i\cdot}(s) = \sum_{j' \neq i} p_{ij'} S_{ij'}(s)$ is the marginal survival probability in state i and $S_{ij}(s) \doteq 1 - F_{ij}(s)$ is the conditional survival probability in state i before transitioning to state j.

When the semi-Markov process under consideration is Markov, the cause-specific hazard function $\eta_{ij}(\cdot)$ reduces to q_{ij}, the corresponding element of the Markov kernel:

$$\eta_{ij}(s) = \frac{p_{ij} \cdot \lambda_i \cdot \exp(-\lambda_i s)}{\exp(-\lambda_i s)} = p_{ij} \cdot \lambda_i = q_{ij}.$$

As we discussed in the first chapter, there are many examples of applications in which continuously observed sojourn times in each state are observed to be poorly approximated by an exponential distribution. Semi-Markov models allow us to relax the assumption that the sojourn times in each state are exponential, and to choose a more appropriate model for them based on the particular application.

2.2.5 Special cases.

In the above description of semi-Markov processes, the embedded Markov chain was allowed to be an m-state chain in which a transition could be made between any two distinct states with some probability, which could be positive or zero. The only assumption about the chain was that a transition could not occur from a state to itself. In a specific application, however, it may be that only a subset of the transitions are possible. For example, if we are modeling an incurable illness and the disease states are (1) healthy, (2) diseased, and (3) dead, then transitions are possible from $1 \to 2$, $1 \to 3$, or $2 \to 3$, but all other transition

probabilities are restricted to be zero. As we will see in the following section, restricting some transition probabilities to be zero can simplify estimation. Some methods of estimation apply to a specific *state model*, or set of transition probabilities of the embedded Markov chain that are positive (see Figure 2.2 for examples). In particular, estimation is simplified when the embedded Markov chain is progressive in some sense. Though the terminology has yet to be standardized in the literature, in this dissertation we will refer to a general *progressive* process as one in which state j is reachable from state i only if $i < j$. Further, we will call a process *simple progressive* if the states occur in a prescribed sequence. For a progressive process with a small number of states, the number of possible trajectories through the states is limited, and in the case of a simple progressive process, there is just one possible trajectory.

Our overarching goal in this dissertation is to develop methods to model progressive chronic diseases when subjects are observed periodically rather than continuously. We begin by examining a simple three-state progressive model, and subsequently relax assumptions on the process to allow for more realistic circumstances, such as left-censored entry, multiple trajectories through the states, and competing risks. Finally, we will approach the case in which the disease process may not be progressive.

2.3 Literature Review

Although Markov models have been used to approach multi-state data for decades (Meira-Machado et al., 2009; Foucher et al., 2007), semi-Markov models, which were introduced independently by Lévy and Smith in 1954 (Lévy, 1954a,b; Smith, 1955), have been utilized only relatively recently and in limited circumstances (De Gruttola and Lagakos, 1989; Kang and Lagakos, 2007). Relative to Markov models, semi-Markov models provide additional flexibility in that they do not impose an exponential assumption on the sojourn times and they allow the sojourn time distribution in a state to vary according to the next state to be visited, but the benefit provided by this flexibility was not widely recognized until semi-Markov models came into use. Semi-Markov models also present considerable difficulty for estimation. Frequently in applications a process is not observed continuously but only at discrete time points, so that the full sequence of visited states is not known. In a semi-

Markov model, this observation scheme can make estimation challenging without extra assumptions about the nature of the process, but in a Markov model estimates are available. These barriers have prevented semi-Markov models from being widely used in biomedical applications. In many applications where a Markov model was used, a semi-Markov model may have been more suitable (e.g. Longini et al. (1989); Marshall and Jones (1995); Chen et al. (1996)).

In this section we begin by briefly summarizing existing semi-Markov approaches to estimation of continuously observed multi-state processes. We then examine available approaches to estimation of intermittently observed processes. As alluded to previously, when a process is observed continuously, we know both the sequence of states and the exact sojourn times; however, when we have only snapshots of the process at certain points in time, we may not observe the full sequence of states and we do not know the sojourn times exactly. If the process is observed only at discrete time points, a homogeneous Markov model has just enough structure so that the transition intensities can still be estimated (Kalbfleisch and Lawless, 1985). However, in a semi-Markov model the transition intensities depend on the elapsed time in the current state, which is unknown when the process is observed only at discrete time points. Estimation is therefore less tractable for semi-Markov models, and methods for a general process under panel observation do not exist. We examine methods that have been developed for specific state models and under various assumptions. In particular, we examine methods for progressive processes.

2.3.1 Methods for continuously observed processes.

When a homogeneous semi-Markov model is used for a continuously observed process, in the absence of covariates, the parameters corresponding to the embedded Markov chain are easily estimated via

$$\hat{p}_{ij} = \frac{n_{ij}}{n_i},$$

where n_{ij} is the number of observed transitions from state i to j and n_i is the total number of transitions from i to any state (Anderson and Goodman, 1957). With the sequence

of states known, it remains to estimate the distributions of the sojourn times conditional on this sequence of states. Regardless of the state model under consideration, the parameters corresponding to the sequence of states and to the sequence of sojourn times can be estimated separately. Assuming the process is observed continuously or is subject to right-censoring only, a variety of approaches may be taken to modeling the conditional sojourn times: fully parametric (Weiss and Zelen, 1965), piecewise exponential (Colvert and Boardman, 1976; Ouhbi and Limnios, 1999), or nonparametric (Voelkel and Crowley, 1984; Kaplan and Meier, 1958). Moreover, tests have been developed to examine the Markov or semi-Markov assumption. Chang, Chuang, and Hsiung (2001) consider an illness-death model (see Figure 2.2b) and propose goodness-of-fit statistics for testing the hypotheses that the underlying process is either (1) homogeneous semi-Markov or (2) nonhomogeneous Markov, and derive asymptotic distributions of the statistics.

Frequently there is reason to account for differences among subgroups of the patient population. To carry this out by covariate adjustment, parametric and semiparametric regression approaches have been taken (Therneau and Grambsch, 2000). Covariates may be included in a regression model via the embedded Markov chain, conditional sojourn distributions, or both. Lawless and Fong (1999) give an overview of methods for modeling sojourn times that account for the presence of covariates and possible dependencies among sojourn times within a subject. They also review methods for dealing with various observation schemes, including left truncation of observations as well as selection mechanisms in observational studies. They discuss the use of random effects to deal with unexplained inter-subject or temporal variability. Since random effects are a modeling device and can introduce computational difficulties, the authors suggest that the use of random effects be avoided when the process is incompletely observed. All of the methods under consideration assume that process is continuously observed, and the authors note the need for methods that address the case of panel data.

There are several classes of methods that apply to certain state models. Some methods are built on the assumption that the embedded Markov chain of the process is *ergodic*, which implies in particular that an absorbing state such as death cannot exist (Ross, 1996). By contrast, other methods assume that the underlying process is progressive. Voelkel and

Crowley (1984) approach semi-Markov processes in a counting processes framework and show that, under some assumptions, a progressive semi-Markov process can be transformed via a random function of the chronological time into the *multiplicative intensity model* introduced by Aalen (1978). Voelkel and Crowley then consider a particular progressive state model and establish asymptotic properties of the estimator of the probability of being in one of the states of this model.

Although a number of methods have addressed censoring, most have focused on right-censoring in the final state or left-censoring in the initial state (e.g. Lagakos, Sommer, and Zelen, 1978). Extending these methods to a panel observation scheme has been elusive, and has not been done to our knowledge.

2.3.2 Panel data: Markov models.

In a seminal paper, Kalbfleisch and Lawless (1985) proposed a method to estimate the instantaneous transition probabilities of a general multi-state process under panel observation assuming the process is Markov. Using the fact that the transition probability and transition intensity matrices are related via $\mathbf{P}(s) = \exp(\mathbf{Q}s) \doteq \sum_{r=0}^{\infty} \frac{\mathbf{Q}^r s^r}{r!}$ for $s \geq 0$ for a homogeneous Markov process, the authors proposed an efficient scoring procedure to estimate $\mathbf{Q}$ via maximum likelihood. Specifically, if subjects are observed at times $t_0, t_1, \ldots, t_m$, and if $\mathbf{Q}$ depends on $\boldsymbol{\theta}$ then the likelihood of $\boldsymbol{\theta}$ is given by

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{l=1}^{m} \prod_{i,j \in \mathcal{S}} p_{ij}(t_l - t_{l-1})^{n_{ijl}}$$

where n_{ijl} is the number of subjects who are observed in state i at t_{l-1} and state j at t_l. The closed-form expression of $\mathbf{P}(s)$ as well as $\frac{\partial}{\partial \theta_u}\mathbf{P}(s)$ enables the application of a scoring rule involving only first derivatives to carry out inference about $\boldsymbol{\theta}$. The algorithm was extended to allow the transition rates to depend on covariates (Kalbfleisch and Lawless, 1985). Moreover, Hubbard (2007) extended the method to nonhomogeneous Markov models.

At its core this method relies on the Markov assumption, and hence sojourn times in each state are modeled as exponential, where the exponential parameters may depend on various factors. However, many disease processes are observed to progress in a way that

exhibits non-constant hazard (Weiss and Zelen, 1965; Kang and Lagakos, 2007). Hence, methods that do not rely on the Markov assumption are needed.

2.3.3 Panel data: semi-Markov models for progressive processes.

Estimation of a semi-Markov process in the case where subjects are observed intermittently is much more difficult in general than the case where they are observed continuously since panel observation of a subject does not necessarily yield the complete sequence of states that the subject has gone through. This missing information complicates estimation of the process.

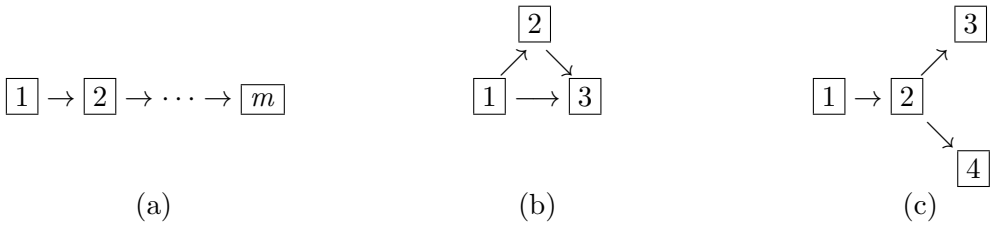


Figure 2.2: Examples of types of state models for which estimation for panel data is simplified: (a) simple progressive, (b) illness-death model, (c) progressive with competing risks.

These difficulties can be overcome in some state models. In a simple progressive model (see Figure 2.2a), which gives rise to chain-of-events data, events are assumed to occur in a prescribed sequence, and without loss of generality these m states can be numbered in the order in which they are assumed to occur. Hence, in a semi-Markov model the transition probability matrix of the embedded Markov chain is degenerate and need not be estimated:

$$p_{ij} = \begin{cases} 1, & \text{for } j = i+1,\ i = 1, \ldots, m-1; \text{ and } i = j = m; \\ 0, & \text{otherwise.} \end{cases}$$

We note that for a simple progressive model, the semi-Markov assumption reduces to the assumption that the sojourn times in each state are conditionally independent. With the sequence of states known, it remains only to estimate the sojourn time distribution in

each of these states. With m states in the model, assuming that each conditional sojourn time distribution is absolutely continuous leads to a likelihood with convolution products of the conditional sojourn densities $f_{i,i+1}(\cdot;\boldsymbol{\theta_i})$ and survival distributions $S_i(\cdot;\boldsymbol{\theta_i})$ for $i = 1,\ldots,m-1$. In an m-state simple progressive process the observation of subject i may be expressed as a sequence of observed states, as it was in Kalbfleisch and Lawless (1985), or equivalently in the "sufficient" form $\mathbf{t}$, a vector of length $2(m-1)$. In this vector, successive pair of components represents the observation times preceding and following a time of transition. For example, considering $m = 3$, the components of $\mathbf{t}$ represent:

$$\begin{aligned} t_1: &\quad \text{last observed time in state 1,} \\ t_2: &\quad \text{first observed time in state 2,} \\ t_3: &\quad \text{last observed time in state 2, and} \\ t_4: &\quad \text{first observed time in state 3,} \end{aligned}$$

where t_2, t_3, and t_4 may or may not be defined, depending on how each subject was censored. With this notation, the likelihood of the parameters given panel observations on N subjects is given by

$$\mathcal{L}(\boldsymbol{\theta_1},\boldsymbol{\theta_2}|\mathbf{t_1},\ldots,\mathbf{t_N}) \;=\; \prod_{i=1}^{N} \mathcal{L}_1^{(1-\delta_i)\cdot(1-\epsilon_i)} \cdot \mathcal{L}_2^{\delta_i\cdot(1-\epsilon_i)} \cdot \mathcal{L}_3^{(1-\delta_i)\cdot\epsilon_i} \cdot \mathcal{L}_4^{\delta_i\cdot\epsilon_i},$$

where the likelihood contributions are given by

$$\begin{aligned} \mathcal{L}_1 &= \int_{t_3}^{t_4}\int_{t_1}^{t_2} f_{12}(u_1)\cdot f_{23}(u_2-u_1)du_1du_2 \\ \mathcal{L}_2 &= \int_{t_1}^{t_4}\int_{t_1}^{u_2} f_{12}(u_1)\cdot f_{23}(u_2-u_1)du_1du_2 \\ \mathcal{L}_3 &= \int_{t_3}^{\infty}\int_{t_1}^{t_2} f_{12}(u_1)\cdot S_{23}(u_2-u_1)du_1du_2 \\ \mathcal{L}_4 &= \int_{t_1}^{\infty} S_{12}(u)du, \end{aligned}$$

and δ_i and ϵ_i are indicators that subject i was not observed in states 2 and 3 respectively.

That is, $\mathcal{L}_1$ is the contribution of a subject who was observed in all three states; $\mathcal{L}_2$ represents a subject observed in states 1 and 3 only; and $\mathcal{L}_3$ and $\mathcal{L}_4$ represent subjects who were right censored in state 2 and state 1 respectively.

Considering the case of a three-state simple progressive model, De Gruttola and Lagakos (1989) proposed a nonparametric approach to estimate the conditional distributions of the sojourn times in states 1 and 2. Their approach, an extension of the self-consistency algorithm of Turnbull (1976) for univariate survival data, involves modeling the two sojourn time distributions as discrete random variables. Assuming the process enters state 1 at time zero, they let Y_1 and $Z = Y_1 + Y_2$ denote the transition times into states 2 and 3 respectively, and (Y_L, Y_R, Z_L, Z_R) be the "sufficient data" for a single realization of the process, i.e. the observation times immediately preceding and following the two transitions. This notation is similar to $\mathbf{t}$ introduced above. The authors choose locations of the mass points of Y_1 and Y_2, $0 \leq y_{11} < \cdots < y_{1r}$ and $0 \leq y_{21} < \cdots < y_{2s}$ respectively, and note that the observation (Y_L, Y_R, Z_L, Z_R) uniquely determines a set of "admissible values" of (y_{1j}, y_{2k}). To recast the data in this format they define α_{jk} as the indicator that (y_{1j}, y_{2k}) is an admissible value of (Y_1, Y_2). With $w_{1j} = P(Y_1 = y_{1j})$ and $w_{2k} = P(Y_2 = y_{2k})$, the likelihood is given by

$$\mathcal{L}(\mathbf{w_1}, \mathbf{w_2}) = \prod_{i=1}^{N} \left(\sum_{j=1}^{r} \sum_{k=1}^{s} \alpha_{jk}^i w_{1j} w_{2k} \right),$$

where $\mathbf{w_1} = (w_{11}, \ldots, w_{1r})'$ and $\mathbf{w_2} = (w_{21}, \ldots, w_{2s})'$. If I_{jk}^i is the indicator that (y_{1j}, y_{2k}) is the true value of (Y_1, Y_2) for subject i, then its expectation is

$$\mu_{jk}^i = \frac{\alpha_{jk}^i w_{1j} w_{2k}}{\sum_l \sum_m \alpha_{lm}^i w_{1l} w_{2m}}. \tag{2.1}$$

Let $\boldsymbol{\mu}^i = [\mu_{jk}^i]$, $i = 1, \ldots, N$. The values of $\mathbf{w_1}$ and $\mathbf{w_2}$ can be expressed in terms of μ_{jk}^i:

$$w_{1j} = \frac{1}{N} \sum_{i=1}^{n} \sum_{k=1}^{s} \mu_{jk}^i \qquad \text{and} \qquad w_{2k} = \frac{1}{N} \sum_{i=1}^{n} \sum_{j=1}^{r} \mu_{jk}^i \tag{2.2}$$

for $j = 1, \ldots, r$, $k = 1, \ldots, s$. These equations suggest an iterative scheme for estimating

$\boldsymbol{\mu}^1, \ldots, \boldsymbol{\mu}^N$, $\mathbf{w_1}$ and $\mathbf{w_2}$. It can be shown (Turnbull, 1976) that the self-consistent estimate is equivalent to the maximum likelihood estimate. We also note that it is a version of the EM algorithm (Dempster, Laird, and Rubin, 1977). Specifically, (2.1) corresponds to the E-step of "filling in" the missing data based on the current parameter estimates and (2.2) corresponds to the M-step of carrying out maximum likelihood estimation of the parameters based on the current estimates of the complete data.

In parallel with De Gruttola and Lagakos (1989), Frydman (1992) extended the algorithm of Turnbull (1976) to a three-state simple progressive model, but assumed the underlying process was nonhomogeneous Markov rather than homogeneous semi-Markov. Frydman additionally assumed that entry into the third state was either observed exactly or right censored. She applied her method to the HIV application in De Gruttola and Lagakos (1989) and obtained results similar to those of De Gruttola and Lagakos, but with some differences in the inference regarding the sojourn time in the infected state before developing AIDS symptoms.

Although method of De Gruttola and Lagakos avoids imposing distributional assumptions on the sojourn times in each state, it has several drawbacks. Most alarmingly, it assumes implicitly that each subject was observed at least once in every visited state. Thus, because subjects are assumed to progress through the states sequentially, a subject who was observed in only states 1 and 3 would need to be discarded from analysis. This dubious practice would inevitably lead to biased estimation. Since the proportion of such illegal observations in a dataset tends to increase as the observation scheme becomes more sparse, the magnitude of the bias would increase as the interval between observations lengthened. Second, the method involves discretizing the two sojourn times. This means that decisions must be made about where to locate the mass points of Y_1 and Y_2. Though certain guidelines may be used to avoid lack of identifiability and loss of information, the final say must be left to the particular dataset under consideration to ensure that each subject has at least one admissible value of $(\mathbf{y_1}, \mathbf{y_2})$. To summarize these disadvantages, as the observation scheme becomes more sparse compared with the rate of the underlying disease process, the algorithm performs worse, and the results depend more on the *ad hoc* decisions that must be made to implement the algorithm.

De Gruttola and Lagakos (1989) note that with some modifications, their algorithm may be applied to a weakly structured parametric approach in which the sojourn times are modeled as piecewise uniform densities, and that this approach may perform better when the number of mass points is small. The authors did not implement this approach, but it seems that it would have many of the drawbacks of the original algorithm.

The self-consistency algorithm of De Gruttola and Lagakos was extended to handle a simple progressive process with m states by Sternberg and Satten (1999). Since a Markov model for this situation is a special case of the semi-Markov model that imposes a truncated geometric structure on the sojourn times, the authors are able to carry out a formal test of the Markov assumption for each of the first $m-1$ states. The authors suggest extending their method to continuous time but do not indicate an approach. Their method is computationally efficient but requires a simple progressive model with no clear extension to other types of models. Sternberg and Satten (1999) propose a method that allows for left-censored entry, in which subjects do not necessarily enter state 1 at time zero.

A slightly more general state model that arises in many applications is the *illness-death* model shown in Figure 2.2b. This model is useful for studying an incurable, potentially fatal disease. Beginning in a state of good health, subjects at risk may progress to illness, or may die from another cause or they may die from the illness. When the third state represents death, it may be assumed that transitions to this state are observed exactly. Alternatively, the three states may represent stages of disease that do not necessarily occur in a prescribed sequence. For example, in developing the goodness-of-fit statistics mentioned previously, Chang, Chuang, and Hsiung (2001) consider modeling the risk of breast cancer over time among women who may or may not have had benign breast disease. Though they assumed the transition times were known exactly, in this application each of these two transitions may be subject to interval censoring. Developing methods for this model is more challenging than for a simple three-state progressive process because a subject with observed trajectory $1, 1, 3$, for example, may or may not have visited state 2.

Methods for estimation and inference exist for more general progressive state models, but since many approaches to panel data are motivated by a particular dataset, they are tailored to a specific state model and impose assumptions specific to the application which

can greatly simplify estimation. Common assumptions are that some of the transition times are known exactly or that subjects are observed at least once in each state they visit. We examine several of these methods here.

Frequently in applications there is a need to consider multiple absorbing states, or competing risks; an example of a state model accounting for this feature is shown in Figure 2.2c. Foucher, Giral, Soulillou, and Daures (2007) considered a slightly more complicated five-state progressive disease process with competing risks to model patients' natural history following kidney transplantation. The authors modeled sojourn times in each state as generalized Weibull, a parametric form that allows the conditional hazard function to be nonmonotonic. However, the assumption that subjects were observed in each visited state implied that estimation of the embedded Markov chain was trivial. Additionally, they did not account for the interval censoring of intermediate states.

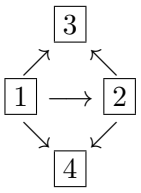


Figure 2.3: State model of Foucher et al. (2010). State numbers indicate (1) baseline value of creatinine clearance (CL); (2) decreased CL; (3) return to dialysis; (4) death.

If the number of states in the process is small in the progressive state model under consideration, it may be feasible to consider all possible trajectories through the state space when formulating the likelihood. Motivated by a prospective study of kidney transplant recipients, Foucher, Giral, Soulillou, and Daures (2010) carry out this procedure for the state model shown in Figure 2.3, where in this example the states 1-4 are defined by the patient's baseline creatinine clearance (CL); decreased CL; return to dialysis; and death with a functional transplant. They assume that the times of each patient's entry into state 1 and entry into state 3 or 4 (if applicable) are known exactly. The time of entry into state 2 is interval-censored, and a patient who is not observed in state 2 may or may not have entered state 2. Similar to Foucher et al. (2007), the authors impose a generalized Weibull

form on the sojourn times, and build the likelihood from each of the four possible trajectories through the state space, using convolution products to deal with censoring of state 2. They obtain maximum likelihood estimates of the generalized Weibull parameters and of the transition probabilities of the embedded Markov chain. The authors additionally incorporate covariates and derive a goodness-of-fit statistic to test homogeneity of the semi-Markov process. Their method for modeling progressive disease could be generalized somewhat: it could be adapted for other fairly simple state diagrams and, as they note, it could be modified to handle interval-censored absorbing states. However, numerical maximization of the likelihood is quite computationally expensive when the likelihood contributions involve more than two interval-censored times.

2.3.4 Panel data: semi-Markov models for more general processes.

A few authors have developed methods for modeling non-progressive processes without the assumption that the process is Markov, but with other strong assumptions. Motivated by an application involving human papillomavirus (HPV) and cervical abnormalities arising from HPV infection, Kang and Lagakos (2007) propose a method to model a nonprogressive process subject to interval censoring as well as misclassification of states. Subjects in the placebo arm of an HPV vaccine trial were tested for HPV and examined for high-grade cervical intraepithelial neoplasia (CIN) at prespecified clinic visits. The transition rates to CIN from the HPV infected and uninfected states were of interest. Since HPV infection may resolve spontaneously and may recur, the state model allowed transitions between the HPV uninfected and infected CIN-free states (states 1 and 2 respectively in Figure 2.4). Transition to CIN was considered from the states defined by HPV infection status (states 3 and 4 respectively).

Considering a general nonprogressive state model, the authors modeled the process as homogeneous semi-Markov, and made the simplifying assumption that transition intensities from at least one state were duration-independent. That is, they assumed there was a set of states $\mathcal{C}$ such that for each $i \in \mathcal{C}$, $\eta_{ij}(s) \equiv \eta_{ij}$ for each $j \neq i$. This means that for each $i \in \mathcal{C}$, sojourn times in state i must be exponential with rate $\sum_{j' \neq i} \eta_{ij'}$, which is

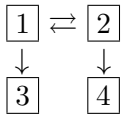


Figure 2.4: State model of Kang and Lagakos (2007). State 3 represents CIN diagnosis after a visit in which the patient was not infected with HPV (state 1), while state 4 represents a CIN diagnosis was HPV infected (state 2).

equivalent to the Markov assumption for state i. This assumption is substantial, and allows for great simplification of the likelihood. The combination of these two assumptions—that the process is homogeneous semi-Markov and that the process is memoryless in at least one state—bestows on the process a stationarity property whenever it visits a Markov state. Specifically, for any times $0 < t_1 < \cdots < t_m < \cdots$, if the process is in a Markov state, z_m, at time t_m, then the likelihood of the evolution of the process after t_m is given by

$$\begin{aligned}
& P(Z(t_{m+1}) = z_{m+1}, Z(t_{m+2}) = z_{m+2}, \ldots | Z(t), t \in [0, t_m];\ Z(t_m) = z_m) \\
= \ & P(Z(t_{m+1}) = z_{m+1}, Z(t_{m+2}) = z_{m+2}, \ldots | Z(t_m) = z_m,\ \phi(t_m)) \\
= \ & P(Z(t_{m+1}) = z_{m+1}, Z(t_{m+2}) = z_{m+2}, \ldots | Z(t_m) = z_m) \\
= \ & P(Z(t_{m+1} - t_m) = z_{m+1}, Z(t_{m+2} - t_m) = z_{m+2}, \ldots | Z(0) = z_m),
\end{aligned}$$

where $\phi(t_m)$ is defined as the most recent time the process entered state z_m prior to time t_m, and the above steps are justified by the semi-Markov assumption, the assumption that state z_m is Markov, and the homogeneity of the semi-Markov process, respectively. This stationarity property implies that each subject's likelihood contribution can be decomposed into a product of more manageable factors, as we illustrate later.

In developing their method, Kang and Lagakos allowed for the possibility that observations of the states were subject to classification error. For example, in their application, the assays for HPV and CIN were allowed to have imperfect specificity. Hence, a subject who was classified as being HPV-negative could have been truly HPV-positive with some probability, and similarly for CIN. The authors assumed that these classification errors oc-

curred independently, but did not go into detail about how they dealt with this aspect in the estimation.

In the HPV application, Kang and Lagakos assumed that transition intensities from state 1 were duration-independent—that is, $\mathcal{C} = \{1\}$—and that subjects began in state 1. By the stationarity property described above, the likelihood contribution of a subject observed in states $1, 2, 2, 1, 1, 2, 1, 3$ at times $0 < t_1 < \cdots < t_7$, for example, could be simplified as

$$
\begin{aligned}
& P(Z(t_1) = 2, Z(t_2) = 2, Z(t_3) = 1, Z(t_4) = 1, Z(t_5) = 2, Z(t_6) = 1, Z(t_7) = 3 | Z(0) = 1) \\
= \;& P(Z(t_7) = 3 | Z(0) = 1, Z(t_1) = 2, Z(t_2) = 2, Z(t_3) = 1, Z(t_4) = 1, Z(t_5) = 2, Z(t_6) = 1) \\
& \cdot P(Z(t_5) = 2, Z(t_6) = 1 | Z(0) = 1, Z(t_1) = 2, Z(t_2) = 2, Z(t_3) = 1, Z(t_4) = 1) \\
& \cdot P(Z(t_4) = 1 | Z(0) = 1, Z(t_1) = 2, Z(t_2) = 2, Z(t_3) = 1) \\
& \cdot P(Z(t_1) = 2, Z(t_2) = 2, Z(t_3) = 1 | Z(0) = 1) \\
= \;& P(Z(t_7) = 3 | Z(t_6) = 1) \\
& \cdot P(Z(t_5) = 2, Z(t_6) = 1 | Z(t_4) = 1) \\
& \cdot P(Z(t_4) = 1 | Z(t_3) = 1) \\
& \cdot P(Z(t_1) = 2, Z(t_2) = 2, Z(t_3) = 1 | Z(0) = 1) \\
= \;& P(Z(t_7 - t_6) = 3 | Z(0) = 1) \\
& \cdot P(Z(t_5 - t_4) = 2, Z(t_6 - t_4) = 1 | Z(0) = 1) \\
& \cdot P(Z(t_4 - t_3) = 1 | Z(0) = 1) \\
& \cdot P(Z(t_1) = 2, Z(t_2) = 2, Z(t_3) = 1 | Z(0) = 1).
\end{aligned}
$$

Intuitively, the observed process can be decomposed into "hops" defined by state 1, since the process is memoryless while in this state. Hence, the particular decomposition of the joint distribution in the first step of the above calculation is chosen to take advantage of the stationarity property in state 1.

We can see that the stationarity property allowed this 7-dimensional joint distribution to be decomposed into the product of four manageable conditional probabilities. In general,

each subject's likelihood contribution can be expressed as a product of elements of the form

$$P(Z(t_1) = 1|Z(0) = 1) \tag{2.3}$$

$$P(Z(t_1) = 3|Z(0) = 1) \tag{2.4}$$

$$P(Z(t_1) = 2, Z(t_2) = 2, \ldots, Z(t_{m-1}) = 2, Z(t_m) = j|Z(0) = 1) \tag{2.5}$$

for $j = 1, 2, 3$, where $0 < t_1 < \cdots < t_m$ are the observation times for this subject. The law of total probability allows the first of these conditional probabilities to be expressed as

$$P(Z(t_1) = 1|Z(0) = 1) = \sum_{k=0}^{\infty} P_{11}(k, t_1),$$

where $P_{11}(k, t_1)$ denotes the probability that the process is in state 1 at t_1 after k visits to state 2. This last conditional probability may be expressed as a k-fold convolution product for each $k \geq 1$. The authors used a similar procedure to deal with the last conditional probability (2.5). They carried out estimation for the HPV application under the additional assumption that state 2 had a "guarantee time" or minimum sojourn time, and that transition intensities were duration-independent thereafter. This additional assumption facilitated the computation in the estimation problem.

The method of Kang and Lagakos is a significant step in the frontier of modeling nonprogressive processes in a semi-Markov framework. Though it imposes the strong assumption that the process satisfies the Markov property for at least one state, and has the potential to be quite computationally intensive, the method is applicable to a variety of state models. In applications involving moderately large sample sizes where the assumption of one Markov state may be reasonable, this method may be a good choice.

Other methods for modeling nonprogressive processes without the Markov assumption have considered models with just two states. Such a process, whose state model is pictured in Figure 2.5a, has a deterministic embedded Markov chain governed by the transition

probability matrix

$$p_{ij} = \begin{cases} 1, & i \neq j; \\ 0, & i = j, \end{cases}$$

for $i, j \in \{1, 2\}$, and the task at hand is to estimate the sojourn time distributions.

Similarly to Kang and Lagakos (2007), Mitchell, Hudgens, King, Cu-Uvin, Lo, Rompalo, Sobel, and Smith (2011) considered an HPV application, but were interested primarily in the duration of HPV infection itself rather than its relationship to incidence of cervical neoplasia. The authors proposed an approach to estimate the "persistence" of HPV infection given panel observations of infection status. Specifically, they modeled the sojourn times in states 1 and 2 (HPV-uninfected and -infected, respectively) as discrete random variables, and developed a maximum likelihood method to estimate the weights, with the assumptions that: (1) all subjects are initially in state 1, (2) the Markov assumption is satisfied when the process is in state 1, and (3) subjects are observed at prespecified, equally-spaced, common visit times (e.g. every six months). They allowed for isolated missing visits, assuming they occurred at random (MAR assumption), but excluded subjects with two or more consecutive missing visits.

(a) (b)

Figure 2.5: State models considered by Mitchell et al. (2011). In the primary method, states (1) and (2) represent the uninfected and infected states respectively, as shown in (a). In the extension, the uninfected state was split into never infected (1*) and previously infected (1), as shown in (b).

Given the assumptions, the sojourn time in state 1 is a geometric random variable with point masses at the scheduled visit times, say $1, 2, \ldots, n_t$, where n_t is the number of possible observed time points after study entry. Let p_{12} denote the associated parameter, so that

the probability of spending t units of time in state 1 before transitioning to state 2 is given by $p_{12} \cdot (1 - p_{12})^{t-1}$ for $t = 1, 2, \ldots$. The sojourn time in state 2 is an unrestricted discrete random variable with point masses at these common scheduled visit times, so that the probability of spending time t in state 2 before transitioning to state is is denoted by $p_{21}(t)$ for $t = 1, 2, \ldots, n_t$ with $\sum_{t=1}^{n_t} p_{21}(t) = 1$. Letting $\mathbf{p} \doteq \{p_{12}; p_{21}(1), \ldots, p_{21}(n_t)\}$ and imposing the above restriction on $p_{21}(1), \ldots, p_{21}(n_t)$, the authors expressed each individual's likelihood contribution as

$$\pi_{y_0,\ldots,y_{n_t}}(\mathbf{p}) \quad = \quad p_{j_0 j_1}(x_1) \cdots p_{j_{m-1} j_m}(x_m) \cdot S_{j_m}(x_m+),$$

where $j_0, j_1, \ldots$ is the sequence of visited states, $x_0 = 0, x_1, x_2, \ldots$ is the sequence of corresponding sojourn times, $y_0, y_1, \ldots$ is the sequence of observed states at each time point, m is the number of states visited by visit n_t, x_m+ is the right-censored time spent in the final state, $S_{J_n}(\cdot)$ is the survival function in state J_n, and $p_{J_{n-1}J_n}(t) = P(J_n = j, X_n = t | J_{n-1} = i)$. Under the MAR assumption, the likelihood for N subjects is given by

$$\mathcal{L}(\mathbf{p}) \quad = \quad \prod_{i=1}^{N} \sum_{y_{n_t} \in \{0,1\}} \cdots \sum_{y_0 \in \{0,1\}} \alpha^i_{y_0,\ldots,y_{n_t}} \cdot \pi_{y_0,\ldots,y_{n_t}}(\mathbf{p}),$$

where $\alpha^i_{y_0,\ldots,y_{n_t}}$ is the indicator that $\{y_0, \ldots, y_{n_t}\}$ is an "admissible" observation in the sense of De Gruttola and Lagakos (1989), given the possibility of missing observations at scheduled visit times. Mitchell et al. maximized the log likelihood over the parameter space via a quasi-Newton algorithm.

Mitchell et al. extended this method to make a distinction between subjects who have not been infected since time zero and those who have been infected and cleared the infection while on the study, to allow for the possibility that previous infection influences the rate by which subjects transition into the infected state. Specifically, they considered the three-state model shown in Figure 2.5b. They assumed that subjects were in state 1^* at time zero, and additionally that both states 1 and 2 were Markov. They extended this method to relax the assumption that state 2 was Markov, but noted that relaxing this assumption for state 1^* would be challenging because times in the first observed state are subject to left

censoring.

Although Mitchell et al. cited neither Turnbull (1976) nor De Gruttola and Lagakos (1989), their primary approach is a self-consistency algorithm, and is very similar to the method of De Gruttola and Lagakos. The present method has advantages such as not imposing distributional assumptions on the sojourn time in the infected state, but has stringent assumptions on the observational scheme: since it models the process as a discrete-time semi-Markov chain, the method in its present form requires that subjects are observed at a common set of evenly-spaced visits. As a result of modeling time discretely, this method is subject to some of the same issues as De Gruttola and Lagakos (1989) is. Mitchell et al. compared their method with that of Kang and Lagakos (2007), and noted that their own method imposes no parametric assumptions or guarantee times on the sojourn time in the infected state. However, the discrete nature of the method, itself, imposes a guarantee time on this sojourn time since there is no point mass at time zero.

In a similar vein, Crespi, Cumberland, and Blower (2005) considered modeling herpes simplex virus type 2 (HSV-2), which is characterized by recurrent lesions. However, the number of lesions is not observable; only the presence or absence of lesions, known as the viral shedding status, can be ascertained. The viral shedding status itself is often asymptomatic and is therefore observed only at clinic visits, giving rise to panel data.

The latent number of lesions at a point in time can be considered a *birth-death process*, a Markov process in which the states, $0, 1, 2, \ldots$ represent the size of the population at each point in time (see Figure 2.6a). Crespi et al. modeled the latent number of lesions as such a process with Markov kernel

$$Q_{ij}(t) = \begin{cases} \lambda, & j = i + 1,\ i = 0, 1, 2, \ldots; \\ \mu, & j = i - 1,\ i = 1, 2, 3, \ldots; \\ 0, & \text{otherwise}, \end{cases}$$

for all $t \geq 0$, for some $\lambda, \mu > 0$, where λ represents the rate at which lesions are formed, and μ represents the rate at which they are cleared. Implicit in this model is the assumption that lesions form independently of each other. If states $1, 2, 3, \ldots$ of this homogeneous Markov

process are collapsed into a single state, denoted 1+, the corresponding process, denoted $Z(\cdot)$, is semi-Markov, as the sojourn time in state 1+ now depends on the elapsed time in this state (Figure 2.6b). While the distribution of the sojourn time in the non-shedding state is exponential with rate λ, the sojourn time in the shedding state does not follow a familiar distribution. The observed viral shedding status at each point in time is an induced semi-Markov model.

(a) $\boxed{0} \rightleftarrows \boxed{1} \rightleftarrows \boxed{2} \rightleftarrows \cdots$ hidden Markov process, $W(\cdot)$

(b) $\boxed{0} \rightleftarrows \boxed{1+}$ semi-Markov process, $Z(\cdot)$

Figure 2.6: State models considered by Crespi et al. (2005). The unobservable number of recurrences at time t is modeled as a birth-death process (a), while the viral shedding status is modeled as the corresponding semi-Markov process (b) defined by collapsing states $1, 2, 3, \ldots$ of the birth-death process.

Crespi et al. express the panel data likelihood via a so-called hidden Markov model approach and carry out inference on the parameters in a Bayesian framework. Moreover, they use a random effects model to accommodate heterogeneity across individuals, and express the mean of each random effect as a function of covariates. They assumed priors on the covariate parameters and variance terms, derived full conditional distributions, and carried out a Gibbs sampling algorithm to obtain estimates. Posterior distributions of λ and μ make inferences on both the hidden Markov process $W(\cdot)$ and the semi-Markov process $Z(\cdot)$ possible.

Motivated by an application involving bronchiolitis obliterans syndrome (BOS), an irreversible lung disease whose assessment is subject to classification error, Titman and Sharples (2010) propose a particular semi-Markov model for a nonprogressive process that involves an underlying hidden Markov model. Specifically, they assume each observable state of the process is composed of several unobservable sub-states or "phases" that the process goes through in a fixed sequence, and that the resulting latent process is Markov (see Figure 2.7). Since the process defined by the phases is Markov and the phases are not observed directly, this is an example of a *hidden Markov model* (HMM). However, unlike the hidden Markov model of Crespi et al. (2005), in this model the phases do not necessarily correspond to an

unobservable physical process; rather, the partition of each state into phases is a modeling device. The induced process is a semi-Markov process in which the conditional distributions of the sojourn times have a particular structure.

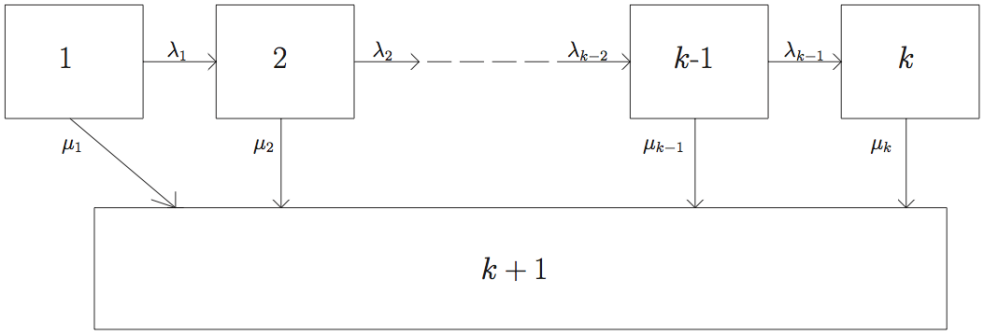


Figure 2.7: General phase-type distribution considered by Titman and Sharples (2010). The upper box represents a single state that is partitioned into k phases.

Titman and Sharples assume that (1) subjects enter each state in phase 1 and (2) there are specific relationships among the transition intensities. The transition intensities between the k phases of a state r are denoted by $\lambda_{r_1}, \ldots, \lambda_{r_{k-1}}$, and the transition intensities from the k phases of state r to state s as $\mu_{r_1 s}, \ldots, \mu_{r_k s}$, as shown in the example in Figure 2.8. Even with these assumptions, the parameters in a particular model are prone to lack of identifiability, and choosing a model to ensure identifiability is not straightforward.

The authors additionally consider the possibility of classification error in the assessment of the state at each observation time. Suppose the underlying multi-state process $Z(\cdot)$ defined by the true states is Markov, and the observed process $O(\cdot)$ is related to $Z(\cdot)$ via the misclassification probabilities $P(O(t) = s|Z(t) = r) = e_{rs}(t)$ for $t \geq 0$, and that $e_{rs}(t) \equiv e_{rs}$. Then the observed process is a hidden Markov process. However, if the process defined by the true states is modeled as the hidden Markov process described above, and classification errors are additionally considered, then the observed process is a *hidden semi-Markov model* (HSMM). The authors extend the original hidden Markov model to a hidden semi-Markov model, assuming that classification errors occur independently, conditional on the true process $Z(\cdot)$. Identifiability issues are exacerbated in this extended model, and

strong assumptions as well as large sample sizes are required in general.

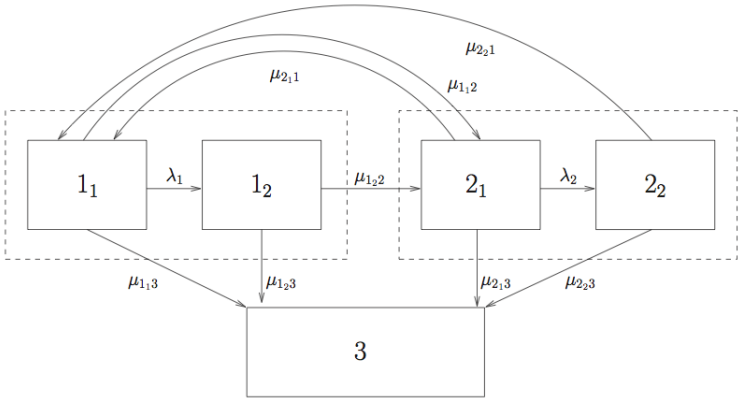


Figure 2.8: State model of latent Markov process and induced phase-type semi-Markov process considered by Titman and Sharples (2010). State 1 represents good health, state 2 is BOS, and state 3 represents death.

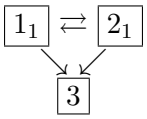


Figure 2.9: State model of Markov process considered by Titman and Sharples (2010).

Titman and Sharples illustrated the extended model on a dataset examining the onset of bronchiolitis obliterans syndrome in patients following lung transplantation. The gold standard for assessing disease status is histological analysis, but forced expiratory volume in one second (FEV-1) is used as a surrogate in practice. FEV-1 is measured at baseline and subsequent measurements are expressed as a percentage thereof. The standard definition of BOS is a drop of 20% or more of FEV-1 from baseline, and measurement error of FEV-1 allows for misclassification of BOS status. Titman and Sharples fit a hidden semi-Markov model allowing for classification error and with $k = 2$ phases for each of the two transient states (see Figure 2.8), assuming $\mu_{1_2 r} = \tau_{1_2}\mu_{1_1 r}$ for $r = 1, 2$ and $\mu_{2_2 r} = \tau_{2_2}\mu_{2_1 r}$ for $r = 1, 3$, for some $\tau_{1_2}, \tau_{2_2} > 0$. They additionally fit a model allowing for classification error and with $k = 1$ phase per transient state, i.e., a hidden Markov model (see Figure 2.9). Though in theory the disease process is progressive, enforcing these assumptions led to relatively poor model fit, so they allowed reversals of the underlying disease process, as shown in the state diagrams. They found, via a modified likelihood ratio test, that the fit of the hidden semi-Markov model was significantly better than that of the hidden Markov model.

2.4 Motivating examples

The methods presented above and those we develop in this dissertation are motivated by studies of intermittently monitored progressive disease. We will illustrate the performance of some of the existing and proposed approaches via two such studies, which we describe next.

2.4.1 HIV infection and progression to AIDS-related symptoms in hemophiliacs receiving blood transfusions.

Development of the method of De Gruttola and Lagakos (1989) was motivated by a retrospective study of a cohort of 262 patients with hemophilia who received transfusions of blood that was later found to be contaminated with HIV. Blood samples were taken at various times from these patients, allowing for intermittent retrospective assessment of HIV infection status. The study population consisted of a cohort of patients with type A or B hemophilia who had received periodic transfusions of HIV-contaminated blood at the Hôpital Kremlin Bicêtre and Hôpital Cœur des Yvelines in France since 1978.

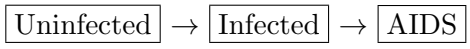


Figure 2.10: State model of De Gruttola and Lagakos (1989).

De Gruttola and Lagakos chose a simple progressive three-state model for this situation, as depicted in Figure 2.10. They split the patients into two groups defined by the amount of blood product they had received, and carried out separate estimation of the times to infection and to progression to AIDS. Of the 262 patients in the cohort, 197 were infected with HIV at the end of follow-up, 43 of whom had progressed to AIDS. All HIV infections were believed to have been caused by receiving contaminated blood. De Gruttola and Lagakos divided the chronological time axis into 6-month intervals, with $Y = 1$ denoting July 1, 1978. Each interval was given a point mass, and the point masses were indexed by $1, 2, 3, \ldots$. De Gruttola and Lagakos presented an incomplete version of the dataset and

graphical results based on applying their proposed method.

In an extension to allow for covariate adjustment, Kim, De Gruttola, and Lagakos (1993) examined an updated version of the same dataset, and redefined the third state, which had previously been progression to AIDS, as the onset of the AIDS-related symptoms: AIDS-related complex, low platelet count, or progression to AIDS itself. Of the 257 hemophilia patients in this updated dataset, 188 were infected with HIV by the end of follow-up, and 41 had progressed to AIDS-related symptoms. Kim et al. (1993) presented the full dataset and results adjusting for heavy versus light treatment as a covariate.

2.4.2 WHO staging of HIV/AIDS among untreated subjects infected with HIV-1 or HIV-2 in Senegal.

Several weeks after infection with HIV, the majority of subjects develop flu-like symptoms, a period known as acute retroviral syndrome. This is followed by a clinical latency period that can last for a number of years in the absence of antiretroviral therapy (World Health Organization, 2007). The end of this latency period is marked by a drop in CD4+ T cell count below a certain threshold and the onset of constitutional symptoms. AIDS is defined by a CD4+ T cell count below 200 cells per μL of blood or the appearance of one or more of a set of diseases associated with HIV infection (Del Rio and Curran, 2009).

It is important to express a measure of a patient's stage of disease, both for the purpose of treating the patient appropriately, and—on a grander epidemiologic scale—tracking the HIV/AIDS pandemic. The World Health Organization (WHO) has a staging system for HIV infection and disease that is based on observed symptoms (World Health Organization, 2007; AETC, 2012):

- primary HIV infection: either asymptomatic or characterized by acute retroviral syndrome
- stage I: asymptomatic; may be characterized by persistent generalized lymphadenopathy
- stage II: mild symptoms (including recurrent respiratory tract infections, mucocutaneous involvement, moderate unexplained weight loss)

- stage III: advanced symptoms (including chronic unexplained diarrhea, persistent unexplained fever, severe unexplained weight loss, severe bacterial infections)
- stage IV (AIDS): severe symtoms (including HIV wasting syndrome, central nervous system toxoplasmosis, pneumonia, Kaposi's sarcoma, lymphoma).

A related staging system involving nine stages, the CDC staging system, involves both CD4+ T cell count and viral load measurements and thus captures more detail. However, the WHO staging system is more useful in practice as it is applicable even in resource-poor settings, where there may be no access to a laboratory (AETC, 2012).

There are two major types of HIV, HIV-1 and HIV-2, both of which are known to cause AIDS. HIV-1 is much more common and pathogenic, whereas HIV-2 is characterized by slower disease progression and is confined mainly to Africa. To gain insight into the biological reasons for the distinct natural histories of these two viruses, researchers studied disease progression among subjects in Senegal who were infected with HIV-1, HIV-2, or both viruses beginning in the 1990s. We have access to longitudinal data on two cohorts, comprised of 611 subjects. For the first cohort, all patients of age 16 years or older who presented at one of three clinics in Senegal between 1994–1998 were offered an HIV test. Those who tested positive for HIV-1 or HIV-2 and who met other criteria were invited to participate in longitudinal studies. For the second cohort, patients who presented at one of two clinics in Senegal between 2000–2005 were offered an HIV test. Those who were found to be positive for HIV-1 or HIV-2 were invited to participate in longitudinal studies. Patients who were positive for HIV-1 only were eligible for inclusion in the study only if their CD4 counts were above some cutoff at screening. Data on subjects in these two cohorts were based on questionnaires administered at baseline and laboratory tests, and were collected by staff at the three clinics in Senegal. Further details are available in Gottlieb et al. (2002).

The 611 subjects contributed information on 4111 clinic visits. Subjects were followed for a mean of 2.5 years after the initial visit, contributing a mean of 6.7 visits. Baseline information on patient demographics, medical and sexual history, virus type, and laboratory values are available. Information on WHO stage of HIV/AIDS, CD4+ T cell count, viral load level, and treatment with antivirals is available at follow-up visits.

The original study (Gottlieb et al., 2002) aimed to examine how CD4 cell count decline was related to baseline viral load, but an analysis of the longitudinal WHO staging information has not been carried out. Existing methods to handle this type of data are not suitable: Sternberg and Satten (1999) assumes the process is simple progressive, while Kalbfleisch and Lawless (1985) assumes exponential sojourn times in each state, and each of these assumptions may not be reasonable. We will apply our proposed approach to the multi-state panel observations of HIV/AIDS stage in this dataset in an attempt to gain further understanding of the differences between the natural histories of HIV-1 and HIV-2.

2.5 Discussion

Existing approaches to modeling progressive disease under panel observation are: (1) to impose the Markov assumption on the process (Kalbfleisch and Lawless, 1985); (2) to model the sojourn times as discrete random variables (De Gruttola and Lagakos, 1989); or (3) to use a flexible parametric model for the sojourn times but to impose strong assumptions on the process or observation scheme (Kang and Lagakos, 2007; Foucher et al., 2007). Though nonparametric approaches have a distinct advantage in terms of flexibility, estimates are subject to lack of identifiability as well as inefficiency. On the other hand, although the method of Kalbfleisch and Lawless is robust and efficient, the Markov assumption is unreasonable in many applications.

There is a need for methods for progressive disease processes that provide a flexible model for sojourn times, yet perform well even for relatively small sample sizes. In the following chapters we develop methods for modeling progressive disease that address some of the limitations of the existing methods. We extend these methods to allow for features commonly encountered in applications. Finally we apply our methods to the WHO staging data of untreated HIV-infected subjects in Senegal.

Chapter 3

A SEMI-MARKOV MODEL FOR SIMPLE PROGRESSIVE DISEASE PROCESSES

3.1 Introduction

In this chapter we propose an approach for modeling a simple progressive state model. As we discussed in the previous chapter, existing methods impose heavy assumptions on the underlying process or on the observational scheme, or model time discretely. In particular, methods commonly impose the Markov assumption, but this assumption is often not appropriate in applications. For example, in the context of HPV infection, the hazard of clearing the infection may decrease as the elapsed time in the infected state grows (Mitchell et al., 2011). Our goal is to develop a method that allows for flexible modeling of the disease process but is not prone to the issues that plague currently available discrete time methods. We first consider an m-state simple progressive disease and propose a parametric modeling approach to model the conditional distribution of the sojourn time in each state. In Section 3.2 we demonstrate how the principle of data augmentation can be used to deal with the unobservable true sojourn times in each state. We discuss possible parametric models for the conditional distributions of the sojourn times in each state. Then in Section 3.3 we discuss Bayesian estimation of the parameters via Markov chain Monte Carlo (MCMC) methods with the Metropolis-Hastings algorithm. In Section 3.4 we compare the performance of existing and proposed approaches for a simple progressive state model with $m = 3$ via simulation study, assuming that subjects were observed until they entered the final state. In Section 3.5 we examine the performance of the proposed approach when right censoring is allowed. In Section 3.6 we illustrate the proposed approach in one of the HIV/AIDS applications. Finally, in Section 3.7, we provide a summary of the performance of the proposed approach and existing methods for modeling a simple progressive state model and guidance on the circumstances under which each may be appropriate.

3.2 Data augmentation

We propose as our primary approach a method that uses data augmentation. For any multi-state process, an intermittent observation scheme potentially creates a missing data problem: the sojourn times in each state, as well as the sequence of states, may not be observed. In the case of a simple progressive process with m states, the sequence of states is known, but the sojourn times in each state are subject to censoring. We can treat the true, unobservable sojourn times as *latent data* (Tanner, 1991) to assist inference about the parameters of interest. This approach to missing data is known as data augmentation. We introduce a particular type of data augmentation and demonstrate how it can be used in the case of an m-state simple progressive model.

A natural question to ask is: how is the introduction of additional parameters into the model helpful? Using a semi-Markov model for the underlying process allows us to model the sojourn time in each state in a flexible manner, but we must ultimately carry out inference about these sojourn times. The observed data contain some information about the sojourn times, though not their exact values. If we knew the sojourn times in each state exactly, then it would be straightforward to carry out inference about the parameters of interest in the model. However, it would be computationally difficult in general to carry out inference about the parameters in the model directly from the observed data. We introduce the unknown values of the sojourn times into the model to serve as a stepping stone from the observed data to the parameters of interest. In the following exposition we make this procedure explicit.

3.2.1 Background: chained data augmentation and extensions.

Let Z denote the observed data generated by a process governed by some parameters of interest $\boldsymbol{\theta}$, where $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^n$. The goal is to make inference about the distribution of $p(\boldsymbol{\theta}|Z)$. Suppose the expression for $p(\boldsymbol{\theta}|Z)$ is very cumbersome, possibly not available in closed form, and difficult to evaluate. Suppose further that if we were able to observe data X, then $p(\boldsymbol{\theta}|Z, X)$ would be in a simpler form that could be evaluated more easily. We approach the problem of estimating $p(\boldsymbol{\theta}|Z)$ by treating the unobservable data X as a

nuisance parameter, which we estimate together with $\boldsymbol{\theta}$ in an iterative scheme described below.

We have the two chained data augmentation equations, the posterior and predictive equations, respectively:

$$\begin{aligned} p(\boldsymbol{\theta}|Z) &= \int_X p(\boldsymbol{\theta}|Z,X)p(X|Z)dX \\ p(X|Z) &= \int_{\boldsymbol{\theta}} p(X|Z,\boldsymbol{\theta})p(\boldsymbol{\theta}|Z)d\boldsymbol{\theta}. \end{aligned}$$

Note that each left-hand side is contained in the integrand of the other equation. Hence, this pair of equations suggests an iterative algorithm to estimate $p(\boldsymbol{\theta}|Z)$. In *data augmentation* as described in Tanner (1991), we iterate between updating $p(\boldsymbol{\theta}|Z)$ (posterior step) and $p(X|Z)$ (imputation step), and each iteration of the imputation step consists of n_c repetitions of generating updated values of θ^* and x^* to get the imputed values $x_1, \ldots, x_{n_c}$. The value n_c is chosen based on practical considerations.

A special case of this algorithm is $n_c = 1$, in which case the algorithm is called *chained data augmentation*. The algorithm now consists of just a single iterative scheme:

1. Given x^*, generate a single $\boldsymbol{\theta}^*$ from $p(\boldsymbol{\theta}|x^*, Z)$.
2. Given $\boldsymbol{\theta}^*$, generate a single x^* from $p(X|\boldsymbol{\theta}^*, Z)$.

Under the following regularity condition, the algorithm converges to the observed posterior $p(\boldsymbol{\theta}|Z)$.

Condition (Tanner and Wong, 1987)**.** *Assume that* $\Theta \subset \mathbb{R}^n$ *is connected and that* $\kappa : \Theta \times \Theta \to \mathbb{R}$ *defined by*

$$\kappa(\boldsymbol{\theta}, \boldsymbol{\varphi}) \doteq \int_X p(\boldsymbol{\theta}|X,Z)p(X|\boldsymbol{\varphi},Z)dX, \quad \boldsymbol{\theta}, \boldsymbol{\varphi} \in \Theta$$

is uniformly bounded and equicontinuous in $\boldsymbol{\theta}$*, and that for any* $\boldsymbol{\theta}_0 \in \Theta$ *there is an open neighborhood* $\mathcal{U}$ *of* $\boldsymbol{\theta}_0$ *such that* $\kappa(\boldsymbol{\theta}, \boldsymbol{\varphi}) > 0$ *for all* $\boldsymbol{\theta}, \boldsymbol{\varphi} \in \mathcal{U}$*.*

The first part of this condition ensures that the family of functions $\kappa(\cdot, \boldsymbol{\varphi})$ over $\boldsymbol{\varphi}$ is smooth in $\boldsymbol{\theta}$ in some sense, while the second part requires that if $\boldsymbol{\theta}$ and $\boldsymbol{\varphi}$ are close, then

it is possible to generate latent data X from $p(X|\boldsymbol{\varphi}, Z)$ such that $p(\boldsymbol{\theta}|X, Z)$ is nonzero. The condition guarantees that the target posterior $p(\boldsymbol{\theta}|Z)$ is unique and that the algorithm converges in law:

$$\boldsymbol{\theta}^{(i)} \to_{\mathcal{L}} \boldsymbol{\theta} \sim p(\boldsymbol{\theta}|Z) \qquad \text{and} \qquad x^{(i)} \to_{\mathcal{L}} X \sim p(X|Z),$$

where $\boldsymbol{\theta}^{(i)}$ and $x^{(i)}$ are the i^{th} samples of the distributions of $\boldsymbol{\theta}$ and x in the above iterative algorithm.

Extension to the multivariate case in which $\boldsymbol{\theta}$ can be partitioned into d independent sub-vectors $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$ is straightforward. The corresponding algorithm is known as the *Gibbs sampler* (Tanner, 1991):

1. Given x^* and $\boldsymbol{\theta}_2^*, \ldots, \boldsymbol{\theta}_d^*$, generate $\boldsymbol{\theta}_1^*$ from $p(\boldsymbol{\theta}_1|\boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_d, X)$.
2. Given x^* and $\boldsymbol{\theta}_1^*, \boldsymbol{\theta}_3^*, \ldots, \boldsymbol{\theta}_d^*$, generate $\boldsymbol{\theta}_2^*$ from $p(\boldsymbol{\theta}_2|\boldsymbol{\theta}_1, \boldsymbol{\theta}_3, \ldots, \boldsymbol{\theta}_d, X)$.

   $\vdots$

d. Given x^* and $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{d-1}^*$, generate $\boldsymbol{\theta}_d^*$ from $p(\boldsymbol{\theta}_d|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{d-1}, X)$.

$d+1$. Given $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_d^*$, generate x^* from $p(X|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_d)$.

Frequently it is not possible to sample from these full conditional distributions directly. In cases where we can express the full conditional distributions in closed form but cannot sample from them easily, we can choose a proposal distribution for each of the $\boldsymbol{\theta}_i$ and use the Metropolis algorithm or the Metropolis-Hastings algorithm (see Gelman et al., 1995, p. 323–326).

3.2.2 Chained data augmentation applied to a simple progressive model.

We now illustrate how chained data augmentation can be used in the case of an m-state simple progressive semi-Markov process (see Figure 3.1). We assume for now that the process enters state 1 at time zero. For each $i = 1, \ldots, m-1$ we let the random variable X_i denote the true sojourn time in state i before proceeding to state $i+1$. Hence $\mathbf{X} = (X_1, \ldots, X_{m-1})$, with $X_i \geq 0$ for each i, denotes the true, unobservable data for a given

subject. We assume an absolutely continuous parametric form for the sojourn time in each state: $X_i \sim f_i$ for $i = 1, \ldots, m-1$, and assume that the densities $f_1, \ldots, f_{m-1}$ collectively depend on the vector of parameters $\boldsymbol{\theta}$.

$$\boxed{1} \rightarrow \boxed{2} \rightarrow \cdots \rightarrow \boxed{m}$$

Figure 3.1: An m-state simple progressive state model.

We observe the process periodically, giving rise to panel observations $\mathbf{Z} = (Z_0 = 1, Z_1, \ldots, Z_n)$, with $Z_i \in \{1, \ldots, m\}$ for each i, corresponding to observation times $\mathbf{s} = (s_0 = 0, s_1, \ldots, s_n)$. Because we are considering a restricted state model, we have $Z_0 \leq \cdots \leq Z_n$. Hence, $\mathbf{Z}$ clearly contains redundant information, and can be expressed equivalently as the vector $\mathbf{t} = (t_1, \ldots, t_{2(m-1)})$, where

$$\begin{aligned} t_1 &= \max\{s_k : Z_k = 1\} \\ t_2 &= \min\{s_k : Z_k = 2\} \\ t_3 &= \max\{s_k : Z_k = 2\} \\ &\vdots \\ t_{2(m-1)} &= \min\{s_k : Z_k = m\}, \end{aligned}$$

if each of these exists. Due to censoring, some elements of this vector may not exist. For example, suppose that $m = 4$ and a given subject is observed in states $\mathbf{z} = (1, 1, 1, 3, 3)$ at times $\mathbf{s} = (0, 1, 2, 3, 4)$. Then the observed states may be expressed as the vector $\mathbf{t} = (2, NA, NA, 3, 3, NA)$. Since this subject was not observed in states 2 or 4, the corresponding elements of $\mathbf{t}$ are not defined.

With the above notation, the posterior and predictive equations become

$$p(\boldsymbol{\theta}|\mathbf{t}) = \int_{\mathbf{X}} p(\boldsymbol{\theta}|\mathbf{t}, \mathbf{X}) p(\mathbf{X}|\mathbf{t}) d\mathbf{X} \tag{3.1}$$

$$p(\mathbf{X}|\mathbf{t}) = \int_{\boldsymbol{\theta}} p(\mathbf{X}|\mathbf{t}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{t}) d\boldsymbol{\theta}. \tag{3.2}$$

From (3.1)–(3.2), the goal is to obtain $p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{t})$ and $p(\mathbf{X}|\mathbf{t}, \boldsymbol{\theta})$. The particular form of each of these expressions depends on the choice of the model for the sojourn time distributions in each state. We examine several choices in the next subsection.

3.2.3 Parametric forms of the sojourn time distribution in each state.

In the framework described above, any choice of parametric model for the sojourn time distributions in each state can be made, in theory. The choice should be made based on scientific reasoning about the particular application and any previous studies or expert opinion. In practice, the number of subjects in the study and the frequency of observation should also be taken into account. For example, if the number of subjects is small and the observation scheme is sparse, choosing a sojourn time distribution with many parameters would likely lead to lack of identifiability of the parameters.

Two parametric forms that are commonly used in survival analysis to describe a time-to-event are the exponential and Weibull distributions, the former being a special case of the latter. Specifically, if $X \sim We(k, \theta)$ with $k, \theta > 0$, then

$$f_X(x) = \frac{k}{\theta}\left(\frac{x}{\theta}\right)^{k-1} \cdot \exp\left(-\left(\frac{x}{\theta}\right)^{k}\right)$$

for $x \geq 0$. The parameter k is known as the *shape* parameter while θ is called the *scale* parameter. This distribution reduces to the exponential distribution, governed solely by scale parameter θ, for $k = 1$. The hazard function corresponding to the Weibull distribution is given by

$$h_X(x) = \frac{k}{\theta}\left(\frac{x}{\theta}\right)^{k-1}$$

for $x \geq 0$; that is, the Weibull hazard spans the set of power functions with power greater than -1. The Weibull distribution is the natural choice for a more flexible alternative to the exponential distribution, for which the hazard is constant.

Other parametric forms can be chosen for each X_i. For example, if X arises from the exponentiated Weibull distribution (Mudholkar and Srivastava, 1993), then the cumulative

distribution function of X is given by

$$F_X(x) = \left[1 - \exp\left(-\left(\frac{x}{\theta}\right)^k\right)\right]^a$$

for $x \geq 0$, where $k > 0$ and $a > 0$ are the first and second shape parameters respectively and $\theta > 0$ is the scale parameter. The corresponding density is

$$f_X(x) = a \cdot \left[1 - \exp\left(-\left(\frac{x}{\theta}\right)^k\right)\right]^{a-1} \cdot \exp\left(-\left(\frac{x}{\theta}\right)^k\right) \cdot \frac{k}{\theta}\left(\frac{x}{\theta}\right)^{k-1}.$$

The standard Weibull distribution is obtained as the special case where $a = 1$. The exponentiated Weibull distribution provides a quite flexible framework for modeling sojourn times as its hazard function is not necessarily monotonic: as the second shape parameter varies, the hazard function varies from the bathtub shape to the unimodal shape (Mudholkar and Srivastava, 1993). Hence, the exponentiated Weibull form is a reasonable choice for modeling a non-monotonic hazard. A random variable that follows this distribution will be denoted $X \sim expWe(k, \theta, a)$ in this dissertation.

For additional flexibility, we can model the sojourn time in each state with a spline function. We follow the approach of Alvarez (2005), who modeled the hazard of continuously observed sojourn times in a multi-state model via a spline function. Specifically, he expressed the log hazard function as a linear spline function of log-transformed time with data-dependent knots. Given $n_k \geq 1$ knots at ordered quantiles of the data on the original scale, $q_1, \ldots, q_{n_k}$, Alvarez modeled $\log h(\cdot)$ as a piecewise linear function $w(\cdot)$ of the transformed time scale:

$$\begin{aligned} w(x) \doteq \log h(x) \quad = \quad & w_1 + b_1 \cdot (\log x - \log q_1) \cdot \mathbf{1}(\log x < \log q_1) \\ & + \sum_{i=1}^{n_k} b_{i+1} \cdot (\min\{\log x, \log q_{i+1}\} - \log q_i) \cdot \mathbf{1}(\log x \geq \log q_i) \end{aligned}$$

for $x > 0$, where $b_1, \ldots, b_{n_k+1}$ give the slopes of the line segments defined by the knots, w_1 is the value of $w(\cdot)$ at the first knot q_1, and $q_{n_k+1} \doteq \infty$. As we will see, we must impose several conditions on the parameters. For the density function to be non-negative everywhere and

integrate to 1, the cumulative hazard function, $H(x) = \int_0^x h(s)ds$, must satisfy the following conditions:

$$H(x) < \infty \quad \text{for all } x \in (0, \infty); \tag{3.3}$$

$$\lim_{x\to\infty} H(x) = \infty. \tag{3.4}$$

The first condition (3.3) implies in particular that $H(q_1) < \infty$. We compute

$$\begin{aligned} H(q_1) &= \int_0^{q_1} h(s)ds = \int_0^{q_1} \exp(w(s))ds = \int_0^{q_1} \exp(w_1 + b_1 \cdot (\log s - \log q_1))ds \\ &= \frac{\exp(w_1)}{q_1^{b_1}} \cdot \int_0^{q_1} s^{b_1} ds, \end{aligned}$$

which is finite if and only if $b_1 > -1$. The second condition (3.4), combined with the first, implies in particular that $\lim_{x\to\infty}[H(x)] - H(q_{n_k}) = \infty$. We let $w_i \doteq w(q_i)$ for $i = 2, \ldots, n_k$; note that w_1 is already so defined. Then

$$\begin{aligned} \lim_{x\to\infty}[H(x) - H(q_{n_k})] &= \int_{q_{n_k}}^{\infty} h(s)ds = \int_{q_{n_k}}^{\infty} \exp(w(s))ds \\ &= \int_{q_{n_k}}^{\infty} \exp(b_{n_k+1} \cdot \log s + w_{n_k} - b_{n_k+1} \cdot q_{n_k})ds \\ &= \exp(w_{n_k} - b_{n_k+1} \cdot q_{n_k}) \cdot \int_{q_{n_k}}^{\infty} s^{b_{n_k+1}} ds, \end{aligned}$$

which is unbounded above if and only if $b_{n_k+1} \geq -1$. A random variable X that follows such a distribution will be denoted in this dissertation by $X \sim spl(w_1, \mathbf{b})$.

There are certain advantages to this choice of parameterization of the hazard function in defining the spline model. If we were to model the hazard function on the original scale, some parameter values would correspond to hazard functions with negative values, thus requiring additional parameter constraints. Hence, modeling the log-transformed hazard avoids these further constraints. Additionally dealing with time on the log-transformed scale means that the linear spline model includes the Weibull hazard function as a special case. In the general case, however, this linear spline model corresponds to a wide variety of shapes of the hazard function.

3.3 Bayesian approach

In this section we illustrate how we can use a version of chained data augmentation as described previously to carry out inference about the parameters governing the sojourn times in each of the $m-1$ states of a simple progressive model. To make the algorithm explicit, we must first choose the parametric model for the sojourn time in each state. Since the goal is to make inference about the posterior distribution of the parameters of interest, we must also choose prior distributions for these parameters.

As we will see, our proposed approach has a number of advantages: (1) it is able to accommodate any chosen parametric model for the sojourn times, including spline models; (2) it is able to incorporate expert opinion on parameters; and (3) it does not rely on asymptotic results, providing valid inference even with small sample sizes.

We examine four parametric models for the sojourn times in each state: exponential, Weibull, exponentiated Weibull, and linear spline. For each of these models we describe the chained data augmentation algorithm in detail. In particular, to carry out the algorithm in each case, we must derive the full conditional distributions of the parameters of interest as well as the latent data.

For simplicity of presentation we assume the same parametric model for each of the $m-1$ states in each case, though this is not necessary in practice. Also for simplicity we take $m = 3$ in the presentation that follows; extension to the general case is conceptually straightforward. In the following development we suppose that each of N subjects is observed periodically, yielding a set of panel observations. Observations on the n^{th} subject can be summarized as the vector $\mathbf{t}^n = (t_1^n, t_2^n, t_3^n, t_4^n)$. We assume that observations occur independently of the underlying disease process, and that subjects are censored independently of the underlying disease process. We do not require the observation times to be evenly spaced or common across subjects. Importantly, we do not assume that subjects are observed in every visited state.

Given this framework, with three states, there are four possible types of observations. Let $\delta^n(i)$ be the indicator that the n^{th} subject was *not* observed in state i for $i = 2, 3$. That is, $\delta^n(2)$ indicates that the subject was not seen in state 2, though he may have visited this

state, while $\delta^n(3)$ indicates that the subject was right–censored before entering state 3. The possible observation types on subject n are shown in Table 5.3.

	$\delta^n(3) = 0$	$\delta^n(3) = 1$
$\delta^n(2) = 0$	{1,2,3}	{1,2}
$\delta^n(2) = 1$	{1,3}	{1}

Table 3.1: Set of observed states as a function of censoring indicators $\delta^n(2)$ (state 2 unobserved) and $\delta^n(3)$ (state 3 unobserved).

For an m-state simple progressive model, there are 2^{m-1} possible types of observations, which can be represented by assigning an indicator to each of the states $2, \ldots, m$ that the state was unobserved. Note, however, that these indicators do not comprise additional data regarding each subject; all information on the n^{th} subject is encapsulated in $\mathbf{t}^n$, and the indicators can be derived from $\mathbf{t}^n$.

3.3.1 Exponential model for the sojourn times.

We first assume an exponential form for each of the sojourn times: $X_i \sim exp(\theta_i)$ for $i = 1, 2$, where θ_i is the scale parameter. Hence $\boldsymbol{\theta} = (\theta_1, \theta_2)$ is the vector of parameters of interest, and X_1^n, X_2^n are the latent data, $\mathbf{t}^n$ are the "sufficient data", and $\delta^n(2)$ and $\delta^n(3)$ are the unobserved state indicators for the n^{th} subject as described before. We assume that $\theta_1, \theta_2 > 0$ are *a priori* independent with prior densities $\pi(\theta_1)$ and $\pi(\theta_2)$. We now derive the full conditional distributions of the parameters of interest $\boldsymbol{\theta}$ as well as the latent data $\mathbf{X}^n = (X_1^n, X_2^n)$, $n = 1, \ldots, N$.

We begin with the full conditional for θ_1. Noting first that θ_1 depends on the other parameters and observed data only through the latent data corresponding to the first sojourn

time, X_1^n for $n = 1, \ldots, N$, then by applying Bayes' rule:

$$\begin{aligned} p(\theta_1|\theta_2, \mathbf{X}^1, \ldots, \mathbf{X}^N, \mathbf{t}^1, \ldots, \mathbf{t}^N) &\propto p(\theta_1|X_1^1, \ldots, X_1^N) \\ &\propto \left[\prod_{n=1}^{N} p(X_1^n|\theta_1)\right] \cdot \pi(\theta_1) \\ &\propto \frac{1}{\theta_1^N} \cdot \exp\left(-\frac{1}{\theta_1} \cdot \sum_{n=1}^{N} X_1^n\right) \cdot \pi(\theta_1). \end{aligned}$$

The prior distribution of θ_1 can be chosen considering the particular application. In our simulation studies, we use a relatively noninformative prior in an effort to reduce the impact of the prior distribution. If $\pi(\theta_1) \sim Unif(a_1, b_1)$, the full conditional for θ_1 becomes

$$p(\theta_1|X_1^1, \ldots, X_1^N) \propto \frac{1}{\theta_1^N} \cdot \exp\left(-\frac{1}{\theta_1} \cdot \sum_{n=1}^{N} X_1^n\right) \cdot \mathbf{1}(\theta_1 \in (a_1, b_1)).$$

Similarly, the full conditional for θ_2 with $\pi(\theta_2) \sim Unif(a_2, b_2)$ is given by

$$p(\theta_2|X_2^1, \ldots, X_2^N) \propto \frac{1}{\theta_2^N} \cdot \exp\left(-\frac{1}{\theta_2} \cdot \sum_{n=1}^{N} X_2^n\right) \cdot \mathbf{1}(\theta_2 \in (a_2, b_2)).$$

We turn now to the full conditional for $\mathbf{X}^n = (X_1^n, X_2^n)$. Note that for a given subject n, X_1^n and X_2^n are not independent parameters, as their sum is constrained. For example, considering a fully observed subject with $\mathbf{t}^n = (2, 3, 5, 6)$, we must have $X_1^n \in [2, 3]$ and $X_1^n + X_2^n \in [5, 6]$. For a subject whose sojourn in state 2 was censored, we may observe $\mathbf{t}^n = (1, NA, NA, 2)$, in which case $X_1^n \in [1, 2]$ and $X_1^n + X_2^n \in [1, 2]$ as well. For fixed $n \in \{1, \ldots, N\}$

$$p(X_1^n, X_2^n|\theta_1, \theta_2, \mathbf{X}^1, \ldots, \mathbf{X}^{n-1}, \mathbf{X}^{n+1}, \ldots, \mathbf{X}^N, \mathbf{t}^1, \ldots, \mathbf{t}^N) \propto p(X_1^n, X_2^n|\theta_1, \theta_2, \mathbf{t}^n).$$

The form of this full conditional depends on the nature of $\mathbf{t}^n$, or equivalently, on the indicators $\delta^n(2)$ and $\delta^n(3)$. We have:

$$p(X_1^n, X_2^n|\theta_1, \theta_2, \mathbf{t}^n) = \frac{1}{\theta_1} \cdot \exp\left(-\frac{X_1^n}{\theta_1}\right) \cdot \frac{1}{\theta_2} \cdot \exp\left(-\frac{X_2^n}{\theta_2}\right) \cdot \mathbf{1}((X_1^n, X_2^n) \in A^n),$$

where $A^n = A^n(\boldsymbol{\delta}^n; \mathbf{t}^n)$ is determined by the observation type and observed data. Specifically:

1. If $(\delta^n(2), \delta^n(3)) = (0, 0)$, then A^n is the parallelogram region in the (X_1, X_2)-plane determined by $t_1 \leq X_1 \leq t_2$ and $t_3 \leq X_1 + X_2 \leq t_4$.
2. If $(\delta^n(2), \delta^n(3)) = (1, 0)$, then A^n is the triangular region in the (X_1, X_2)-plane determined by $t_1 \leq X_1 \leq t_4$ and $t_1 \leq X_1 + X_2 \leq t_4$ with $X_2 \geq 0$.
3. If $(\delta^n(2), \delta^n(3)) = (0, 1)$, then A^n is the unbounded region in the (X_1, X_2)-plane determined by $t_1 \leq X_1 \leq t_2$ and $X_1 + X_2 \geq t_3$ with $X_2 \geq 0$.
4. If $(\delta^n(2), \delta^n(3)) = (1, 1)$, then A^n is the unbounded region in the (X_1, X_2)-plane determined by $X_1 \geq t_1$ and $X_2 \geq 0$.

See Figure 3.2 for a graphical view of these regions. In the above, the dependencies of $\mathbf{t}^n$ and (X_1^n, X_2^n) on n have been suppressed for clarity. In the general case of m states, the 2^{m-1} allowable regions are subsets of $\mathbb{R}^{m-1}$. Although it is straightforward to write the inequalities that determine these regions for general m, it is challenging to represent them graphically for $m > 4$.

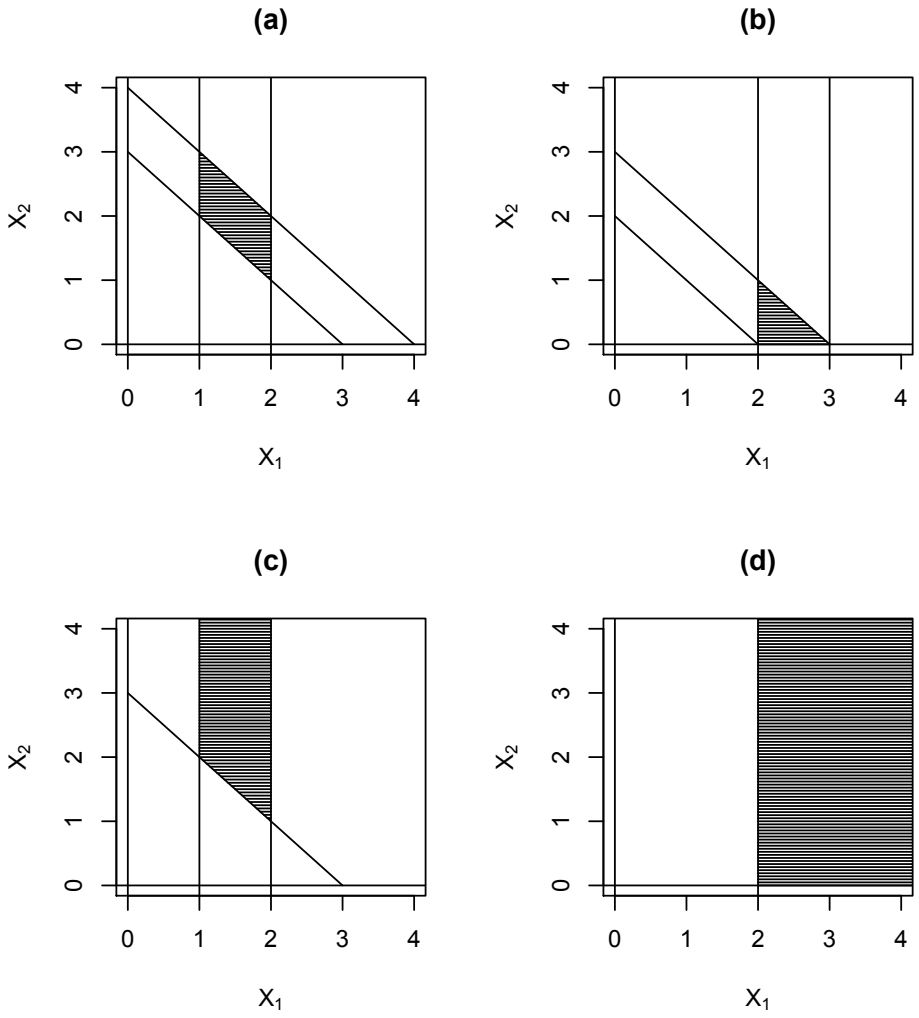


Figure 3.2: Sets of allowable pairs of true sojourn times (X_1^n, X_2^n) given sufficient data $\mathbf{t}^n$ on a subject. The form of the allowable region A^n depends on the observation type, summarized by $(\delta^n(2), \delta^n(3))$. In (a), the region for a fully observed subject is shown; this subject has sufficient data $\mathbf{t}^n = (1, 2, 3, 4)$, so that $(\delta^n(2), \delta^n(3)) = (0, 0)$. (b) corresponds to a subject with $\mathbf{t}^n = (2, NA, NA, 3)$ and $(\delta^n(2), \delta^n(3)) = (1, 0)$ whose sojourn in state 2 was unobserved. (c) shows the region for a subject with $\mathbf{t}^n = (1, 2, 3, NA)$ and $(\delta^n(2), \delta^n(3)) = (0, 1)$ whose sojourn in state 2 was right censored. (d) corresponds to a subject with $\mathbf{t}^n = (2, NA, NA, NA)$ and $(\delta^n(2), \delta^n(3)) = (1, 1)$ whose sojourn in state 1 was right censored.

Since we now have the full conditional distributions for the parameters of interest (θ_1, θ_2)

as well as for the latent data (X_1^n, X_2^n) for each n, we can proceed with carrying out the algorithm. Conceptually, we would like to choose an initial value for (θ_1, θ_2) and iteratively update the parameters of interest and latent data by sampling from the full conditionals until convergence, but for several reasons it may be difficult to generate values directly from these full conditionals.

Under the uniform prior chosen here, the full conditionals for θ_1 and θ_2 follow the Inverse Gamma distribution, so we could update these parameters using Gibbs sampling. However, for a general choice of priors, it may be difficult to sample directly from the full conditionals for θ_1 and θ_2, so in general one may use a Metropolis-Hastings step to update those parameters. In the present case, we use a Metropolis-Hastings step to update θ_1 and θ_2. Since these parameters are positive, we let the proposal distribution J be a normal distribution truncated at zero, with mean equal to the current parameter value. That is,

$$J(\theta_i) = \mathcal{N}_{trunc}(\theta_i, \sigma^2, 0)$$

for $i = 1, 2$ and for some $\sigma > 0$. To update $\theta_i^{(n-1)}$, the current value of θ_i, we generate a candidate value θ_i^* by taking a draw from the proposal J: $\theta_i^* \sim J(\theta_i^{(n-1)})$, and compute

$$r = \frac{p(\theta_i^*|X_1^i, \ldots, X_N^i)}{p(\theta_i^{(n-1)}|X_1^i, \ldots, X_N^i)} \Bigg/ \frac{p_J(\theta_i^*|\theta_i^{(n-1)})}{p_J(\theta_i^{(n-1)}|\theta_i^*)} .$$

The numerator of this fraction is the ratio of the density of the full conditional distribution at the current and candidate values of θ_i, while the denominator is the ratio of the density of the proposal distribution at these values. At the n^{th} step,

$$\theta_i^{(n)} = \begin{cases} \theta_i^*, & \text{with probability } \min\{r, 1\}; \\ \theta_i^{(n-1)}, & \text{else.} \end{cases}$$

As for the latent data, since the constrained support of (X_1^n, X_2^n) for each n makes direct sampling difficult, we use a Metropolis-Hastings step to update these parameters for each subject. The choice of proposal distribution for (X_1^n, X_2^n) given the observed data $\mathbf{t}^n$

should be appropriate for the observation type, and should allow for efficient sampling. For subjects observed in state 3 ($\delta^n(3) = 0$), the support of (X_1^n, X_2^n), A^n, is bounded, so a uniform proposal distribution is reasonable. We carry this out separately for fully observed subjects (Figure 3.2(a)) and partially observed subjects (Figure 3.2(b)):

1. If $(\delta^n(2), \delta^n(3)) = (0, 0)$, sample $X_1 \sim Unif(t_1, t_2)$ and $X_2|X_1 \sim Unif(t_4 - X_1, t_3 - X_1)$.
2. If $(\delta^n(2), \delta^n(3)) = (1, 0)$, sample $\frac{X_1 - t_1}{t_4 - t_1} \sim Beta(1, 2)$ and $X_2|X_1 \sim Unif(0, t_4 - X_1)$.

That is, in the second case, X_1 is sampled from a shifted beta distribution. We have again suppressed the dependencies of X_1^n, X_2^n, and $\mathbf{t}^n$ on n for clarity.

For subjects not observed in state 3 ($\delta^n(3) = 1$), the support of (X_1^n, X_2^n) is unbounded, we choose a shifted exponential proposal distribution with scale ξ that can be chosen appropriately. We carry this out separately for subjects censored in state 1 (Figure 3.2(c)) and state 2 (Figure 3.2(d)):

3. If $(\delta^n(2), \delta^n(3)) = (0, 1)$, sample $X_1 \sim Unif(t_1, t_2)$ and $X_2|X_1 \sim (t_3 - X_1) + \exp(\xi)$.
4. If $(\delta^n(2), \delta^n(3)) = (1, 1)$, sample $X_1 \sim t_1 + \exp(\xi)$ and $X_2|X_1 \sim \exp(\xi)$.

We consider right censoring in Section 3.5.

3.3.2 Weibull model for the sojourn times.

Here we assume a Weibull model for the sojourn times in states 1 and 2: $X_i \sim We(k_i, \theta_i)$ for $i = 1, 2$. Hence $(k_1, \theta_1, k_2, \theta_2)$ are now the parameters of interest, the latent data are (X_1^n, X_2^n), and observed data are as before for $n = 1, \ldots, N$. We assume that $k_1, \theta_1, k_2, \theta_2 > 0$ are *a priori* independent with priors $\pi(k_1)$, $\pi(\theta_1)$, $\pi(k_2)$, and $\pi(\theta_2)$. In this subsection we present the full conditional distributions for parameters of interest and latent data under the Weibull model.

We begin with the full conditional for (k_1, θ_1). Proceeding as before:

$$p(k_1, \theta_1 | k_2, \theta_2, \mathbf{X}^1, \ldots, \mathbf{X}^N, \mathbf{t}^1, \ldots, \mathbf{t}^N)$$

$$\begin{aligned}
&\propto \quad p(k_1, \theta_1 | X_1^1, \ldots, X_1^N) \\
&\propto \quad \left[\prod_{n=1}^{N} p(X_1^i | k_1, \theta_1) \right] \cdot \pi(k_1) \cdot \pi(\theta_1) \\
&\propto \quad \left[\frac{k_1^N}{\theta_1^{Nk_1}} \cdot \prod_{n=1}^{N} (X_1^i)^{k_1 - 1} \cdot \exp\left(-\frac{1}{\theta_1^{k_1}} \cdot \sum_{n=1}^{N} (X_1^n)^{k_1} \right) \right] \cdot \pi(k_1) \cdot \pi(\theta_1).
\end{aligned}$$

Similarly, the full conditional for (k_2, θ_2) is given by

$$p(k_2, \theta_2 | X_2^1, \ldots, X_2^N) \quad \propto \quad \left[\frac{k_2^N}{\theta_2^{Nk_2}} \cdot \prod_{n=1}^{N} (X_2^n)^{k_2 - 1} \cdot \exp\left(-\frac{1}{\theta_2^{k_2}} \cdot \sum_{n=1}^{N} (X_2^n)^{k_2} \right) \right] \cdot \pi(k_2) \cdot \pi(\theta_2).$$

The full conditionals for the latent data for each subject (X_1^n, X_2^n) are:

$$\begin{aligned}
p(X_1^n, X_2^n | k_1, \theta_1, k_2, \theta_2, \mathbf{t}^n) \quad \propto \quad & \frac{k_1}{\theta_1} \cdot \left(\frac{X_1^n}{\theta_1} \right)^{k_1 - 1} \cdot \exp\left(- \left(\frac{X_1^n}{\theta_1} \right)^{k_1} \right) \\
& \cdot \frac{k_2}{\theta_2} \cdot \left(\frac{X_2^n}{\theta_2} \right)^{k_2 - 1} \cdot \exp\left(- \left(\frac{X_2^n}{\theta_2} \right)^{k_2} \right) \\
& \cdot \mathbf{1}((X_1^n, X_2^n) \in A^n),
\end{aligned}$$

where A^n is defined by $\mathbf{t}^n$ exactly as we found previously in Section 3.3.1.

To carry out inference we implement a Metropolis-Hastings algorithm, choosing proposal distributions for (k_1, θ_1) and (k_2, θ_2) and the latent data and prior distributions on the parameters of interest. We set initial values for (k_1, θ_1) and (k_2, θ_2) and iteratively update the parameters of interest and latent data until convergence.

Specifically, since k_i and θ_i are positive parameters for $i = 1, 2$, we choose a truncated normal proposal for each of these parameters:

$$J(k_i) = \mathcal{N}_{trunc}(k_i, \sigma_1^2, 0) \qquad \text{and} \qquad J(\theta_i) = \mathcal{N}_{trunc}(\theta_i, \sigma_2^2, 0)$$

for $i = 1, 2$. We generate a candidate pair of parameters, (k_i^*, θ_i^*), and compute

$$r = \frac{p(k_i^*, \theta_i^* | X_1^i, \ldots, X_N^i)}{p(k_i^{(n-1)}, \theta_i^{(n-1)} | X_1^i, \ldots, X_N^i)} \Bigg/ \frac{p_J(k_i^*, \theta_i^* | k_i^{(n-1)}, \theta_i^{(n-1)})}{p_J(k_i^{(n-1)}, \theta_i^{(n-1)} | k_i^*, \theta_i^*)},$$

and accept the candidate (k_i^*, θ_i^*) as $(k_i^{(n)}, \theta_i^{(n)})$ with probability $\min\{r, 1\}$.

We update the latent data (X_1^n, X_2^n) for each subject exactly as we did in the exponential case.

3.3.3 Exponentiated Weibull model for the sojourn times.

We now assume $X_i \sim expWe(k_i, \theta_i, a_i)$ for $i = 1, 2$. The parameters of interest are $(k_1, \theta_1, a_1, k_2, \theta_2, a_2)$, and as before we assume they are *a priori* independent. The full conditional for (k_1, θ_1, a_1) is given by

$$\begin{aligned}
& p(k_1, \theta_1, a_1 | k_2, \theta_2, a_2, \mathbf{X}^1, \ldots, \mathbf{X}^N, \mathbf{t}^1, \ldots, \mathbf{t}^N) \\
\propto\ & p(k_1, \theta_1, a_1 | X_1^1, \ldots, X_1^N) \\
\propto\ & \left[\prod_{n=1}^{N} p(X_1^i | k_1, \theta_1, a_1) \right] \cdot \pi(k_1) \cdot \pi(\theta_1) \cdot \pi(a_1) \\
\propto\ & a_1^N \cdot \prod_{n=1}^{N} \left[1 - \exp\left(- \left(\frac{X_1^n}{\theta_1} \right)^{k_1} \right) \right]^{a_1 - 1} \cdot \left[\frac{k_1^N}{\theta_1^{N k_1}} \cdot \prod_{n=1}^{N} (X_1^n)^{k_1 - 1} \cdot \exp\left(- \frac{1}{\theta_1^{k_1}} \cdot \sum_{n=1}^{N} (X_1^n)^{k_1} \right) \right] \\
& \cdot \pi(k_1) \cdot \pi(\theta_1) \cdot \pi(a_1).
\end{aligned}$$

Similarly, the full conditional for (k_2, θ_2, a_2) is given by

$$\begin{aligned}
p(k_2, \theta_2, a_2 | X_2^1, \ldots, X_2^N) \ \propto\ & a_2^N \cdot \prod_{n=1}^{N} \left[1 - \exp\left(- \left(\frac{X_2^n}{\theta_2} \right)^{k_2} \right) \right]^{a_2 - 1} \\
& \cdot \left[\frac{k_2^N}{\theta_2^{N k_2}} \cdot \prod_{n=1}^{N} (X_2^n)^{k_2 - 1} \cdot \exp\left(- \frac{1}{\theta_2^{k_2}} \cdot \sum_{n=1}^{N} (X_2^i)^{k_2} \right) \right] \cdot \pi(k_2) \\
& \cdot \pi(\theta_2) \cdot \pi(a_2).
\end{aligned}$$

The full conditionals for the latent data (X_1^n, X_2^n) are given by

$$\begin{aligned}
p(X_1^n, X_2^n | k_1, \theta_1, a_1, k_2, \theta_2, a_2, \mathbf{t}^n) \ \propto\ & f_1(X_1; k_1, \theta_1, a_1) \cdot f_2(X_2; k_2, \theta_2, a_2) \\
& \cdot \mathbf{1}((X_1^n, X_2^n) \in A^n),
\end{aligned}$$

where $f_i(x)$ is the exponentiated Weibull density of X_i for $i = 1, 2$ and A^n defined as before.

Updating the parameters of interest and latent data in this case is a straightforward

extension of the Weibull case. Since k_i, θ_i, and a_i are positive parameters, we choose a truncated normal proposal for each, and update the latent data exactly as before.

3.3.4 Linear spline model for the sojourn times.

Finally, we assume a linear spline model with n_k knots for the sojourn times in each state: $X_1 \sim spl(w_1^1, \mathbf{b}^1)$ and $X_2 \sim spl(w_1^2, \mathbf{b}^2)$, where $\mathbf{b}^1$ and $\mathbf{b}^2$ are vectors of length $n_k + 1$ with $b_1^1, b_1^2 > -1$ and $b_{n_k+1}^1, b_{n_k+1}^2 \geq -1$ and all other parameters are unconstrained real numbers. Hence $(w_1^1, \mathbf{b}^1, w_1^2, \mathbf{b}^2)$ are the parameters of interest, and the latent and observed data are as before. We assume that the parameters w_1^1, $b_1^1, \ldots, b_{n_k+1}^1$, w_1^2, and $b_1^2, \ldots, b_{n_k+1}^2$ are *a priori* independent with priors $\pi(w_1^1)$, $\pi(b_j^1)$, $\pi(w_1^2)$, and $\pi(b_j^2)$ for $j = 1, \ldots, n_k + 1$. We denote $\pi(w_1^i, \mathbf{b}^i) = \pi(w_1^i) \cdot \pi(b_1^i) \cdots \pi(b_{n_k+1}^i)$ for $i = 1, 2$. We present the full conditionals for these parameters and the latent data under the linear spline model.

We begin with the full conditional for $(w_1^1, \mathbf{b}^1)$. Since the spline model is in terms of the hazard, we express the density of X_1 as $f(x) = h(x) \cdot \exp(-H(x))$ for $x > 0$:

$$\begin{aligned} & p(w_1^1, \mathbf{b}^1 | w_1^2, \mathbf{b}^2, \mathbf{X}^1, \ldots, \mathbf{X}^N, \mathbf{t}^1, \ldots, \mathbf{t}^N) \\ & \quad \propto \quad p(w_1^1, \mathbf{b}^1 | X_1^1, \ldots, X_1^N) \\ & \quad \propto \quad \left[\prod_{n=1}^{N} h(X_1^n | w_1^1, \mathbf{b}^1) \cdot \exp\left(-H(X_1^n | w_1^1, \mathbf{b}^1)\right) \right] \cdot \pi(w_1^1, \mathbf{b}^1). \end{aligned}$$

Because of the piecewise nature of the model for the hazard, $h(x | w_1^1, \mathbf{b}^1)$ is given by

$$\log h(x | w_1^1, \mathbf{b}^1) = \begin{cases} w_1^1 + b_1^1 \cdot (\log x - \log q_1), & x < q_1; \\ w_1^1 + b_2^1 \cdot (\log x - \log q_1), & q_1 \leq x < q_1; \\ w_2^1 + b_3^1 \cdot (\log x - \log q_2), & q_2 \leq x < q_3; \\ \quad \vdots & \\ w_{n_k}^1 + b_{n_k+1}^1 \cdot (\log x - \log q_{n_k}), & q_{n_k} \leq x, \end{cases}$$

where w_j is as defined previously for $j = 2, \ldots, n_k$. Note that the w_js can be computed iteratively via $w_{j+1} = w_j + b_{j+1} \cdot (\log q_{j+1} - \log q_j)$ for $j = 1, \ldots, n_k - 1$. Evaluating

$H(x|w_1^1, \mathbf{b}^1)$ involves integrating $h(x)$ appropriately.

The full conditional for $(w_1^2, \mathbf{b}^2)$ is similar:

$$p(w_1^2, \mathbf{b}^2 | X_2^1, \ldots, X_2^N) \quad \propto \quad \left[\prod_{n=1}^{N} h(X_2^n | w_1^2, \mathbf{b}^2) \cdot \exp\left(-H(X_2^n | w_1^2, \mathbf{b}^2)\right) \right] \cdot \pi(w_1^2, \mathbf{b}^2).$$

Finally, the full conditional for (X_1^n, X_2^n) is

$$p(X_1^n, X_2^n | w_1^1, \mathbf{b}^1, w_1^2, \mathbf{b}^2, \mathbf{t}^n) \quad = \quad p(X_1^n | w_1^1, \mathbf{b}^1) \cdot p(X_2^n | w_1^2, \mathbf{b}^2) \cdot \mathbf{1}((X_1^n, X_2^n) \in A^n),$$

where A^n is the region of the (X_1, X_2)-plane defined by $\mathbf{t}^n$ defined before. We again use the Metropolis–Hastings algorithm to carry out posterior sampling.

Note that for $i = 1, 2$ the parameters b_1^i and $b_{n_k+1}^i$ have restricted support while the others—w_1^i and $b_2^i, \ldots, b_{n_k}^i$—are unrestricted. For the first set of parameters we choose a truncated normal proposal:

$$J(b_1^i) \sim \mathcal{N}_{trunc}(b_1^i, \sigma_1^2, -1) \qquad \text{and} J(b_{n_k+1}^i) \sim \mathcal{N}_{trunc}(b_{n_k+1}^i, \sigma_{n_k+1}^2, -1),$$

and for the second set we choose a normal proposal:

$$J(b_i^i) \sim \mathcal{N}(b_j^i, \sigma_j^2)$$

for $i = 1, 2$ and $j = 2, \ldots, n_k$. The computation of the acceptance probability, r, is analogous to the previous cases, and we update the latent data in the same way.

3.3.5 Alternative approach: integration over latent sojourn times

Finally we note that it is possible to take an alternative approach to handling the true, unobservable sojourn times in each state. In the data augmentation approach we treat the true sojourn times as nuisance parameters. Another possibility, which avoids estimating these many parameters, is to integrate over the true sojourn times. Specifically, we again choose a parametric model for the sojourn times $X_1, \ldots, X_{m-1}$ in each state, but we now express the likelihood of the corresponding parameters $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{m-1}$ given the panel data.

For example, supposing $m = 3$ and letting $\mathbf{t}^n$, $\delta^n(2)$, and $\delta^n(3)$ be defined as above, we can express the panel data likelihood as

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 | \mathbf{t}^1, \ldots, \mathbf{t}^N) &= \prod_{n=1}^{N} \left[\int_{t_3^n}^{t_4^n} \int_{t_1^n}^{t_2^n} f_1(u_1) \cdot f_2(u_2 - u_1) du_1 du_2 \right]^{(1-\delta^n(2)) \cdot (1-\delta^n(3))} \\
&\quad \times \left[\int_{t_3^n}^{t_4^n} \int_{t_1^n}^{u_2} f_1(u_1) \cdot f_2(u_2 - u_1) du_1 du_2 \right]^{\delta^n(2) \cdot (1-\delta^n(3))} \\
&\quad \times \left[\int_{t_3^n}^{\infty} \int_{t_1^n}^{t_2^n} f_1(u_1) \cdot f_2(u_2 - u_1) du_1 du_2 \right]^{(1-\delta^n(2)) \cdot \delta^n(3)} \\
&\quad \times \left[\int_{t_1^n}^{\infty} \int_{t_1^n}^{u_2} f_1(u_1) \cdot f_2(u_2 - u_1) du_1 du_2 \right]^{\delta^n(2) \cdot \delta^n(3)} \\
&= \prod_{n=1}^{N} \left[\int_{t_3^n}^{t_4^n} \int_{t_1^n}^{t_2^n} f_1(u_1) \cdot f_2(u_2 - u_1) du_1 du_2 \right]^{(1-\delta^n(2)) \cdot (1-\delta^n(3))} \\
&\quad \times \left[\int_{t_3^n}^{t_4^n} \int_{t_1^n}^{u_2} f_1(u_1) \cdot f_2(u_2 - u_1) du_1 du_2 \right]^{\delta^n(2) \cdot (1-\delta^n(3))} \\
&\quad \times \left[\int_{t_1^n}^{t_2^n} f_1(u) \cdot S_2(t_3^n - u) du \right]^{(1-\delta^n(2)) \cdot \delta^n(3)} \\
&\quad \times \left[S_1(t_1^n) \right]^{\delta^n(2) \cdot \delta^n(3)}.
\end{aligned}$$

Although these integrals cannot be expressed in closed form in general, it is possible to evaluate them numerically and to optimize the log likelihood function with respect to the parameters.

Writing the likelihood function in the general case of m states is straightforward, though maximizing the corresponding log likelihood becomes extremely computationally intensive as the number of states, m, grows. Due to the increased computational burden relative to our proposed approach, we do not explore this approach further.

3.4 Simulation studies: no right censoring

In this section we examine the performance of the proposed approach and several existing approaches via simulation study. Specifically, in Section 3.4.1 we examine the performance of the proposed data augmentation approach in the case where subjects are observed frequently

relative to the progression rate of the disease process. With relatively rich information on the disease progression, we can consider a variety of models for the sojourn time in each state; otherwise, complicated models may not be feasible. In Section 3.4.2, considering the case in which subjects are sparsely observed, we compare the performance of the proposed data augmentation approach with that of a naïve approach, Kalbfleisch and Lawless (1985), and De Gruttola and Lagakos (1989) under the exponential and Weibull models and under a variety of conditions. In each case we assume that subjects are followed until they enter the final state. We examine the parameter estimates and compare the methods via two derived quantities.

3.4.1 Performance of data augmentation approach when subjects are observed frequently.

Here we consider the case in which the observation scheme is frequent, so that we have relatively detailed information about the progression rate of the underlying disease process. We explore the performance of the proposed data augmentation approach under each choice of model for the sojourn times under several circumstances.

Table 3.2: Scenarios for frequent observation scheme.

Scenario	X_1, X_2
1^*	$exp(2)$
2^*	$We(4, 2)$
3^*	$expWe(2, 5, 0.4)$
4^*	$log\mathcal{N}(0.5, 1)$

Specifically, we simulate data under the scenarios in Table 3.2 and illustrate the performance of the data augmentation approach with the exponential, Weibull, exponentiated Weibull, and spline models for the sojourn times. For simplicity we suppose that each subject is observed at times $t = 0.00, 0.25, 0.50, \ldots$, though the proposed approach does not require this assumption. Hence, if the unit of time is years, subjects would be observed every three months. We assume throughout this dissertation that subjects progress through the states independently of one another, and in this section we assume that subjects enter state 1 at time zero. As stated previously, we do not consider right censoring in this section; that is, we take $\delta^n(3) = 0$ for $n = 1, \ldots, N$. We illustrate the performance of the proposed approach under these circumstances for $N = 400$ subjects.

For Scenario 3*, informative priors were needed for the exponentiated Weibull model to ensure convergence. The linear spline model is implemented with two knots placed at crude estimates of quantiles of the true sojourn times.

Since presenting estimates of the parameters for each scenario and model does not shed much light on how the proposed approach with each model performs, we instead present a graphical representation of how the models perform in each of the four scenarios. In Figures 3.3–3.4 we show the true and estimated hazard functions corresponding to X_1 and X_2 from each model in each scenario. Though not a quantitative measure, these plots give an intuition for how well each of the models performs in each scenario under consideration. In particular, the plots reveal the inadequacy of the exponential model under model misspecification (Scenarios 2*–4*). Though the Weibull model is fairly flexible, as expected, the corresponding hazard cannot accommodate non-monotonic hazards, as we can see in Scenarios 3* and 4*, when the hazard is bathtub-shaped and unimodal, respectively. The exponentiated Weibull and the linear spline model with two knots performed well in all scenarios, though the latter was more stable. Both of these models were able to accommodate the variety of hazard shapes explored in these scenarios.

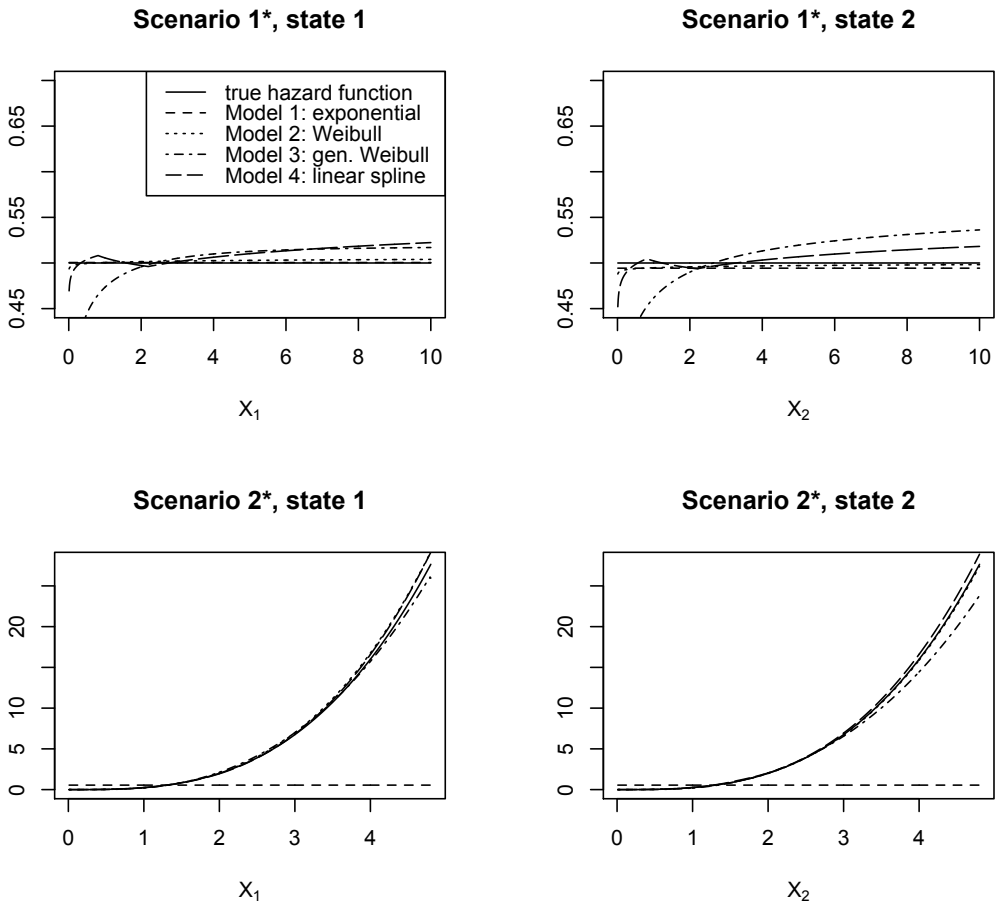


Figure 3.3: True and estimated hazard functions for Scenarios 1* (top panels) and 2* (bottom panels). Results for the sojourn times in state 1 (left panels) and state 2 (right panels) are shown. In each plot, the solid line represents the true hazard, and the other lines show the hazard estimated by the data augmentation approach assuming each of the four models: exponential, Weibull, exponentiated Weibull, and linear spline.

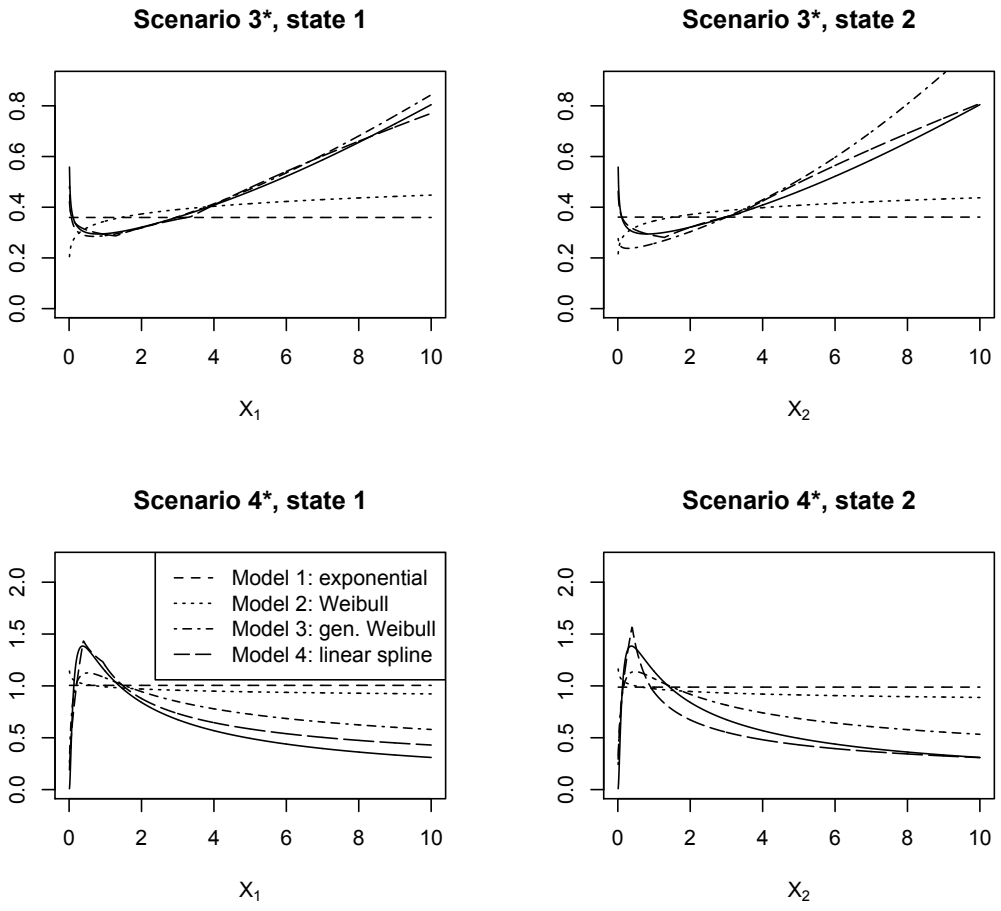


Figure 3.4: True and estimated hazard functions for Scenarios 3* (top panels) and 4* (bottom panels).

3.4.2 Comparison of all methods when subjects are observed infrequently.

To illustrate the performance of the existing and proposed approaches, we simulate data under several scenarios and examine the estimates of the parameters and a specific measure of interest. If the application under consideration were an infection that can progress to disease, then states 1–3 could represent the uninfected, infected, and diseased states respectively. We may be interested in estimating the magnitude of the infection in the population of interest at a certain point in time; the probability of being in the infected state at time s

after baseline would be the corresponding parameter of interest. If the assumed densities for the sojourn time in states 1 and 2 are f_1 and f_2 respectively, with corresponding survival functions S_1 and S_2, then the probability of being in state 1 at a time $s \geq 0$ is given by simply $p_{11}(s) = S_1(s)$, while the probability of being in state 2 at time $s \geq 0$ is given by

$$p_{12}(s) = \int_0^s f_1(u) \cdot S_2(s-u)du.$$

The latter quantity gives a rough measure of model fit, as it reflects both the first and second transition times. Note that since $p_{11}(s) + p_{12}(s) + p_{13}(s) = 1$ for any s, the probability of being in state 3 at a fixed time can be calculated from the above probabilities.

For the methods that model sojourn times in each state as a continuous quantity, computing estimates of these probabilities is straightforward. However, since the method of De Gruttola and Lagakos (1989) models the sojourn time in each state as a discrete random variable, for each state i the function $f_i(\cdot)$ is the corresponding probability mass function and $S_i(\cdot)$ is the corresponding survival function which decreases only by jump discontinuities. Hence, in the notation of Section 2.3.3 of Chapter 2, the above probabilities are given by

$$\begin{aligned} p_{11}(s) &= S_i(s) = 1 - \sum_{j:\, y_{ij} \leq s} f(y_{ij}) \\ p_{12}(s) &= \sum_{j:\, y_{1j} \leq s} f_1(y_{1j}) \cdot S_2(s - y_{1j}). \end{aligned}$$

Because of the discrete nature of the sojourn times, the values of $f_i(s)$ and $S_i(s)$ at a given point s—and therefore the values of $p_{11}(s)$ and $p_{12}(s)$—depend heavily on the choice of locations of the point masses for each of the sojourn times.

Recall from Section 2.3.3 that implementing the method of De Gruttola and Lagakos (1989) involves the crucial choice of the locations of the point masses for the sojourn times in states 1 and 2. We note that in the case where subjects are observed at a common set of evenly-spaced visit times, it is clear that the point masses should be placed between these visit times with the same spacing. Placing the point masses further apart would lead

to loss of valuable information, while placing them closer together would lead to lack of identifiability of the parameters. Hence, using the same spacing seems to avoid both of these computational pitfalls. The exact location of the point mass within each interval between two successive visit times is immaterial. In the present simulation study, since the visit times occur at $0, 1, 2, \ldots$, we arbitrarily place the point masses at $0.5, 1.5, 2.5, \ldots$ for the sojourn time in state 1, and at $1, 2, 3, \ldots$ for the sojourn time in state 2. Based on this choice, both of the transition times must occur at positive half-integers, so that we preserve identifiability without losing information.

In this simulation study we generate the true sojourn times in states 1 and 2, X_1 and X_2, from the exponential and Weibull distributions. We examine four scenarios, shown in Table 3.3, and for simplicity we let X_1 and X_2 have a common distribution as we did in Section 3.4.1, though this is not necessary. For example, in a particular application we could choose to model X_1 as exponential and X_2 as a spline function. Hence $X_i \sim We(k_i, \theta_i)$ for $i = 1, 2$ for some parameters k_i, θ_i given in Table 3.3.

We consider datasets with $N = 50, 100, 200, 400$ subjects, and for simplicity suppose that each subject is observed at times $t = 0, 1, 2, \ldots$, though the proposed approaches do not require this assumption.

For the data augmentation method we assume $Uniform(a, b)$ priors on the parameters of interest. For the exponential and Weibull models we used $(a, b) = (0, 10)$ for each parameter. Given the sparse observations of the subjects considered here, we do not consider models more flexible than the Weibull model.

Table 3.3: Scenarios for sparse observation scheme.

Scenario	k_1, k_2	θ_1, θ_2
1	1.000	2.000
2	1.000	0.500
3	2.000	2.000
4	4.000	1.000

For comparison to the existing and new methods we also consider a naïve approach which is not correct, but is straightforward to implement. Specifically, under the assumption that subjects enter state 1 at time zero, the sojourn time in state 1 is interval-censored on the right end, but the sojourn time in state 2 is interval-censored on both the left and right ends. The situation for the sojourn time in state 2 is referred to in the literature as "double censoring".

Since we can readily carry out a parametric survival model for the sojourn time in state 1 in most statistical software packages, it may seem reasonable to perform a similar procedure to carry out estimation for the sojourn time in state 2 by treating the doubly-censored data as right-censored data. As noted by De Gruttola and Lagakos (1989) as well as Lawless and Yan (1993), applying this method to these transformed data is not correct: not only does the separate estimation of the sojourn times in states 1 and 2 ignore the dependence between them, but the transformation of the sojourn times in state 2 leads to biased estimation of this distribution.

We illustrate the performance of the naïve approach (Tables 3.4–3.7), the method of Kalbfleisch and Lawless (1985) (Table 3.8), and the proposed data augmentation approach (Tables 3.11–3.13) under the exponential and Weibull models. For the existing approaches the means maximum likelihood estimates, mean estimated asymptotic standard errors based on the Fisher information matrix, and empirical standard errors are presented. For the data augmentation approach we report the average posterior means of the parameters, average posterior standard deviations, and empirical standard errors, computed across simulations for each scenario and sample size. Results are based on $M = 100$ simulated datasets for each scenario and sample size.

The naïve approach uses a standard procedure for estimating the parameters corresponding to the sojourn time in state 1, and thus these estimates are consistent, as we would expect (see Tables 3.4–3.7). Since the estimation procedure used for the sojourn time in state 1 is sound, we would expect that derived quantities that depend only on the first transition, such as $p_{11}(s)$, would be consistently estimated in this approach. However, the naïve approach uses a flawed procedure for estimating the parameters corresponding to the sojourn time in state 2, and the result is evident in the parameter estimates, which are severely biased. The more sparse the observation scheme is relative to the progression rate, the more severe is the bias in estimates of the parameters that correspond to state 2, as we can see by comparing the results for Scenarios 1 and 2.

When the Weibull model was used, the procedure very frequently failed to converge. That is, for a number of simulated datasets, the procedure failed to pass the criterion for convergence in a set number of iterations, and thus did not produce estimates of the

parameters. Table 3.5 shows the proportion of simulated datasets for which at least one of the two regressions failed to converge. The results presented in the following tables are based on runs that were deemed to converge. However, in several cases the regression was deemed to converge yet produced a very large point estimate for a parameter. These regressions inflated the overall point estimate of the parameter (see Table 3.7). To streamline the presentation, point estimates above 10000 are reported as such.

Table 3.4: Naïve method, exponential model.

Scenario 1. $\theta_1 = 2.000$, $\theta_2 = 2.000$.

N	θ_1	SE_{model}	SE_{emp}	θ_2	SE_{model}	SE_{emp}
50	1.959	0.280	0.265	1.999	0.293	0.324
100	1.981	0.200	0.185	1.925	0.200	0.214
200	1.989	0.142	0.128	1.966	0.144	0.123
400	1.987	0.100	0.087	1.950	0.101	0.096

Scenario 2. $\theta_1 = 0.500$, $\theta_2 = 0.500$.

N	θ_1	SE_{model}	SE_{emp}	θ_2	SE_{model}	SE_{emp}
50	0.480	0.081	0.085	0.385	0.083	0.105
100	0.492	0.058	0.056	0.356	0.057	0.064
200	0.498	0.041	0.037	0.372	0.040	0.042
400	0.499	0.029	0.024	0.362	0.028	0.031

The method of Kalbfleisch and Lawless (1985), as expected, does very well in Scenarios 1 and 2, in which the sojourn times arise from an exponential distribution (see Table 3.8). Unfortunately, the method does not accommodate other parametric forms for the sojourn times.

Since the method of De Gruttola and Lagakos (1989) is nonparametric and treats time as a discrete quantity, the form of the results is different from that of the other methods we have discussed. Rather than parameters of an assumed model for the sojourn times, we present estimated weights associated with the point masses at each discrete time point for the sojourn times in states 1 and 2. In Tables 3.9–3.10 we present the estimated weights, $\hat{w}_{11}, \ldots, \hat{w}_{1r}$ and $\hat{w}_{21}, \ldots, \hat{w}_{2s}$, corresponding to each point mass, $y_{11}, \ldots, y_{1r}$ and $y_{21}, \ldots, y_{2s}$, respectively, for each of the two sojourn times. Standard errors are presented in Appendix B.

To generate the "true" value of the weight at the k^{th} mass point for each of the sojourn

Table 3.5: Naïve method, Weibull model: proportion of datasets in which at least one regression failed to converge.

	Scenario			
N	1	2	3	4
50	0.14	0.80	0.27	0.94
100	0.17	0.64	0.23	0.95
200	0.15	0.31	0.26	0.90
400	0.15	0.10	0.23	0.96

Table 3.6: Naïve method, Weibull model (scenarios 1–2).

Scenario 1. $k_1 = 1.000$, $k_2 = 1.000$, $\theta_1 = 2.000$, $\theta_2 = 2.000$.

N	k_1	SE_{model}	SE_{emp}	θ_1	SE_{model}	SE_{emp}
50	1.055	0.145	0.160	1.983	0.299	0.272
100	1.028	0.100	0.128	1.977	0.214	0.274
200	0.989	0.069	0.085	1.971	0.158	0.143
400	0.995	0.049	0.050	1.987	0.111	0.093

N	k_2	SE_{model}	SE_{emp}	θ_2	SE_{model}	SE_{emp}
50	1.360	0.149	1.813	1.951	0.336	0.390
100	0.966	0.110	0.128	1.866	0.238	0.250
200	0.964	0.076	0.072	1.915	0.171	0.146
400	0.955	0.054	0.084	1.902	0.119	0.115

Scenario 2. $k_1 = 1.000$, $k_2 = 1.000$, $\theta_1 = 0.500$, $\theta_2 = 0.500$.

N	k_1	SE_{model}	SE_{emp}	θ_1	SE_{model}	SE_{emp}
50	1.011	35.941	0.708	0.601	2.283	0.214
100	0.962	0.235	0.182	0.511	0.415	0.168
200	1.017	0.174	0.162	0.498	0.119	0.083
400	1.011	0.119	0.108	0.510	0.091	0.074

N	k_2	SE_{model}	SE_{emp}	θ_2	SE_{model}	SE_{emp}
50	1.362	8.034	0.758	0.284	0.179	0.200
100	1.294	4.449	0.713	0.197	0.146	0.100
200	1.008	2.262	0.649	0.214	0.111	0.089
400	0.771	0.352	0.322	0.214	0.088	0.067

times, we integrated the true density between $k-1$ and k. This approach, though *ad hoc*, seems to be a logical way to translate the true density, on the continuous scale, to the true values, on the discrete scale. Estimates of the weights when the sojourn times are exponential, in Scenarios 1–2, appear to be consistent for these true values; however, when the sojourn times are Weibull, estimated weights appear to converge to other values.

Table 3.7: Naïve method, Weibull model (scenarios 3–4).

Scenario 3. $k_1 = 2.000$, $k_2 = 2.000$, $\theta_1 = 2.000$, $\theta_2 = 2.000$.

N	k_1	SE_{model}	SE_{emp}	θ_1	SE_{model}	SE_{emp}
50	2.069	0.268	0.285	1.985	0.157	0.225
100	2.030	0.184	0.207	2.020	0.114	0.157
200	2.010	0.129	0.141	1.997	0.078	0.078
400	2.018	0.092	0.108	1.989	0.055	0.069

N	k_2	SE_{model}	SE_{emp}	θ_2	SE_{model}	SE_{emp}
50	2.304	0.413	0.524	1.982	0.161	0.129
100	2.393	0.285	0.725	1.974	0.114	0.116
200	2.224	0.190	0.337	1.990	0.082	0.078
400	2.201	0.130	0.131	2.009	0.059	0.046

Scenario 4. $k_1 = 4.000$, $k_2 = 4.000$, $\theta_1 = 1.000$, $\theta_2 = 1.000$.

N	k_1	SE_{model}	SE_{emp}	θ_1	SE_{model}	SE_{emp}
50	4.750	1903.255	0.505	$\geq$10000	2.795	$\geq$10000
100	4.561	1889.382	0.165	$\geq$10000	0.927	$\geq$10000
200	4.552	1273.615	0.138	0.972	1.085	0.070
400	5.226	734.769	1.067	0.959	0.417	0.106

N	k_2	SE_{model}	SE_{emp}	θ_2	SE_{model}	SE_{emp}
50	4.549	4356.976	0.666	1.048	93.300	0.043
100	4.469	1812.460	0.471	1.020	9.685	0.040
200	4.966	192.946	0.921	1.031	0.895	0.034
400	4.242	224.089	0.301	1.001	1.475	0.037

Since implementation of the method involves discarding observations of subjects who were not observed in state 2, we would expect the estimates for all scenarios to be biased to varying degrees. All discarded observations have a sojourn time in state 2, X_2, that is less than one unit of time, so we would expect that the bias for the parameters corresponding to X_2 would be particularly severe, and in the direction of longer sojourn times in this state (refer to Appendix A to see the impact of excluding these observations). Therefore, we would not expect the estimated weights to converge to the "true" values discussed previously. In Scenarios 1–2, however, the estimates appear to be consistent in spite of the dubious way in which they were obtained. This coincidence is a consequence of the discretization of the time axis, and does not hold in general.

As we have seen, the method of De Gruttola and Lagakos is not able to handle observations for which state 2 is unobserved, but the bias introduced by excluding such observations

Table 3.8: Method of Kalbfleisch and Lawless (1985).

Scenario 1. $\theta_1 = 2.000$, $\theta_2 = 2.000$.

N	θ_1	SE_{model}	SE_{emp}	θ_2	SE_{model}	SE_{emp}
50	1.961	0.280	0.265	2.070	0.298	0.320
100	1.982	0.200	0.186	1.993	0.203	0.211
200	1.989	0.142	0.127	2.033	0.146	0.122
400	1.987	0.100	0.087	2.016	0.103	0.096

Scenario 2. $\theta_1 = 0.500$, $\theta_2 = 0.500$.

N	θ_1	SE_{model}	SE_{emp}	θ_2	SE_{model}	SE_{emp}
50	0.481	0.081	0.083	0.514	0.090	0.109
100	0.493	0.058	0.055	0.497	0.062	0.063
200	0.498	0.041	0.037	0.510	0.044	0.042
400	0.498	0.029	0.024	0.500	0.031	0.027

is difficult to assess. Later we examine the performance of the method from a different perspective, via the probabilities of being in states 1 and 2 at various times, in Tables 3.14 and 3.15.

Table 3.9: Method of De Gruttola and Lagakos (1989) (scenarios 1–2).
For each sample size N we present the estimated weights $\hat{w}_{1,1}, \ldots, \hat{w}_{1,r}$ corresponding to point masses $(y_{1,1}, \ldots, y_{1,r}) = (\frac{1}{2}, \ldots, \frac{2r-1}{2})$ for state 1, and $\hat{w}_{2,1}, \ldots, \hat{w}_{2,s}$ corresponding to point masses $(y_{2,1}, \ldots, y_{2,s}) = (1, \ldots, s)$ for state 2.

Scenario 1. True weights: $\mathbf{w}_1 = (0.393, 0.239, 0.145, 0.088, 0.053, 0.032, 0.020, 0.012, 0.007, 0.004, 0.003, 0.002, 0.001, 0.001, 0.000, 0.000, 0.000, 0.000, 0.000)$.

N	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$	$y_{1,5}$	$y_{1,6}$	$y_{1,7}$	$y_{1,8}$	$y_{1,9}$	$y_{1,10}$	$y_{1,11}$	$y_{1,12}$	$y_{1,13}$	$y_{1,14}$	$y_{1,15}$	$y_{1,16}$	$y_{1,17}$	$y_{1,18}$	$y_{1,19}$
50	0.391	0.261	0.138	0.087	0.046	0.033	0.016	0.011	0.007	0.002	0.003	0.003	0.001	0.001	0.000	0.000	0.000	0.001	0.000
100	0.396	0.237	0.149	0.086	0.051	0.032	0.019	0.012	0.006	0.005	0.003	0.002	0.001	0.001	0.000	0.000	0.000	0.000	0.000
200	0.401	0.238	0.139	0.089	0.050	0.033	0.020	0.011	0.008	0.005	0.003	0.002	0.000	0.001	0.000	0.000	0.000	0.000	0.000
400	0.396	0.235	0.147	0.087	0.053	0.034	0.019	0.011	0.007	0.005	0.002	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000

$\mathbf{w}_2 = (0.393, 0.239, 0.145, 0.088, 0.053, 0.032, 0.020, 0.012, 0.007, 0.004, 0.003, 0.002, 0.001, 0.001, 0.000, 0.000, 0.000, 0.000, 0.000)$.

N	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$	$y_{2,5}$	$y_{2,6}$	$y_{2,7}$	$y_{2,8}$	$y_{2,9}$	$y_{2,10}$	$y_{2,11}$	$y_{2,12}$	$y_{2,13}$	$y_{2,14}$	$y_{2,15}$	$y_{2,16}$	$y_{2,17}$	$y_{2,18}$	$y_{2,19}$
50	0.387	0.240	0.143	0.092	0.050	0.033	0.020	0.013	0.007	0.006	0.002	0.003	0.002	0.002	0.000	0.000	0.000	0.000	0.000
100	0.399	0.235	0.138	0.089	0.055	0.037	0.018	0.013	0.006	0.004	0.001	0.001	0.002	0.001	0.001	0.000	0.000	0.000	0.000
200	0.390	0.239	0.143	0.087	0.056	0.032	0.020	0.013	0.009	0.005	0.002	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000
400	0.389	0.238	0.149	0.087	0.054	0.033	0.022	0.011	0.007	0.005	0.003	0.001	0.001	0.000	0.000	0.001	0.000	0.000	0.000

Scenario 2.
$\mathbf{w}_1 = (0.865, 0.117, 0.016, 0.002, 0.000)$.

N	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$	$y_{1,5}$
50	0.882	0.105	0.013	0.000	0.000
100	0.866	0.112	0.019	0.003	0.000
200	0.872	0.109	0.017	0.001	0.000
400	0.866	0.118	0.015	0.002	0.000

$\mathbf{w}_2 = (0.865, 0.117, 0.016, 0.002, 0.000, 0.000, 0.000, 0.000)$.

N	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$	$y_{2,5}$	$y_{2,6}$	$y_{2,7}$	$y_{2,8}$
50	0.851	0.123	0.023	0.003	0.000	0.000	0.000	0.000
100	0.866	0.117	0.015	0.002	0.000	0.000	0.000	0.000
200	0.855	0.126	0.015	0.004	0.000	0.000	0.000	0.000
400	0.861	0.120	0.015	0.003	0.000	0.000	0.000	0.000

Table 3.10: Method of De Gruttola and Lagakos (1989) (scenarios 3–4).

Scenario 3. $\mathbf{w}_1 = (0.221, 0.411, 0.262, 0.087, 0.016, 0.002, 0.000, 0.000)$.

N	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$	$y_{1,5}$	$y_{1,6}$	$y_{1,7}$	$y_{1,8}$
50	0.227	0.413	0.260	0.082	0.017	0.002	0.000	0.000
100	0.228	0.408	0.261	0.087	0.015	0.001	0.000	0.000
200	0.233	0.406	0.258	0.085	0.016	0.002	0.000	0.000
400	0.232	0.408	0.259	0.086	0.013	0.002	0.000	0.000

$\mathbf{w}_2 = (0.221, 0.411, 0.262, 0.087, 0.016, 0.002, 0.000, 0.000)$.

N	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$	$y_{2,5}$	$y_{2,6}$	$y_{2,7}$	$y_{2,8}$
50	0.378	0.390	0.177	0.047	0.008	0.001	0.000	0.000
100	0.392	0.379	0.179	0.044	0.007	0.000	0.000	0.000
200	0.381	0.386	0.177	0.048	0.007	0.000	0.000	0.000
400	0.387	0.375	0.181	0.049	0.008	0.001	0.000	0.000

Scenario 4.
$\mathbf{w}_1 = (0.632, 0.368)$.

N	$y_{1,1}$	$y_{1,2}$
50	0.732	0.268
100	0.731	0.269
200	0.737	0.263
400	0.737	0.263

$\mathbf{w}_2 = (0.632, 0.368, 0.000)$.

N	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$
50	0.890	0.110	0.000
100	0.897	0.103	0.000
200	0.903	0.097	0.000
400	0.899	0.101	0.000

The proposed data augmentation approach consistently estimates parameters in each of the scenarios. Under the exponential model (Table 3.11) the scale parameters are consistently estimated in Scenarios 1 and 2. The model for the sojourn times is misspecified in Scenarios 3 and 4, so true values for the scale parameters do not exist. The shape and scale parameters are consistently estimated in each scenario under the Weibull model (Tables 3.12–3.13).

We note that subjects are observed quite infrequently relative to the progression rate of the process in each of the scenarios. In Scenario 2, for example, the probability of state 2 being unobserved is $1 - \frac{\frac{1}{\theta}\cdot\exp(-\frac{1}{\theta})}{1-\exp(-\frac{1}{\theta})}$ with $\theta = 0.5$, or 68.7% (see Appendix A for derivation).

Table 3.11: Proposed data augmentation method, exponential model.

Scenario 1. $\theta_1 = 2.000$, $\theta_2 = 2.000$.

N	θ_1	SD_{model}	SD_{emp}	θ_2	SD_{model}	SD_{emp}
50	2.042	0.300	0.277	2.156	0.319	0.333
100	2.021	0.207	0.190	2.034	0.210	0.215
200	2.008	0.144	0.128	2.054	0.149	0.123
400	1.997	0.101	0.087	2.026	0.104	0.095

Scenario 2. $\theta_1 = 0.500$, $\theta_2 = 0.500$.

N	θ_1	SD_{model}	SD_{emp}	θ_2	SD_{model}	SD_{emp}
50	0.502	0.086	0.086	0.537	0.095	0.113
100	0.504	0.060	0.056	0.508	0.064	0.063
200	0.503	0.042	0.037	0.515	0.045	0.042
400	0.501	0.029	0.024	0.503	0.031	0.027

Hence, for Scenario 2, we would expect that the sojourn in state 2 would not be observed for over two-thirds of the subjects.

To provide a more direct comparison of the methods, we examine their estimates of the probabilities of being in states 1 and 2 respectively at chronological time s, $p_{11}(s)$ and $p_{12}(s)$, for $s = 0.5$, 1.0, and 2.0. Results are shown for a single sample size, $N = 400$, in Tables 3.14 and 3.15. When the model is correctly specified, the data augmentation (DA) approach performs well. As we would expect, the exponential model yields more precise estimates than the Weibull model does under correct specification (Scenarios 1 and 2), but is biased when the model is misspecified (Scenarios 3 and 4). As expected, the method of Kalbfleisch and Lawless (K&L) performs very similarly to the data augmentation approach under the exponential model.

Since for each s, $p_{11}(s)$ depends on only the first transition time, whereas $p_{12}(s)$ depends on both transition times, the naïve approach estimates $p_{11}(s)$ consistently but is biased for $p_{12}(s)$. This observation applies to both models. The method of De Gruttola and Lagakos (DGL) estimates some of these probabilities well but is severely biased for others, and no patterns are evident. We believe that the departures in the estimates from the true values can be attributed mainly to (1) the necessary discarding of observations to implement the method and (2) consequences of discretizing the time axis. When subjects are observed frequently, the method of De Gruttola and Lagakos performs well, but the two above issues

Table 3.12: Proposed data augmentation method, Weibull model (scenarios 1–2).

Scenario 1. $k_1 = 1.000$, $k_2 = 1.000$, $\theta_1 = 2.000$, $\theta_2 = 2.000$.

N	k_1	SD_{model}	SD_{emp}	θ_1	SD_{model}	SD_{emp}
50	1.058	0.143	0.156	2.055	0.315	0.267
100	1.023	0.099	0.126	2.018	0.222	0.206
200	0.992	0.068	0.083	1.992	0.159	0.145
400	1.002	0.048	0.055	1.995	0.111	0.096

N	k_2	SD_{model}	SD_{emp}	θ_2	SD_{model}	SD_{emp}
50	1.053	0.140	0.159	2.166	0.333	0.345
100	1.016	0.096	0.109	2.032	0.227	0.232
200	1.005	0.067	0.064	2.054	0.163	0.145
400	1.002	0.047	0.044	2.026	0.114	0.106

Scenario 2. $k_1 = 1.000$, $k_2 = 1.000$, $\theta_1 = 0.500$, $\theta_2 = 0.500$.

N	k_1	SD_{model}	SD_{emp}	θ_1	SD_{model}	SD_{emp}
50	1.457	0.536	1.012	0.537	0.138	0.157
100	1.203	0.321	0.665	0.509	0.106	0.125
200	1.057	0.160	0.216	0.509	0.075	0.071
400	1.023	0.107	0.122	0.504	0.054	0.048

N	k_2	SD_{model}	SD_{emp}	θ_2	SD_{model}	SD_{emp}
truth	1.000			0.500		
50	1.433	0.422	1.102	0.515	0.136	0.132
100	1.152	0.258	0.479	0.510	0.098	0.118
200	1.008	0.128	0.097	0.509	0.070	0.062
400	1.005	0.090	0.113	0.497	0.049	0.050

increasingly plague the method as the observation scheme becomes more sparse.

Based on the estimates of the parameters and of the probabilities of being in each state at the various times, it appears that the proposed data augmentation approach performs better than the existing approaches we examined when the sojourn times arise from a Weibull distribution.

3.5 Simulation studies: right censoring

In the previous section we illustrated the performance of the proposed and existing approaches under an ideal scenario to characterize the performance of the data augmentation approach and to give a clear picture of how the proposed and existing approaches compare

Table 3.13: Proposed data augmentation method, Weibull model (scenarios 3–4).

Scenario 3. $k_1 = 2.000$, $k_2 = 2.000$, $\theta_1 = 2.000$, $\theta_2 = 2.000$.

N	k_1	SD_{model}	SD_{emp}	θ_1	SD_{model}	SD_{emp}
50	2.068	0.267	0.285	2.013	0.158	0.192
100	2.031	0.185	0.206	2.020	0.112	0.118
200	2.006	0.127	0.135	2.003	0.079	0.086
400	2.018	0.090	0.108	1.995	0.055	0.071

N	k_2	SD_{model}	SD_{emp}	θ_2	SD_{model}	SD_{emp}
50	2.074	0.290	0.270	2.021	0.165	0.129
100	2.048	0.200	0.184	1.983	0.114	0.112
200	2.016	0.138	0.147	1.992	0.082	0.080
400	1.993	0.096	0.093	1.999	0.058	0.056

Scenario 4. $k_1 = 4.000$, $k_2 = 4.000$, $\theta_1 = 1.000$, $\theta_2 = 1.000$.

N	k_1	SD_{model}	SD_{emp}	θ_1	SD_{model}	SD_{emp}
50	4.572	1.140	1.315	0.997	0.049	0.065
100	4.525	0.907	1.070	1.003	0.033	0.032
200	4.335	0.689	0.774	1.001	0.023	0.024
400	4.166	0.459	0.546	0.999	0.016	0.021

N	k_2	SD_{model}	SD_{emp}	θ_2	SD_{model}	SD_{emp}
50	4.819	1.146	1.475	1.014	0.059	0.064
100	4.673	0.887	1.328	1.003	0.041	0.035
200	4.432	0.692	0.812	0.993	0.029	0.030
400	4.224	0.457	0.509	0.999	0.021	0.020

in the absence of complicating features of real datasets. In the vast majority of longitudinal studies, participants are subject to right censoring due to administrative reasons and dropout. We examine in the current section the performance of the proposed data augmentation approach in the presence of noninformative loss to follow-up.

We consider the case in which subjects are observed frequently, and explore the performance of the proposed data augmentation approach when various sojourn time models are used. We assume the observation scheme and scenarios used in Section 3.4.1. The scenarios under consideration are shown in Table 3.2. We assume subjects are observed at times $t = 0.00, 0.25, 0.50, \ldots$, for simplicity. For each scenario we model the two sojourn times as exponential, Weibull, exponentiated Weibull, and linear spline with two knots.

We assume, as before, that subjects enter state 1 at time zero. In this study we allow for noninformative loss to follow-up due to both administrative censoring and dropout.

Table 3.14: Probability of being in state 1 at various timepoints s. $N = 400$.

$s = 0.5$.

	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
Method	Mean	SE_{emp}	Mean	SE_{emp}	Mean	SE_{emp}	Mean	SE_{emp}
truth	0.779	–	0.368	–	0.939	–	0.939	–
Naive, exp	0.777	0.009	0.367	0.018	0.747	0.008	0.517	0.016
Naive, Weibull	0.776	0.020	0.369	0.039	0.940	0.011	0.964	0.009
K&L	0.777	0.009	0.366	0.018	0.751	0.008	0.546	0.014
DGL	0.802	0.013	0.567	0.014	0.884	0.013	0.631	0.014
DA, exp	0.778	0.009	0.368	0.018	0.752	0.008	0.548	0.014
DA, we	0.778	0.020	0.373	0.036	0.940	0.011	0.942	0.019

$s = 1.0$.

	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
Method	Mean	SE_{emp}	Mean	SE_{emp}	Mean	SE_{emp}	Mean	SE_{emp}
truth	0.607	–	0.135	–	0.779	–	0.368	–
Naive, exp	0.604	0.013	0.135	0.013	0.558	0.012	0.268	0.017
Naive, Weibull	0.604	0.022	0.136	0.014	0.779	0.024	0.370	0.039
K&L	0.604	0.013	0.134	0.013	0.565	0.012	0.298	0.015
DGL	0.604	0.026	0.134	0.027	0.768	0.026	0.263	0.027
DA, exp	0.606	0.013	0.136	0.013	0.566	0.012	0.300	0.015
DA, we	0.606	0.021	0.135	0.013	0.779	0.024	0.367	0.031

$s = 2.0$.

	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
Method	Mean	SE_{emp}	Mean	SE_{emp}	Mean	SE_{emp}	Mean	SE_{emp}
truth	0.368	–	0.018	–	0.368	–	0.000	–
Naive, exp	0.365	0.016	0.018	0.003	0.312	0.013	0.072	0.009
Naive, Weibull	0.366	0.017	0.019	0.005	0.364	0.025	0.000	0.000
K&L	0.365	0.016	0.018	0.003	0.319	0.013	0.089	0.009
DGL	0.369	0.024	0.017	0.014	0.360	0.032	0.000	0.000
DA, exp	0.367	0.016	0.019	0.004	0.321	0.013	0.090	0.009
DA, we	0.367	0.018	0.018	0.005	0.366	0.026	0.000	0.000

Each subject has a potential administrative censoring time $T_1 \sim \mathcal{N}(10, 1)$, and a potential dropout time $T_2 \sim Unif(0, T_1)$ random variable with a dropout indicator $W \sim Bernoulli(0.10)$. Each subject's true censoring time T is a mixture of these two potential censoring times:

$$T = \begin{cases} T_1, & W = 0; \\ T_2, & W = 1. \end{cases}$$

Table 3.15: Probability of being in state 2 at various timepoints s. $N = 400$.

$s = 0.5$.

	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
Method	Mean	SE_{emp}	Mean	SE_{emp}	Mean	SE_{emp}	Mean	SE_{emp}
truth	0.195	–	0.368	–	0.060	–	0.061	–
Naive, exp	0.195	0.008	0.306	0.018	0.216	0.007	0.306	0.015
Naive, Weibull	0.194	0.017	0.213	0.040	0.060	0.011	0.036	0.009
K&L	0.196	0.008	0.368	0.014	0.215	0.007	0.339	0.012
DGL	0.396	0.026	0.866	0.027	0.232	0.026	0.737	0.027
DA, exp	0.196	0.008	0.368	0.014	0.214	0.007	0.339	0.012
DA, we	0.196	0.017	0.366	0.031	0.060	0.011	0.058	0.019

$s = 1.0$.

	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
Method	Mean	SE_{emp}	Mean	SE_{emp}	Mean	SE_{emp}	Mean	SE_{emp}
truth	0.303	–	0.271	–	0.212	–	0.620	–
Naive, exp	0.303	0.011	0.189	0.020	0.320	0.010	0.285	0.016
Naive, Weibull	0.297	0.017	0.137	0.029	0.214	0.023	0.624	0.039
K&L	0.306	0.011	0.270	0.015	0.323	0.010	0.381	0.012
DGL	0.396	0.026	0.866	0.027	0.232	0.026	0.737	0.027
DA, exp	0.305	0.011	0.272	0.015	0.322	0.010	0.381	0.012
DA, we	0.305	0.016	0.272	0.022	0.211	0.023	0.622	0.031

$s = 2.0$.

	Scenario 1		Scenario 2		Scenario 3		Scenario 4	
Method	Mean	SE_{emp}	Mean	SE_{emp}	Mean	SE_{emp}	Mean	SE_{emp}
truth	0.368	–	0.073	–	0.519	–	0.307	–
Naive, exp	0.364	0.012	0.038	0.007	0.351	0.010	0.126	0.010
Naive, Weibull	0.355	0.016	0.036	0.007	0.533	0.022	0.297	0.010
K&L	0.370	0.012	0.073	0.008	0.364	0.010	0.240	0.007
DGL	0.477	0.023	0.238	0.039	0.551	0.029	0.337	0.024
DA, exp	0.370	0.012	0.074	0.009	0.365	0.010	0.241	0.007
DA, we	0.371	0.014	0.073	0.009	0.520	0.022	0.305	0.021

This setup is a model for the situation in which subjects are accrued over 1–2 years and are followed up for roughly ten years, and each subject drops out prematurely with probability 10%. We illustrate the performance of this approach for $N = 400$ subjects. For Scenario 3^* informative priors were needed for the exponentiated Weibull model.

As in Section 3.4.1 we present a graphical representation of how the models perform in each of the four scenarios, since the parameter estimates do not convey how well the proposed approach with each model performs in cases of model misspecification. In Figures 3.5–3.6

we show the true and estimated hazard functions corresponding to X_1 and X_2 from each model in each scenario. The results are similar to those we saw in Section 3.4.1. We see that the exponential model is inadequate in the non-exponential scenarios and that the Weibull model is reasonably flexible but cannot accommodate unimodal and bathtub-shaped hazards. The exponentiated Weibull model and linear spline model with two knots performed well in all scenarios.

We demonstrate the impact of more extreme right censoring on inference for selected scenarios in Appendix C. Specifically, we consider the potential administrative censoring time $T_1 \sim \mathcal{N}(5, 1)$ and potential dropout time $T_2 \sim Unif(0, T_1)$ with dropout indicator $W \sim Bernoulli(0.50)$. That is, subjects are followed for a mean of five years with a 50% probability of premature dropout. As expected, there is more uncertainty associated with inference on the parameters associated with the sojourn times X_1 and X_2 under this severe right censoring than under the cases considered in this section. Additionally, as expected, the impact of right censoring is greater for inference about X_2 than about X_1.

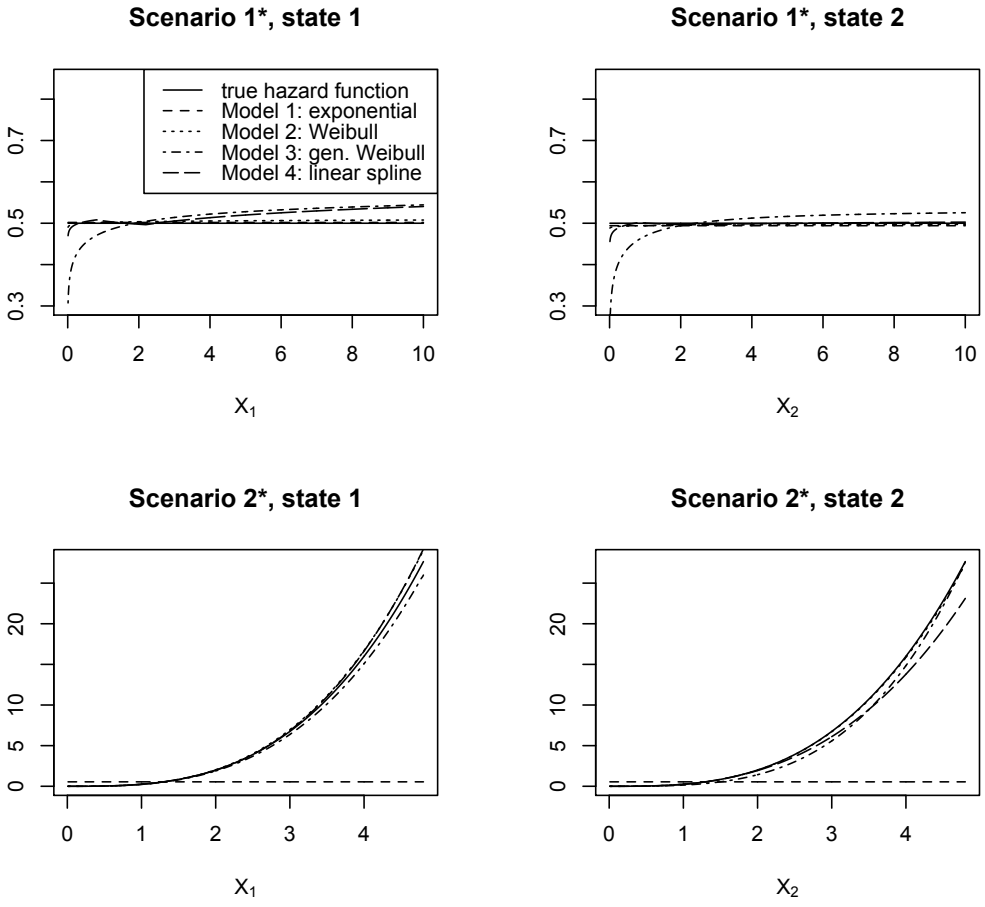


Figure 3.5: True and estimated hazard functions for Scenarios 1* (top panels) and 2* (bottom panels) under right censoring with dropout probability 10%. Results for the sojourn times in state 1 (left panels) and state 2 (right panels) are shown. In each plot, the solid line represents the true hazard, and the other lines show the hazard estimated by the data augmentation approach assuming each of the four models: exponential, Weibull, exponentiated Weibull, and linear spline.

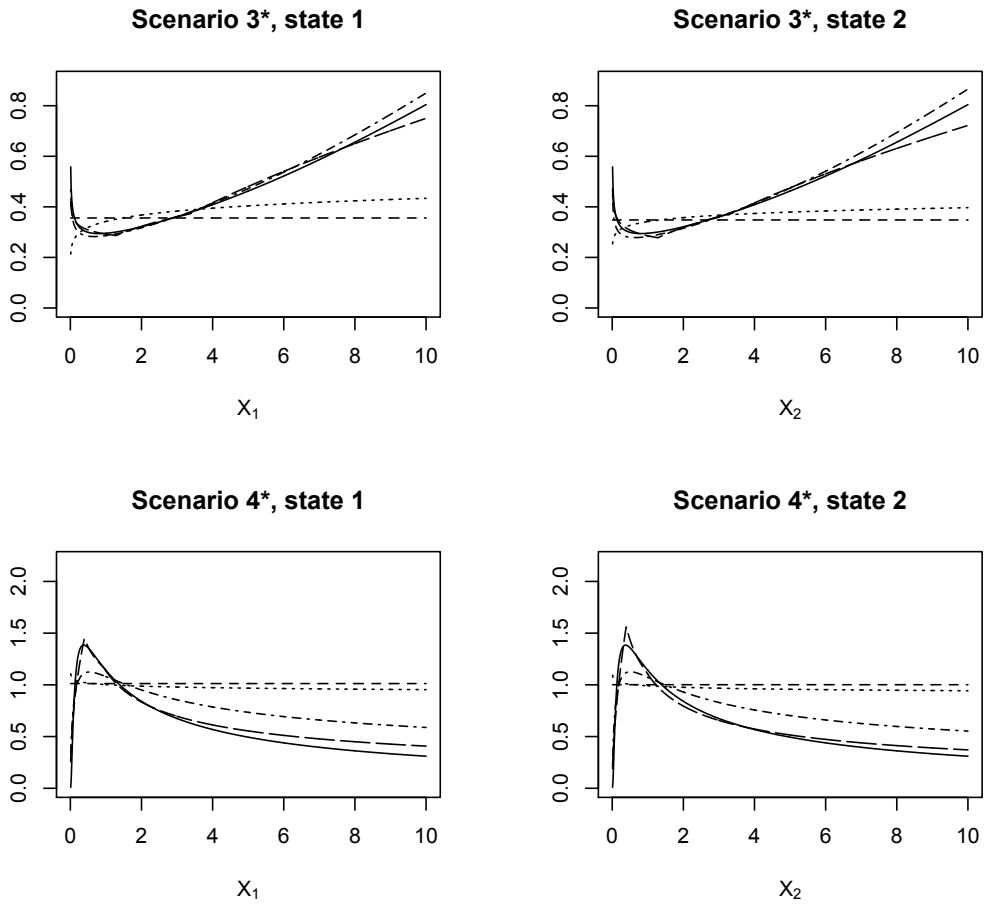


Figure 3.6: True and estimated hazard functions for Scenarios 3* (top panels) and 4* (bottom panels) under right censoring with dropout probability 10%. See Figure 3.5 for legend and explanation.

3.6 Application

We apply the proposed data augmentation method to the study of HIV infection in hemophiliacs described in Section 2.4.1. In this application, subjects progress through three states sequentially. We do not explore the second application described in Section 2.4.2 here, as subjects in that study do not necessarily visit all of the states.

3.6.1 HIV infection in hemophiliacs.

We estimate the sojourn times in the HIV-uninfected and -infected states among hemophiliacs who received contaminated blood products based on the incomplete dataset presented in De Gruttola and Lagakos (1989). The authors presented the discretized sufficient data for each subject, but for the subjects who were observed to progress to AIDS (state 3) during the study, the first observed time in this state was not presented. As the authors did, we frame the situation in terms of the state model in Figure 3.7, and assume all subjects began in the uninfected state.

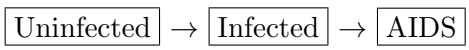


Figure 3.7: State model of De Gruttola and Lagakos (1989).

We first present the results of De Gruttola and Lagakos and compare them to those we obtained in our implementation of their methods (see Figures 3.8–3.9). Though the authors did not publish their numerical estimates, the figures portray good correspondence between the estimates from the original implementation of the algorithm and our own, except over the last interval, where the masses are not identifiable. As the authors note, some masses may not be identifiable if a condition fails (De Gruttola and Lagakos, 1989, p. 4). This lack of identifiability is analogous to the non-uniqueness of the Kaplan-Meier curve when the longest survival time is right–censored.

Based on these estimated cumulative distribution functions, there appears to be evidence that the underlying hazard function for the sojourn time in the uninfected state is increasing, so we would expect that the exponential model would not perform well. The estimates of the sojourn time in the infected state appears to be increasing as well, but since only 43 subjects were observed to progress to AIDS symptoms across the two treatment strata, evidence for this trend is weak.

As we noted in Section 2.3.3 of Chapter 2, the locations of the mass points for each of the sojourn times has a large impact on the results for a given dataset. If more than one mass point is located in an interval in which there is no sufficient data for any subject, then these mass points will not be distinguishable to the estimation procedure, and the corresponding

weights will not be uniquely identifiable. In particular, all mass points located to the right of the longest observed time in any state will not be identifiable. Hence, for each sojourn time, there is a point to the right of which we should ignore estimates of the corresponding weights.

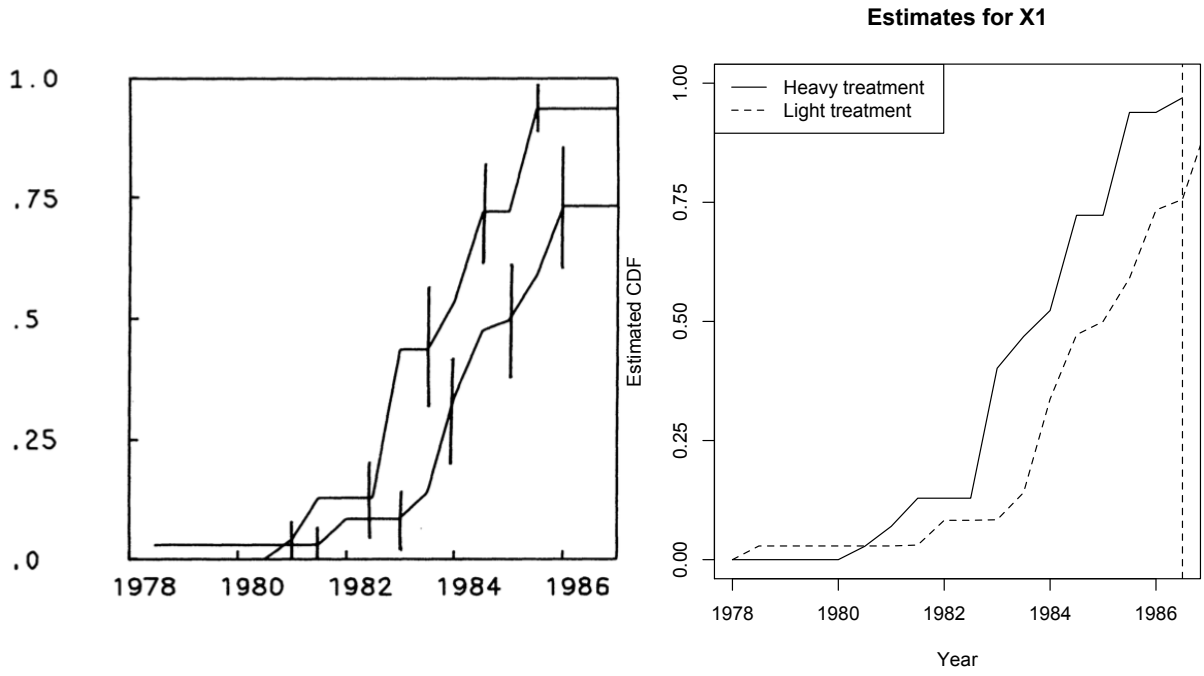


Figure 3.8: Estimated cumulative distribution function corresponding to time-to-infection based on method of De Gruttola and Lagakos (1989), as obtained originally (left) and from our own implementation (right). In the right panel, the solid line represents heavily-treated, while the dashed line represents lightly-treated patients. Also, in this panel, the vertical dashed line shows time beyond which parameters are not uniquely identifiable.

We apply the proposed data augmentation approach to this dataset and compare the results to those obtained by the method of De Gruttola and Lagakos. Specifically, we consider several models for the sojourn times in the HIV-uninfected and -infected states, and carry out inference about the corresponding parameters separately for the heavily and lightly treated patients. We consider exponential, Weibull, and linear spline models for the

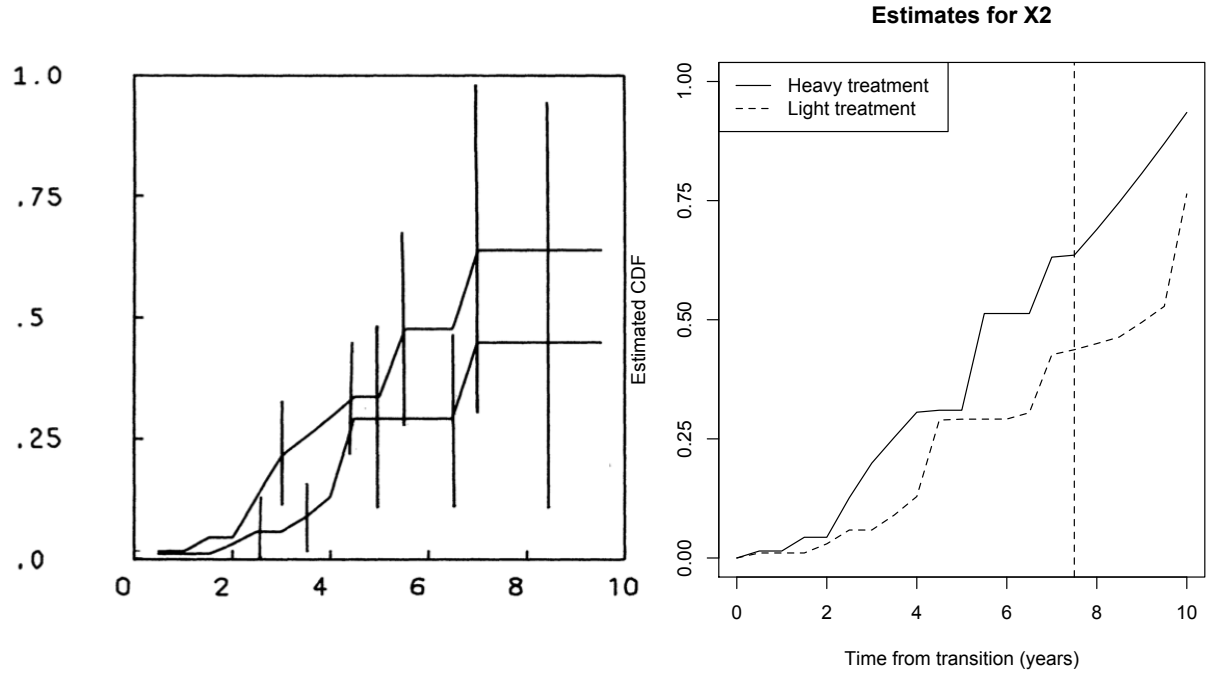


Figure 3.9: Estimated cumulative distribution function corresponding to time-to-infection based on method of De Gruttola and Lagakos (1989), as obtained originally (left) and from our own implementation (right). In the right panel, the solid line represents heavily-treated, while the dashed line represents lightly-treated patients. Also, in this panel, the vertical dashed line shows time beyond which parameters are not uniquely identifiable.

sojourn time in each state. For the linear spline model we use just one knot, since the data are relatively sparse and sample sizes are small: there are 105 and 157 subjects in the heavily and lightly treated strata, respectively.

In this application, "time zero" is defined as the first half of the year 1978 for all subjects, regardless of the elapsed time receiving treatment for hemophilia up to this point. In the original paper De Gruttola and Lagakos presented the data with the event times discretized into six-month intervals; we use this version of the data, but treat the times as if they had been measured on a continuous scale. For the subjects who were observed to develop AIDS symptoms during the study, the authors did not present the times at which these subjects

were first observed to display these symptoms. We carried out the method of De Gruttola and Lagakos on two versions of the dataset: an "optimistic" and a "pessimistic" one, in which patients were observed to enter the third state as late and as early as possible, respectively. Since our results from the "pessimistic" version of the dataset corresponded with the original results much more closely than those from the "optimistic" version did, we use the "pessimistic" version for all analyses here, including the results presented in Figures 3.8–3.9.

Results are presented for heavily and lightly treated subjects in Figure 3.10. Results are presented, as above, in the form of estimated cumulative distribution functions so that they may be compared with those obtained originally by De Gruttola and Lagakos. For each model we present the estimated median cumulative distribution function as well as the corresponding 95% posterior credible interval. As we can see from Figure 3.10, the exponential model yields an estimate that is not close to the one based on the method of De Gruttola and Lagakos (1989), as it is not able to capture change in the underlying hazard function in each state. The Weibull model, on the other hand, has enough flexibility to accommodate each of the underlying hazard functions, as it produced estimated cumulative distribution functions similar to those of the method of De Gruttola and Lagakos (1989). Results from the linear spline model were very close to those from the Weibull model, but since it did not seem to improve model fit, the extra flexibility provided by this model was perhaps not necessary in this case. Thus, we focus on the exponential and Weibull models in our analysis of convergence. Specifically, we carried out the stationarity test and the halfwidth test of Heidelberger and Welch on the MCMC chains of all the parameters of interest in each of the models we used for the heavily-treated and lightly-treated patients. Each of the parameters passed both tests (see Table 3.16).

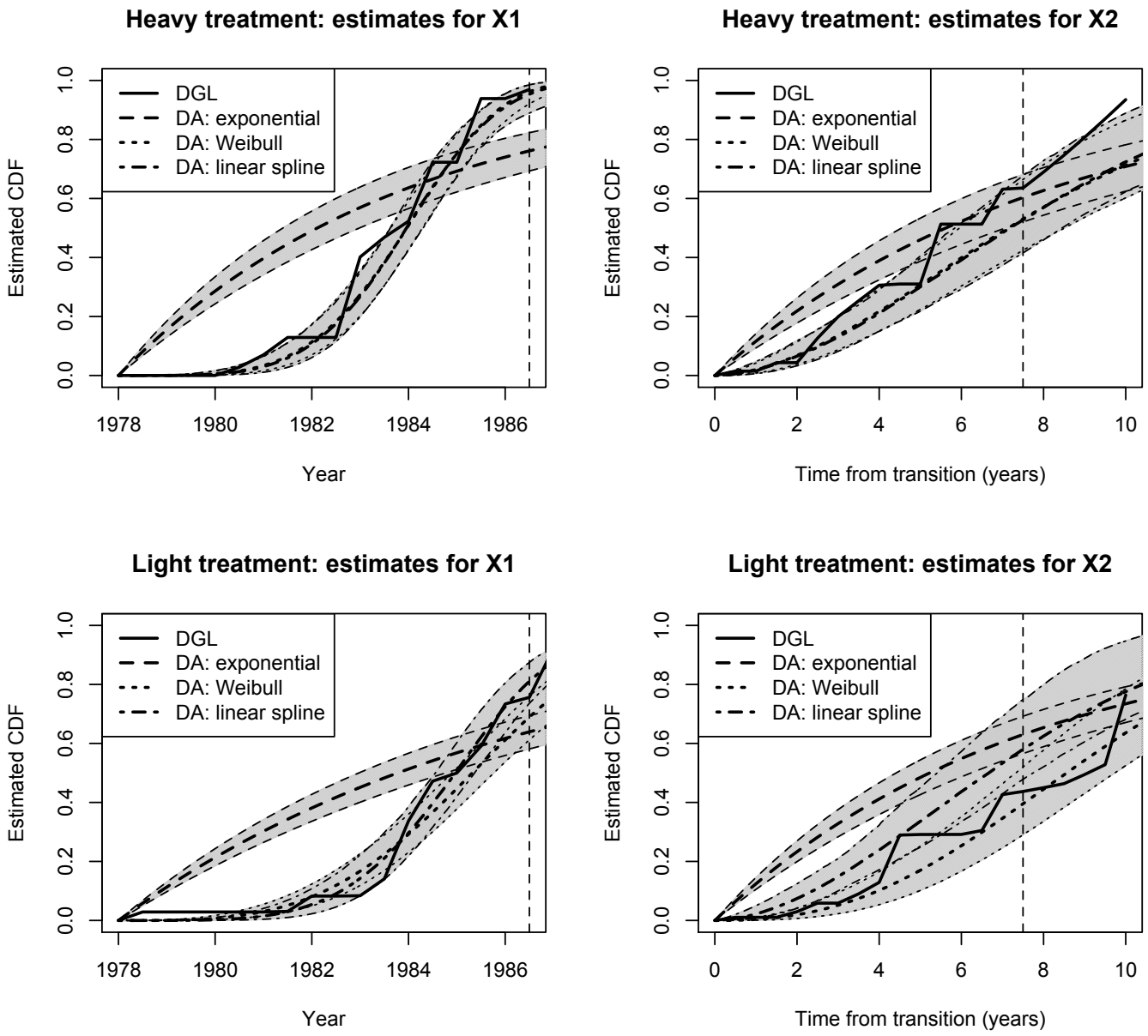


Figure 3.10: Estimated cumulative distribution functions (CDFs) of the sojourn times in the uninfected and infected states based on the proposed data augmentation method. Results are shown for heavily treated (upper panels) and lightly treated subjects (lower panels), based on exponential, Weibull, and linear spline models of the sojourn times in each state. For each model, the estimated median CDF is given and 95% posterior credible interval is shown as a grey band. Results from the method of De Gruttola and Lagakos are shown for reference.

Table 3.16: Convergence diagnostics for exponential and Weibull models: stationarity and interval halfwidth tests of Heidelberger and Welch.

			Stationarity		Halfwidth		
Model	Treatment	Parameter	Test	p-value	Test	Mean	Halfwidth
exponential	heavy	θ_1	passed	0.2782	passed	11.9	0.150
		θ_2	passed	0.0962	passed	16.4	0.356
	light	θ_1	passed	0.268	passed	16.8	0.144
		θ_2	passed	0.210	passed	15.2	0.158
Weibull	heavy	k_1	passed	0.5213	passed	4.38	0.0251
		θ_1	passed	0.2052	passed	12.98	0.0151
		k_2	passed	0.0612	passed	1.80	0.0365
		θ_2	passed	0.0707	passed	18.27	0.2121
	light	k_2	passed	0.3113	passed	3.50	0.0478
		θ_1	passed	0.0503	passed	16.28	0.0522
		k_2	passed	0.1052	passed	2.40	0.0467
		θ_2	passed	0.1842	passed	20.25	0.2096

To provide a measure of how well each model fits the data, we carry out tests of fit at several fixed times. Specifically, we consider the probability that a subject chosen at random has progressed to state 2 at chronological times 2.25, 4.25, and 6.25 years, and compare the observed proportions of subjects who have progressed to the expected proportion based on the estimates from each model. Due to right censoring, we cannot compute the observed proportions exactly, but we can put bounds on the true proportion. Note that we are interested here in the probability of progressing to state 2, rather than being in state 2, and hence we are not considering transitions from state 2 to 3. To compute the expected proportion of subjects in state 2 at each fixed time, we calculate the value of the estimated cumulative distribution function for the first sojourn time based on the parameter estimates.

For each time point and each model, for the heavily and lightly treated strata, the null hypothesis is that the number of subjects who have progressed to state 2 is binomially distributed with probability equal to the estimated cumulative distribution function at that time point. We use the normal approximation to the binomial distribution to compute the probability of observing the observed number of subjects, or a number more extreme, under the null hypothesis.

In Table 3.17 we present the z-statistics associated with each test for each model, consid-

Table 3.17: z-statistics corresponding to goodness-of-fit tests for each model.

Time points		2.25 years			4.25 years			6.25 years		
Treatment	Model	exp	Weibull	spline	exp	Weibull	spline	exp	Weibull	spline
heavy	optimistic	-6.946	-1.022	-0.914	-8.706	-1.747	-1.509	-5.580	-3.726	-3.801
	pessimistic	-2.744	18.743	21.132	-4.411	4.335	4.750	0.764	2.385	2.317
light	optimistic	-6.749	-0.525	-0.149	-9.504	-3.005	-2.599	-9.156	-4.451	-4.253
	pessimistic	-0.910	23.547	28.749	-2.815	8.093	9.289	-1.006	4.235	4.493

ering the minimum ("optimistic") and maximum ("pessimistic") possible observed numbers of subjects progressing to state 2 at each time. If the z-statistics have opposite signs, then the expected proportion based on the model is between the minimum and maximum observed proportions, so there is no indication that the model is a poor fit. We can see that there is strong evidence at 2.25 and 4.25 years that the exponential model is a poor fit to the observed data. This observation is reflected in the left panels of Figure 3.10.

3.7 Discussion

In this chapter we have proposed an approach for modeling a simple progressive process that is intermittently observed. This approach is able to accommodate a simple progressive process with any number of states. The proposed approach for a simple progressive disease process allows for a flexible modeling procedure of the time spent in each state in which any level of structure may be imposed on each of these sojourn times.

In any given situation, the choice of which approach to take depends on characteristics of the observed data such as the frequency of observation relative to the speed of the process, as well as on the assumptions about the underlying process. If subjects are observed infrequently relative to the speed of the process and the Markov assumption is appropriate, then an approach such as that of Kalbfleisch and Lawless (1985) that utilizes this assumption may be the best choice. For example, in our simulation studies in which subjects were observed infrequently, the method of Kalbfleisch and Lawless (1985) performed well (see subsection 3.4.2). The method of De Gruttola and Lagakos (1989) was prone to biased estimation since subjects not observed in the intermediate state had to be excluded from analysis. If, on the other hand, subjects are observed quite frequently relative to the speed

of the disease process, then several approaches may be suitable. If a nonparametric model is desired and *if every subject is observed in every visited state*, then the approach of De Gruttola and Lagakos (1989) or its extension Sternberg and Satten (1999) may be appropriate. Otherwise, if imposing some structure on the sojourn times is acceptable, then the proposed approach may be chosen. The final possibility is the intermediate case, in which subjects are observed moderately frequently relative to the speed of the process. If the Markov assumption is still appropriate, then the method of Kalbfleisch and Lawless (1985) may be used. Otherwise, the proposed approach may be applied, with an appropriate choice of parametric model for the sojourn time in each state. We note that the naïve approach presented in this chapter should never be used, as it is not methodologically sound.

In this development we have made several assumptions about the underlying process which may not be realistic for many applications. We will relax these assumptions in following chapters. We assumed in the present development that each patient visits each of the stages of disease in a prescribed sequence, but in an application, a patient may skip intermediate stages. In the following chapter we take that possibility into consideration.

Chapter 4

A SEMI-MARKOV MODEL FOR GENERAL PROGRESSIVE PROCESSES

4.1 Introduction

In the previous chapter we proposed an approach to model sojourn times in each state of a disease process under panel observation assuming that subjects visit each of these states in a sequence. In the current chapter we extend the approach to the case in which there is an order in which patients visit the states, but they may skip intermediate states as they progress. We begin by considering in Section 4.2 the simplest case of a general progressive process, the illness-death model, which is a very common and useful state model in applications (Frydman, 1992; Chang et al., 2001; Commenges et al., 2004). We lay the groundwork for carrying out inference on both the probability of visiting the illness state and on the sojourn times in each state. In Section 4.3 we extend the approach for the illness-death model to accommodate the more general progressive process in which it is possible to skip any number of intermediate states, and develop methodology to make inference on both the probability of taking a given path through the states and on the sojourn times in each of these states. In Section 4.4 we illustrate the performance of our proposed approach for a general progressive state model via simulation study, considering the illness-death model. We examine a number of scenarios in this simulation study. In Section 4.5, we briefly illustrate the performance of the approach for a general progressive state model with four states. Finally in Section 4.6 we discuss modeling choices in using a general progressive model.

4.2 Illness-death model

Here we consider the illness-death model, shown in Figure 4.1, which is the special case of a general progressive state model in which there are three states. Unlike the simple

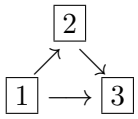


Figure 4.1: Illness-death model.

progressive state model we considered previously, in the illness-death model there is more than one possible trajectory through the states, since from state 1 it is possible to progress to state 2 or directly to state 3. We let X_{12} and X_{13} denote the sojourn times in state 1 and next transitioning to states 2 and 3, respectively. Similarly, we let X_{23} denote the sojourn time in state 2.

The embedded Markov chain $\mathbf{P} = [p_{ij}]$ is governed by the transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0 & p_{12} & 1-p_{12} \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix},$$

where $p_{12} \in (0,1)$ is the probability that a patient in state 1 makes a transition to state 2. We choose a parametric model for each of the conditional sojourn times. That is, we assume $X_{12} \sim f_{12}(\cdot)$, where f_{12} is some density that depends on parameters $\boldsymbol{\theta}_{12}$, and similarly for X_{23} and X_{13}. The statistical problem is to carry out inference on the parameters of interest, p_{12} and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{12}, \boldsymbol{\theta}_{23}, \boldsymbol{\theta}_{13}\}$.

We assume throughout this chapter that subjects enter state 1 at time zero. The n^{th} subject is assessed at times $\mathbf{s}^n = (s_0^n = 0, s_1^n, \ldots, s_{n_t}^n)$, which are assumed independent of the subject's disease progression but may differ across subjects. These periodic assessments give rise to the panel observations $\mathbf{Z} = (Z_0^n, Z_1^n, \ldots, Z_{n_t}^n)$ with $Z_i^n \in \{1,2,3\}$ for each i and n. As in Chapter 3, the progressive nature of the state model implies that $Z_0^n \leq \cdots \leq Z_{n_t}^n$, and in particular, the set of observations on each subject can be expressed equivalently as

$\mathbf{t}^n = (t_1^n, t_2^n, t_3^n, t_4^n)$ where each of these elements is defined as before:

$$\begin{aligned} t_1^n &= \max\{t_k^n : \ Z_k = 1\} \\ t_2^n &= \min\{t_k^n : \ Z_k = 2\} \\ t_3^n &= \max\{t_k^n : \ Z_k = 2\} \\ t_4^n &= \min\{t_k^n : \ Z_k = 3\}, \end{aligned}$$

whenever each exists. We let $\delta^n(i)$ be the indicator that a subject was not observed in state i for $i = 2, 3$. Subjects may not be observed in states 2 or 3 due to right censoring. Note, however, that for a subject who is not observed in state 2, there are two possible underlying trajectories through the states: the subject may have either truly visited state 2, or made a transition directly from state 1 to state 3.

Let v^n be the latent indicator that the n^{th} subject visited state 2. The sum of these indicators over all subjects gives the number of subjects with underlying paths that include a visit to state 2. Thus, a subject who is seen in state 2 has $\delta^n(2) = 0$ and $v^n = 1$. On the other hand, a subject who is not seen in state 2 has $\delta^n(2) = 1$ but v^n may be equal to either 0 or 1, as the subject may or may not have visited state 2. The collection of latent sojourn times depends on the latent true path of disease progression. If the subject's true path includes state 2, that is, if $v^n = 1$, then the latent sojourn times are represented by $\mathbf{X}^n = (X_{12}^n, X_{23}^n)$. Otherwise, the latent sojourn time is given by $\mathbf{X}^n = (X_{13}^n)$. Using the above notation, a subject who is observed in state 2 has $\delta^n(2) = 0$ and the set of latent sojourn times is given by $\mathbf{X}^n(v^n = 1)$. Conversely, a subject who is not observed in state 2 has $\delta^n(2) = 1$, and the set of latent sojourn times are given by $\mathbf{X}^n(v^n = 1)$ or $\mathbf{X}^n(v^n = 0)$ depending on the underlying latent path.

Given observed data on N subjects, $\mathbf{t}^1, \ldots, \mathbf{t}^N$, we must carry out inference on the parameters of interest, p_{12} and $\boldsymbol{\theta}$, as well as the latent data: the conditional sojourn times $\mathbf{X}^1, \ldots, \mathbf{X}^N$ as well as the visit indicators $v^1, \ldots, v^N$. As in Chapter 3, we alternately update the parameters of interest and latent data for each of the subjects. Next we discuss our approach to each of these two parts of the algorithm, beginning with the updating of the

parameters of interest.

For the simple progressive state model considered previously, we updated parameters corresponding to each sojourn time separately. In the present case, however, it is simpler to update all the parameters corresponding to the sojourn times simultaneously via block updating. We derive the full conditional distribution of the transition probability p_{12} as well as the joint full conditional distribution of the sojourn time parameters, given the observed data and fixed values of the latent conditional sojourn times.

First, we let $A^n = A^n(v^n, \boldsymbol{\delta}^n; \mathbf{t}^n)$ denote the set of "allowable" sojourn times corresponding to the path defined by v^n and observed data $(\mathbf{t}^n, \boldsymbol{\delta}^n)$. This region has one or two dimensions depending on v^n, and does not exist for some combinations of v^n and $\boldsymbol{\delta}^n$:

$$A^n = \begin{cases} \{(x_{12}, x_{23}) : \ x_{12} \in [t_1^n, t_2^n], \ x_{12} + x_{23} \in [t_3^n, t_4^n]\}, & v^n = 1, \ \boldsymbol{\delta}^n = (0,0); \\ \{(x_{12}, x_{23}) : \ x_{12} \in [t_1^n, t_2^n], \ x_{12} + x_{23} \in [t_3^n, \infty)\}, & v^n = 1, \ \boldsymbol{\delta}^n = (0,1); \\ \{(x_{12}, x_{23}) : \ x_{12} \in [t_1^n, t_4^n], \ x_{12} + x_{23} \in [t_1^n, t_4^n]\}, & v^n = 1, \ \boldsymbol{\delta}^n = (1,0); \\ \{x_{13} : \ x_{13} \in [t_1^n, t_4^n]\}, & v = 0, \ \boldsymbol{\delta}^n = (1,0); \\ \{(x_{12}, x_{23}) : \ x_{12} \in [t_1^n, \infty), \ x_{12} + x_{23} \in [t_1^n, \infty)\}, & v^n = 1, \ \boldsymbol{\delta}^n = (1,1); \\ \{x_{13} : \ x_{13} \in [t_1^n, \infty)\}, & v^n = 0, \ \boldsymbol{\delta}^n = (1,1), \end{cases}$$

for $x_{12}, x_{23}, x_{13} \geq 0$. The cases for which $v^n = 1$ correspond to the cases in Chapter 3: trapezoidal, triangular, and two unbounded regions, respectively.

Using this notation for the allowable region, we can express the joint posterior of the

parameters of interest and latent data given the observed data as

$$
\begin{aligned}
& \prod_{n=1}^{N} [p(v^n, \mathbf{X}^n | p_{12}, \boldsymbol{\theta}) \cdot p(\mathbf{t}^n, \boldsymbol{\delta}^n | p_{12}, \boldsymbol{\theta}, \mathbf{X}^n, v^n)] \cdot \pi(p_{12}, \boldsymbol{\theta}) \\
\propto \; & \prod_{n=1}^{N} [[p_{12} \cdot f_{12}(X_{12}^n) \cdot f_{23}(X_{23}^n)]^{v^n} \cdot [(1-p_{12}) \cdot f_{13}(X_{13}^n)]^{1-v^n} \\
& \cdot \mathbf{1}(X_{12}^n \in [t_1^n, t_2^n],\ X_{12}^n + X_{23}^n \in [t_3^n, t_4^n])^{(1-\delta^n(2)) \cdot (1-\delta^n(3))} \\
& \cdot \mathbf{1}(X_{13}^n \in [t_1^n, t_4^n],\ X_{12}^n + X_{23}^n \in [t_1^n, t_4^n])^{\delta^n(2) \cdot (1-\delta^n(3))} \\
& \cdot \mathbf{1}(X_{12}^n \in [t_1^n, t_2^n],\ X_{12}^n + X_{23}^n \in [t_3^n, \infty))^{(1-\delta^n(2)) \cdot \delta^n(3)} \\
& \cdot \mathbf{1}(X_{12}^n \in [t_1^n, \infty),\ X_{12}^n + X_{23}^n \in [t_1^n, \infty),\ X_{13}^n \in [t_1^n, \infty))^{\delta^n(2) \cdot \delta^n(3)} \cdot \pi(p_{12}, \boldsymbol{\theta}) \\
\propto \; & \prod_{n=1}^{N} [[p_{12} \cdot f_{12}(X_{12}^n) \cdot f_{23}(X_{23}^n)]^{v^n} \cdot [(1-p_{12}) \cdot f_{13}(X_{13}^n)]^{1-v^n} \\
& \cdot \mathbf{1}((X_{12}^n, X_{23}^n) \in A^n(v^n = 1, \boldsymbol{\delta}^n) \cdot \mathbf{1}(X_{13}^n \in A^n(v^n = 0, \boldsymbol{\delta}^n)) \cdot \pi(p_{12}, \boldsymbol{\theta}).
\end{aligned}
$$

From now on we abbreviate $\mathbf{1}((X_{12}^n, X_{23}^n) \in A^n(v^n = 1, \boldsymbol{\delta}^n)) \cdot \mathbf{1}(X_{13}^n \in A^n(v^n = 0, \boldsymbol{\delta}^n))$ as $\mathbf{1}(\mathbf{X}^n \in A^n)$.

Based on this joint posterior, we can express the full conditional distribution of the parameter p_{12}, conditional on the latent sojourn times, as

$$
\begin{aligned}
& p(p_{12} | \boldsymbol{\theta}, v^1, \ldots, v^N, \mathbf{X}^1(v^1), \ldots, \mathbf{X}^N(v^N), \mathbf{t}^1, \ldots, \mathbf{t}^N, \boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^N) \qquad (4.1) \\
& \quad \propto \prod_{n=1}^{N} [[p_{12} \cdot f_{12}(X_{12}^n) \cdot f_{23}(X_{23}^n)]^{v^n} \cdot [(1-p_{12}) \cdot f_{13}(X_{13}^n)]^{1-v^n}] \cdot \pi(p_{12}) \\
& \quad \propto p_{12}^{N_{12}} \cdot (1-p_{12})^{N-N_{12}} \cdot \pi(p_{12}),
\end{aligned}
$$

where $N_{12} \doteq \sum_{n=1}^{N} v^n$, the number of subjects who visited state 2. Assuming $p_{12} \sim Beta(a, b)$ for $a, b > 0$, this full conditional becomes

$$
\begin{aligned}
& p(p_{12} | \boldsymbol{\theta}, v^1, \ldots, v^N, \mathbf{X}^1(v^1), \ldots, \mathbf{X}^N(v^N), \mathbf{t}^1, \ldots, \mathbf{t}^N, \boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^N) \\
& \qquad \propto \; p_{12}^{N_{12}+a-1} \cdot (1-p_{12})^{N-N_{12}+b-1}.
\end{aligned}
$$

That is, conditional on the observed and latent data and the other parameters of interest,

p_{12} is $Beta(N_{12}+a, N-N_{12}+b)$. Thus, we use a Gibbs step to sample p_{12}.

The full conditional distribution of the parameters corresponding to the sojourn times, given p_{12} and the latent and observed data, can be expressed as

$$\begin{aligned} p(\boldsymbol{\theta}|p_{12}, v^1, \ldots, v^N, \mathbf{X}^1(v^1), \ldots, \mathbf{X}^N(v^N), \mathbf{t}^1, \ldots, \mathbf{t}^N, \boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^N) \qquad (4.2) \\ \propto \prod_{n=1}^{N} [[p_{12} \cdot f_{12}(X_{12}^n) \cdot f_{23}(X_{23}^n)]^{v^n} \cdot [(1-p_{12}) \cdot f_{13}(X_{13}^n)]^{1-v^n}] \cdot \pi(\boldsymbol{\theta}) \\ \propto \prod_{n=1}^{N} [[f_{12}(X_{12}^n) \cdot f_{23}(X_{23}^n)]^{v^n} \cdot [f_{13}(X_{13}^n)]^{1-v^n}] \cdot \pi(\boldsymbol{\theta}). \end{aligned}$$

The particular form of this distribution depends on the chosen model for each of the conditional sojourn times X_{12}, X_{23}, and X_{13}. Since it is not straightforward to sample from this distribution in general, we use a Metropolis-Hastings step to update these parameters. We use a normal or truncated normal proposal distribution as appropriate to generate a candidate value of each individual parameter, with mean equal to the current parameter value and a fixed standard deviation.

The other half of the algorithm deals with the latent data. Specifically, we must update the latent visit indicator and sojourn times for each of the subjects. We assume that subjects progress through the health states independently of one another, conditional on the embedded Markov chain and the distributions of the sojourn times. That is,

$$p(v^1, \ldots, v^N, \mathbf{X}^1(v^1), \ldots, \mathbf{X}^N(v^N)|p_{12}, \boldsymbol{\theta}) \;=\; \prod_{n=1}^{N} p(v^n, \mathbf{X}^n(v^n)|p_{12}, \boldsymbol{\theta}).$$

Therefore we can update the latent data for each subject individually, as we did in the original proposed approach in the previous chapter.

As we have noted, a subject not seen in state 2 has two possible trajectories through the states. The models corresponding to these two trajectories are of different dimension. We must address this uncertainty in the dimension of the model. A common approach to address dimension uncertainty in the model is to implement a reversible jump Markov chain Monte Carlo (RJMCMC) algorithm (Green, 1995). However, the performance of the RJMCMC algorithm is often plagued with slow mixing as it can be quite challenging to devise

good proposal distributions for the variable that matches the dimensions between different trajectories and a one-to-one, differentiable, transformation function that generates proposals consistent with the data. A related approach, proposed by Carlin and Chib (1995), involves consideration of the composite space of all candidate models. Godsill (1997) proposed a generalization of the Carlin-Chib algorithm known as the *Metropolized Carlin-Chib* algorithm. We apply the Metropolized Carlin-Chib algorithm to our problem. Specifically, for the n^{th} subject we consider the joint full conditional distribution $p(v^n, \mathbf{X}^n | p_{12}, \boldsymbol{\theta}, \mathbf{t}^n)$ of the model indicator v^n and all latent sojourn times $\mathbf{X}^n$ that are defined. In the Metropolized Carlin-Chib algorithm, if the current state of the chain is $(v, \mathbf{X})$, then

- Generate $(v^*, \mathbf{X}^*)$ from a proposal transition kernel $q(v^*, \mathbf{X}^* | v, \mathbf{X}, p_{12}, \boldsymbol{\theta}, \mathbf{t})$.
- Accept candidate $(v^*, \mathbf{X}^*)$ with probability

$$\alpha \quad = \quad \min \left\{ 1, \frac{p(v^*, \mathbf{X}^* | p_{12}, \boldsymbol{\theta}, \mathbf{t}) \cdot q(v, \mathbf{X} | v^*, \mathbf{X}^*, p_{12}, \boldsymbol{\theta}, \mathbf{t})}{p(v, \mathbf{X} | p_{12}, \boldsymbol{\theta}, \mathbf{t}) \cdot q(v^*, \mathbf{X}^* | v, \mathbf{X}, p_{12}, \boldsymbol{\theta}, \mathbf{t})} \right\},$$

where, from the above point on, we suppress dependence on n for readability. We now make the form of the joint full conditional explicit and discuss how to choose the proposal transition kernel.

First, we let $p(\mathbf{X}(v)|v)$ denote the "prior" distribution of the latent sojourn times corresponding to model v. Analogously, we let $p(\mathbf{X}(1-v)|\mathbf{X}(v), v)$ represent the distribution of the latent sojourn times that correspond to the other model, $1-v$, conditional on model v and $\mathbf{X}(v)$. This distribution is often referred to as a *pseudo-prior*.

The joint full conditional distribution of the model indicator v and the suite of all latent sojourn times $\mathbf{X}$ that are defined is given by

$$\begin{aligned} p(v, \mathbf{X} | p_{12}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) &\propto p(\mathbf{X} | v, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \cdot p(v | p_{12}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \\ &\propto p(\mathbf{X}(v) | v, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \cdot p(\mathbf{X}(1-v) | \mathbf{X}(v), v, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \cdot p(v | p_{12}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \\ &\propto f_{12}(X_{12}) \cdot f_{23}(X_{23}) \cdot [f_{13}(X_{13})]^{\delta(2)} \cdot \mathbf{1}(\mathbf{X} \in A) \cdot p_v^v \cdot (1-p_v)^{1-v}, \end{aligned}$$

where p_v is the full conditional probability that a subject visited state 2, which we now compute.

Using Bayes' rule we have:

$$\begin{aligned} p(v|p_{12},\boldsymbol{\theta},\mathbf{t},\boldsymbol{\delta}) &\propto p(\mathbf{t}|v,p_{12},\boldsymbol{\theta},\boldsymbol{\delta})\cdot p(v|p_{12},\boldsymbol{\theta},\boldsymbol{\delta}) \\ &\propto p(\mathbf{t}|v,\boldsymbol{\theta})\cdot p(v|p_{12},\boldsymbol{\delta}). \end{aligned}$$

We examine these two factors for a single subject:

$\boldsymbol{\delta}$	v	$p(\mathbf{t}\|v,\boldsymbol{\theta}) \propto$:	$p(n\|p_{12},\boldsymbol{\delta})$
(0,0)	1	$P(X_{12}\in[t_1,t_2],\ X_{12}+X_{23}\in[t_3,t_4])$	1
	0	0	0
(1,0)	1	$P(X_{12}\in[t_1,t_4],\ X_{12}+X_{23}\in[t_1,t_4])$	p_{12}
	0	$P(X_{13}\in[t_1,t_4])$	$1-p_{12}$
(0,1)	1	$P(X_{12}\in[t_1,t_2],\ X_{12}+X_{23}\in[t_3,\infty))$	1
	0	0	0
(1,1)	1	$P(X_{12}\in[t_1,\infty))$	p_{12}
	0	$P(X_{13}\in[t_1,\infty))$	$1-p_{12}$

Expressing each subject's contribution to this full conditional in another way, we have:

$$p(v|p_{12},\boldsymbol{\theta},\mathbf{t},\boldsymbol{\delta}) = p_v^v\cdot(1-p_v)^{1-v}$$

where

$$p_v = \begin{cases} 1, & \boldsymbol{\delta}=(0,0); \\ \frac{p_{12}\cdot P(X_{12}\in[t_1,t_4],\ X_{12}+X_{23}\in[t_1,t_4])}{p_{12}\cdot P(X_{12}\in[t_1,t_4],\ X_{12}+X_{23}\in[t_1,t_4])+(1-p_{12})\cdot P(X_{13}\in[t_1,t_4])}, & \boldsymbol{\delta}=(1,0); \\ 1, & \boldsymbol{\delta}=(0,1); \\ \frac{p_{12}\cdot P(X_{12}\in[t_1,\infty))}{p_{12}\cdot P(X_{12}\in[t_1,\infty))+(1-p_{12})\cdot P(X_{13}\in[t_1,\infty))}, & \boldsymbol{\delta}=(1,1). \end{cases}$$

Hence, for a subject who is observed in state 2, v is equal to one with probability one, which is what we would expect. For a subject not observed in state 2, v is Bernoulli with the probability given in the above fraction that corresponds to whether the subject was observed in state 3. The probabilities involved in the fraction in the second case respectively correspond to the "relative probabilities" that a subject who was not observed in state 2

did or did not visit state 2, and can be calculated as:

$$\begin{aligned} P(X_{12} = u \in [t_1, t_4],\ X_{23} \in [0, t_4 - u]) &= \int_{t_1}^{t_4} f_{12}(u) \cdot F_{23}(t_4 - u)du \\ P(X_{13} = u \in [t_1, t_4]) &= \int_{t_1}^{t_4} f_{13}(u)du \ = \ F_{13}(t_4) - F_{13}(t_1) \\ P(X_{12} \in [t_1, \infty)) &= \int_{t_1}^{\infty} f_{12}(u)du \ = \ 1 - F_{12}(t_1) \\ P(X_{13} \in [t_1, \infty)) &= \int_{t_1}^{\infty} f_{13}(u)du \ = \ 1 - F_{13}(t_1). \end{aligned}$$

In this light, this Bernoulli probability is intuitive. In our implementation of the algorithm we use stochastic integration to evaluate the first of these four quantities for each subject. We update p_v every ten iterations to reduce the computational burden.

Now we discuss how to generate values for the visit indicator v and latent sojourn times $\mathbf{X}$ for the n^{th} subject—that is, how to choose the proposal transition kernel $q(\cdot)$. Although, in theory, any choice of $q(\cdot)$ is acceptable, a poor choice can lead to very slow mixing of the chain. We first consider a very simple choice for $q(\cdot)$:

$$q_1(v^*, \mathbf{X}^*|v, \mathbf{X}, p_{12}, \boldsymbol{\theta}, \mathbf{t}) \ \propto \ p_\delta^{v^*} \cdot (1 - p_\delta)^{1-v^*} \cdot q_X(\mathbf{X}^*|\mathbf{t}) \cdot \mathbf{1}(\mathbf{X} \in A),$$

where p_δ is equal to one if $\delta(2) = 0$ and to a relatively small probability, such as 20%, if $\delta(2) = 1$. Hence, for a subject who was not seen in state 2, the proposed trajectory involves having visited state 2 with probability 20%. The distribution $q_X(\mathbf{X}^*|\mathbf{t})$ represents the proposal for the latent sojourn times. One simple approach is the following:

$$q_X(\mathbf{X}^*|\mathbf{t}) \ = \ \begin{cases} (X_{12}, X_{23}) \sim Unif(A(v = 1)), & \boldsymbol{\delta} = (0, 0); \\ (X_{12}, X_{23}) \sim Unif(A(v = 1));\ X_{13} \sim Unif(A(v = 0)), & \boldsymbol{\delta} = (1, 0); \\ X_{12} \sim Unif(t_1, t_2);\ X_{23} \sim t_2 + exp(\lambda), & \boldsymbol{\delta} = (0, 1); \\ X_{12}, X_{13} \sim t_1 + exp(\lambda);\ X_{23} \sim exp(\lambda), & \boldsymbol{\delta} = (1, 1), \end{cases}$$

for some scale parameter $\lambda > 0$. Note that we are using uniform and exponential proposals for regions that are respectively bounded and unbounded in a given dimension. This

proposal for the latent sojourn times ensures that a variety of allowable sojourn times are considered.

Another choice for the proposal transition kernel involves the best guess we have available for the probability that a subject visited state 2. That is, for each subject we propose a trajectory with $v = 1$ with probability p_v, the full conditional probability that the subject visited state 2:

$$q_2(v^*, \mathbf{X}^*|v, \mathbf{X}, p_{12}, \boldsymbol{\theta}, \mathbf{t}) \quad \propto \quad p_v^{v^*}(1-p_v)^{1-v^*} \cdot q_X(\mathbf{X}^*|\mathbf{t}) \cdot \mathbf{1}(\mathbf{X}^* \in A),$$

where $q_X(\mathbf{X}^*|\mathbf{t})$ is the proposal for the latent sojourn times, as before.

Given the full conditional distribution of a subject's trajectory and $q_1(\cdot)$ as the choice of proposal transition kernel, assuming that the candidate $\mathbf{X}^*$ is in the allowable region A, the acceptance probability is given by

$$\begin{aligned}
\alpha &= \min\left\{1, \frac{\pi^*(v^*, \mathbf{X}^*|p_{12}, \boldsymbol{\theta}, \mathbf{t}^n) \cdot q_1(v, \mathbf{X}|v^*, \mathbf{X}^*)}{\pi^*(v, \mathbf{X}|p_{12}, \boldsymbol{\theta}, \mathbf{t}^n) \cdot q_1(v^*, \mathbf{X}^*|v, \mathbf{X})}\right\} \\
&= \min\left\{1, \frac{p_v^{v^*}(1-p_v)^{1-v^*} \cdot f_{12}(X_{12}^*) \cdot f_{23}(X_{23}^*) \cdot [f_{13}(X_{13}^*)]^{\delta^n(2)} \cdot p_\delta^v(1-p_\delta)^{1-v} \cdot q_X(\mathbf{X}|\mathbf{t})}{p_v^v(1-p_v)^{1-v} \cdot f_{12}(X_{12}) \cdot f_{23}(X_{23}) \cdot [f_{13}(X_{13})]^{\delta^n(2)} \cdot p_\delta^{v^*}(1-p_\delta)^{1-v^*} \cdot q_X(\mathbf{X}^*|\mathbf{t})}\right\}.
\end{aligned}$$

If, instead, $q_2(\cdot)$ is chosen, then if $\mathbf{X}^* \in A$, the acceptance probability is given by

$$\begin{aligned}
\alpha &= \min\left\{1, \frac{\pi^*(v^*, \mathbf{X}^*|p_{12}, \boldsymbol{\theta}, \mathbf{t}^n) \cdot q_2(v, \mathbf{X}|v^*, \mathbf{X}^*)}{\pi^*(v, \mathbf{X}|p_{12}, \boldsymbol{\theta}, \mathbf{t}^n) \cdot q_2(v^*, \mathbf{X}^*|v, \mathbf{X})}\right\} \\
&= \min\left\{1, \frac{p_{12}^{v^*}(1-p_{12})^{1-v^*} \cdot f_{12}(X_{12}^*) \cdot f_{23}(X_{23}^*) \cdot [f_{13}(X_{13}^*)]^{\delta^n(2)} \cdot p_v^v(1-p_v)^{1-v} \cdot q_X(\mathbf{X}|\mathbf{t})}{p_{12}^v(1-p_{12})^{1-v} \cdot f_{12}(X_{12}) \cdot f_{23}(X_{23}) \cdot [f_{13}(X_{13})]^{\delta^n(2)} \cdot p_v^{v^*}(1-p_v)^{1-v^*} \cdot q_X(\mathbf{X}^*|\mathbf{t})}\right\} \\
&= \min\left\{1, \frac{f_{12}(X_{12}^*) \cdot f_{23}(X_{23}^*) \cdot [f_{13}(X_{13}^*)]^{\delta^n(2)} \cdot q_X(\mathbf{X}|\mathbf{t})}{f_{12}(X_{12}) \cdot f_{23}(X_{23}) \cdot [f_{13}(X_{13})]^{\delta^n(2)} \cdot q_X(\mathbf{X}^*|\mathbf{t})}\right\}.
\end{aligned}$$

We note that this acceptance probability does not involve the full conditional probability that a subject visited state 2 (p_v), or even the current value of p_{12}, and its form is quite simple. Additionally, making use of p_v in the proposal for v leads to better mixing than using a naïve proposal for v. Because of these computational advantages, we choose to use the proposal $q_2(\cdot)$ to produce a candidate trajectory for each subject in the simulations. Note that for the other proposal transition kernel under consideration, $q_1(\cdot)$, choosing a reasonable p_δ depends on the observation interval, and a poor choice may lead to slow

mixing of the chain.

Hence, in this approach, for a subject not seen in state 2, at each iteration of the algorithm we update the model indicator v as well as the latent sojourn times X_{12}, X_{23}, and X_{13}. For a subject who was seen in state 2, X_{13} is not defined and the algorithm reduces to a standard Metropolis-Hastings step. Specifically, if the subject was seen in state 2, the joint full conditional of $(v, \mathbf{X})$ becomes

$$p(v, \mathbf{X}|p_{12}, \mathbf{t}, \boldsymbol{\theta}) \quad = \quad f_{12}(X_{12}) \cdot f_{23}(X_{23}) \cdot \mathbf{1}(\mathbf{X}(v=1) \in A(v=1)).$$

We update the latent data for this subject as follows:

- Generate (X_{12}^*, X_{23}^*).
- If $\mathbf{X}^*(v=1) \in A(v=1)$, accept this set of candidate latent sojourn times with probability

$$\alpha \quad = \quad \min\left\{1, \frac{f_{12}(X_{12}^*) \cdot f_{23}(X_{23}^*) \cdot q(\mathbf{X}|\mathbf{t})}{f_{12}(X_{12}) \cdot f_{23}(X_{23}) \cdot q(\mathbf{X}^*|\mathbf{t})}\right\}.$$

To summarize the algorithm, the first half deals with the parameters of interest: we update p_{12} via Gibbs sampling, and the sojourn time parameters $\boldsymbol{\theta}$ via Metropolis-Hastings with appropriate normal or truncated normal proposal distributions on each parameter. The second half of the algorithm deals with the latent data: we update the indicator that each subject visited state 2 and the conditional sojourn times using the Metropolized Carlin-Chib algorithm. We use the "best guess" proposal $q_2(\cdot)$ for the indicators and a uniform or exponential proposal for the sojourn times as appropriate.

4.3 General progressive model

In this section we consider the case of a general progressive model with m states where $m \geq 3$. As we have discussed, in a general progressive model it is possible to transition directly from state i to state j only if $i < j$. The transition probability matrix governing

the embedded Markov chain is given by

$$\mathbf{P} = \begin{bmatrix} 0 & p_{12} & p_{13} & \cdots & p_{1m} \\ 0 & 0 & p_{23} & \cdots & p_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix},$$

where $p_{ij} \in (0,1)$ for each (i,j) and $\sum_{j=i+1}^{m} p_{ij} = 1$ for each $i \in \{1,\ldots,m-1\}$. Hence, for a given m, there are $\frac{1}{2} \cdot (m-1) \cdot (m-2)$ independent probabilities, with the above restrictions. There are $\frac{1}{2} \cdot m \cdot (m-1)$ latent sojourn times, X_{ij} with $j > i$ for $i \in \{1,\ldots,m\}$. There are 2^{m-2} possible complete trajectories through the states—that is, paths beginning in state 1 and ending in state m—since each intermediate state is either visited or skipped. A trajectory can thus be defined by indicators $v_2,\ldots,v_{m-1}$ that a subject visited each of states $2,\ldots,m-1$ respectively. We denote $\mathbf{v} = (v_2,\ldots,v_{m-1})$ and treat these state visit indicators as latent data. We have suppressed the dependence on the subject index, n, for simplicity.

We will show how the proposed approach to the illness-death model can be extended to the general progressive model. First, we note that each subject's panel observations $\mathbf{Z} = (Z_0, Z_1, \ldots, Z_{n_t})$ with $Z_i \in \{1,2,\ldots,m\}$ may be expressed equivalently as $\mathbf{t} = (t_1, t_2, \ldots, t_{2(m-1)})$, where these elements are defined as

$$\begin{aligned} t_1 &= \max\{t_k : \ Z_k = 1\} \\ t_2 &= \min\{t_k : \ Z_k = 2\} \\ &\vdots \\ t_{2(m-1)} &= \min\{t_k : \ Z_k = m\}, \end{aligned}$$

whenever each of these exists. We let $\delta(i)$ be the indicator that a subject was not observed in state i for $i = 2,\ldots,m$.

Recall that we are currently assuming that each subject enters state 1 at time zero. For

the n^{th} subject we have observed data $\mathbf{t}^n = (t_1^n, \ldots, t_{2(m-1)}^n)$ and vector of derived indicators $\boldsymbol{\delta}^n = (\delta^n(2), \ldots, \delta^n(m))$ that the subject was not observed in states $2, \ldots, m$. Each subject has an associated collection of latent conditional sojourn times corresponding to the path defined by the set of all possible state visit indicators: $\mathbf{X}^n = \{\mathbf{X}^n(\mathbf{v}) : \ v_i \in \{1, 0\}, \ i = 2, \ldots, m-1\}$, where $\mathbf{X}^n(\mathbf{v})$ is the set of path-specific sojourn times corresponding to $\mathbf{v}$. For example, with $m = 4$, we have

$$\begin{aligned}
\mathbf{X}^n &= \{\mathbf{X}^n(1,1), \mathbf{X}^n(1,0), \mathbf{X}^n(0,1), \mathbf{X}^n(0,0)\} \\
&= \{(X_{12}^n(1,1), X_{23}^n(1,1), X_{34}^n(1,1)), (X_{12}^n(1,0), X_{24}^n(1,0)), \\
&\quad (X_{13}^n(0,1), X_{34}^n(0,1)), (X_{14}^n(0,0))\}.
\end{aligned}$$

The existence of each path-specific sequence of sojourn times depends on the subject's observed data. In the example, if the subject was observed in state 2 but not state 3, then only $\mathbf{X}^n(1,1)$ and $\mathbf{X}^n(1,0)$ would exist, since these are the only two path-specific sequence of sojourn times that involve visiting state 2. We note that the sequence of sojourn times corresponding to visiting every state—$\mathbf{X}^n(\mathbf{v})$ with $v_i = 1$ for each i—always exists.

Given observed data on N subjects, $\mathbf{t}^1, \ldots, \mathbf{t}^N$, we must carry out inference on the parameters of interest: $\mathbf{p} = \{p_{12}, \ldots, p_{1m}, p_{23}, \ldots, p_{2m}, \ldots, p_{m-1,m}\}$ and $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{12}, \ldots, \boldsymbol{\theta}_{1m}, \boldsymbol{\theta}_{23}, \ldots, \boldsymbol{\theta}_{2m}, \ldots, \boldsymbol{\theta}_{m-1,m}\}$, as well as the latent data: the collection of path-specific sojourn times $\mathbf{X}^1, \ldots, \mathbf{X}^N$ and the vectors of latent visit indicators $\mathbf{v}^1, \ldots, \mathbf{v}^N$.

For the n^{th} subject we let $A^n = A^n(\mathbf{v}, \boldsymbol{\delta}^n; \mathbf{t}^n)$ denote the "allowable" region for $\mathbf{X}^n(\mathbf{v})$. This region has between one and $m-1$ dimensions, since the dimension is given by the number of latent sojourn times for the path defined by $\mathbf{v}$. The form of A^n depends on the type of observation $\boldsymbol{\delta}^n$ as well as $\mathbf{v}$. For $m > 3$, the allowable region is the logical extension of what it was in the illness-death case. For example, for $m = 4$, A^n is a region of 1–3

dimensions given by

$$A^n = \begin{cases}
\{(x_{12}, x_{23}, x_{34}) : \ x_{12} \in [t_1, t_2], \ x_{12} + x_{23} \in [t_3, t_4], \ x_{12} + x_{23} + x_{34} \in [t_5, t_6]\}, & \mathbf{v}^n = (1,1), \ \boldsymbol{\delta}^n = (0,0,0); \\ \hline
\{(x_{12}, x_{23}, x_{34}) : \ x_{12} \in [t_1, t_2], \ x_{12} + x_{23} \in [t_3, t_4], \ x_{12} + x_{23} + x_{34} \in [t_5, \infty)\}, & \mathbf{v}^n = (1,1), \ \boldsymbol{\delta}^n = (0,0,1); \\ \hline
\{(x_{12}, x_{23}, x_{34}) : \ x_{12} \in [t_1, t_2], \ x_{12} + x_{23} \in [t_3, t_6], \ x_{12} + x_{23} + x_{34} \in [t_3, t_6]\}, & \mathbf{v}^n = (1,1), \ \boldsymbol{\delta}^n = (0,1,0); \\
\{(x_{12}, x_{24}) : \ x_{12} \in [t_1, t_2], \ x_{12} + x_{24} \in [t_3, t_6]\}, & \mathbf{v}^n = (1,0), \ \boldsymbol{\delta}^n = (0,1,0); \\ \hline
\{(x_{12}, x_{23}, x_{34}) : \ x_{12} \in [t_1, t_2], \ x_{12} + x_{23} \in [t_3, \infty)\}, & \mathbf{v}^n = (1,1), \ \boldsymbol{\delta}^n = (0,1,1); \\
\{(x_{12}, x_{24}) : \ x_{12} \in [t_1, t_2], \ x_{12} + x_{24} \in [t_3, \infty)\}, & \mathbf{v}^n = (1,0), \ \boldsymbol{\delta}^n = (0,1,1); \\ \hline
\{(x_{12}, x_{23}, x_{34}) : \ x_{12} \in [t_1, t_4], \ x_{12} + x_{23} \in [t_1, t_4], \ x_{12} + x_{23} + x_{34} \in [t_5, t_6]\}, & \mathbf{v}^n = (1,1), \ \boldsymbol{\delta}^n = (1,0,0); \\
\{(x_{13}, x_{34}) : \ x_{13} \in [t_1, t_4], \ x_{13} + x_{34} \in [t_5, t_6]\}, & \mathbf{v}^n = (1,1), \ \boldsymbol{\delta}^n = (1,0,0); \\ \hline
\{(x_{12}, x_{23}, x_{34}) : \ x_{12} \in [t_1, t_4], \ x_{12} + x_{23} \in [t_1, t_4], \ x_{12} + x_{23} + x_{34} \in [t_5, \infty)\}, & \mathbf{v}^n = (1,1), \ \boldsymbol{\delta}^n = (1,0,1); \\
\{(x_{13}, x_{34}) : \ x_{13} \in [t_1, t_4], \ x_{13} + x_{34} \in [t_5, \infty)\}, & \mathbf{v}^n = (0,1), \ \boldsymbol{\delta}^n = (1,0,1); \\ \hline
\{(x_{12}, x_{23}, x_{34}) : \ x_{12} \in [t_1, t_6], \ x_{12} + x_{23} \in [t_1, t_6], \ x_{12} + x_{23} + x_{34} \in [t_1, t_6]\}, & \mathbf{v}^n = (1,1), \ \boldsymbol{\delta}^n = (1,1,0); \\
\{(x_{12}, x_{24}) : \ x_{12} \in [t_1, t_6], \ x_{12} + x_{24} \in [t_1, t_6]\}, & \mathbf{v}^n = (1,0), \ \boldsymbol{\delta}^n = (1,1,0); \\
\{(x_{13}, x_{34}) : \ x_{13} \in [t_1, t_6], \ x_{13} + x_{34} \in [t_1, t_6]\}, & \mathbf{v}^n = (0,1), \ \boldsymbol{\delta}^n = (1,1,0); \\
\{x_{14} : \ x_{14} \in [t_1, t_6]\}, & \mathbf{v}^n = (0,0), \ \boldsymbol{\delta}^n = (1,1,0); \\ \hline
\{(x_{12}, x_{23}, x_{34}) : \ x_{12} \in [t_1, t_6], \ x_{12} + x_{23} \in [t_1, t_6], \ x_{12} + x_{23} + x_{34} \in [t_1, \infty)\}, & \mathbf{v}^n = (1,1), \ \boldsymbol{\delta}^n = (1,1,1); \\
\{(x_{12}, x_{24}) : \ x_{12} \in [t_1, \infty)\}, & \mathbf{v}^n = (1,0), \ \boldsymbol{\delta}^n = (1,1,1); \\
\{(x_{13}, x_{34}) : \ x_{13} \in [t_1, \infty)\}, & \mathbf{v}^n = (0,1), \ \boldsymbol{\delta}^n = (1,1,1); \\
\{x_{14} : \ x_{14} \in [t_1, \infty)\}, & \mathbf{v}^n = (0,0), \ \boldsymbol{\delta}^n = (1,1,1).
\end{cases}$$

Similarly to the illness-death case, A^n does not exist for some combinations of $\mathbf{v}^n$ and $\boldsymbol{\delta}^n$. The form of A^n for larger m can be expressed as the obvious extension of the form given above.

The joint posterior of the parameters of interest and latent data given the observed data can be written as

$$\begin{aligned}
&\prod_{n=1}^{N} [p(\mathbf{v}^n, \mathbf{X}^n(\mathbf{v}^n) | \mathbf{p}, \boldsymbol{\theta}) \cdot p(\mathbf{t}^n, \boldsymbol{\delta}^n | \mathbf{p}, \boldsymbol{\theta}, \mathbf{v}^n, \mathbf{X}^n(\mathbf{v}^n))] \cdot \pi(\mathbf{p}, \boldsymbol{\theta}) \\
&\quad \propto \prod_{n=1}^{N} \prod_{i=1}^{m-1} \prod_{j=i+1}^{m} [p_{ij} \cdot f_{ij}(X_{ij}^n(\mathbf{v}^n))]^{v_i^n \cdot v_j^n \cdot \prod_{k=i+1}^{j-1} (1 - v_k^n)} \cdot \mathbf{1}(\mathbf{X}^n(\mathbf{v}^n) \in A^n) \cdot \pi(\mathbf{p}, \boldsymbol{\theta}),
\end{aligned}$$

where we define $v_1^n = 1$ and $v_m^n = 1$ for notational convenience.

We can express the full conditional distribution of the transition probabilities $\mathbf{p}$, conditional on the latent sojourn times, as

$$\begin{aligned} p(\mathbf{p}|\boldsymbol{\theta}, \mathbf{v}^1, \ldots, \mathbf{v}^N, \mathbf{X}^1(\mathbf{v}^1), \ldots, \mathbf{X}^N(\mathbf{v}^N), \mathbf{t}^1, \ldots, \mathbf{t}^N, \boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^N) \\ \propto \prod_{n=1}^{N} \prod_{i=1}^{m-1} \prod_{j=i+1}^{m} [p_{ij} \cdot f_{ij}(X_{ij}^n(\mathbf{v}^n))]^{v_i^n \cdot \prod_{k=i+1}^{j-1}(1-v_k^n) \cdot v_j^n} \cdot \pi(\mathbf{p}) \\ \propto \prod_{n=1}^{N} \prod_{i=1}^{m-1} \prod_{j=i+1}^{m} p_{ij}^{N_{ij}} \cdot \pi(\mathbf{p}), \end{aligned}$$

where $N_{ij} \doteq \sum_{n=1}^{N} v_i^n \cdot \prod_{k=i+1}^{j-1}(1-v_k^n) \cdot v_j^n$ is the number of subjects who transition directly from state i to state j, conditional on the state visit indicators. Assuming a Dirichlet prior for the transition probabilities from each state: $\pi(\mathbf{p}_i) \sim Dir(a_{i,i+1}, \ldots, a_{i,m})$ where each $a_{i,j} > 0$ for $i = 1, \ldots, m-1$ and $j = i+1, \ldots, m$, this full conditional becomes

$$\begin{aligned} p(\mathbf{p}|\boldsymbol{\theta}, \mathbf{v}^1, \ldots, \mathbf{v}^N, \mathbf{X}^1(\mathbf{v}^1), \ldots, \mathbf{X}^N(\mathbf{v}^N), \mathbf{t}^1, \ldots, \mathbf{t}^N, \boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^N) \\ \propto \prod_{n=1}^{N} \prod_{i=1}^{m-1} \prod_{j=i+1}^{m} p_{ij}^{N_{ij}+a_{ij}-1}. \end{aligned}$$

That is, the full conditional distribution of $\mathbf{p}_i$ for $i = 1, \ldots, m-1$ is also Dirichlet.

The full conditional distribution of the parameters corresponding to the sojourn times is given by

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{p}, \mathbf{v}^1, \ldots, \mathbf{v}^N, \mathbf{X}^1(\mathbf{v}^1), \ldots, \mathbf{X}^N(\mathbf{v}^N), \mathbf{t}^1, \ldots, \mathbf{t}^N, \boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^N) \\ \propto \prod_{n=1}^{N} \prod_{i=1}^{m-1} \prod_{j=i+1}^{m} [p_{ij} \cdot f_{ij}(X_{ij}^n(\mathbf{v}^n))]^{v_i^n \cdot v_j^n \cdot \prod_{k=i+1}^{j-1}(1-v_k^n)} \cdot \pi(\boldsymbol{\theta}) \\ \propto \prod_{n=1}^{N} \prod_{i=1}^{m-1} \prod_{j=i+1}^{m} f_{ij}(X_{ij}^n(\mathbf{v}^n))^{N_{ij}} \cdot \pi(\boldsymbol{\theta}). \end{aligned}$$

As in the illness-death model, the particular form of this distribution depends on the chosen model for each of the sojourn times. We again use a Metropolis-Hastings step to update these parameters, choosing a normal or truncated normal proposal distribution as appropriate to

generate a candidate value of each individual parameter with mean equal to the current parameter value and a fixed standard deviation.

We update the vector of latent visit indicators and sojourn times for each of the subjects as in the illness-death model, assuming that subjects progress through the health states independently of one another conditional on the embedded Markov chain and the distributions of the sojourn times:

$$p(\mathbf{v}^1, \ldots, \mathbf{v}^N, \mathbf{X}^1(\mathbf{v}^1), \ldots, \mathbf{X}^N(\mathbf{v}^N)|\mathbf{p}, \boldsymbol{\theta}) \quad = \quad \prod_{n=1}^{N} p(\mathbf{v}^n, \mathbf{X}^n(\mathbf{v}^n)|\mathbf{p}, \boldsymbol{\theta}).$$

Hence, as before, we can update the latent data for each subject individually.

A subject who is not seen in one or more intermediate states $2, \ldots, m-1$ has multiple possible trajectories through the states, and each possible trajectory corresponds to a different model. Since these candidate models involve different numbers of visited states, they are of different dimensions. As in the illness-death model, we address this uncertainty in the dimension of the model via the Metropolized Carlin-Chib algorithm. For the n^{th} subject we consider the joint full conditional distribution $p(\mathbf{v}^n, \mathbf{X}^n|\mathbf{p}, \boldsymbol{\theta}, \mathbf{t}^n)$ of the vector of state visit indicators $\mathbf{v}^n$ and the collection of all sequences of latent sojourn times $\mathbf{X}^n = \{\mathbf{X}^n(\mathbf{v}) : \ v_i \in \{0, 1\}, \ i = 2, \ldots, m-1\}$ that are defined.

Since the vector of state visit indicators uniquely defines a model, this is the joint full conditional distribution of a given model being the "correct" one and of the suite of all latent sojourn times—those that correspond to the model under consideration and those that do not. Using the Metropolized Carlin-Chib algorithm, if the current state of the chain is $(\mathbf{v}, \mathbf{X})$, then

- Generate $(\mathbf{v}^*, \mathbf{X}^*)$ from a proposal transition kernel $q(\mathbf{v}^*, \mathbf{X}^*|\mathbf{v}, \mathbf{X}, \mathbf{p}, \boldsymbol{\theta}, \mathbf{t})$.
- Accept candidate $(\mathbf{v}^*, \mathbf{X}^*)$ with probability

$$\alpha \quad = \quad \min\left\{1, \frac{p(\mathbf{v}^*, \mathbf{X}^*|\mathbf{p}, \boldsymbol{\theta}, \mathbf{t}) \cdot q(\mathbf{v}, \mathbf{X}|\mathbf{v}^*, \mathbf{X}^*, \mathbf{p}, \boldsymbol{\theta}, \mathbf{t})}{p(\mathbf{v}, \mathbf{X}|\mathbf{p}, \boldsymbol{\theta}, \mathbf{t}) \cdot q(\mathbf{v}^*, \mathbf{X}^*|\mathbf{v}, \mathbf{X}, \mathbf{p}, \boldsymbol{\theta}, \mathbf{t})}\right\},$$

where we have suppressed dependence on n for readability.

Similarly to the procedure for the illness-death model, we let $p(\mathbf{X}(\mathbf{v})|\mathbf{v})$ denote the "prior" distribution of the latent sojourn times corresponding to the model defined by $\mathbf{v}$. For example, with $m = 4$ there are four possibilities for $\mathbf{v}$, with the corresponding sets of latent sojourn times:

$$\begin{aligned}
\mathbf{X}(\mathbf{v}) = (X_{12}(\mathbf{v}), X_{23}(\mathbf{v}), X_{34}(\mathbf{v})), \qquad & \mathbf{v} = (1,1); \\
\mathbf{X}(\mathbf{v}) = (X_{12}(\mathbf{v}), X_{24}(\mathbf{v})), \qquad & \mathbf{v} = (1,0); \\
\mathbf{X}(\mathbf{v}) = (X_{13}(\mathbf{v}), X_{34}(\mathbf{v})), \qquad & \mathbf{v} = (0,1); \\
\mathbf{X}(\mathbf{v}) = X_{14}(\mathbf{v}), \qquad & \mathbf{v} = (0,0).
\end{aligned}$$

We let $p(\mathbf{X}(-\mathbf{v})|\mathbf{X}(\mathbf{v}), \mathbf{v})$ represent the distribution of the latent sojourn times that correspond to all models other than $\mathbf{v}$, conditional on model $\mathbf{v}$ and $\mathbf{X}(\mathbf{v})$. This is the pseudo-prior for the latent sojourn times not included in model $\mathbf{v}$.

The joint full conditional distribution of model $\mathbf{v}$ and the collection of all path-specific latent sojourn times $\mathbf{X}$ that are defined is given by

$$\begin{aligned}
p(\mathbf{v}, \mathbf{X}|\mathbf{p}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) &= p(\mathbf{v}|\mathbf{p}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \cdot p(\mathbf{X}|\mathbf{v}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \\
&= \prod_{k_2=0}^{1} \cdots \prod_{k_{m-1}=0}^{1} (p_{vk_2\cdots k_{m-1}} \cdot \xi(\boldsymbol{\theta}; (k_2, \cdots, k_{m-1}), \boldsymbol{\delta}))^{\delta(2)^{1-k_2} \cdots \delta(m-1)^{1-k_{m-1}}} \\
&\quad \cdot \mathbf{1}(\mathbf{X} \in A),
\end{aligned}$$

where $\xi(\boldsymbol{\theta}; \mathbf{k}, \boldsymbol{\delta})$ with $\mathbf{k} = (k_2, \cdots, k_{m-1})$ denotes the contribution from the path in which states l with $k_l = 1$ were visited, and $p_{vk_2\ldots k_{m-1}}$ is the full conditional probability that a subject's trajectory involved visiting exactly those states. Specifically, if $\mathcal{S} = \{l : \ k_l = 1\} \cup \{1, m\}$ and $n_{\mathcal{S}}$ is the number of elements in $\mathcal{S}$, then

$$\xi(\boldsymbol{\theta}; \mathbf{k}, \boldsymbol{\delta}) = \prod_{\tau=1}^{n_{\mathcal{S}}-1} f_{\mathcal{S}_{(\tau)}, \mathcal{S}_{(\tau+1)}}(X_{\mathcal{S}_{(\tau)}, \mathcal{S}_{(\tau+1)}}(\mathbf{v} = \mathbf{k})).$$

In this computation we have suppressed dependence on n for readability, but since it is still

not entirely transparent, we illustrate the form of the joint full conditional for $m = 4$:

$$\begin{aligned}
p(\mathbf{v}, \mathbf{X}|\mathbf{p}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) &= \prod_{k_2=0}^{1}\prod_{k_3=0}^{1} [p_{vk_2k_3} \cdot \xi(\boldsymbol{\theta}; (k_2, k_3), \boldsymbol{\delta})]^{\delta(2)^{1-k_2}\cdot\delta(3)^{1-k_3}} \\
&= [p_{v11} \cdot \xi(\boldsymbol{\theta}; (1,1), \boldsymbol{\delta})]^{\delta(2)^0\cdot\delta(3)^0} \cdot [p_{v10} \cdot \xi(\boldsymbol{\theta}; (1,0), \boldsymbol{\delta})]^{\delta(2)^0\cdot\delta(3)^1} \\
&\quad \cdot [p_{v01} \cdot \xi(\boldsymbol{\theta}; (0,1), \boldsymbol{\delta})]^{\delta(2)^1\cdot\delta(3)^1} \cdot [p_{v00} \cdot \xi(\boldsymbol{\theta}; (0,0), \boldsymbol{\delta})]^{\delta(2)^1\cdot\delta(3)^1} \\
&= [p_{v11} \cdot f_{12}(X_{12}(1,1)) \cdot f_{23}(X_{23}(1,1)) \cdot f_{34}(X_{34}(1,1))] \\
&\quad \cdot [p_{v10} \cdot f_{12}(X_{12}(1,0)) \cdot f_{24}(X_{24}(1,0))]^{\delta(3)} \\
&\quad \cdot [p_{v01} \cdot f_{13}(X_{13}(0,1)) \cdot f_{34}(X_{34}(0,1))]^{\delta(2)} \\
&\quad \cdot [p_{v00} \cdot f_{14}(X_{14}(0,0))]^{\delta(2)\cdot\delta(3)}.
\end{aligned}$$

We see that the trajectory in which both states 2 and 3 are visited are included in this distribution for every subject, but that each of the other three trajectories are included only when it is not in conflict with the observed data. Also, we can see more clearly in this example that the full conditional for $\mathbf{v}$, viewed simply as a random variable taking on four values, has a categorical distribution, where the number of categories depends on the observed data.

We now turn to the full conditional probability that a subject took a particular path through the states. Applying Bayes' rule:

$$p(\mathbf{v}|\mathbf{p}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \quad \propto \quad p(\mathbf{t}|\mathbf{v}, \boldsymbol{\theta}, \boldsymbol{\delta}) \cdot p(\mathbf{v}|\mathbf{p}, \boldsymbol{\delta}).$$

For each subject we compute this quantity for each of the 2^{m-2} values of $\mathbf{v}$. If the subject is observed in state i, then $\boldsymbol{\delta}(i) = 0$, so $v_i = 1$. Hence, any $\mathbf{v}$ with $v_i = 0$ will have zero probability. This observation simplifies the calculation somewhat. For the remaining $\mathbf{v}$, the probabilities $p(\mathbf{t}|\mathbf{v}, \boldsymbol{\theta}, \boldsymbol{\delta})$ can be computed using the necessary convolutions and $p(\mathbf{v}|\mathbf{p}, \boldsymbol{\delta}) = \prod_{i=1}^{m-1}\prod_{j=i+1}^{m} p_{ij}^{N_{ij}}$, where $N_{ij} \doteq v_i \cdot \prod_{k=i+1}^{j-1}(1 - v_k) \cdot v_j$ is the indicator that the subject transitioned directly from state i to j, given $\mathbf{v}$. That is, N_{ij} indicates whether the corresponding p_{ij} is included in the product. For example, with $m = 4$ and $\mathbf{v} = (1, 0)$, we have $p(\mathbf{v}|\mathbf{p}, \boldsymbol{\delta}) = p_{12} \cdot p_{24}$ since this subject visited state 2 but not state 3. Also, we have

$N_{12} = N_{24} = 1$ while $N_{13} = N_{14} = N_{23} = N_{34} = 0$.

We see that $(\mathbf{v}|\mathbf{p}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta})$ has a categorical distribution, and the corresponding probabilities remain to be computed. The explicit expression of $p(\mathbf{v}|\mathbf{p}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta})$ for a general m, though not conceptually difficult, is quite cumbersome. We shall illustrate the form of this probability for $m = 4$.

We examine the factors $p(\mathbf{t}|\mathbf{v}, \boldsymbol{\theta}, \boldsymbol{\delta})$ and $p(\mathbf{v}|\mathbf{p}, \boldsymbol{\delta})$ in the table below:

$\boldsymbol{\delta}$	$\mathbf{v}$	$p(\mathbf{t}\|\mathbf{v}, \boldsymbol{\theta}, \boldsymbol{\delta}) \propto$:	$p(\mathbf{v}\|\mathbf{p}, \boldsymbol{\delta}) \propto$:
(0,0,0)	(1,1)	$P(X_{12} \in [t_1, t_2],\ X_{12} + X_{23} \in [t_3, t_4],\ X_{12} + X_{23} + X_{34} \in [t_5, t_6])$	$p_{12}p_{23}p_{34}$
	(1,0)	0	0
	(0,1)	0	0
	(0,0)	0	0
(1,0,0)	(1,1)	$P(X_{12} \in [t_1, t_4],\ X_{12} + X_{23} \in [t_1, t_4],\ X_{12} + X_{23} + X_{34} \in [t_5, t_6])$	$p_{12}p_{23}p_{34}$
	(1,0)	0	0
	(0,1)	$P(X_{13} \in [t_1, t_4],\ X_{13} + X_{34} \in [t_5, t_6])$	$p_{13}p_{34}$
	(0,0)	0	0
(0,1,0)	(1,1)	$P(X_{12} \in [t_1, t_2],\ X_{12} + X_{23} \in [t_3, t_6],\ X_{12} + X_{23} + X_{34} \in [t_3, t_6])$	$p_{12}p_{23}p_{34}$
	(1,0)	$P(X_{12} \in [t_1, t_2],\ X_{12} + X_{24} \in [t_3, t_6])$	$p_{12}p_{24}$
	(0,1)	0	0
	(0,0)	0	0
(1,1,0)	(1,1)	$P(X_{12} \in [t_1, t_2],\ X_{12} + X_{23} \in [t_3, t_4],\ X_{12} + X_{23} + X_{34} \in [t_5, t_6])$	$p_{12}p_{23}p_{34}$
	(1,0)	$P(X_{12} \in [t_1, t_6],\ X_{12} + X_{24} \in [t_1, t_6])$	$p_{12}p_{24}$
	(0,1)	$P(X_{13} \in [t_1, t_4],\ X_{13} + X_{34} \in [t_5, t_6])$	$p_{13}p_{34}$
	(0,0)	$P(X_{14} \in [t_1, t_6])$	p_{14}
(0,0,1)	(1,1)	$P(X_{12} \in [t_1, t_2],\ X_{12} + X_{23} \in [t_3, t_4],\ X_{12} + X_{23} + X_{34} \in [t_5, \infty))$	$p_{12}p_{23}p_{34}$
	(1,0)	0	0
	(0,1)	0	0
	(0,0)	0	0
(1,0,1)	(1,1)	$P(X_{12} \in [t_1, t_4],\ X_{12} + X_{23} \in [t_1, t_4],\ X_{12} + X_{23} + X_{34} \in [t_5, \infty))$	$p_{12}p_{23}p_{34}$
	(1,0)	0	0
	(0,1)	$P(X_{13} \in [t_1, t_4],\ X_{13} + X_{34} \in [t_5, \infty))$	$p_{13}p_{34}$
	(0,0)	0	0
(0,1,1)	(1,1)	$P(X_{12} \in [t_1, t_2],\ X_{12} + X_{23} \in [t_3, \infty))$	$p_{12}p_{23}p_{34}$
	(1,0)	$P(X_{12} \in [t_1, t_2],\ X_{12} + X_{24} \in [t_3, \infty))$	$p_{12}p_{24}$
	(0,1)	0	0
	(0,0)	0	0
(1,1,1)	(1,1)	$P(X_{12} \in [t_1, \infty))$	$p_{12}p_{23}p_{34}$
	(1,0)	$P(X_{12} \in [t_1, \infty))$	$p_{12}p_{24}$
	(0,1)	$P(X_{13} \in [t_1, \infty))$	$p_{13}p_{34}$
	(0,0)	$P(X_{14} \in [t_1, \infty))$	p_{14}

We evaluate each probability in the third column of the above table via stochastic integration. Similarly to the illness-death case, we update these probabilities at every tenth iteration of the algorithm. For a given subject with observation type $\boldsymbol{\delta}$, the full conditional probability of each trajectory is given by the ratio of the product of the corresponding factors in the table above to the sum of the products of all factors. For example, for $\boldsymbol{\delta} = (0, 1, 0)$, in which a subject is observed in states 1, 2, and 4, the full conditional probability that the subject visited all four states is given by

$$p_{v11} = \frac{p_{1-2-3-4}}{p_{1-2-3-4} + p_{1-2-4} + p_{1-3-4} + p_{1-4}},$$

where the probabilities in this fraction are given by

$$\begin{aligned}
p_{1-2-3-4} &= p_{12}p_{23}p_{34} \cdot P(X_{12} \in [t_1, t_2],\ X_{12} + X_{23} \in [t_3, t_6],\ X_{12} + X_{23} + X_{34} \in [t_3, t_6]) \\
p_{1-2-4} &= p_{12}p_{24} \cdot P(X_{12} \in [t_1, t_2],\ X_{12} + X_{24} \in [t_3, t_6]) \\
p_{1-3-4} &= 0 \\
p_{1-4} &= 0.
\end{aligned}$$

Given this full conditional distribution for each subject's vector of visit indicators, the joint full conditional distribution $p(\mathbf{v}, \mathbf{X}|\mathbf{p}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta})$ of the model $\mathbf{v}$ and latent sojourn times $\mathbf{X}$ is now available. To complete specification of the Metropolized Carlin-Chib algorithm, we must choose a proposal transition kernel. We consider the "best guess" proposal, the natural extension of the proposal introduced for the illness-death model in the previous section:

$$\begin{aligned}
q_2(\mathbf{v}^*, \mathbf{X}^*|\mathbf{v}, \mathbf{X}, \mathbf{p}, \boldsymbol{\theta}, \mathbf{t}) &= q_2(\mathbf{v}^*, \mathbf{X}^*|\mathbf{p}, \boldsymbol{\theta}, \mathbf{t}) \\
&\propto \prod_{k_2=0}^{1} \cdots \prod_{k_{m-1}=0}^{1} (p_{vk_2\cdots k_{m-1}})^{\prod_{l=2}^{m-1} v_l^{k_l} \cdot (1-v_l)^{1-k_l}} \cdot q_X(\mathbf{X}^*|\mathbf{t}) \cdot \mathbf{1}(\mathbf{X}^* \in A),
\end{aligned}$$

where $q_X(\mathbf{X}^*|\mathbf{t})$ is the proposal for the full suite of latent sojourn times that are defined. As a simple approach, we consider the extension of $q_X(\cdot)$ that was introduced for the illness-death model. Given these choices, the probability of accepting candidate $(\mathbf{v}^*, \mathbf{X}^*)$, if $\mathbf{X}^* \in A$, is

given by

$$\begin{aligned}\alpha &= \min\left\{1, \frac{\pi^*(\mathbf{v}^*, \mathbf{X}^*|\mathbf{p}, \boldsymbol{\theta}, \mathbf{t}^n) \cdot q_2(\mathbf{v}, \mathbf{X}|\mathbf{v}^*, \mathbf{X}^*)}{\pi^*(\mathbf{v}, \mathbf{X}|\mathbf{p}, \boldsymbol{\theta}, \mathbf{t}^n) \cdot q_2(\mathbf{v}^*, \mathbf{X}^*|\mathbf{v}, \mathbf{X})}\right\} \\ &= \min\left\{1, \frac{\prod_{i=1}^{m-1}\prod_{j=i+1}^{m} f_{ij}(X_{ij}^*)^{(1-\delta(j))\cdot\prod_{k=i+1}^{j-1}\delta(k)} \cdot q_X(\mathbf{X}|\mathbf{t})}{\prod_{i=1}^{m-1}\prod_{j=i+1}^{m} f_{ij}(X_{ij})^{(1-\delta(j))\cdot\prod_{k=i+1}^{j-1}\delta(k)} \cdot q_X(\mathbf{X}^*|\mathbf{t})}.\right\}\end{aligned}$$

We may now employ the algorithm to update the model indicator $\mathbf{v}$ and full suite of latent sojourn times $\mathbf{X}$ that are defined for each subject.

Hence, to summarize the algorithm, we update the transition probabilities $\mathbf{p}$ via Gibbs sampling, and the sojourn time parameters $\boldsymbol{\theta}$ via Metropolis-Hastings with appropriate normal, truncated normal, or exponential proposal distributions on each parameter. Using the Metropolized Carlin-Chib algorithm, we update the latent trajectory for each subject, which consists of the vector of state visit indicators and the corresponding collection of sequences of latent sojourn times. To generate candidate trajectories for each subject, we use the "best guess" proposal $q_2(\cdot)$ for the vector of visit indicators and a simple uniform or exponential proposal for the sojourn times, as appropriate.

4.4 Simulation studies: illness-death model

In this section we examine the performance of the proposed approach to the illness-death model via simulation study. Specifically, in Section 4.4.1 we consider the case in which subjects are observed frequently to demonstrate that the approach yields consistent inference for the embedded Markov chain as well as the sojourn times under ideal circumstances. The frequent observation scheme allows us to explore the method's performance for a variety of sojourn time models. Additionally we demonstrate the performance of the method for a range of true probabilities, and in cases where the sojourn time distributions differ. In Section 4.4.2 we consider the case in which subjects are observed infrequently, so that a subject who was not seen in the intermediate state could well have visited that state between observations. Under this sparse observation scheme, we examine how the method performs for a range of true probabilities of visiting the intermediate state.

4.4.1 Performance of proposed approach when subjects are observed frequently.

In this first set of simulations we consider a frequent observation scheme, and demonstrate the performance of the approach under a variety of circumstances. We focus on examining estimation only under correct model specification.

Specifically, we generate trajectories from the scenarios in Table 4.1 and use the corresponding model for the sojourn times for each scenario. In the first group of scenarios, we consider a variety of models for the sojourn times, but assume that the sojourn times have a common distribution and that the probability of visiting state 2 is 50% in each case. In Scenario 1* we consider an exponential data-generating distribution, whereas in Scenarios 2*–4* we consider a Weibull distribution and evaluate the performance of the method when the the sojourn time model is Weibull, generalized Weibull, and linear spline. In the second group of scenarios, we vary the probability of visiting state 2, keeping the sojourn time distributions constant. We allow the two sojourn times in state 1—X_{12} and X_{13}—to differ in the third group of scenarios.

Table 4.1: Scenarios for frequent observation scheme.

Scenario	X_{12}, X_{23}, X_{13}	model	p_{12}
1*	$X_{ij} \sim_{iid} exp(2)$	exponential	0.50
2*	$X_{ij} \sim_{iid} We(4,2)$	Weibull	0.50
3*	$X_{ij} \sim_{iid} We(4,2)$	exp. Weibull	0.50
4*	$X_{ij} \sim_{iid} exp(2)$	linear spline	0.50
5*	$X_{ij} \sim_{iid} We(2,2)$	Weibull	0.10
6*	$X_{ij} \sim_{iid} We(2,2)$	Weibull	0.50
7*	$X_{ij} \sim_{iid} We(2,2)$	Weibull	0.90
8*	$X_{12}, X_{23} \sim_{iid} We(2,3);\ X_{13} \sim We(2,1)$	Weibull	0.50
9*	$X_{12}, X_{23} \sim_{iid} We(2,1);\ X_{13} \sim We(2,3)$	Weibull	0.50

In each of these nine scenarios, subjects are observed at times $0.00, 0.25, 0.50, \ldots$, i.e. every three months if the unit of time is years. Note, however, that as in Chapter 3, the proposed approach assumes neither that the observation scheme is evenly-spaced nor that

it is common across subjects. We continue to assume in this chapter that subjects enter state 1 at time zero. We assume each subject is followed until he enters the absorbing state. We put noninformative uniform priors on each of the sojourn time parameters and on the transition probability p_{12} in each of the scenarios except Scenario 3*, in which informative normal priors were used for the sojourn time parameters. We illustrate the performance of the method for $N = 400$ subjects. The results, based on $M = 100$ simulated datasets, are shown in Tables 4.2–4.4. Posterior means as well as model-based and empirical posterior standard deviations are reported for each of the parameters of interest.

From Table 4.2 we see that under correct model specification, the proposed approach gives consistent inference about the probability of visiting state 2 as well as the parameters of the sojourn times, for a variety of sojourn time models. The approach performs very well for the exponential and Weibull sojourn time models. For Scenario 3*, informative priors were needed to ensure convergence of the chain, but the approach performs well when such priors are used. Similarly, for Scenario 4*, informative priors were used, and the approach performs well in this case.

Table 4.3 shows that the proposed approach yields consistent inference for the parameters of interest over a wide range of probabilities that subjects truly visited state 2. Specifically, the approach consistently estimates p_{12} even when its true value is close to zero or one. The approach also consistently estimates the sojourn time parameters. As we would expect, the posterior standard deviations for the parameters corresponding to X_{12} and X_{23} decrease as p_{12} grows, since increasing the expected number of subjects who visit state 2 leads to greater precision for estimating the corresponding sojourn times. Conversely, the posterior standard deviations for the parameters corresponding to X_{13} increase as p_{12} grows.

From Table 4.4 we can see that, under a frequent observation scheme, the approach yields consistent inference for all parameters of interest when X_{12} and X_{13} are distributed quite differently.

4.4.2 Performance of proposed approach when subjects are observed infrequently.

In this second set of simulations we consider an infrequent observation scheme. We demonstrate the performance of the proposed method in various scenarios, and compare our semi-Markov modeling approach to a Markov modeling approach. For an infrequent observation scheme, there can be great uncertainty about both the underlying trajectories and the sojourn times themselves. We thus first consider the case in which all three sojourn times share a common Weibull distribution and examine the performance of the proposed approach over a range of probabilities of visiting state 2. Secondly, we want to compare the performance of the proposed approach and the standard approach to a general progressive process. Specifically, we want to investigate: (1) the performance of a Markov model when the Markov assumption is not satisfied, and (2) the degree of efficiency loss when the proposed approach is used unnecessarily.

The scenarios under consideration and model for each conditional sojourn time are given in Table 4.5. Subjects are observed at times $0, 1, 2, \ldots$, i.e. annually, until they enter the absorbing state.

We examine cases in which the rate of the disease process is moderate in comparison to the observation rate, so that we have moderate uncertainty about subjects' true trajectories and sojourn times in each state. In Scenarios 1–3 we investigate how the approach performs when the true probability of visiting state 2 is 10%, 50%, and 90%. We assume uniform priors on each of the sojourn time parameters as well as on the transition probability p_{12}. Scenarios 4–6 collectively address the questions of model misspecification. Specifically, in Scenario 4 the underlying process that generates trajectories violates the Markov assumption, but a Markov model is used in the analysis. In Scenario 5 the situation is reversed: the data-generating mechanism satisfies the Markov assumption, but we use the proposed approach, which is unnecessarily flexible. For Scenarios 4 and 6, we use the proposed approach but imposed the Markov assumption. We will compare the results to those in Scenario 6, in which we correctly impose the Markov assumption. In each of these six scenarios we illustrate the performance of the approach for $N = 400$ subjects, using $M = 100$ simulated datasets.

Table 4.5: Scenarios for infrequent observation scheme.

Scenario	True Sojourn Model	p_{12}	Fitted Sojourn Model
1	$X_{ij} \sim_{iid} We(2,2)$	0.10	Weibull
2	$X_{ij} \sim_{iid} We(2,2)$	0.50	Weibull
3	$X_{ij} \sim_{iid} We(2,2)$	0.90	Weibull
4	$X_{ij} \sim_{iid} We(2,2)$	0.50	exponential (Markov)
5	$X_{ij} \sim_{iid} exp(2)$	0.50	Weibull (semi-Markov)
6	$X_{ij} \sim_{iid} exp(2)$	0.50	exponential (Markov)

Table 4.6 shows that the proposed approach performs well and consistently estimates the parameters of interest in each of Scenarios 1–3. Note that Scenarios 1–3 are parallel to Scenarios 5*–7*, as only the observation scheme is different. We observe the same trends in the posterior standard deviations as we did in Table 4.3. As we would expect, the posterior standard deviations in Table 4.6 are larger than the corresponding ones in Table 4.3. There is some bias for finite samples when the expected number of subjects taking a given trajectory is small—in the shape parameter for X_{13} in Scenario 3, for example—but the posterior standard deviation is large.

In Table 4.7 we present the results for Scenarios 4–6. Since the model is misspecified in Scenario 4, true values for the parameters of interest do not exist. Hence, in addition to presenting the results for each of the parameters of interest, we also present results for a derived quantity so that the performance of the two approaches under consideration may be compared. Specifically, we examine the probability of being in the intermediate state at one and two years, $p_{12}(t)$ with $t = 1$ and $t = 2$, and consider results for Scenarios 2 and 4–6. To address question (1) above, we compare results from Scenarios 2 and 4. In each of these scenarios, the true sojourn times arise from a Weibull distribution, but we account for this only in Scenario 2. We see that using a Markov model leads to a biased estimate of the probability of being in state 2 at a fixed time (Scenario 4), while using our proposed approach without the Markov assumption leads to an unbiased estimate (Scenario 2). To address question (2), we compare results from Scenarios 5 and 6. In each scenario, the true

trajectories satisfy the Markov assumption. Scenario 5 does not make use of the Markov assumption, and by comparing results of the probability of being in state 2 at each fixed time from Scenarios 5 and 6 we see that there is a modest loss of precision in the estimate. That is, the cost of allowing for additional flexibility in the model is quite small.

4.5 Simulation studies: general progressive model with four states

In this section we examine the performance of the proposed approach for a general progressive process with four states. With four states, there are six conditional sojourn times included in the model. Additionally, there are five transition probabilities, three of which are free parameters. We consider an infrequent observation scheme only and demonstrate the performance of the approach for a single scenario in which both the data-generating distribution and the sojourn time model are Weibull.

In the scenario under consideration, $X_{ij} \sim Weibull(2,2)$ for each pair (i,j), $i = 1,2,3$, $j = i+1,\ldots,4$. The embedded Markov chain is given by

$$\mathbf{P} = \begin{bmatrix} 0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Subjects are observed at times $0,1,2,\ldots$ until they enter the absorbing state.

We illustrate the performance of the proposed approach for $N = 400$ subjects. We assume uniform priors on each of the sojourn time parameters as well as on the parameters of the transition probability matrix. The results ($M = 100$) are given in Table 4.8.

The algorithm yields consistent inference for each of the sojourn time parameters in this scenario. There was a slight upward bias for p_{12} and corresponding downward bias for p_{14}. However, this issue did not have a discernible impact on inference for the sojourn time parameters.

4.6 Discussion

In this chapter we have presented a framework for modeling progressive multi-state processes that allows for skipping of intermediate states. Our proposed methodology can accommodate progressive state models with any number of states. A general progressive model having three states is an important special case known as the illness-death model. Our proposed approach allows us to carry out inference about the probability of visiting the intermediate state and about the sojourn times in each state using any chosen parametric model. We showed through simulation studies that the proposed approach for a general progressive process performs well in a variety of scenarios, and demonstrated its value relative to an approach that imposes the Markov assumption.

In each of the simulation studies we imposed a noninformative prior distribution on the parameters governing the path that subjects take through the disease states. That is, in the case of the illness-death model, our prior belief was that all values of the probability of visiting the intermediate state were equally likely. We note, however, that our approach allows us to incorporate informative prior information about the embedded Markov chain. The chosen distribution could be based on results of previous studies or expert opinion. For example, in the illness-death model, previous studies may indicate that the intermediate disease stage is visited by the vast majority of patients. We could use this information by assuming a prior distribution for the probability of visiting the intermediate state that favored higher values. Using existing information about parameters of the embedded Markov chain has the potential to enhance our ability to make inference about all parameters in the model.

As an extreme instance of prior information about the embedded Markov chain, we may believe that each patient visits each of the disease stages with very high probability. In this case it may be wise to use a simple progressive state model for the situation. Using this choice of a model, we can avoid estimating the parameters of the embedded Markov chain and can potentially improve the quality of inference about the sojourn time parameters.

We have assumed up to this point in the development that the chronological time at which a patient enters the process is known exactly. However, this is rarely the case in

applications. In the following chapter we address the possiblity that each patient entered the first stage of disease sometime before being observed.

Table 4.2: Scenarios 1*–4*: Sojourn time model varies.

Scenario	(i,j)	$\hat{\theta}_{ij}$	SE_{model}	SE_{emp}
truth		2.000	–	–
1*	$(1,2)$	2.056	0.149	0.128
	$(2,3)$	2.020	0.147	0.139
	$(1,3)$	2.047	0.149	0.154

Scenario	(i,j)	$\hat{k}_{ij}$	SD_{model}	SD_{emp}	$\hat{\theta}_{ij}$	SD_{model}	SD_{emp}
truth		4.000	–	–	2.000	–	–
2*	$(1,2)$	3.989	0.216	0.252	2.000	0.037	0.038
	$(2,3)$	4.017	0.230	0.298	2.005	0.038	0.034
	$(1,3)$	4.015	0.221	0.270	2.000	0.038	0.034

Scenario	(i,j)	$\hat{k}_{ij}$	SD_{model}	SD_{emp}	$\hat{\theta}_{ij}$	SD_{model}	SD_{emp}	$\hat{a}_{ij}$	SD_{model}	SD_{emp}
truth		4.000	–	–	2.000	–	–	1.000	–	–
3*	$(1,2)$	3.996	0.227	0.212	2.003	0.050	0.043	1.007	0.082	0.042
	$(2,3)$	3.962	0.235	0.208	1.996	0.052	0.041	1.006	0.084	0.040
	$(1,3)$	4.004	0.233	0.171	2.001	0.050	0.036	1.007	0.083	0.037

Scenario	(i,j)	$\hat{v}_{1,ij}$	SD_{model}	SD_{emp}	$\hat{b}_{1,ij}$	SD_{model}	SD_{emp}	$\hat{b}_{2,ij}$	SD_{model}	SD_{emp}
truth		-0.693	–	–	0.000	–	–	0.000	–	–
4*	$(1,2)$	-0.686	0.115	0.109	0.013	0.092	0.097	0.010	0.146	0.135
	$(2,3)$	-0.735	0.155	0.164	0.045	0.291	0.301	0.017	0.070	0.078
	$(1,3)$	-0.676	0.114	0.111	0.026	0.092	0.096	-0.031	0.146	0.151

Scenario	true p_{12}	$\hat{p}_{12}$	SD_{model}	SD_{emp}
1*	0.500	0.503	0.029	0.021
2*	0.500	0.503	0.025	0.032
3*	0.500	0.503	0.025	0.032
4*	0.500	0.511	0.031	0.036

Table 4.3: Scenarios 5*–7*: embedded Markov chain varies.

Scenario	(i,j)	true k_{ij}	$\hat{k}_{ij}$	SD_{model}	SD_{emp}	true θ_{ij}	$\hat{\theta}_{ij}$	SD_{model}	SD_{emp}
5*	$(1,2)$	2.000	2.087	0.261	0.288	2.000	2.022	0.175	0.171
	$(2,3)$	2.000	2.047	0.274	0.287	2.000	2.016	0.182	0.155
	$(1,3)$	2.000	1.999	0.083	0.082	2.000	1.998	0.056	0.041
6*	$(1,2)$	2.000	2.006	0.111	0.120	2.000	2.001	0.074	0.071
	$(2,3)$	2.000	2.017	0.117	0.132	2.000	2.012	0.076	0.067
	$(1,3)$	2.000	2.014	0.113	0.124	2.000	2.002	0.075	0.071
7*	$(1,2)$	2.000	1.994	0.083	0.088	2.000	2.009	0.056	0.048
	$(2,3)$	2.000	2.014	0.087	0.100	2.000	2.022	0.057	0.063
	$(1,3)$	2.000	2.078	0.260	0.304	2.000	2.020	0.176	0.164

Scenario	true p_{12}	$\hat{p}_{12}$	SD_{model}	SD_{emp}
5*	0.100	0.101	0.015	0.014
6*	0.500	0.503	0.025	0.032
7*	0.900	0.898	0.015	0.015

Table 4.4: Scenarios 8*–9*: embedded Markov chain varies.

Scenario	(i,j)	true k_{ij}	$\hat{k}_{ij}$	SD_{model}	SD_{emp}	true θ_{ij}	$\hat{\theta}_{ij}$	SD_{model}	SD_{emp}
8^*	$(1,2)$	2.000	1.997	0.110	0.116	3.000	2.995	0.113	0.106
	$(2,3)$	2.000	2.029	0.115	0.132	3.000	3.024	0.115	0.097
	$(1,3)$	2.000	1.983	0.115	0.142	1.000	1.004	0.040	0.036
9^*	$(1,2)$	2.000	1.999	0.113	0.124	1.000	1.000	0.038	0.037
	$(2,3)$	2.000	2.012	0.129	0.152	1.000	1.003	0.040	0.038
	$(1,3)$	2.000	2.025	0.117	0.126	3.000	3.012	0.116	0.108

Scenario	true p_{12}	$\hat{p}_{12}$	SD_{model}	SD_{emp}
8^*	0.500	0.502	0.025	0.032
9^*	0.500	0.504	0.026	0.031

Table 4.6: Scenarios 1–3: embedded Markov chain varies.

Scenario	(i,j)	true k_{ij}	$\hat{k}_{ij}$	SD_{model}	SD_{emp}	true θ_{ij}	$\hat{\theta}_{ij}$	SD_{model}	SD_{emp}
1	$(1,2)$	2.000	2.057	0.313	0.338	2.000	2.006	0.190	0.185
	$(2,3)$	2.000	2.103	0.454	0.477	2.000	1.973	0.240	0.244
	$(1,3)$	2.000	2.007	0.097	0.103	2.000	1.998	0.059	0.042
2	$(1,2)$	2.000	2.001	0.133	0.144	2.000	1.996	0.082	0.086
	$(2,3)$	2.000	2.042	0.185	0.204	2.000	2.011	0.099	0.099
	$(1,3)$	2.000	2.019	0.136	0.153	2.000	2.007	0.083	0.073
3	$(1,2)$	2.000	1.996	0.097	0.103	2.000	2.011	0.062	0.053
	$(2,3)$	2.000	1.971	0.131	0.167	2.000	1.995	0.074	0.097
	$(1,3)$	2.000	2.387	0.487	0.968	2.000	2.008	0.316	0.444

Scenario	true p_{12}	$\hat{p}_{12}$	SD_{model}	SD_{emp}
1	0.100	0.103	0.018	0.016
2	0.500	0.503	0.029	0.034
3	0.900	0.908	0.025	0.030

Table 4.7: Scenarios 4–6: model misspecification.

Scenario	i	true θ_i	$\hat{\theta}_i$	SD_{model}	SD_{emp}
4	1	—	1.746	0.089	0.046
	2	—	1.383	0.101	0.069
6	1	2.000	2.027	0.103	0.096
	2	2.000	2.030	0.163	0.165

Scenario	(i,j)	true k_{ij}	$\hat{k}_{ij}$	SD_{model}	SD_{emp}	true θ_{ij}	$\hat{\theta}_{ij}$	SD_{model}	SD_{emp}
5	$(1,2)$	1.000	0.992	0.081	0.079	2.000	0.992	0.081	0.079
	$(2,3)$	1.000	1.023	0.112	0.134	2.000	1.023	0.112	0.134
	$(1,3)$	1.000	1.044	0.071	0.048	2.000	1.044	0.071	0.048

Scenario	true p_{12}	$\hat{p}_{12}$	SD_{model}	SD_{emp}
4	—	0.675	0.040	0.045
5	0.500	0.509	0.048	0.051
6	0.500	0.504	0.033	0.036

Scenario	true $p_{12}(1)$	$\hat{p}_{12}(1)$	SD_{emp}	true $p_{12}(2)$	$\hat{p}_{12}(2)$	SD_{emp}
2	0.106	0.107	0.015	0.259	0.262	0.018
4	0.106	0.202	0.013	0.259	0.212	0.013
5	0.152	0.153	0.016	0.184	0.185	0.015
6	0.152	0.152	0.011	0.184	0.185	0.013

Table 4.8: Four-state process under infrequent observation.

(i,j)	true k_{ij}	$\hat{k}_{ij}$	SD_{model}	SD_{emp}	true θ_{ij}	$\hat{\theta}_{ij}$	SD_{model}	SD_{emp}
$(1,2)$	2.000	2.026	0.159	0.183	2.000	2.015	0.100	0.105
$(1,3)$	2.000	2.049	0.174	0.199	2.000	1.978	0.105	0.122
$(1,4)$	2.000	2.004	0.191	0.193	2.000	2.006	0.116	0.104
$(2,3)$	2.000	2.094	0.310	0.365	2.000	2.004	0.174	0.225
$(2,4)$	2.000	2.010	0.339	0.451	2.000	1.894	0.204	0.332
$(3,4)$	2.000	1.983	0.181	0.172	2.000	1.986	0.101	0.093

(i,j)	true p_{ij}	$\hat{p}_{ij}$	SD_{model}	SD_{emp}
$(1,2)$	0.333	0.374	0.058	0.057
$(1,3)$	0.333	0.329	0.036	0.036
$(1,4)$	0.333	0.296	0.058	0.051
$(2,3)$	0.500	0.512	0.068	0.058
$(2,4)$	0.500	0.488	0.068	0.058

Chapter 5

LEFT CENSORING

5.1 Introduction

In our previous chapters we assumed *known* chronological time at which each subject enters the first state of the process. In the current chapter we allow for initiating observation of subjects after they have made an initial transition in the disease process. Specifically, we address left censoring. In Section 5.2 we define "left censoring" and discuss computational issues that arise when the length of time since each subject has entered the process, which we call *total lead time*, is unknown. We propose an approach to carry out inference on the subjects' trajectories through the state model as well as on the total lead times. We consider a model with m states. For clarity, we begin in Section 5.3 by presenting the proposed approach in the special case of the illness-death model with $m = 3$ before presenting it in the general case in Section 5.4. In Section 5.5 we discuss modeling choices. We examine the performance of our proposed approach via simulation study in Section 5.6. Finally in Section 5.7 we discuss modeling issues that arise when left censoring is present in applications, distinguishing between those that our proposed approach can and cannot address.

5.2 Left-censored observation

In the survival analysis literature, a subject's survival time is said to be *left-censored* when it is only known that the survival time is less than some fixed value (Hosmer et al., 2008, p. 7). In the current context we use the term instead to describe the situation in which it is only known that the time of entry into a state is before some fixed chronological time. That is, if we imagine a subject's sojourn in a given state as a line segment on the axis of chronological time, then our observation of this time is *left-censored* if the left end of this segment is known only to lie at or before a fixed time (see Figure 5.1 below).

In this chapter we consider the possibility that we are not able to observe each subject

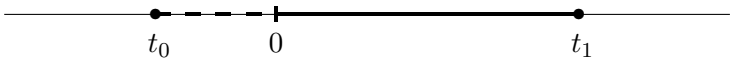


Figure 5.1: A left-censored observation of a survival time in a given state. The subject entered this state at some unknown time $t_0 \leq 0$ and left the state at known time $t_1 \geq 0$.

from the moment he enters state 1. That is, each subject's trajectory is subject to left censoring. In this dissertation we refer to the length of time from a subject's entry into state 1 to the first observation as the *total lead time*. Note that a subject's total lead time may be greater than his sojourn time in state 1, since his first observation may occur when he is in a state more advanced than state 1.

The total lead times are not observable. We will treat them as latent data just as the path indicators and sojourn times. The challenging problem is that the observed data provide very little information about the total lead times on individual subjects. As we will discuss below, we will make use of an assumption that will allow inferences about the total lead times for individual subjects and hence about the parameters of interest.

5.3 Proposed approach to address left-censored observation: Illness-death model

We propose an extension to our existing approach by which we accommodate left-censored panel observations of subjects. In this section we illustrate this extension in the case of an illness-death model, that is, the special case where $m = 3$ (shown for reference in Figure 5.2).

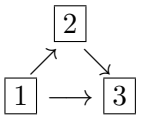


Figure 5.2: Illness-death model.

We maintain the existing notation for parameters in the model and now define some additional quantities. Up to this point, we have used "time zero" for a given subject to

refer to both the time at which the subject entered state 1 and the first observation of this subject without ambiguity, since we assumed these were identical. Since we now allow these two events to be distinct, from now on we will use "time zero" for a given subject to refer to the time of the first observation of this subject.

Each subject enters state 1 at some unknown time prior to the first observation. We let $T \geq 0$ denote the length of this interval of time, the *total lead time* for this subject. Because of the way in which we have defined "time zero", each subject enters state 1 at time $-T$. Further, if the subject was first observed in state i, then we let $L_i \geq 0$ denote the length of time the subject spent in state i before being observed. We refer to L_i as the *state lead time*. For example, for a subject who was first seen in state 1, we have $T = L_1$. By contrast, for a subject who was first seen in state 2, we have $T = X_{12} + L_2$. That is, the total lead time is the sum of the sojourn time in state 1 and the time spent in state 2 before the first observation. Let $\phi = \phi(n)$ denote the first observed state for the n^{th} subject. We define the *excess time* in the first observed state $\phi = i$ before making a transition to state j, $E_\phi = E_i$, as the difference between the sojourn time X_{ij} and the state lead time L_i. That is, $E_i \doteq X_{ij} - L_i$. For example, supposing a subject was first observed in state 1 and proceeded to state 2, we would have $E_1 = X_{12} - L_1$.

We maintain the existing definitions for the observed data with several modifications. Specifically, as before, each subject is observed to be in states $Z_0, Z_1, \ldots$ at times $s_0 = 0$, $s_1, \ldots$. We continue to let $\mathbf{t}$ denote the "sufficient" form of the observed data for one subject, whose components are defined as follows:

$$\begin{aligned} t_1 &= \max\{s_k : Z_k = 1\} \\ t_2 &= \min\{s_k : Z_k = 2\} \\ t_3 &= \max\{s_k : Z_k = 2\} \\ t_4 &= \min\{s_k : Z_k = 3\}. \end{aligned}$$

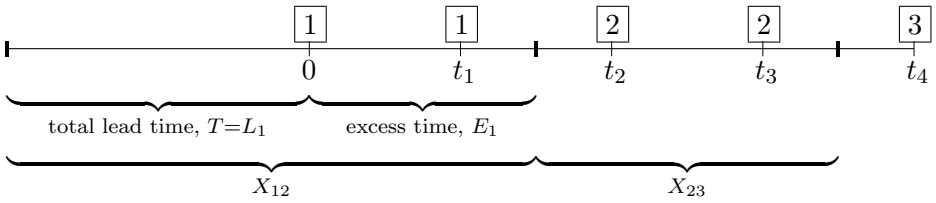
That is, t_1 is the last time the subject was observed in state 1, t_2 and t_3 are the first and last times the subject was observed in state 2, and t_4 is the first time the subject was observed

in state 3, whenever each of these exists. By our previous assumption that each subject was observed in state 1 at time zero, t_1 always existed. Under left censoring, however, t_1 may or may not exist since a subject may or may not be observed in state 1. We define $\boldsymbol{\delta} = (\delta(1), \delta(2), \delta(3))$ for each subject as the vector of indicators that the subject was *not* observed in states 1, 2, and 3, respectively. Note that the first component of $\boldsymbol{\delta}$ has been added to the notation used in previous chapters.

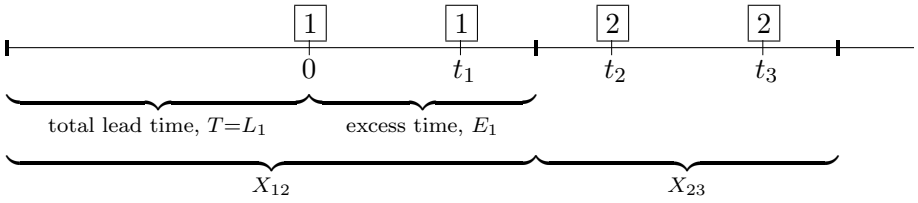
Since each subject may or may not be observed in each of states 1, 2, and 3, eight types of observations are now theoretically possible. However, subjects who are observed only in state 3 or are not observed in any state do not contribute to the likelihood function. Hence, in extending the proposed approach to accommodate left censoring we consider six types of observations, pictured in Figure 5.3. This figure shows the relationship between the latent and observed data for each observation type. For the observation types for which state 2 is not observed ($\delta(2) = 1$), two trajectories are possible. Separate diagrams are shown in each case.

Figure 5.3: In the following diagrams we illustrate how the observed data and latent data are related for each observation type and possible trajectory. The long horizontal line represents the axis of chronological time, where the point on the left extremity is the point at which the subject entered state 1. The thick tick marks represent times at which the subject made a transition, and the thin tick marks represent the times at which the subject was observed. Each boxed number above the axis represents the observed state at that time. Only the times of the first and last observation in a state are shown. If a subject is seen only once in some state, then some components of the "sufficient data" may be equal (e.g. $0 = t_1$ or $t_2 = t_3$). The latent sojourn times as well as the total lead and excess times are shown below the axis.

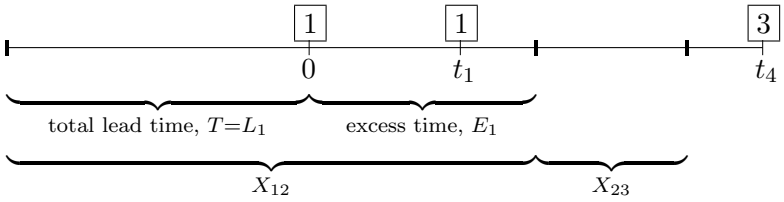
$\boldsymbol{\delta} = (0, 0, 0)$ (states 1, 2, 3 observed):



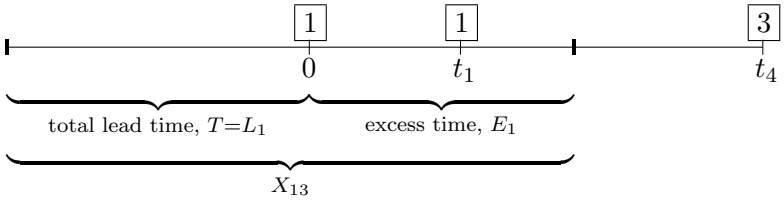
$\boldsymbol{\delta} = (0, 0, 1)$ (states 1, 2 observed):



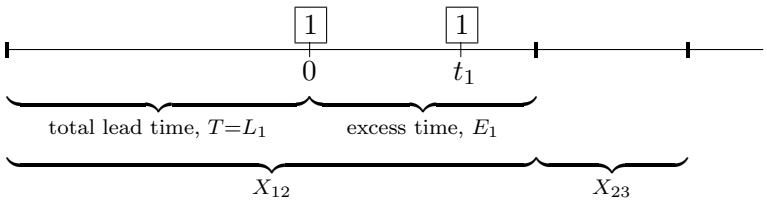
$\boldsymbol{\delta} = (0, 1, 0),\ v = 1$ (states 1, 3 observed; state 2 visited):



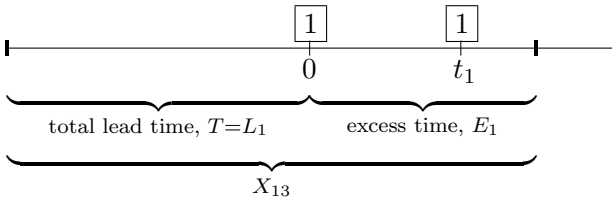
$\boldsymbol{\delta} = (0, 1, 0),\ v = 0$ (states 1, 3 observed; state 2 skipped):



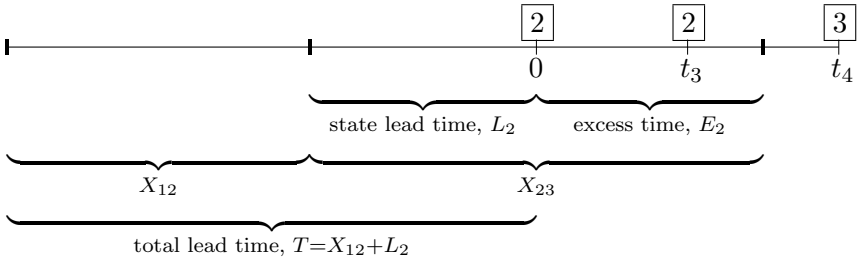
$\boldsymbol{\delta} = (0, 1, 1),\ v = 1$ (state 1 observed only; state 2 visited):



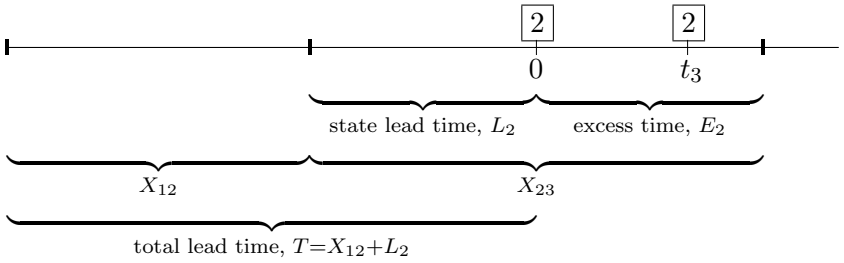
$\boldsymbol{\delta} = (0, 1, 1),\ v = 0$ (state 1 observed only; state 2 skipped):



$\boldsymbol{\delta} = (1, 0, 0)$ (states 2, 3 observed):



$\boldsymbol{\delta} = (1, 0, 1)$ (state 2 observed only):



5.3.1 Assumption about the total lead times

Our goal is to carry out inference about the parameters of interest in the model: those governing the embedded Markov chain, and those governing the sojourn times. Recall that under left censoring we can decompose the sojourn time in the first observed state i before proceeding to state j as $X_{ij} = L_i + E_i$. This expression does not lead to identifiable quantities unless there is knowledge about two of them or if the sojourn time in state i has a constant hazard function. Thus, in our context of semi-Markov models, for identifiability of the model, it is not sufficient to make a modeling assumption about the sojourn times—we must also make an assumption about either the total lead times or the excess times. We choose to make an asumption about the former. One option is the strict assumption that the total lead times arise from a degenerate distribution: a point mass at some unknown value. Another option is the more relaxed assumption that the total lead times arise from some distribution for which the mean is unknown and the standard deviation is known but small compared to the means of the sojourn times. In the next sections we describe the estimation approach under each of these options.

5.3.2 Case 1: unknown common total lead times

Here we present the algorithm for carrying out inference about the parameters of interest in the illness-death state model under the assumption, to which we refer as Case 1, that the total lead times are unknown but identical across all subjects. We focus on the ways in which the algorithm is modified from the one presented in Section 4.2 in Chapter 4. Given observed data on N subjects, $\mathbf{t}^1, \ldots, \mathbf{t}^N$, we must carry out inference about the parameters of interest, p_{12} and $\boldsymbol{\theta}$, as well as the latent data: the sojourn times $\mathbf{X}^1, \ldots, \mathbf{X}^N$, the indicators of visiting state 2, $v^1, \ldots, v^N$, and the total lead times, $T^1, \ldots, T^N$. As before, we update the parameters of interest in the first half of the algorithm and the latent trajectories for each subject in the second half.

First we note that the set of "allowable" sojourn times for each subject now depends on the subject's total lead time in addition to the observed data and indicator of visiting

state 2. Specifically, $A^n = A^n(T^n, \mathbf{t}^n; v^n, \boldsymbol{\delta}^n)$ is given by

$$A^n = \begin{cases} \{(x_{12}, x_{23}) : \ x_{12} \in [T^n + t_1^n, T + t_2^n], \ x_{12} + x_{23} \in [T + t_3^n, T + t_4^n]\}, & v^n = 1, \ \boldsymbol{\delta}^n = (0,0,0); \\ \{(x_{12}, x_{23}) : \ x_{12} \in [T^n + t_1^n, T + t_2^n], \ x_{12} + x_{23} \in [T + t_3^n, \infty)\}, & v^n = 1, \ \boldsymbol{\delta}^n = (0,0,1); \\ \{(x_{12}, x_{23}) : \ x_{12} \in [T^n + t_1^n, T^n + t_4^n], \ x_{12} + x_{23} \in [T^n + t_1^n, T^n + t_4^n]\}, & v^n = 1, \ \boldsymbol{\delta}^n = (0,1,0); \\ \{x_{13} : \ x_{13} \in [T^n + t_1^n, T^n + t_4^n]\}, & v^n = 0, \ \boldsymbol{\delta}^n = (0,1,0); \\ \{(x_{12}, x_{23}) : \ x_{12} \in [T^n + t_1^n, \infty), \ x_{12} + x_{23} \in [T^n + t_1^n, \infty)\}, & v^n = 1, \ \boldsymbol{\delta}^n = (0,1,1); \\ \{x_{13} : \ x_{13} \in [T^n + t_1^n, \infty)\}, & v^n = 0, \ \boldsymbol{\delta}^n = (0,1,1); \\ \{(x_{12}, x_{23}) : \ x_{12} \in [0, T^n], \ x_{12} + x_{23} \in [T^n + t_3^n, T^n + t_4^n], & v^n = 1, \ \boldsymbol{\delta}^n = (1,0,0); \\ \{(x_{12}, x_{23}) : \ x_{12} \in [0, T^n], \ x_{12} + x_{23} \in [T^n + t_3^n, \infty), & v^n = 1, \ \boldsymbol{\delta}^n = (1,0,1); \end{cases}$$

for $x_{12}, x_{23}, x_{13} \geq 0$. For the cases not included above, A^n does not exist. We do not consider observations for which the subject was observed neither in state 1 nor 2.

Now we can express the joint distribution of the parameters of interest and the latent trajectories for each subject conditional on the observed data as

$$\begin{aligned} & p(p_{12}, \boldsymbol{\theta}, v^1, \ldots, v^N, \mathbf{X}^1, \ldots, \mathbf{X}^N, T_0 | \mathbf{t}^1, \ldots, \mathbf{t}^N, \boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^N) \qquad (5.1) \\ \propto & \left[\prod_{n=1}^{N} [p_{12} \cdot f_{12}(X_{12}^n) \cdot f_{23}(X_{23}^n)]^{v^n} \cdot [(1 - p_{12}) \cdot f_{13}(X_{13}^n)]^{1 - v^n} \cdot \mathbf{1}(\mathbf{X}^n \in A^n) \right] \cdot \pi(p_{12}, \boldsymbol{\theta}, T_0), \end{aligned}$$

where $\pi(p_{12}, \boldsymbol{\theta}, T_0)$ gives the prior distribution of the parameters in the model and T_0 is the assumed common total lead time for all subjects, that is, $T^n = T_0$ for $n = 1, \ldots, N$. Based on this joint distribution we derive the full conditional distribution of the parameters of interest.

The implied full conditional distributions for the parameters of interest, p_{12} and $\boldsymbol{\theta}$, are conditionally independent of the common total lead time T_0 and the observed data, and thus unchanged from those derived in Chapter 4. Specifically, the full conditional distributions of p_{12} and $\boldsymbol{\theta}$ are given by Eq. (4.1) and Eq. (4.2), respectively.

We turn our attention to the second half of the algorithm, in which we update the latent trajectory for each subject. Specifically, we must update the common total lead time T_0 as well as the trajectories for all subjects: the indicators of visiting state 2, $v^1, \ldots, v^N$, and the full suite of sojourn times, $\mathbf{X}^1, \ldots, \mathbf{X}^N$. We compute the conditional distribution of T_0

by applying Bayes' rule, using the assumption that subjects progress independently of each other, and using the relationship between the sojourn times and the total lead times and excess times:

$$
\begin{aligned}
&p(T_0|v^1,\ldots,v^N,\mathbf{X}^1,\ldots,\mathbf{X}^N,\mathbf{t}^1,\ldots,\mathbf{t}^N,\boldsymbol{\delta}^1,\ldots,\boldsymbol{\delta}^N,p_{12},\boldsymbol{\theta}) \\
&\quad\propto\ p(\mathbf{X}^1,\ldots,\mathbf{X}^N|T_0,v^1,\ldots,v^N,\mathbf{t}^1,\ldots,\mathbf{t}^N,\boldsymbol{\delta}^1,\ldots,\boldsymbol{\delta}^N,p_{12},\boldsymbol{\theta})\cdot\pi(T_0) \\
&\quad\propto\ \left[\prod_{n=1}^{N} p(\mathbf{X}^n|T_0,v^n,\mathbf{t}^n,\boldsymbol{\delta}^n,p_{12},\boldsymbol{\theta})\right]\cdot\pi(T_0) \\
&\quad\propto\ \left[\prod_{n=1}^{N}[f_{12}(X_{12}^n|T_0,v^n,\mathbf{t}^n,\boldsymbol{\delta}^n,p_{12},\boldsymbol{\theta})\cdot f_{23}(X_{23}^n|T_0,v^n,\mathbf{t}^n,\boldsymbol{\delta}^n,p_{12},\boldsymbol{\theta})]^{v^n}\right. \\
&\qquad\left.\cdot\,[f_{13}(X_{13}^n|T_0,v^n,\mathbf{t}^n,\boldsymbol{\delta}^n,p_{12},\boldsymbol{\theta})]^{1-v^n}\right]\cdot\pi(T_0) \\
&\quad\propto\ \prod_{n=1}^{N}\left[f_{12}\left(T_0+E_{\phi(n)}^n(v^n)\right)\cdot f_{23}(X_{13}^n)\right]^{(1-\delta^n(1))\cdot(1-\delta^n(2))} \\
&\qquad\cdot\ \left[f_{12}\left(T_0+E_{\phi(n)}^n(v^n)\right)\cdot f_{23}(X_{23}^n)\right]^{v^n\cdot(1-\delta^n(1))\cdot\delta^n(2)\cdot(1-\delta^n(3))} \\
&\qquad\cdot\ \left[f_{13}\left(T_0+E_{\phi(n)}^n(v^n)\right)\right]^{(1-v^n)\cdot(1-\delta^n(1))\cdot\delta^n(2)\cdot(1-\delta^n(3))} \\
&\qquad\cdot\ [f_{12}\left(X_{12}^n\right)\cdot f_{23}(X_{23}^n)]^{v^n\cdot(1-\delta^n(1))\cdot\delta^n(2)\cdot\delta^n(3)} \\
&\qquad\cdot\ [f_{13}\left(X_{13}^n\right)]^{(1-v^n)\cdot(1-\delta^n(1))\cdot\delta^n(2)\cdot\delta^n(3)} \\
&\qquad\cdot\ \left[f_{12}(X_{12}^n)\cdot f_{23}(T_0-X_{12}^n+E_{\phi(n)}^n(v^n))\right]^{\delta^n(1)\cdot(1-\delta^n(2))\cdot(1-\delta^n(3))} \\
&\qquad\cdot\ [f_{12}(X_{12}^n)\cdot f_{23}(X_{23})]^{\delta^n(1)\cdot(1-\delta^n(2))\cdot\delta^n(3)}\cdot\pi(T_0) \\
&\quad\propto\ \prod_{n=1}^{N}\left[f_{12}\left(T_0+E_{\phi(n)}^n(v^n)\right)\right]^{(1-\delta^n(1))\cdot(1-\delta^n(2))} \\
&\qquad\cdot\ \left[f_{12}\left(T_0+E_{\phi(n)}^n(v^n)\right)\right]^{v^n\cdot(1-\delta^n(1))\cdot\delta^n(2)\cdot(1-\delta^n(3))} \\
&\qquad\cdot\ \left[f_{13}\left(T_0+E_{\phi(n)}^n(v^n)\right)\right]^{(1-v^n)\cdot(1-\delta^n(1))\cdot\delta^n(2)\cdot(1-\delta^n(3))} \\
&\qquad\cdot\ \left[f_{23}(T_0-X_{12}^n+E_{\phi(n)}^n(v^n))\right]^{\delta^n(1)\cdot(1-\delta^n(2))\cdot(1-\delta^n(3))}\cdot\pi(T_0).
\end{aligned}
$$

In the fifth line we have expressed the sojourn time densities, conditional on the other quantities, according to observation type and path through the states. For some observation

types, the observed data provide useful information about the common total lead time, and in such cases, the sojourn time in the first observed state may be re-expressed in terms of the total lead time. For example, consider a subject who was observed in states 1 and 2. Given his total lead time, $T^n = T_0$, his sojourn time in state 1, X_{12}, may be expressed as $T_0 + E_1$. On the other hand, a subject who was observed in state 1 only does not provide useful information about the common total lead time. In the final line, we have removed all elements that do not contribute in the estimation of T_0. We see that subjects who are observed in one state only do not provide information about the common total lead time. We update T_0 via a Metropolis-Hastings step.

For a given subject, the joint full conditional distribution of the full suite of latent sojourn times $\mathbf{X}$ and the model indicator v is given by

$$
\begin{aligned}
& p(\mathbf{X}, v | T_0, p_{12}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \qquad (5.3) \\
& \quad \propto \quad p(\mathbf{X} | v, T_0, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \cdot p(v | T_0, p_{12}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \\
& \quad \propto \quad p(\mathbf{X}(v) | v, T_0, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \cdot p(\mathbf{X}(1-v) | \mathbf{X}(v), v, T_0, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \cdot p(v | T_0, p_{12}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \\
& \quad \propto \quad f_{12}(X_{12}) \cdot f_{23}(X_{23}) \cdot [f_{13}(X_{13})]^{\delta(2)} \cdot \mathbf{1}(\mathbf{X} \in A) \cdot p_v^v \cdot (1 - p_v)^{1-v},
\end{aligned}
$$

where $A \doteq A(T, \mathbf{t}; v, \boldsymbol{\delta})$ as before, and $p_v \doteq p_v(p_{12}, \boldsymbol{\theta}, T, \mathbf{t}; \boldsymbol{\delta})$ is the full conditional probability that this subject visited state 2. This probability is given by

$$
p_v \quad = \quad \begin{cases}
1, & \boldsymbol{\delta} = (0,0,0); \\
1, & \boldsymbol{\delta} = (0,0,1); \\
\frac{p_{12} \cdot P(X_{12} \in [T+t_1, T+t_4],\ X_{12}+X_{23} \in [T+t_1, T^n+t_4])}{p_{12} \cdot P(X_{12} \in [T+t_1, T+t_4],\ X_{12}+X_{23} \in [t_1, t_4]) + (1-p_{12}) \cdot P(X_{13} \in [T+t_1, T+t_4])}, & \boldsymbol{\delta} = (0,1,0); \\
\frac{p_{12} \cdot P(X_{12} \in [T+t_1, \infty))}{p_{12} \cdot P(X_{12} \in [T+t_1, \infty)) + (1-p_{12}) \cdot P(X_{13} \in [T+t_1, \infty))}, & \boldsymbol{\delta} = (0,1,1); \\
1, & \boldsymbol{\delta} = (1,0,0); \\
1, & \boldsymbol{\delta} = (1,0,1).
\end{cases}
$$

Under the current assumption, the total lead time for this subject is equal to the common total lead time: $T = T_0$. The first four cases are analogous to those presented in Section 4.2.

The only difference is that the subject's total lead time must be taken into account. In the last two cases, the subject in question was observed in state 2, so the result is immediate. We update the indicator of visiting state 2 and the full suite of sojourn times $\mathbf{X}$ for each subject via the Metropolized Carlin-Chib algorithm as described in Section 4.2 of Chapter 4.

To summarize the algorithm for the illness-death model in Case 1, the first half deals with the parameters of interest: we update p_{12} via Gibbs sampling, and the sojourn time parameters $\boldsymbol{\theta}$ via Metropolis-Hastings with appropriate normal or truncated normal proposal distributions on each parameter. The second half of the algorithm deals with the common total lead time and latent trajectories for each subject: we update the common total lead time via a Metropolis-Hastings step with a truncated normal proposal distribution and a truncated normal prior, and for each subject we update the indicator that the subject visited state 2 and the full suite of sojourn times using the Metropolized Carlin-Chib algorithm. We use the "best guess" proposal for the indicators and a uniform or exponential proposal for the sojourn times as appropriate.

5.3.3 Case 2: homogeneous total lead times

Here we present the algorithm for carrying out inference about the parameters of interest in the illness-death state model under the assumption, to which we refer as Case 2, that the total lead times are identically distributed across all subjects, where the distribution has unknown mean but standard deviation small in comparison to the means of the sojourn times. Specifically, we assume that the total lead times are from a truncated normal distribution. Define

$$\begin{aligned} T' &\sim \mathcal{N}(\mu_T, \sigma_T^2); \\ T &\doteq T' \mid T' \geq 0, \end{aligned}$$

for some $\mu_T \in \mathbb{R}$ and some small and known $\sigma_T > 0$. Note, $\sigma_T = 0$ yields Case 1. Also note that μ_T is the mean of the original normally distributed random variable T' rather than of T. For simplicity we choose to parameterize the model in terms of the mean of the normal random variable. For ease of terminology we refer to μ_T as the "mean total lead time",

though it is slightly different from the mean of T. Under Case 2, we have the parameter μ_T and the total lead times for all subjects, $T^1, \ldots, T^N$. The standard deviation σ_T of the normal distribution must be small in comparison to μ_T since deviations from μ_T are not identifiable in the algorithm. As a rough guide, $\sigma_T \leq \frac{\mu_T}{2}$.

The first half of the algorithm is unchanged from the presentation in the previous subsection. In the second half of the algorithm we must update the mean total lead time and the total lead times for all subjects as well as the indicators of visiting state 2 and the full suite of sojourn times.

For a given subject, the joint distribution of the suite of all latent sojourn times $\mathbf{X}$, the model indicator v, and the total lead time T, conditional on the other quantities in the model, is given by

$$
\begin{aligned}
& p(\mathbf{X}, v, T | \mu_T, \sigma_T, p_{12}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \qquad (5.4) \\
& \quad \propto \quad p(\mathbf{X}|v, T, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \cdot p(v|T, p_{12}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \cdot p(T|\mu_T, \sigma_T) \\
& \quad \propto \quad p(\mathbf{X}(v)|v, T, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \cdot p(\mathbf{X}(1-v)|\mathbf{X}(v), v, T, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \cdot p(v|T, p_{12}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \\
& \qquad \cdot \frac{\exp\left(-\frac{(T-\mu_T)^2}{2\sigma_T^2}\right)}{P(T \geq 0)} \cdot \mathbf{1}(T \geq 0) \\
& \quad \propto \quad f_{12}(X_{12}) \cdot f_{23}(X_{23}) \cdot [f_{13}(X_{13})]^{\delta(2)} \cdot \mathbf{1}(\mathbf{X} \ \in A) \cdot p_v^v \cdot (1-p_v)^{1-v} \\
& \qquad \cdot \frac{\exp\left(-\frac{(T-\mu_T)^2}{2\sigma_T^2}\right)}{1 - \Phi(-\frac{\mu_T}{\sigma_T})} \cdot \mathbf{1}(T \geq 0).
\end{aligned}
$$

Here, $\Phi(x)$ denotes the standard normal cumulative distribution function at $x \in \mathbb{R}$. The set of "allowable" sojourn times for this subject, A, is a function of each subject's total lead time and observed data, as it was before: $A \doteq A(T, \mathbf{t}; v, \boldsymbol{\delta})$.

We first consider the updating of the total lead times, $T^1, \ldots, T^N$. Based on Eq. (5.4), we derive the full conditional distribution of the total lead time for the n^{th} subject:

$$
p(T^n | \mu_T, \sigma_T, v^1, \ldots, v^N, \mathbf{X}^1, \ldots, \mathbf{X}^N, \mathbf{t}^1, \ldots, \mathbf{t}^N, \boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^N, p_{12}, \boldsymbol{\theta})
$$

$$
\begin{aligned}
&\propto \quad p(T^n|\mu_T, \sigma_T, v^n, \mathbf{X}^n(v^n), \mathbf{t}^n, \boldsymbol{\delta}^n, p_{12}, \boldsymbol{\theta}) \\
&\propto \quad p(\mathbf{X}^n(v^n)|T^n, v^n, \mathbf{t}^n, \boldsymbol{\delta}^n, \boldsymbol{\theta}) \cdot p(T^n|\mu_T, \sigma_T) \\
&\propto \quad \mathbf{1}(\mathbf{X}^n(v^n) \in A^n(T^n, \mathbf{t}^n; v^n, \boldsymbol{\delta}^n)) \cdot \frac{\exp\left(-\frac{(T^n-\mu_T)^2}{2\sigma_T^2}\right)}{1-\Phi\left(-\frac{\mu_T}{\sigma_T}\right)} \cdot \mathbf{1}(T^n \geq 0) \\
&\propto \quad \mathbf{1}(\mathbf{X}^n(v^n) \in A^n(T^n, \mathbf{t}^n; v^n, \boldsymbol{\delta}^n)) \cdot \exp\left(-\frac{(T^n-\mu_T)^2}{2\sigma_T^2}\right) \cdot \mathbf{1}(T^n \geq 0),
\end{aligned}
$$

where the second line follows from the assumption of conditional independent disease progression of subjects. In the fourth line we have noted that, based on the observed data and the latent path indicator v^n and total lead time T^n for this subject, we learn about the path-specific sojourn times $\mathbf{X}^n(v^n)$ via the extent of the "allowable" region A^n. We update each of the total lead times via a Metropolis-Hastings step.

Next we consider the updating of the mean total lead time, μ_T. This parameter depends on the total lead times as well as σ_T. Based on Eq. (5.4), we derive the full conditional distribution:

$$
\begin{aligned}
&p(\mu_T|\sigma_T, v^1, \ldots, v^N, \mathbf{X}^1, \ldots, \mathbf{X}^N, T^1, \ldots, T^N, \mathbf{t}^1, \ldots, \mathbf{t}^N, \boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^N, p_{12}, \boldsymbol{\theta}) \\
&\qquad\qquad \propto \quad p(\mu_T|\sigma_T, T^1, \ldots, T^N) \\
&\qquad\qquad \propto \quad p(T^1, \ldots, T^N|\mu_T, \sigma_T) \cdot \pi(\mu_T) \\
&\qquad\qquad \propto \quad \left[\prod_{n=1}^{N} p(T^n|\mu_T, \sigma_T)\right] \cdot \pi(\mu_T) \\
&\qquad\qquad \propto \quad \left[\prod_{n=1}^{N} \frac{\exp\left(-\frac{(T^n-\mu_T)^2}{2\sigma_T^2}\right)}{1-\Phi\left(-\frac{\mu_T}{\sigma_T}\right)}\right] \cdot \pi(\mu_T),
\end{aligned}
$$

where $\pi(\mu_T)$ is the prior distribution for the mean total lead time. We have invoked Bayes' rule in the third line. We update the mean total lead time via a Metropolis step.

Finally we consider the updating of the indicator of visiting state 2 and the full suite of sojourn times for each subject. The joint distribution of the path indicator v and the suite of sojourn times $\mathbf{X}$ for a given subject, conditional on this subject's total lead time

and other quantities in the model, is given by

$$
\begin{aligned}
& p(\mathbf{X}, v|T, p_{12}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \\
& \qquad \propto \quad p(\mathbf{X}|v, T, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \cdot p(v|T, p_{12}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \\
& \qquad \propto \quad p(\mathbf{X}(v)|v, T, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \cdot p(\mathbf{X}(1-v)|\mathbf{X}(v), v, T, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \cdot p(v|T, p_{12}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \\
& \qquad \propto \quad f_{12}(X_{12}) \cdot f_{23}(X_{23}) \cdot [f_{13}(X_{13})]^{\delta(2)} \cdot \mathbf{1}(\mathbf{X} \ \in A) \cdot p_v^v \cdot (1-p_v)^{1-v},
\end{aligned}
$$

where p_v is again the full conditional probability that this subject visited state 2. Note the similarity to Eq. (5.3). We update the path indicator and full suite of sojourn times for each subject via the Metropolized Carlin-Chib algorithm.

To summarize, the first half of the algorithm proceeds exactly as described in Section 5.3.2 under Case 1. In the second half, we update the latent total lead times for all subjects via a Metropolis-Hastings step, using a uniform or exponential proposal distribution as appropriate. Given the updated total lead times we update the mean total lead time via a Metropolis step, using a symmetrical normal proposal distribution and a normal prior. We update the path indicator and suite of sojourn times via the Metropolized Carlin-Chib algorithm.

5.4 Proposed approach to address left-censored observation: General progressive model

In this section we present the left censoring approach for a general progressive model with m states where $m \geq 3$. We present the algorithm first under the strict assumption that the total lead times are common but unknown across subjects (Case 1) and then under the more relaxed assumption that the total lead times are relatively homogeneous across subjects with unknown mean and known variance (Case 2).

We use the notation for a general progressive model introduced in Section 4.3 of Chapter 4 and the left censoring notation introduced in the previous Section 5.3. We note that 2^m types of observations are possible, of which $2^m - 2$ involve observing the subject before the absorbing state. For a fixed m the relationship between the observed and latent data can be represented for each observation type as a natural extension of the diagrams shown

in Figure 5.3.

5.4.1 Case 1: unknown common total lead times

Here we present the algorithm for carrying out inference about the parameters of interest in the general progressive state model under the Case 1 assumption. That is, given observed data on N subjects, $\mathbf{t}^1, \ldots, \mathbf{t}^N$, we must carry out inference about the parameters of interest—those governing the embedded Markov chain, $\mathbf{p} = \{p_{12}, \ldots, p_{1m}, p_{23}, \ldots, p_{2m}, \ldots, p_{m-1,m}\}$, and those governing the sojourn times, $\boldsymbol{\theta}$—as well as the latent trajectories for all subjects—the path indicators $\mathbf{v}^1, \ldots, \mathbf{v}^N$, suites of sojourn times $\mathbf{X}^1, \ldots, \mathbf{X}^N$, and the common total lead time T_0.

We examine the set of "allowable" sojourn times for each subject, $A = A(T, \mathbf{t}; \mathbf{v}, \boldsymbol{\delta})$, where $T = T_0$ under the current assumption. For each observation type and path, this region has between one and $m-1$ dimensions. This is because the dimension is given by the number of latent sojourn times for the path defined by $\mathbf{v}$, which is between one and $m-1$. The form of A also depends on the type of observation $\boldsymbol{\delta}$ as well as the path indicator $\mathbf{v}$. For $m > 3$, the allowable region is the logical extension of what it was in the illness-death case. For $m = 4$, A is a region having between one and three dimensions depending on $\boldsymbol{\delta}$ and $\mathbf{v}$. Since there are 64 cases, however, we omit the illustration.

As always, in the first half of the algorithm we update of the parameters of interest. The full conditional distribution of the transition probabilities $\mathbf{p}$ is given by

$$p(\mathbf{p}|\boldsymbol{\theta}, \mathbf{v}^1, \ldots, \mathbf{v}^N, \mathbf{X}^1(\mathbf{v}^1), \ldots, \mathbf{X}^N(\mathbf{v}^N), \mathbf{t}^1, \ldots, \mathbf{t}^N, \boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^N) \propto \prod_{n=1}^{N} \prod_{i=1}^{m-1} \prod_{j=i+1}^{m} p_{ij}^{N_{ij}} \cdot \pi(\mathbf{p}),$$

where $N_{ij} \doteq \sum_{n=1}^{N} v_i^n \cdot \prod_{k=i+1}^{j-1}(1 - v_k^n) \cdot v_j^n$ is the latent number of subjects who transition directly from state i to j. Imposing a Dirichlet prior on $\mathbf{p}$, $\pi(\mathbf{p}_i) \sim Dir(a_{i,i+1}, \ldots, a_{i,m})$ where each $a_{i,j} > 0$ for $i = 1, \ldots, m-1$ and $j = i+1, \ldots, m$, the full conditional becomes

$$p(\mathbf{p}|\boldsymbol{\theta}, \mathbf{v}^1, \ldots, \mathbf{v}^N, \mathbf{X}^1(\mathbf{v}^1), \ldots, \mathbf{X}^N(\mathbf{v}^N), \mathbf{t}^1, \ldots, \mathbf{t}^N, \boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^N)$$

$$\propto \prod_{n=1}^{N} \prod_{i=1}^{m-1} \prod_{j=i+1}^{m} p_{ij}^{N_{ij}+a_{ij}-1}.$$

We use a Gibbs step to update $\mathbf{p}$.

The full conditional distribution of the parameters governing the sojourn times for all subjects is given by

$$p(\boldsymbol{\theta}|\mathbf{p}, \mathbf{v}^1, \ldots, \mathbf{v}^N, \mathbf{X}^1(\mathbf{v}^1), \ldots, \mathbf{X}^N(\mathbf{v}^N), \mathbf{t}^1, \ldots, \mathbf{t}^N, \boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^N)$$
$$\propto \prod_{n=1}^{N} \prod_{i=1}^{m-1} \prod_{j=i+1}^{m} f_{ij}(X_{ij}^n(\mathbf{v}^n))^{N_{ij}} \cdot \pi(\boldsymbol{\theta}).$$

As in the illness-death model, the particular form of this distribution depends on the chosen model for each of the sojourn times. We use a Metropolis-Hastings step to update these parameters, choosing a normal or truncated normal proposal distribution as appropriate. Refer to Section 4.3 in Chapter 4 for derivations of these full conditional distributions.

Next we consider the updating of each subject's latent trajectory. We first update the common total lead time T_0. As in the illness-death model, we compute the full conditional distribution of T_0 using Bayes' rule, the assumption of conditionally independent disease progression of subjects, and the relationship between sojourn times and excess and total lead times:

$$\begin{aligned} & p(T_0|\mathbf{v}^1, \ldots, \mathbf{v}^N, \mathbf{X}^1, \ldots, \mathbf{X}^N, \mathbf{t}^1, \ldots, \mathbf{t}^N, \boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^N, p_{12}, \boldsymbol{\theta}) \\ & \quad \propto \quad p(\mathbf{X}^1(\mathbf{v}^1), \ldots, \mathbf{X}^N(\mathbf{v}^N)|T_0, \mathbf{v}^1, \ldots, \mathbf{v}^N, \mathbf{t}^1, \ldots, \mathbf{t}^N, \boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^N, p_{12}, \boldsymbol{\theta}) \cdot \pi(T_0) \\ & \quad \propto \quad \left[\prod_{n=1}^{N} p(\mathbf{X}^n(\mathbf{v}^n)|T_0, \mathbf{v}^n, \mathbf{t}^n, \boldsymbol{\delta}^n, p_{12}, \boldsymbol{\theta})\right] \cdot \pi(T_0) \\ & \quad \propto \quad \left[\prod_{n=1}^{N} \xi(T_0, \boldsymbol{\theta}; \mathbf{v}, \boldsymbol{\delta})\right] \cdot \pi(T_0), \end{aligned}$$

where $\xi(T, \boldsymbol{\theta}; \mathbf{v}, \boldsymbol{\delta})$ denotes the sojourn time likelihood contribution from a subject who took the path defined by $\mathbf{v}$ and has observation indicator $\boldsymbol{\delta}$, given the current parameter values $\boldsymbol{\theta}$ and total lead time T. Specifically, if $\mathcal{S} = \{l : \ v_l = 1\} \cup \{1, m\}$ and $n_{\mathcal{S}}$ is the number of

elements in $\mathcal{S}$, then we define

$$\xi(T, \boldsymbol{\theta}; \mathbf{v}, \boldsymbol{\delta}) \;=\; \prod_{\tau=1}^{n_{\mathcal{S}}-1} f_{\mathcal{S}_{(\tau)},\mathcal{S}_{(\tau+1)}}(X_{\mathcal{S}_{(\tau)},\mathcal{S}_{(\tau+1)}}(\mathbf{v});\ \mathbf{v}, T, \boldsymbol{\delta})),$$

where $\mathcal{S}_{(\tau)}$ denotes the τ^{th} smallest element of $\mathcal{S}$, and $f_{ij}(X_{ij};\ \mathbf{v}, T, \boldsymbol{\delta})$ denotes the density of the path-specific sojourn time X_{ij} expressed in terms of T. For example, considering a subject who was observed in all states of a four-state progressive process, we have $\mathcal{S} = \{1, 2, 3, 4\}$ and

$$\begin{aligned}
\xi(T, \boldsymbol{\theta}; \mathbf{v}, \boldsymbol{\delta}) &= \prod_{\tau=1}^{3} f_{\mathcal{S}_{(\tau)},\mathcal{S}_{(\tau+1)}}(X_{\mathcal{S}_{(\tau)},\mathcal{S}_{(\tau+1)}}(\mathbf{v});\ \mathbf{v}, T, \boldsymbol{\delta})) \\
&= f_{12}(X_{12}(\mathbf{v});\ \mathbf{v}, T, \boldsymbol{\delta})) \cdot f_{23}(X_{23}(\mathbf{v});\ \mathbf{v}, T, \boldsymbol{\delta})) \cdot f_{34}(X_{34}(\mathbf{v});\ \mathbf{v}, T_0, \boldsymbol{\delta})) \\
&= f_{12}(T + E_1(\mathbf{v})) \cdot f_{23}(X_{23}(\mathbf{v})) \cdot f_{34}(X_{34}(\mathbf{v})),
\end{aligned}$$

where in the last line we have expressed the sojourn time X_{12} in terms of the total lead time T.

For a given subject, the joint distribution of the path $\mathbf{v}$ and the collection of all path-specific latent sojourn times $\mathbf{X}$ that are defined, conditional on the common total lead time T_0 and all other quantities in the model, is given by

$$\begin{aligned}
&p(\mathbf{v}, \mathbf{X} | T_0, \mathbf{p}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \qquad (5.7) \\
&= p(\mathbf{v} | T_0, \mathbf{p}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \cdot p(\mathbf{X} | T_0, \mathbf{v}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \\
&= \prod_{k_2=0}^{1} \cdots \prod_{k_{m-1}=0}^{1} (p_{v_{k_2 \cdots k_{m-1}}} \cdot \xi(T_0, \boldsymbol{\theta}; \mathbf{k}, \boldsymbol{\delta}))^{\delta(2)^{1-k_2} \cdots \delta(m-1)^{1-k_{m-1}}} \cdot \mathbf{1}(\mathbf{X} \in A),
\end{aligned}$$

where, as before, $\xi(T_0, \boldsymbol{\theta}; \mathbf{k}, \boldsymbol{\delta})$ with $\mathbf{k} = (k_2, \cdots, k_{m-1})$ denotes the contribution from the path in which states l with $k_l = 1$ were visited. Additionally, $A = A(T, \mathbf{t}; \mathbf{v}, \boldsymbol{\delta})$ denotes the "allowable" region of $\mathbf{X}(\mathbf{v})$, and $p_{v_{k_2 \cdots k_{m-1}}} = p_{v_{k_2 \cdots k_{m-1}}}(\mathbf{p}, \boldsymbol{\theta}, T, \mathbf{t}; \boldsymbol{\delta})$ is the full conditional probability that a subject's trajectory involved visiting the set of states l with $k_l = 1$. Refer to Section 4.3 in Chapter 4 for details. Note that this joint distribution is very close to the one in Section 4.3. The two differences are that the full conditional probability that the

subject took a specific path through the states, and the set of "allowable" sojourn times, are now calculated using the latent common total lead time T_0. We use the Metropolized Carlin-Chib algorithm to update the path $\mathbf{v}$ and full suite of latent sojourn times $\mathbf{X}$ for each subject as we did in Section 4.3.

To summarize the algorithm, we update the transition probabilities $\mathbf{p}$ via Gibbs sampling, and the sojourn time parameters $\boldsymbol{\theta}$ via Metropolis-Hastings with appropriate proposal distributions on each parameter. We update the common total lead time T_0 via Metropolis-Hastings. Using the Metropolized Carlin-Chib algorithm, we update the latent trajectory for each subject—the vector of state visit indicators and the collection of sequences of latent sojourn times. To generate candidate trajectories for each subject, we use the "best guess" proposal $q_2(\cdot)$ for the vector of visit indicators and an appropriate proposal distribution for the sojourn times.

5.4.2 Case 2: homogeneous total lead times

Now we present the algorithm in the general progressive state model under the Case 2 assumption. We focus on changes from the algorithm in Case 1.

Under Case 2, we will carry out inference about the subject-specific total lead times $T^1, \ldots, T^N$. The set of "allowable" sojourn times for each subject is $A^n = A^n(\mathbf{v}^n, \boldsymbol{\delta}^n; T^n, \mathbf{t}^n)$, that is, it depends on the subject's total lead time T^n. Thus, the first half of the algorithm proceeds similarly as for Case 1 with T_0 replaced by T^n.

In the second half of the algorithm, we must update the mean total lead time μ_T as well as the latent trajectories for each subject. For a given subject, the joint distribution of the suite of all latent sojourn times $\mathbf{X}$, the path $\mathbf{v}$, and the total lead time T, conditional on the other quantities in the model, is given by

$$p(\mathbf{v}, \mathbf{X}, T | \mu_T, \sigma_T, \mathbf{p}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta})$$

$$
\begin{aligned}
&= p(\mathbf{v}|T,\mathbf{p},\boldsymbol{\theta},\mathbf{t},\boldsymbol{\delta}) \cdot p(\mathbf{X}|T,\mathbf{v},\boldsymbol{\theta},\mathbf{t},\boldsymbol{\delta}) \cdot p(T|\mu_T,\sigma_T) \\
&= \prod_{k_2=0}^{1} \cdots \prod_{k_{m-1}=0}^{1} (p_{vk_2\cdots k_{m-1}} \cdot \xi(T,\boldsymbol{\theta};\mathbf{k},\boldsymbol{\delta}))^{\delta(2)^{1-k_2}\cdots\delta(m-1)^{1-k_{m-1}}} \\
&\quad \cdot \mathbf{1}(\mathbf{X} \in A) \cdot \frac{\exp\left(-\frac{(T-\mu_T)^2}{2\sigma_T^2}\right)}{1-\Phi(-\frac{\mu_T}{\sigma_T})} \cdot \mathbf{1}(T \geq 0),
\end{aligned}
$$

where σ_T is known.

It follows from the above that the full conditional distributions of a given subject's total lead time T and of the mean total lead time μ_T are the same to those derived under the illness-death case and thus are omitted here. Finally, the joint distribution of a given subject's path $\mathbf{v}$ and collection of sojourn times $\mathbf{X}$ conditional on other quantities in the model is identical to that from Eq. (5.7) presented in the previous subsection except that T_0 is now replaced by T^n for the n^{th} subject.

Thus, to summarize the estimation algorithm, we update the transition probabilities $\mathbf{p}$ via Gibbs sampling, and the sojourn time parameters $\boldsymbol{\theta}$ via Metropolis-Hastings with appropriate proposal distributions on each parameter just as in Case 1. We update the each of the total lead times $T^1,\ldots,T^N$ via Metropolis-Hastings, then update the mean total lead time via Metropolis with a symmetric proposal distribution. Using the Metropolized Carlin-Chib algorithm, we update each subject's path $\mathbf{v}$ and collection of latent sojourn times $\mathbf{X}$ as in Case 1.

5.5 Defining the model

We have developed an approach to handle left-censored observations, but our approach may need to be tailored to each application. We discuss some choices and modifications here.

First, in many applications it may not be reasonable to assume that subjects who were first observed in an early state have approximately the same total lead times as those first observed in an advanced state. It is very likely that for subjects who were first observed in an advanced state, the elapsed time since entry into the first state is, on average, longer than for those who were first observed in an early state. Additionally, we may expect more variation in the total lead times among subjects first observed in an advanced state than

those first observed in an early state. We can relax our Case 2 assumption that all subjects have a common mean total lead time in the general progressive case so that subjects first observed in state i have mean total lead time $\mu_{T,i}$ and standard deviation $\sigma_{T,i} > 0$ for $i = 1, \ldots, m-1$, that is, allowing for initial state-dependence. Generally we would expect $\mu_{T,i}$ to increase with i, but this is not a restriction of the model. Likewise, we can relax the Case 1 assumption with an initial state-dependent common total lead time so that subjects first observed in state i have common total lead time T_i for $i = 1, \ldots, m-1$.

Under the more relaxed assumption, the model includes parameters $\mu_{T,1}, \ldots, \mu_{T,m-1}$ and "known" quantities $\sigma_{T,1}, \ldots, \sigma_{T,m-1}$. The estimation procedure requires only minor modifications. Specifically, a subject's latent total lead time is updated in light of the corresponding mean, and the state-specific mean total lead times themselves are updated in light of the total lead times of the subjects first observed in that state. Specifically, the first half of the algorithm proceeds in exactly the same way as in Section 5.4.2. For the second half of the algorithm, note that for a given subject, the joint distribution of the collection of all latent sojourn times $\mathbf{X}$, the path indicator $\mathbf{v}$, and total lead time T, conditional on all other quantities in the model, is given by

$$
\begin{aligned}
& p(\mathbf{v}, \mathbf{X}, T | \mu_{T,1}, \ldots, \mu_{T,m-1}, \sigma_{T,1}, \ldots, \sigma_{T,m-1}, \mathbf{p}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \\
&\quad = p(\mathbf{v}|T, \mathbf{p}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \cdot p(\mathbf{X}|T, \mathbf{v}, \boldsymbol{\theta}, \mathbf{t}, \boldsymbol{\delta}) \cdot p(T|\mu_{T,\phi}, \sigma_{T,\phi}) \\
&\quad = \prod_{k_2=0}^{1} \cdots \prod_{k_{m-1}=0}^{1} (p_{vk_2\cdots k_{m-1}} \cdot \xi(T, \boldsymbol{\theta}; \mathbf{k}, \boldsymbol{\delta}))^{\delta(2)^{1-k_2} \cdots \delta(m-1)^{1-k_{m-1}}} \\
&\qquad \cdot \mathbf{1}(\mathbf{X} \in A) \cdot \frac{\exp\left(-\frac{(T-\mu_{T,\phi})^2}{2\sigma_{T,\phi}^2}\right)}{1 - \Phi(-\frac{\mu_{T,\phi}}{\sigma_{T,\phi}})} \cdot \mathbf{1}(T \geq 0),
\end{aligned}
$$

where $\mu_{T,\phi}$ is the parameter upon which T depends, as this subject was first observed in state ϕ. The full conditional distribution of this subject's total lead time is therefore

$$
p(T^n | \mu_{T,1}, \ldots, \mu_{T,m-1}, \sigma_{T,1}, \ldots, \sigma_{T,m-1}, \mathbf{v}^1, \ldots, \mathbf{v}^N, \mathbf{X}^1, \ldots, \mathbf{X}^N, \mathbf{t}^1, \ldots, \mathbf{t}^N, \boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^N, p_{12}, \boldsymbol{\theta})
$$

$$\begin{aligned} &\propto \quad p(T^n|\mu_{T,\phi}, \sigma_{T,\phi}, \mathbf{v}^n, \mathbf{X}^n(\mathbf{v}^n), \mathbf{t}^n, \boldsymbol{\delta}^n, \mathbf{p}, \boldsymbol{\theta}) \\ &\propto \quad \mathbf{1}(\mathbf{X}^n(\mathbf{v}^n) \in A^n(T^n, \mathbf{t}^n; \mathbf{v}^n, \boldsymbol{\delta}^n)) \cdot \exp\left(-\frac{(T^n - \mu_{T,\phi})^2}{2\sigma^2_{T,\phi}}\right) \cdot \mathbf{1}(T^n \geq 0). \end{aligned}$$

Finally we derive the full conditional distribution of each of the mean total lead times $\mu_{T,i}$, $i = 1, \ldots, m-1$. For a fixed state i, $\mu_{T,i}$ is updated in light of the latent total lead times of subjects first observed in state i. We express this in the second line of the following derivation:

$$p(\mu_{T,i}|\sigma_{T,1}, \ldots, \sigma_{T,m-1}, \mathbf{v}^1, \ldots, \mathbf{v}^N, \mathbf{X}^1, \ldots, \mathbf{X}^N, T^1, \ldots, T^N, \mathbf{t}^1, \ldots, \mathbf{t}^N, \boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^N, \mathbf{p}, \boldsymbol{\theta}) \tag{5.8}$$

$$\begin{aligned} &\propto \quad p(\mu_{T,i}|\sigma_{T,i}, \{T^n : \phi(n) = i\}) \\ &\propto \quad \left[\prod_{n=1}^{N} p(T^n|\mu_T, \sigma_{T,i})^{\mathbf{1}(\phi(n)=i)}\right] \cdot \pi(\mu_{T,i}) \\ &\propto \quad \left[\prod_{n=1}^{N} \left[\frac{\exp\left(-\frac{(T^n-\mu_{T,i})^2}{2\sigma^2_{T,i}}\right)}{1 - \Phi\left(-\frac{\mu_{T,i}}{\sigma_{T,i}}\right)}\right]^{\mathbf{1}(\phi(n)=i)}\right] \cdot \pi(\mu_{T,i}). \end{aligned}$$

Each of these parameters is updated in the same way as in Section 5.4.2.

An important issue to consider is how to choose the assumed variation in total lead times among each group of subjects, $\sigma_{T,1}, \ldots, \sigma_{T,m-1}$. This choice should be made primarily based on knowledge of the process. For example, if there is existing knowledge about the sojourn times in each state, then this should be used to assign an appropriate standard deviation for each state. Consider, for example, a simple progressive process. We may assume that the standard deviations should be chosen such that $\sigma_{T,1} < \cdots < \sigma_{T,m-1}$ implying that the amount of variability in the time until a subject reaches state i increases with i.

5.6 Simulation studies

In this section we examine the performance of the proposed approach to accommodate left-censored observations. For simplicity we use the illness-death model and investigate the impact of factors specific to our modeling of left censoring observations. In Section 5.6.1 we demonstrate the performance of the approach for various scenarios under the Case 1

assumption that the total lead times arise from a degenerate distribution. Then in Section 5.6.2 we investigate the impact of the degree of variability in the total lead times under the Case 2 assumption that the total lead times arise from a truncated normal distribution.

5.6.1 Performance of the proposed approach under the Case 1 assumption.

In this set of simulations we first assume that the total lead times are identical across all subjects regardless of first observed state. We demonstrate the performance of the approach when this common total lead time is zero—that is, subjects are observed exactly when they enter state 1, as we have considered previously—and when it is two years and four years. We repeat these simulations, allowing the assumed common total lead time to differ among subjects first observed in states 1 and 2. That is, we carry out inference separately on the assumed common total lead times, denoted by T_1 and T_2, among subjects first observed in states 1 and 2 respectively. The scenarios under consideration are shown in Table 5.1. In each of the first six scenarios the true total lead times are all equal to the value indicated in the second column of the table with the corresponding σ_T equal to zero. Considering a single scenario we carry out several explorations to assess the impact of several factors. We examine one additional scenario in which the total lead times are generated from a truncated normal distribution in which the original normal random variable had mean 2.0 and standard deviation 1.0.

Table 5.1: Case 1 scenarios.

Scenario	μ_T	σ_T	first observed state assumption
1	0.0	0.0	present
2	2.0	0.0	present
3	4.0	0.0	present
4	0.0	0.0	absent
5	2.0	0.0	absent
6	4.0	0.0	absent
7	2.0	1.0	absent

Throughout these simulations we assume that each subject visits the intermediate state 2 with 90% probability ($p_{12} = 0.9$). We assume that each of the sojourn times follows a Weibull distribution with common shape and scale parameters: $X_{ij} \sim Weibull(k_{ij}, \theta_{ij})$ with $k_{ij} = 2.0$ and $\theta_{ij} = 4.0$ for each $(i, j) \in \{(1, 2), (2, 3), (1, 3)\}$. We assume that subjects are observed frequently: every three months ($\Delta = 0.25$). For simplicity we follow each subject until he enters the absorbing state. We assume informative truncated normal priors on the sojourn time parameters and a noninformative uniform prior on the transition probability p_{12}. For each of the common total lead time parameters, it was necessary to impose an informative truncated normal prior centered at the true value to ensure convergence of the chain. When a noninformative prior is used, the lack of information about the total lead times in the observed data leads to lack of convergence of the chain.

We illustrate the performance of the approach for $N = 400$ subjects. Results, based on $M = 100$ simulated datasets, are shown in Tables 5.2–5.8. Posterior means as well as model-based and empirical standard deviations are given for each of the parameters of interest. Although the true sojourn times were generated from the Weibull distribution given in the previous paragraph, we note that the "true" values of the parameters given in the tables are different from these. For each simulated dataset, subjects who do not contribute useful longitudinal information are not included. Specifically, in the current setting, subjects who are not observed before state 3 due to left censoring are not included in the final simulated dataset. Hence, subjects with short sojourn times—particularly those who skip state 2—are truncated. The impact of the resulting selection process increases as the degree of left censoring becomes more severe. The "true" parameter values presented in the tables take this source of truncation into account.

Additionally, we present combined results for the common total lead time parameters based on the primary simulation studies in Table 5.12. Specifically, we give posterior means and model-based and empirical standard deviations for the assumed common total lead time parameters: T_0 for Scenarios 1–3, and T_1 and T_2 for Scenarios 4–7.

The proposed approach performs well in the Case 1 scenarios under investigation, consistently estimating the parameters of interest in almost every case when the true parameter values are adjusted to reflect the impact of truncation (see Tables 5.2–5.8). The approach

unfailingly estimates the probability of visiting state 2 in every scenario, even when the impact of left censoring is severe (Tables 5.4 and 5.7). Sojourn time parameters were consistently estimated in most cases. When the impact of left censoring is severe, the effective true values of the parameters are quite different from the data-generating parameter values, especially for the parameters corresponding to X_{13}. The prior distribution for the parameters was centered at the data-generating value. In these scenarios, the posterior mean lies between the data-generating and effective true parameter values, which is what we would expect. This effect can be seen to a small degree in Tables 5.3 and 5.6, and to a greater degree in Tables 5.4 and 5.7.

We observe that, as the common total lead time increases, the posterior standard deviation of the parameters associated with the sojourn time in state 1 increases correspondingly. This is to be expected, since we have less information about the sojourn time in state 1 as the impact of left censoring is increased. As we have observed in previous simulation studies, when p_{12} is close to one, the posterior standard errors of the parameters associated with X_{13} are larger than those associated with X_{12} and X_{23}, since the observed data contains less information about subjects who skipped state 2. In each scenario, the proposed approach consistently estimates the common total lead times (Table 5.12), though we note that inference reflects substantial prior information.

We note that the effective true value of p_{12} increases as the effect of left censoring becomes more severe: when the common total lead time is equal to 4.0, the effective true probability of visiting state 2 is equal to 0.955 though the data-generating probability is 0.900. Under left censoring, subjects who truly skip state 2 have a higher probability of being truncated than those who visit state 2. Hence, the effective p_{12} is greater than the data-generating p_{12}, and the difference grows as the degree of left censoring increases. Since truncation disproportionately affects subjects who skip truly state 2, the true values of the parameters corresponding to X_{13} are affected more than those coresonding to X_{12} and X_{23} by the adjustment for truncation.

When the assumption that subjects first observed in states 1 and 2 have the same common total lead time is relaxed, the posterior standard deviations associated with a given parameter increases in many cases (compare Tables 5.2 and 5.5, 5.3 and 5.6, 5.4 and

5.7). This is what expected, since there is an additional parameter to be estimated when the assumption is relaxed.

Finally, when substantial variation truly exists in the total lead times and we incorrectly assume that no variation exists, we observe downward bias in the estimation of the shape parameter associated with X_{12}. We note that our incorrect assumption about the total lead times—that no variation exists—is consistent with underestimation of the shape parameter and overestimation of the scale parameter. That is, assuming common total lead times among subjects first observed in states 1 and 2 when there was substantial variation in true total lead times leads to latent sojourn times X_{12} whose distribution is slightly "flatter" than the correspdonding data-generating distribution. The observed underestimation of the shape parameter could be explained by this phenomenon. The bias resulting from our incorrect assumption would also have an impact on the parameters governing X_{13}, but the impact is less evident here since a number of factors make estimation of these parameters difficult.

To illustrate that our proposed approach is unbiased for estimating the true parameter values in the absence of truncation, we show results in which, for each simulated dataset, even subjects who do not contribute useful longitudinal information are included. Specifically, we repeat the simulation of Scenario 2, with all true parameters and conditions kept the same except that subjects who are first observed in state 3 are retained in the simulated dataset. While they do not provide longitudinal information within the study, given that they were first observed in the absorbing state 3, it is known that either $X_{12}+X_{13} < T$ if the subject had visited state 2, or that $X_{13} < T$ otherwise. Using this information, unlike the original Scenario 2 simulation study, the effective true parameter values are the same as the data-generating true values. Results, presented in Table 5.9, demonstrate that our proposed approach correctly estimates the underlying true values in the absence of truncation.

We used informative priors for the sojourn time parameters in the original simulation studies to replicate the situation where expert opinion about these parameters is incorporated into the estimation, but such information may not be available in an application. To demonstrate that our proposed approach can perform well in the absence of informative priors, we show results where noninformative priors are used for all the parameters of in-

terest. Specifically, we repeat the simulation of Scenario 2 with all true parameters and conditions kept the same except that we assume noninformative uniform priors, rather than truncated normal priors centered at the true values, for the sojourn time parameters. We continue to assume an informative truncated normal prior for the common total lead time parameter for reasons that we discussed previously. We carried out this simulation study for two sample sizes to illustrate that the precision of our estimates improves as the amount of observed information increases. Results are presented in Table 5.10.

In the original studies we used an informative prior for the common total lead time, T_0, that was centered at the true value in each case. To investigate the impact on inference when the prior for the common total lead time is centered at the wrong value, we repeat the study for Scenario 2, using a prior centered at 1.0 rather than the correct value 2.0. Results, shown in Table5.11, indicate that assuming an incorrect prior for the common total lead time had the greatest impact on the sojourn time parameters in state 1, X_{12} and $X_{1,3}$. Inference for the probability of visiting state 2 and the parameters corresponding to the sojourn time in state 2, X_{23}, was relatively insensitive to the choice of prior for the lead time parameter.

5.6.2 Performance of the proposed approach under the Case 2 assumption.

In this set of simulations we demonstrate the performance of the proposed estimation method under the Case 2 assumption. When the assumed amount of variability in the total lead times across subjects is small relative to the means of the sojourn times, our approach under Case 2 has similar performance to that described under Case 1 and the simulation results are omitted. Instead, in this section we explore the impact of the degree of the assumed variability in the total lead times. The scenarios are shown in the first two rows of Table 5.13. The true total lead times for all subjects are generated from a truncated normal distribution where the original normal random variable had mean and standard deviation as indicated in the table. We consider just one mean total lead time, 2 years, with two levels of variability in the total lead times: a standard deviation of 0.1 and 1.0 years. For the first two studies we assume that the value of σ_T has been correctly specified, that

is, that the assumed and true values of σ_T are equal. In the algorithm we allow separate estimation of the mean total lead times for subjects first observed in states 1 and 2, $\mu_{T,1}$ and $\mu_{T,2}$.

Table 5.13: Case 2 scenarios.

Scenario	μ_T	true σ_T	assumed σ_T	first observed state assumption
8	2.0	0.1	0.1	absent
9	2.0	1.0	1.0	absent
10	2.0	0.1	1.0	absent
11	2.0	1.0	0.1	absent

The setup for the current simulation studies is very similar to that for the Case 1 studies in the previous subsection. Specifically, for each of these two simulations we assume that each subject visits the intermediate state 2 with 90% probability ($p_{12} = 0.9$). We assume that each of the sojourn times follows a Weibull distribution with common shape and scale parameters: $X_{ij} \sim Weibull(k_{ij}, \theta_{ij})$ with $k_{ij} = 2.0$ and $\theta_{ij} = 4.0$ for each $(i, j) \in \{(1, 2), (2, 3), (1, 3)\}$. We assume that subjects are observed frequently: every three months ($\Delta = 0.25$). We follow each subject until he enters the absorbing state. We assume informative normal priors on the sojourn time parameters and a noninformative uniform prior on the transition probability p_{12}. We imposed an informative normal prior centered at the true value on the mean total lead time parameters $\mu_{T,1}$ and $\mu_{T,2}$.

We illustrate the performance of the approach for $N = 400$ subjects. Results, based on $M = 100$ simulated datasets, are shown in Tables 5.14–5.15. Posterior means as well as model-based and empirical standard deviations are given for each of the parameters of interest. Additionally, posterior means and model-based and empirical standard deviations are given for the mean total lead time parameters: $\mu_{T,1}$ and $\mu_{T,2}$ in Table 5.16. In each scenario we assume the same degree of variability in the total lead times as was used to generate the data. That is, we do not explore misspecification of σ_T here.

When the assumed degree of variability in the total lead times is small, the approach

consistently estimates the parameters of interest (see Table 5.14). The apparent bias for estimation of p_{12} and some of the parameters governing the sojourn times may result from truncation, as discussed in the previous subsection. When the assumed degree of variability in the total lead times is large, the precision of the estimated parameters is reduced, but the results are similar to the case where σ_T was small (compare Tables 5.14 and 5.15). Table 5.13 shows some downward bias for $\mu_{T,1}$ and upward bias for $\mu_{T,2}$, especially in Scenario 9, for which there is substantial variation in the true total lead times. The bias for these mean total lead times likely contributed to the observed bias for the parameters of interest.

Finally, since the value of σ_T is unknown in practice, we explore the impact of misspecification of σ_T. That is, we repeat the studies of Scenarios 8 and 9, but we assume a value of σ_T other than the true value in each case. Refer to the last two rows of Table 5.13. Results, shown in Tables 5.17–5.18, demonstrate that inference for the parameters of interest is not sensitive to misspecification of the degree of variability in the total lead times.

5.7 Discussion

We have proposed an approach to address left censoring under the assumption of relative homogeneity of total lead times among subjects with the same initial state. This approach allows for modeling panel observations of a multi-state progressive process without imposing the exponential assumption on the sojourn times.

As we have discussed, although our proposed approach can handle left-censored observations, one limitation is that it requires the assumption that the total lead times among subjects with the same observed initial state are relatively homogeneous and additionally requires the use of informative priors.

Our approach may not always be applicable. In such cases, an alternative approach is to directly model the excess time in the first observed state. That is, we would model $E_\phi = X_\phi - L_\phi$ instead of modeling the components X_ϕ and L_ϕ. Specifically, consider a simple progressive three-state process in which observations are subject to left censoring. We choose a model for the sojourn times for states as well as the excess times. Let $X_i \sim f_i(\cdot; \boldsymbol{\theta}_i)$ and $E_i \sim g_i(\cdot; \boldsymbol{\omega}_i)$ for $i = 1, 2$ denote the models for sojourn and excess times, respectively. Note that, for any given subject, we have either E_i or X_i depending on whether the observation

in state i is left-censored or not. Thus, a subject who is observed in all three states will have a latent excess time E_1 in state 1 and a latent sojourn time X_2 in state 2, while a subject who is observed in states 2 and 3 will have only a latent excess time E_2 in state 2. While the goal is still to carry out inference about the parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, we must also carry out inference about the nuisance parameters $\boldsymbol{\omega}_1$ and $\boldsymbol{\omega}_2$.

The above approach has been taken, for example, by Mitchell et al. (2011) to deal with the first observed state in the context of an alternating two-state process under panel observation. The approach has the advantage that it does not impose an assumption on the unobservable lead time. However, it has one drawback in that not all observations provide information about the parameters governing the disease progression. Consider the information provided by observations of state 2. In this alternative approach, only observations from subjects who are not left-censored in state 2 contribute to the learning about $\boldsymbol{\theta}_2$. In our approach, all observations of state 2 contribute to the learning about the sojourn time parameters $\boldsymbol{\theta}_2$, regardless of whether they are left-censored.

We also note that the alternative approach we sketched above could be implemented using the version of our method presented in Chapter 4, which did not account for left censoring. We would need only to expand the state model to account for the fact that left-censored observations must be treated differently. Specifically, in the example of a simple progressive process with three states, we would add states 1^* and 2^* for those who are left-censored in state 1 or 2, respectively. The corresponding expanded state model is shown in Figure 5.4. The addition of states 1^* and 2^* is similar to the approach taken by Sternberg and Satten (1999).

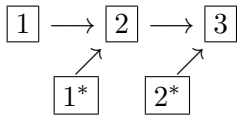


Figure 5.4: Expanded state model, including states 1^* and 2^* for subjects who are left-censored in states 1 and 2, respectively.

Finally, a third approach would be to consider the method proposed by Ware and DeMets

(1976). Specifically, their method reverses the time axis and treats the left-censored state observations analogously to right-censored observations. That is, a latent sojourn time contributes the density of the corresponding distribution, while a latent excess time contributes the cumulative distribution function. Under right censoring, we assume that subjects are censored independently of their disease status. When we apply this approach in the case of left censoring, however, the corresponding assumption is that we begin observing subjects independently of their disease status. This approach may be useful depending on the validity of this assumption.

In our proposed approach we chose to model the total lead time for each subject since it may be natural to express the degree of left censoring in terms of the length of time between entering the initial state and the first observation. However, inference about the parameters associated with the total lead times for advanced states has more uncertainty than those for the early states. Considering a simple progressive model, for example, a subject who is first observed in the third state has a total lead time equal to the sum of his sojourn times in states 1 and 2 and his state lead time in state 3. Uncertainty about each of these latent times contributes to uncertainty about the total lead time. A variation of our proposed approach involves modeling the state lead times directly, that is, we choose some model for the state lead times in each state, such as a truncated normal model and inference proceeds analogously.

We emphasize that our approach addresses *only* left censoring, and not issues of truncation and sampling bias (Hosmer et al. (2008, pp. 8, 228); Klein and Moeschberger (1997, pp. 17, 72–74)). For example, for subjects who progress quickly through the disease stages, we may be less likely to observe these subjects in the early stages of disease. The sojourns in the early stages are subject to left truncation, and ignoring this may result in length-biased sampling. Observation of the patients themselves may also be subject to left truncation. For example, consider the illness-death model and suppose that the first two states are stages of illness and that the absorbing state is death. Suppose that all subjects in the population of interest had an identical waiting period, $T_0 > 0$, from the time they entered the initial disease stage to the time they were first observed. Thus, subjects who take less than T_0 units of time to progress to death do not have a chance to be observed. These subjects

are left-truncated. An implicit assumption of our proposed approach is that the dataset consists of observations from a simple random sample of subjects from the population of interest.

In many applications where left censoring is a potential concern, so is left truncation. For example, suppose that some subjects in a dataset are first observed late in the process while others are first observed in an early stage. We may be suspicious that the first group tend to be faster progressors than the second group, that is, that observations of the early states for subjects first observed in a relatively late disease stage are truncated. Applications in which left censoring is present but left truncation is not a potential concern exist, however. Consider a study of mother-to-child transmission of HIV. A cohort of HIV-positive pregnant women are followed longitudinally and we are interested in characterizing the natural history of HIV/AIDS in the infants. Although transmission of HIV from mother to child can occur during pregnancy, delivery, or through breastfeeding (Tóth et al., 2001), the child's HIV infection status may not be certain until several months after birth (HHS, 2012). If a child is found to be HIV-positive at the first test, then the time of infection is left-censored. Since inclusion of the children in the study seems unrelated to HIV/AIDS progression, left truncation may not be a potential concern in this example.

Our proposed approach is able to estimate the parameters of the semi-Markov process in the absence of truncation. If truncation is present, then the proposed approach estimates the parameters of the underlying process conditional on being observed and the results must be carefully interpreted to reflect this condition. For example, suppose we are interested in characterizing the natural history of a potentially fatal progressive disease that frequently occurs in children, and that our dataset consists of observations that begin at age five. Suppose that we model this disease as a general progressive process in which the absorbing state is death. Children who died of the disease before reaching age five would not be included in the dataset and so the observations would be left-truncated. Hence, when directly applying our proposed methods to such data, our interpretation of the parameter estimates would be conditional on surviving to age five. Based on the observed data, we cannot possibly carry out inference about the entire population of interest having this disease, that is, regardless of survival to age five, without knowledge of how many children

died from the disease before turning five.

Addressing the issues of truncation in a general case is challenging. To carry out inference on the population of interest given a truncated sample, we must make assumptions about the destiny of the subjects whose observations are truncated. We note that, however, that for a variety of cases, our proposed approach may be adapted to accommodate left truncation of observations of entire subjects or of early disease states. Next, we examine the existing methods for dealing with truncation and discuss how one can adapt our proposed methods to handle truncation.

Turnbull (1976) considers the situation in which observations of a time-to-event variable are subject to truncation, and proposes a nonparametric approach to estimate the cumulative distribution function using the concept of "ghost" observations, values that we would observe were it not for the truncation process. Sternberg and Satten (1999) extends the approach of Turnbull (1976) to the case of a simple progressive process. An alternative approach to carrying out inference about a time-to-event variable whose observations are subject to truncation is to make a parametric assumption as in Blight (1970). In many applications, the assumption of Turnbull (1976) may not apply. Blight (1970)'s approach may be reasonable when motivated by scientific knowledge of the truncation process. Building on the idea of "ghost" observations (Turnbull (1976)), in the above children's example, suppose that the number of children in the population of interest who died before age five were known from other records or could be estimated. We could incorporate this information to carry out inference about the parameters of the entire population. Specifically, we add "ghost" observations. For each of these "ghost" observations, although we have no information about the sequence of states visited, we know that the sum of the sojourn times in each state is less than five years. This information could be used in the estimation of the corresponding latent sojourn times in the algorithm. The latent path through the states would be updated in light of the current values of the suite of latent sojourn times and parameters of interest, as usual, and the algorithm would proceed as before.

We have presented an approach to handle the presence of left-censored observation and have outlined approaches to address left-truncated observation. However, each approach makes assumptions and has limitations. For any particular application the choice of ap-

proach, as well as the interpretation of results, should acknowledge these limitations.

Table 5.2: Results for Scenario 1 with true common total lead times = 0.0.

(i,j)	truth	$\hat{k}_{ij}$	$SD_{model}(k_{ij})$	$SD_{emp}(k_{ij})$
$(1,2)$	2.002	2.056	0.090	0.081
$(2,3)$	1.999	1.997	0.083	0.085
$(1,3)$	1.997	2.067	0.221	0.221

(i,j)	truth	$\hat{\theta}_{ij}$	$SD_{model}(\theta_{ij})$	$SD_{emp}(\theta_{ij})$
$(1,2)$	3.995	4.120	0.125	0.122
$(2,3)$	3.999	3.986	0.109	0.117
$(1,3)$	3.989	4.066	0.275	0.253

(i,j)	truth	$\hat{p}_{ij}$	$SD_{model}(p_{ij})$	$SD_{emp}(p_{ij})$
$(1,2)$	0.900	0.899	0.015	0.014

Table 5.3: Results for Scenario 2 with true common total lead times = 2.0.

(i,j)	truth	$\hat{k}_{ij}$	$SD_{model}(k_{ij})$	$SD_{emp}(k_{ij})$
$(1,2)$	2.027	2.054	0.101	0.093
$(2,3)$	2.039	2.026	0.084	0.095
$(1,3)$	2.821	2.549	0.268	0.256

(i,j)	truth	$\hat{\theta}_{ij}$	$SD_{model}(\theta_{ij})$	$SD_{emp}(\theta_{ij})$
$(1,2)$	4.041	4.087	0.147	0.121
$(2,3)$	4.037	4.011	0.108	0.119
$(1,3)$	4.705	4.454	0.270	0.215

(i,j)	truth	$\hat{p}_{ij}$	$SD_{model}(p_{ij})$	$SD_{emp}(p_{ij})$
$(1,2)$	0.918	0.919	0.014	0.011

Table 5.4: Results for Scenario 3 with true common total lead times = 4.0.

(i,j)	truth	$\hat{k}_{ij}$	$SD_{model}(k_{ij})$	$SD_{emp}(k_{ij})$
$(1,2)$	2.232	2.389	0.124	0.124
$(2,3)$	2.233	2.245	0.102	0.106
$(1,3)$	4.175	2.633	0.345	0.301

(i,j)	truth	$\hat{\theta}_{ij}$	$SD_{model}(\theta_{ij})$	$SD_{emp}(\theta_{ij})$
$(1,2)$	4.291	4.446	0.136	0.113
$(2,3)$	4.285	4.204	0.111	0.103
$(1,3)$	6.022	5.039	0.317	0.216

(i,j)	truth	$\hat{p}_{ij}$	$SD_{model}(p_{ij})$	$SD_{emp}(p_{ij})$
$(1,2)$	0.955	0.954	0.010	0.010

Table 5.5: Results for Scenario 4 with common total lead times = 0.0.

(i,j)	truth	$\hat{k}_{ij}$	$SD_{model}(k_{ij})$	$SD_{emp}(k_{ij})$
$(1,2)$	2.002	2.057	0.090	0.083
$(2,3)$	1.999	1.996	0.082	0.083
$(1,3)$	1.997	2.064	0.220	0.218

(i,j)	truth	$\hat{\theta}_{ij}$	$SD_{model}(\theta_{ij})$	$SD_{emp}(\theta_{ij})$
$(1,2)$	3.995	4.120	0.125	0.117
$(2,3)$	3.999	3.987	0.108	0.118
$(1,3)$	3.989	4.071	0.273	0.248

(i,j)	truth	$\hat{p}_{ij}$	$SD_{model}(p_{ij})$	$SD_{emp}(p_{ij})$
$(1,2)$	0.900	0.899	0.015	0.014

Table 5.6: Results for Scenario 5 with common total lead times = 2.0.

(i,j)	truth	$\hat{k}_{ij}$	$SD_{model}(k_{ij})$	$SD_{emp}(k_{ij})$
$(1,2)$	2.027	2.052	0.096	0.094
$(2,3)$	2.039	2.028	0.084	0.095
$(1,3)$	2.821	2.561	0.279	0.261

(i,j)	truth	$\hat{\theta}_{ij}$	$SD_{model}(\theta_{ij})$	$SD_{emp}(\theta_{ij})$
$(1,2)$	4.041	4.086	0.140	0.118
$(2,3)$	4.037	4.011	0.108	0.121
$(1,3)$	4.705	4.470	0.262	0.223

(i,j)	truth	$\hat{p}_{ij}$	$SD_{model}(p_{ij})$	$SD_{emp}(p_{ij})$
$(1,2)$	0.918	0.919	0.014	0.011

Table 5.7: Results for Scenario 6 with common total lead times = 4.0.

(i,j)	truth	$\hat{k}_{ij}$	$SD_{model}(k_{ij})$	$SD_{emp}(k_{ij})$
$(1,2)$	2.232	2.436	0.150	0.124
$(2,3)$	2.233	2.244	0.101	0.106
$(1,3)$	4.175	2.660	0.333	0.313

(i,j)	truth	$\hat{\theta}_{ij}$	$SD_{model}(\theta_{ij})$	$SD_{emp}(\theta_{ij})$
$(1,2)$	4.291	4.408	0.125	0.101
$(2,3)$	4.285	4.202	0.112	0.105
$(1,3)$	6.022	5.000	0.306	0.225

(i,j)	truth	$\hat{p}_{ij}$	$SD_{model}(p_{ij})$	$SD_{emp}(p_{ij})$
$(1,2)$	0.955	0.954	0.010	0.010

Table 5.8: Results for Scenario 7 where true total lead times are generated from truncated normal distribution with mean 2.0, standard deviation 1.0.

(i,j)	truth	$\hat{k}_{ij}$	$SD_{model}(k_{ij})$	$SD_{emp}(k_{ij})$
$(1,2)$	2.050	1.940	0.093	0.088
$(2,3)$	2.061	2.024	0.084	0.086
$(1,3)$	2.615	2.462	0.274	0.226

(i,j)	truth	$\hat{\theta}_{ij}$	$SD_{model}(\theta_{ij})$	$SD_{emp}(\theta_{ij})$
$(1,2)$	4.057	4.093	0.144	0.130
$(2,3)$	4.070	4.002	0.107	0.113
$(1,3)$	4.732	4.538	0.278	0.217

(i,j)	truth	$\hat{p}_{ij}$	$SD_{model}(p_{ij})$	$SD_{emp}(p_{ij})$
$(1,2)$	0.921	0.921	0.014	0.012

Table 5.9: Results for Scenario 2 in the absence of truncation. True common total lead times = 2.0.

(i,j)	truth	$\hat{k}_{ij}$	$SD_{model}(k_{ij})$	$SD_{emp}(k_{ij})$
$(1,2)$	2.000	2.010	0.100	0.094
$(2,3)$	2.000	2.005	0.087	0.090
$(1,3)$	2.000	1.822	0.225	0.212

(i,j)	truth	$\hat{\theta}_{ij}$	$SD_{model}(\theta_{ij})$	$SD_{emp}(\theta_{ij})$
$(1,2)$	4.000	4.037	0.149	0.091
$(2,3)$	4.000	4.038	0.113	0.113
$(1,3)$	4.000	3.937	0.302	0.193

(i,j)	truth	$\hat{p}_{ij}$	$SD_{model}(p_{ij})$	$SD_{emp}(p_{ij})$
$(1,2)$	0.900	0.899	0.016	0.014

Table 5.10: Results for Scenario 2 under noninformative priors for $N = 200$ (first three tables) and $N = 400$ (last three tables). True common total lead times = 2.0.

$N = 200$.

(i,j)	truth	$\hat{k}_{ij}$	$SD_{model}(k_{ij})$	$SD_{emp}(k_{ij})$
$(1,2)$	2.027	2.051	0.140	0.153
$(2,3)$	2.039	2.029	0.121	0.101
$(1,3)$	2.821	2.985	0.493	0.652

(i,j)	truth	$\hat{\theta}_{ij}$	$SD_{model}(\theta_{ij})$	$SD_{emp}(\theta_{ij})$
$(1,2)$	4.041	4.074	0.197	0.165
$(2,3)$	4.037	4.064	0.158	0.159
$(1,3)$	4.705	4.801	0.464	0.549

(i,j)	truth	$\hat{p}_{ij}$	$SD_{model}(p_{ij})$	$SD_{emp}(p_{ij})$
$(1,2)$	0.918	0.916	0.019	0.021

$N = 400$.

(i,j)	truth	$\hat{k}_{ij}$	$SD_{model}(k_{ij})$	$SD_{emp}(k_{ij})$
$(1,2)$	2.027	2.070	0.106	0.097
$(2,3)$	2.039	2.028	0.086	0.097
$(1,3)$	2.821	2.954	0.375	0.453

(i,j)	truth	$\hat{\theta}_{ij}$	$SD_{model}(\theta_{ij})$	$SD_{emp}(\theta_{ij})$
$(1,2)$	4.041	4.112	0.155	0.132
$(2,3)$	4.037	4.015	0.113	0.123
$(1,3)$	4.705	4.773	0.327	0.352

(i,j)	truth	$\hat{p}_{ij}$	$SD_{model}(p_{ij})$	$SD_{emp}(p_{ij})$
$(1,2)$	0.918	0.919	0.014	0.011

Table 5.11: Results for Scenario 2 with prior for common total lead times centered at 1.0. True common total lead times = 2.0.

(i,j)	truth	$\hat{k}_{ij}$	$SD_{model}(k_{ij})$	$SD_{emp}(k_{ij})$
$(1,2)$	2.027	1.530	0.094	0.069
$(2,3)$	2.039	1.994	0.084	0.095
$(1,3)$	2.821	2.215	0.263	0.237

(i,j)	truth	$\hat{\theta}_{ij}$	$SD_{model}(\theta_{ij})$	$SD_{emp}(\theta_{ij})$
$(1,2)$	4.041	3.038	0.166	0.123
$(2,3)$	4.037	3.956	0.111	0.120
$(1,3)$	4.705	3.781	0.284	0.249

(i,j)	truth	$\hat{p}_{ij}$	$SD_{model}(p_{ij})$	$SD_{emp}(p_{ij})$
$(1,2)$	0.918	0.919	0.014	0.011

Table 5.12: Results for common total lead times for Scenarios 1–7. In Scenarios 4–7, common total lead times for subjects first observed in states 1 and 2 are estimated separately.

Scenario	parameter	truth	mean	SD_{model}	SD_{emp}
1	T_0	0.000	0.070	0.055	0.020
2	T_0	2.000	2.016	0.104	0.054
3	T_0	4.000	4.084	0.100	0.061
4	T_1	0.000	0.071	0.056	0.021
	T_2	0.000	0.078	0.065	0.013
5	T_1	2.000	2.017	0.108	0.043
	T_2	2.000	2.003	0.108	0.043
6	T_1	4.000	3.981	0.110	0.036
	T_2	4.000	4.097	0.108	0.058
7	T_1	2.000	1.987	0.114	0.039
	T_2	2.000	2.029	0.107	0.043

Table 5.14: Results for Scenario 8 where true total lead times are generated from truncated normal distribution with mean 2.0, standard deviation 0.1.

(i,j)	truth	$\hat{k}_{ij}$	$SD_{model}(k_{ij})$	$SD_{emp}(k_{ij})$
$(1,2)$	2.035	2.089	0.091	0.093
$(2,3)$	2.040	2.029	0.085	0.091
$(1,3)$	2.790	2.514	0.270	0.238

(i,j)	truth	$\hat{\theta}_{ij}$	$SD_{model}(\theta_{ij})$	$SD_{emp}(\theta_{ij})$
$(1,2)$	4.029	4.082	0.108	0.114
$(2,3)$	4.040	4.021	0.109	0.123
$(1,3)$	4.725	4.443	0.267	0.231

(i,j)	truth	$\hat{p}_{ij}$	$SD_{model}(p_{ij})$	$SD_{emp}(p_{ij})$
$(1,2)$	0.920	0.919	0.014	0.011

Table 5.15: Results for Scenario 9 where true total lead times are generated from truncated normal distribution with mean 2.0, standard deviation 1.0.

(i,j)	truth	$\hat{k}_{ij}$	$SD_{model}(k_{ij})$	$SD_{emp}(k_{ij})$
$(1,2)$	2.050	2.112	0.129	0.126
$(2,3)$	2.061	2.055	0.087	0.093
$(1,3)$	2.615	2.306	0.284	0.179

(i,j)	truth	$\hat{\theta}_{ij}$	$SD_{model}(\theta_{ij})$	$SD_{emp}(\theta_{ij})$
$(1,2)$	4.057	4.070	0.129	0.143
$(2,3)$	4.070	4.063	0.108	0.115
$(1,3)$	4.732	4.543	0.291	0.218

(i,j)	truth	$\hat{p}_{ij}$	$SD_{model}(p_{ij})$	$SD_{emp}(p_{ij})$
$(1,2)$	0.921	0.922	0.013	0.012

Table 5.16: Results for mean total lead times for Scenarios 8–9.

Scenario	parameter	truth	mean	SD_{model}	SD_{emp}
8	$\mu_{T,1}$	2.000	1.988	0.014	0.025
	$\mu_{T,2}$	2.000	2.133	0.061	0.058
9	$\mu_{T,1}$	2.000	1.784	0.104	0.061
	$\mu_{T,2}$	2.000	2.387	0.105	0.051

Table 5.17: Results for Scenario 8 model misspecification: true $\sigma_T = 0.1$, model $\sigma_T = 1.0$.

(i,j)	truth	$\hat{k}_{ij}$	$SD_{model}(k_{ij})$	$SD_{emp}(k_{ij})$
$(1,2)$	2.035	2.232	0.140	0.110
$(2,3)$	2.040	2.049	0.086	0.095
$(1,3)$	2.790	2.347	0.294	0.215

(i,j)	truth	$\hat{\theta}_{ij}$	$SD_{model}(\theta_{ij})$	$SD_{emp}(\theta_{ij})$
$(1,2)$	4.029	4.027	0.123	0.134
$(2,3)$	4.040	4.072	0.109	0.122
$(1,3)$	4.725	4.467	0.294	0.222

(i,j)	truth	$\hat{p}_{ij}$	$SD_{model}(p_{ij})$	$SD_{emp}(p_{ij})$
$(1,2)$	0.920	0.919	0.014	0.011

Table 5.18: Results for Scenario 8 model misspecification: true $\sigma_T = 0.1$, model $\sigma_T = 1.0$.

(i,j)	truth	$\hat{k}_{ij}$	$SD_{model}(k_{ij})$	$SD_{emp}(k_{ij})$
$(1,2)$	2.050	1.971	0.089	0.090
$(2,3)$	2.061	2.029	0.084	0.086
$(1,3)$	2.615	2.463	0.277	0.218

(i,j)	truth	$\hat{\theta}_{ij}$	$SD_{model}(\theta_{ij})$	$SD_{emp}(\theta_{ij})$
$(1,2)$	4.057	4.105	0.116	0.126
$(2,3)$	4.070	4.008	0.107	0.113
$(1,3)$	4.732	4.520	0.263	0.218

(i,j)	truth	$\hat{p}_{ij}$	$SD_{model}(p_{ij})$	$SD_{emp}(p_{ij})$
$(1,2)$	0.921	0.921	0.014	0.012

Chapter 6

ACCOUNTING FOR INTER-SUBJECT DIFFERENCES IN DISEASE PROGRESSION RATE VIA COVARIATE ADJUSTMENT

6.1 Introduction

Throughout our previous chapters we have treated subjects interchangeably so that each had the same chance of maintaining a given health state for a certain time period, for example. Often, it is of interest to investigate how observed differences in disease progression rate depend on demographic or clinical characteristics of the subjects. As an important example, we may wish to determine whether a new treatment slows disease progression, and to characterize the treatment effect. In this chapter we present an extension to our previous approach that quantifies the differences in rates of disease progression. In Section 6.2 we discuss two ways in which we may account for differences in disease progression rates between subjects: through the visited disease stages, and through the times spent in each stage. We propose an extension to our approach that allows for variable rates of disease progression across subjects via covariate adjustment in Section 6.3. We discuss modeling choices in Section 6.4. In Section 6.5 we demonstrate the performance of the model via simulation study. Finally in Section 6.6 we discuss the assumptions and model choices in our approach.

6.2 Covariate adjustment

Suppose, as an example, that we wish to compare the disease progression in two groups of patients. If we are using an illness-death model, then we may be interested in whether one group tends to skip the intermediate disease stage more often than the other. Alternatively, we may be interested in whether one group tends to remain in the first stage of disease longer than the other. In each case, to address the question of interest, we can quantify differences in disease progression rate across subjects via covariate adjustment. The first situation corresponds to adjusting for covariates through the embedded Markov chain, while

the second situation corresponds to adjusting for covariates through the conditional sojourn times. In the following sections we present the extension of our existing algorithm to adjust for covariates through one or both of these parts of the model. We build on the existing development, considering a general progressive model with m states in which observations are subject to left censoring. We assume that observations of subjects are left censored in such a way that the total lead times are relatively homogeneous, but we assume no relationship between the total lead times of subjects with different observed initial states.

6.3 Proposed approach to account for inter-subject differences in disease progression rates via covariate adjustment

Here we propose an extension to our existing approach to account for inter-subject differences via covariate adjustment through the embedded Markov chain (Section 6.3.1) or the sojourn times (Section 6.3.2). We maintain the existing notation for the parameters in the model and define some additional quantities. Let $Z_1, \ldots, Z_r$ be covariates in which we are interested in, either as predictors of interest or as potential confounders of the association between a predictor and disease progression. Each Z_i may be binary or continuous. Since a categorical covariate with k levels may be expressed equivalently using $k-1$ binary dummy variables, the following development applies to categorical covariates as well. Suppose that measurements of each of these covariates are available for all N subjects, giving rise to observed data $\mathbf{z}^1, \ldots, \mathbf{z}^N$ where $\mathbf{z}^n = (z_1^1, \ldots, z_r^N)'$ for $n = 1, \ldots, N$.

6.3.1 Adjustment through embedded Markov chain

Recall that the embedded Markov chain for a general progressive process having m states is governed by the $m \times m$ transition probability matrix $\mathbf{P}$, given by

$$\mathbf{P} = \begin{bmatrix} 0 & p_{12} & p_{13} & \cdots & p_{1m} \\ 0 & 0 & p_{23} & \cdots & p_{2m} \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}$$

where $\sum_{j'=i+1}^{m} p_{ij'} = 1$ for $i = 1, \ldots, m-1$. Suppose we are interested in the impact of the covariates on the subjects' paths through the states. For each (i, j) with $i = 1, \ldots, m-1$, $j = i+1 \ldots, m$ we model the dependence of the transition probability p_{ij} on the covariates $\mathbf{Z} = (Z_1, \ldots, Z_r)'$ in the following way:

$$\text{logit}(p_{ij}|\mathbf{Z}, \boldsymbol{\gamma}_{ij}) = \log\left(\frac{p_{ij}|\mathbf{Z}, \boldsymbol{\gamma}_{ij}}{1 - p_{ij}|\mathbf{Z}, \boldsymbol{\gamma}_{ij}}\right) \quad = \quad \boldsymbol{\gamma}_{ij}'\mathbf{Z} \tag{6.1}$$

and

$$p_{im}|\mathbf{Z}, \boldsymbol{\gamma}_{ij} \quad = \quad 1 - \sum_{j'=i+1}^{m-1} p_{ij'}|\mathbf{Z}, \boldsymbol{\gamma}_{ij'}, \tag{6.2}$$

where $\boldsymbol{\gamma}_{ij} = (\gamma_{ij;0}, \ldots, \gamma_{ij;r})'$, with $\gamma_{ij;k} \in \mathbb{R}$ for each appropriate (i, j) and k, is the set of coefficients corresponding to the covariates $\mathbf{Z}$ and $\gamma_{ij;0}$ is the intercept term. Note that we have incorporated the restriction that $\sum_{j'=i+1}^{m} p_{ij'} = 1$ for each $i \in \{1, \ldots, m-1\}$, but that the estimate of this final probability will reflect the dependence on the covariates. Since each p_{ij} is expressed in terms of $\mathbf{Z}$ and $\boldsymbol{\gamma}_{ij}$ as $p_{ij} = \text{expit}(\boldsymbol{\gamma}_{ij}'\mathbf{Z})$, we estimate $\boldsymbol{\gamma}_{ij}$. We denote the set of all covariate coefficients by $\boldsymbol{\gamma} = \{\boldsymbol{\gamma}_{12}, \ldots, \boldsymbol{\gamma}_{1m}, \ldots, \boldsymbol{\gamma}_{m-1,m}\}$. The quantity $\exp(\gamma_{ij;k})$ can be interpreted as the multiplicative increase in the odds of making a transition from state i to j corresponding to a one-unit increase in the covariate Z_k, while all other covariates are held constant. Also, the quantity $\exp(\gamma_{ij;0})$ can be interpreted as the odds of making a transition from state i to j in the case where all covariates are equal to zero.

Our goal is to carry out inference about the parameters of interest: $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, as well as the means of the total lead times among subjects first observed in each state, $\boldsymbol{\mu}_T = (\mu_{T,1}, \ldots, \mu_{T,m-1})$. We use Markov Chain Monte Carlo methods to do this given the observed data on N subjects, $\mathbf{t}^1, \ldots, \mathbf{t}^N$ and $\mathbf{z}^1, \ldots, \mathbf{z}^N$, as well as our assumed standard deviations of the total lead times among subjects first observed in each state, $\boldsymbol{\sigma}_T = (\sigma_{T,1}, \ldots, \sigma_{T,m-1})$. The algorithm proceeds as in Chapter 5, except that instead of updating each p_{ij} directly, we update the regression coefficients, $\boldsymbol{\gamma}_{ij}$. Then, we use those to compute p_{ij} and update the remaining components of the model.

Specifically, the updating of the regression coefficients $\boldsymbol{\gamma}$, follows from the full conditional

distribution given by

$$p(\boldsymbol{\gamma}|\boldsymbol{\theta}, \mathbf{v}^1, \ldots, \mathbf{v}^N, \mathbf{X}^1(\mathbf{v}^1), \ldots, \mathbf{X}^N(\mathbf{v}^N), T^1, \ldots, T^N, \mathbf{t}^1, \ldots, \mathbf{t}^N, \boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^N, \mathbf{z}^1, \ldots, \mathbf{z}^N, \boldsymbol{\mu}_T, \boldsymbol{\sigma}_T)$$
$$\propto \left[\prod_{n=1}^{N} \prod_{i=1}^{m-1} \prod_{j=i+1}^{m} \operatorname{expit}(\boldsymbol{\gamma}_{ij}' \mathbf{z}^n)^{N_{ij}} \right] \cdot \pi(\boldsymbol{\gamma}),$$

where $N_{ij} \doteq \sum_{n=1}^{N} v_i^n \cdot \prod_{k=i+1}^{j-1} (1 - v_k^n) \cdot v_j^n$ is the latent number of subjects who make a transition directly from state i to j and $\pi(\boldsymbol{\gamma})$ is the prior distribution for the coefficients $\boldsymbol{\gamma}$. Since $\gamma_{ij,k} \in \mathbb{R}$ for each appropriate (i, j) and k, we can use a symmetric proposal distribution for $\gamma_{ij,k}$—a normal proposal, for example—and use a Metropolis step to update $\boldsymbol{\gamma}$.

The remaining full conditionals are omitted since they are similar to those developed under Chapter 5 where $\mathbf{p}$ is obtained from the regression equations (6.1) and (6.2) using the updated values of $\boldsymbol{\gamma}$ and the covariate values $\mathbf{z}$.

To summarize the estimation procedure, we update the coefficients $\boldsymbol{\gamma}$ via a Metropolis step and a symmetric proposal distribution, and use the values of the covariates to compute the transition probabilities $\mathbf{p}$. We update the sojourn time parameters $\boldsymbol{\theta}$ via Metropolis-Hastings with appropriate proposal distributions on each parameter. We update the each of the total lead times $T^1, \ldots, T^N$ via Metropolis-Hastings, then update each of the means of the total lead times among subjects first observed in each state via Metropolis with a symmetric proposal distribution. Using the Metropolized Carlin-Chib algorithm, we update each subject's path $\mathbf{v}$ and collection of latent sojourn times $\mathbf{X}$.

We present this approach here simply to discuss the possibility of capturing inter-subject differences via the embedded Markov chain. We do not explore it further in the simulation studies.

6.3.2 Adjustment through sojourn times

Alternatively, we may be interested in the impact that the covariates have on the sojourn times. We allow each covariate to have a different impact on each of the conditional sojourn times. Specifically, for each $\{(i, j) : \ i = 1, \ldots, m-1, \ j = i+1, \ldots, m\}$, we model the

sojourn time $X_{ij}(\mathbf{v})$ using the proportional hazards assumption:

$$h_{ij}(s;\mathbf{Z}) \quad = \quad \exp(\boldsymbol{\beta}_{ij}\mathbf{Z}) \cdot h_{ij,0}(s),$$

where $h_{ij}(\cdot;\mathbf{Z})$ is the hazard function corresponding to the sojourn time X_{ij}, given covariates $\mathbf{Z} = (Z^1, \ldots, Z^r)$, $h_{ij,0}(\cdot)$ is the baseline hazard function corresponding to X_{ij}, and $\boldsymbol{\beta}_{ij} = (\beta_{ij}^1, \ldots, \beta_{ij}^r)$ is the vector of coefficients that corresponds to $\mathbf{Z}$ where $\beta_{ij}^k \in \mathbb{R}$ for $k = 1, \ldots, r$. Using relationships from standard survival analysis we can express the density of sojourn time X_{ij} as

$$\begin{aligned} f_{ij}(s;\mathbf{Z}) &= h_{ij}(s;\mathbf{Z}) \cdot S_{ij}(s;\mathbf{Z}) \\ &= \exp(\boldsymbol{\beta}_{ij}\mathbf{Z}) \cdot h_{ij,0}(s) \cdot S_{ij,0}(s)^{\exp(\boldsymbol{\beta}_{ij}\mathbf{Z})}, \end{aligned}$$

where $S_{ij}(\cdot;\mathbf{Z})$ and $S_{ij,0}$ are, respectively, the covariate-specific and baseline survivor functions corresponding to X_{ij}. We denote the set of all covariate coefficients by $\boldsymbol{\beta} = \{\boldsymbol{\beta}_{12}, \ldots, \boldsymbol{\beta}_{1m}, \ldots, \boldsymbol{\beta}_{m-1,m}\}$. The quantity $\exp(\beta_{ij;k})$ can be interpreted as the multiplicative increase in the hazard of making the transition from state i to j at any time s corresponding to a one-unit increase in Z_k, while all other covariates are held constant, conditional on j being the next state to be visited. In this model, the parameters $\boldsymbol{\theta}_{ij}$ describe the baseline sojourn time X_{ij} in the case where all the covariates are equal to zero.

Our goal is to carry out inference about the parameters of interest—now $\mathbf{p}$, $\boldsymbol{\theta}$, and $\boldsymbol{\beta}$, as well as $\boldsymbol{\mu}_T$—given the observed data $\mathbf{t}^1, \ldots, \mathbf{t}^N$ and $\mathbf{z}^1, \ldots, \mathbf{z}^N$ as well as our assumed standard deviations of the total lead times among subjects first observed in each state, $\boldsymbol{\sigma}_T = (\sigma_{T,1}, \ldots, \sigma_{T,m-1})$.

The joint distribution of the parameters of interest and the latent trajectories for each subject conditional on the observed data is given by

$$p(\mathbf{p}, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{X}^1, \ldots, \mathbf{X}^N, T^1, \ldots, T^N, \boldsymbol{\mu}_T | \mathbf{t}^1, \ldots, \mathbf{t}^N, \boldsymbol{\delta}^1, \ldots, \boldsymbol{\delta}^N, \mathbf{z}^1, \ldots, \mathbf{z}^N, \phi(1), \ldots, \phi(N), \boldsymbol{\sigma}_T)$$

$$\propto \left[\prod_{n=1}^{N}\left[\prod_{k_2=0}^{1}\cdots\prod_{k_{m-1}=0}^{1}(p_{vk_2\cdots k_{m-1}^n}\cdot\sigma(\mathbf{X}(\mathbf{k}),\mathbf{k},\boldsymbol{\delta}^n;\mathbf{z}^n,T^n,\boldsymbol{\theta},\boldsymbol{\beta}))^{(\delta(2)^n)^{1-k_2}\cdots(\delta(m-1)^n)^{1-k_{m-1}}}\right]\right.$$
$$\left.\cdot\,\mathbf{1}(\mathbf{X}^n\in A^n)\cdot\frac{\exp\left(-\frac{(T^n-\mu_{T,\phi(n)})^2}{2\sigma^2_{T,\phi(n)}}\right)}{1-\Phi(-\frac{\mu_{T,\phi(n)}}{\sigma_{T,\phi(n)}})}\cdot\mathbf{1}(T^n\geq 0)\right]\cdot\pi(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\mu}_T),$$

where $\pi(\boldsymbol{\theta},\boldsymbol{\beta},\boldsymbol{\mu}_T)$ gives the prior distribution of the parameters in the model. Here, the likelihood contribution of a subject taking a specified path, $\sigma(\mathbf{X}(\mathbf{k}),\mathbf{k},\boldsymbol{\delta}^n;\mathbf{z}^n,T^n,\boldsymbol{\theta},\boldsymbol{\beta})$, depends on $\boldsymbol{\beta}$. All other notation is the same as before. Based on this distribution we derive the full conditional distributions of the parameters of interest.

The full conditional distribution of $\boldsymbol{\theta}$, the parameters governing the baseline sojourn times when the covariates are equal to zero, is given by

$$p(\boldsymbol{\theta}|\mathbf{p},\boldsymbol{\beta},\mathbf{v}^1,\ldots,\mathbf{v}^N,\mathbf{X}^1(\mathbf{v}^1),\ldots,\mathbf{X}^N(\mathbf{v}^N),T^1,\ldots,T^N,\mathbf{t}^1,\ldots,\mathbf{t}^N,\boldsymbol{\delta}^1,\ldots,\boldsymbol{\delta}^N,\mathbf{z}^1,\ldots,\mathbf{z}^N,\boldsymbol{\mu}_T,\boldsymbol{\sigma}_T)$$
$$\propto \prod_{n=1}^{N}\sigma(\mathbf{X}^n(\mathbf{v}^n),\mathbf{v}^n,\boldsymbol{\delta}^n;\boldsymbol{\theta},\mathbf{z}^n,\boldsymbol{\beta})\cdot\pi(\boldsymbol{\theta})$$
$$\propto \prod_{n=1}^{N}\prod_{i=1}^{m-1}\prod_{j=i+1}^{m}f_{ij}(X_{ij}^n(\mathbf{v}^n);\boldsymbol{\theta}_{ij},\mathbf{z}^n,\boldsymbol{\beta}_{ij})^{N_{ij}}\cdot\pi(\boldsymbol{\theta}),$$

where $\pi(\boldsymbol{\theta})$ is the prior distribution for $\boldsymbol{\theta}$ and $f_{ij}(X_{ij}^n(\mathbf{v}^n);\boldsymbol{\theta}_{ij},\mathbf{z}^n,\boldsymbol{\beta}_{ij})$ is the density of the path-specific sojourn time $X_{ij}^n(\mathbf{v}^n)$ for this subject. We use a Metropolis-Hastings step to update $\boldsymbol{\theta}$.

We derive the full conditional distribution of each of the vectors of coefficients $\boldsymbol{\beta}_{ij}$ for the covariates $Z_1,\ldots,Z_r$ using Bayes' rule in the third line:

$$p(\boldsymbol{\beta}_{ij}|\mathbf{p},\boldsymbol{\theta},\mathbf{v}^1,\ldots,\mathbf{v}^N,\mathbf{X}^1(\mathbf{v}^1),\ldots,\mathbf{X}^N(\mathbf{v}^N),T^1,\ldots,T^N,\mathbf{t}^1,\ldots,\mathbf{t}^N,\boldsymbol{\delta}^1,\ldots,\boldsymbol{\delta}^N,\mathbf{z}^1,\ldots,\mathbf{z}^N,\boldsymbol{\mu}_T,\boldsymbol{\sigma}_T)$$

$$
\begin{aligned}
&\propto p(\boldsymbol{\beta}_{ij}|\boldsymbol{\theta}, \mathbf{v}^1, \ldots, \mathbf{v}^N, X_{ij}^1(\mathbf{v}^1), \ldots, X_{ij}^N(\mathbf{v}^N), T^1, \ldots, T^N, \mathbf{z}^1, \ldots, \mathbf{z}^N) \\
&\propto \prod_{n=1}^{N} p(X_{ij}^n(\mathbf{v}^n)|\boldsymbol{\theta}_{ij}, \mathbf{z}^1, \ldots, \mathbf{z}^N) \cdot \pi(\boldsymbol{\beta}_{ij}) \\
&\propto \prod_{n=1}^{N} f_{ij}(X_{ij}^n(\mathbf{v}^n)|\boldsymbol{\theta}_{ij}, \mathbf{z}^1, \ldots, \mathbf{z}^N, \boldsymbol{\beta}_{ij}) \cdot \pi(\boldsymbol{\beta}_{ij}) \\
&\propto \prod_{n=1}^{N} \exp(\boldsymbol{\beta}_{ij}\mathbf{Z}) \cdot h_{ij,0}(X_{ij}(\mathbf{v}^n)|\boldsymbol{\theta}_{ij}) \cdot S_{ij,0}(X_{ij}(\mathbf{v}^n)|\boldsymbol{\theta}_{ij})^{\exp(\boldsymbol{\beta}_{ij}\mathbf{Z})} \cdot \pi(\boldsymbol{\beta}_{ij}),
\end{aligned}
$$

where $\pi(\boldsymbol{\beta})$ is the prior distribution for $\boldsymbol{\beta}$. The form of the baseline hazard and survivor functions depends on the model for the sojourn times. Since $\beta_{ij,k} \in \mathbb{R}$ for each appropriate (i, j) and k, we can use a symmetric proposal distribution for $\gamma_{ij,k}$, such as a normal distribution, and use a Metropolis step to update $\boldsymbol{\beta}$.

The remaining full conditionals are the same to those developed in Chapter 5 and thus are omitted here.

To summarize the estimation procedure, we update the transition probabilities $\mathbf{p}$ via Gibbs, and update the sojourn time parameters $\boldsymbol{\theta}$ via Metropolis-Hastings with appropriate proposal distributions on each parameter. Using a Metropolis step, we update each vector of sojourn time coefficients $\boldsymbol{\beta}_{ij}$ with a symmetric proposal distribution for each individual coefficent β_{ij}^k. We update the each of the total lead times $T^1, \ldots, T^N$ via Metropolis-Hastings, then update each of the means of the total lead times among subjects first observed in each state via Metropolis with a symmetric proposal distribution. Using the Metropolized Carlin-Chib algorithm, we update each subject's path $\mathbf{v}$ and collection of latent sojourn times $\mathbf{X}$.

6.4 Defining the model

We have developed a framework that allows us to estimate the impact of clinical or demographic factors on the rate of disease progression, in terms of either the trajectory through the stages of disease or the time spent in each stage of disease. It is straightforward to extend this framework to allow estimation of the impact of these factors on disease progression rate through both the trajectory through the stages and the time spent in each one. That

is, we can define the model such that one group of covariates act on the embedded Markov chain, another group acts on the sojourn times, and possibly a third group acts on both.

Although this framework gives us a great deal of flexibility in defining the model, it is wise to be parsimonious in our choice of covariates and in our assumptions about how they affect the disease progression rate since often there is sparse observation of each subject and including additional parameters increases uncertainty.

In the development of the model in which the covariates act on the sojourn times, we allowed the covariates to impact each of the conditional sojourn times differently. That is, we defined the parameters $\beta_{i,i+1}, \ldots, \beta_{i,m}$ separately. In many applications, it may not be necessary or desirable to allow these covariate effects to be different, as they all pertain to the sojourn time in state i. Instead we may define a single coefficient, β_i, that acts on each of the sojourn times in state i. If the sparsity of information available from the observed data is a major concern, then we may make the analogous assumptions for the parameters of the "baseline" sojourn times, $\boldsymbol{\theta}$.

6.5 Simulation studies

In this section we examine the performance of our approach to quantify differing rates of disease progression between groups of subjects. To illustrate the extension, we consider the illness-death model and a single covariate, and we assume that this covariate impacts the rate of disease progression via the sojourn times. We consider two cases: one in which the covariate is binary-valued, and one in which it is continuous-valued (refer to Table 6.1). We examine whether the proposed approach is able to carry out consistent inference about these coefficients as well as the other parameters of interest.

Table 6.1: Scenarios under examination.

Scenario	model for covariate	parameter values
1	Bernoulli(p)	$p = 0.5$
2	Normal(μ, σ)	$(\mu, \sigma) = (0, 1)$

In both of these simulations we assume that each subject visits the intermediate state 2

with 90% probability ($p_{12} = 0.9$). We assume that each of the baseline sojourn times follows a Weibull distribution with common shape and scale parameters: $(X_{ij}|Z = 0) \sim Weibull(k_{ij}, \theta_{ij})$ with $k_{ij} = 2.0$ and $\theta_{ij} = 4.0$ for each $(i, j) \in \{(1, 2), (2, 3), (1, 3)\}$. We assume that subjects are observed frequently: every three months ($\Delta = 0.25$). We follow each subject until he enters the absorbing state. We assume informative normal priors for the sojourn time parameters and a noninformative uniform prior for the transition probability p_{12}. We use the Case 2 assumption from the previous chapter, and make no assumption about a relationship between the means of the total lead times among subjects with different observed initial states. That is, $\mu_{T,1}$ and $\mu_{T,2}$ are estimated separately. The true total lead times are generated from a truncated normal distribution with mean 2.0 years and standard deviation 0.1 years. For each of the mean total lead times we impose an informative normal prior centered at the true value to ensure convergence of the chain.

In the first simulation study, we consider a binary covariate with 50% prevalence, and we assume that the presence of this covariate is associated with a two-fold multiplicative increase in the hazard of making a transition to the next state. That is, the true value of β_{ij} is $\log(2.0)$ for $(i, j) \in \{(1, 2), (2, 3), (1, 3)\}$, so $\exp(\beta_{ij}) = 2.0$ gives the multiplicative increase in the conditional hazard of making a transition from state i to j for any time $s \geq 0$. In the second simulation study, we consider a continuous-valued covariate that has a standard normal distribution. We assume that a one-unit increase in the value of this covariate is associated with a 1.2-fold multiplicative increase in the hazard of making a transition to the next state. That is, the true value of each β_{ij} is $\log(1.2)$. In each of the two simulation studies, we assume a fairly diffuse normal prior on each of the coefficients centered at the null value: $\pi(\beta_{ij}) \sim \mathcal{N}(0, 2)$.

We illustrate the performance of the approach for $N = 400$ subjects. The results, based on $M = 100$ simulated datasets, are shown in Tables 6.2–6.3. Posterior means as well as model-based and empirical standard deviations are given for each of the parameters of interest: those governing the embedded Markov chain and baseline sojourn times, and the coefficients of the covariates. For the parameters of interest, the true values given in the tables are corrected for the impact of truncation, as we discussed in the previous chapter.

The proposed approach performed well in carrying out inference about both the pa-

rameters governing the embedded Markov chain and sojourn times, and the regression coefficients. Both p_{12} and the sojourn time parameters were consistently estimated when the effect of truncation was taken into account. As we would expect, the precision of the inference about parameters and coefficients corresponding to X_{13} is diminished due to the small probability (10%) that a subject makes a transition directly from state 1 to 3. The approach performed well for the coefficients corresponding to the other sojourn times, X_{12} and X_{23}, since the observed data contained ample information about these transitions.

6.6 Discussion

We have presented an extension to the existing proposed approach that allows us to quantify the impact of clinical or demographic characteristics of the patients on the rate of disease progression, through either the sequence of disease stages visited, the time spent in each stage of disease, or both. Although the framework we presented gives us flexibility in building a model and including covariates for adjustment, a parsimonious model is advisable: the applications for which the proposed approach was built offer, by their nature, limited information about the underlying process. The scientific questions of interest should drive the choice of a model, and parameters or coefficients that may be assumed to be common should be defined as such, as described in Section 6.4.

Despite the flexibility of the framework presented here for allowing disease progression rate to differ depending on patient factors, we have made certain assumptions about how the covariates may affect disease progression. In Section 6.4, for example, we have made the proportional hazards assumption. We believe this assumption is reasonable in this context, and the fact that the panel observation scheme limits the amount of information available from the observed data forces us to make some assumptions about the underlying process. However, if there is reason to believe that the proportional hazards assumption is not appropriate in a particular application, then an alternative formulation may be needed.

Table 6.2: Results for Scenario 1: binary covariate.

(i,j)	truth	$\hat{k}_{ij}$	$SD_{model}(k_{ij})$	$SD_{emp}(k_{ij})$
$(1,2)$	2.035	2.136	0.101	0.109
$(2,3)$	2.040	2.075	0.086	0.078
$(1,3)$	2.790	2.689	0.290	0.206

(i,j)	truth	$\hat{\theta}_{ij}$	$SD_{model}(\theta_{ij})$	$SD_{emp}(\theta_{ij})$
$(1,2)$	4.029	4.010	0.134	0.156
$(2,3)$	4.040	4.033	0.140	0.151
$(1,3)$	4.725	4.284	0.297	0.292

(i,j)	truth	$\hat{\beta}_{ij}$	$SD_{model}(\beta_{ij})$	$SD_{emp}(\beta_{ij})$
$(1,2)$	0.693	0.676	0.105	0.128
$(2,3)$	0.693	0.699	0.106	0.113
$(1,3)$	0.693	0.547	0.345	0.357

(i,j)	truth	$\hat{p}_{ij}$	$SD_{model}(p_{ij})$	$SD_{emp}(p_{ij})$
$(1,2)$	0.920	0.925	0.013	0.011

Table 6.3: Results for Scenario 2: continuous covariate.

(i,j)	truth	$\hat{k}_{ij}$	$SD_{model}(k_{ij})$	$SD_{emp}(k_{ij})$
$(1,2)$	2.035	2.075	0.092	0.084
$(2,3)$	2.040	2.052	0.086	0.089
$(1,3)$	2.790	2.532	0.269	0.222

(i,j)	truth	$\hat{\theta}_{ij}$	$SD_{model}(\theta_{ij})$	$SD_{emp}(\theta_{ij})$
$(1,2)$	4.029	4.009	0.107	0.112
$(2,3)$	4.040	4.027	0.108	0.095
$(1,3)$	4.725	4.451	0.274	0.234

(i,j)	truth	$\hat{\beta}_{ij}$	$SD_{model}(\beta_{ij})$	$SD_{emp}(\beta_{ij})$
$(1,2)$	0.182	0.186	0.053	0.057
$(2,3)$	0.182	0.194	0.052	0.054
$(1,3)$	0.182	0.139	0.193	0.204

(i,j)	truth	$\hat{p}_{ij}$	$SD_{model}(p_{ij})$	$SD_{emp}(p_{ij})$
$(1,2)$	0.920	0.918	0.014	0.012

Chapter 7

APPLICATION

7.1 Introduction

In this chapter we apply our proposed approach to the World Health Organization (WHO) staging data of untreated HIV-infected subjects in Senegal.

Background information on this dataset is in Section 2.4.2 of Chapter 2, and further details about the studies are available in Gottlieb et al. (2002). We provide a brief overview here. Specifically, the dataset consists of baseline information and longitudinal measurements of 611 patients who presented at clinics in Senegal between 1994 and 2005. Baseline information includes HIV serostatus (HIV-1, HIV-2, or dual infection) and information about demographics, education, behavior, and sexual history. For the baseline visit and each follow-up visit, the WHO stage of HIV/AIDS and the CD4+ T cell count are available. The WHO staging system of HIV/AIDS consists of primary HIV infection (stage 0) and stages 1–4 that are each defined by observed symptoms, as described in Section 2.4.2. The stages are progressive by definition, since once a patient has exhibited a stage-2–defining symptom, for example, he remains in stage 2 even if the symptom disappears.

Existing studies of the natural history of infection with HIV-1, HIV-2, and both viruses have focused on overall survival rather than time spent in each stage of disease. In the absence of antiretroviral therapy (ARV), patients infected with HIV-2 have been reported to have longer, but perhaps more variable, survival than those infected with HIV-1 (Jaffar et al., 2004). However, much less is known about the natural history of patients with HIV-2, since its incidence is quite low (Jaffar et al., 2004). Moreover, there are conflicting reports on the natural history of dual infection: some suggest that survival is similar to infection with HIV-1 (Whittle et al., 1992) while others suggest longer survival than infection with just HIV-1 (Esbjornsson et al., 2012). Our goal in this application is to characterize the dependence of the rate of disease progression on viral type in the absence of antiretroviral

therapy.

We carry out descriptive analyses of the data in Section 7.2. In Section 7.3 we discuss our modeling choices, and in Section 7.4 we present the results of our analysis. Finally in Section 7.5 we discuss the results as well as future analyses for this dataset.

7.2 Descriptive analysis

7.2.1 Data Pre-Processing

Given our goal of analysis, we excluded data from the clinic visits at which the patient was on ARV, thus eliminating 472 clinic visits from 106 patients. Only one of the 106 patients was on ARV from the baseline visit.

Further exploration of the dataset revealed that some patients had reversals in WHO staging over time, which should not be possible since the WHO stages are progressive by definition. We relabeled the staging data under the assumption that any observed reversals resulted from the disappearance of symptoms rather than misdiagnosis. Moreover, since the dataset included just one visit at which the patient was categorized as stage 0, we combined disease stages 0 and 1 into a single stage labeled "1".

Patients who did not contribute useful longitudinal information in the context of our state model, which we will discuss in Section 7.3, were also excluded. After 67 such patients were excluded, the remaining dataset consists of 543 patients who contributed information from 3365 clinic visits.

7.2.2 Results

Of the 543 patients in the analysis dataset, 356 (65.6%) tested positive for HIV-1 only at baseline, 136 (25.0%) for HIV-2 only, and the remaining 51 (9.4%) tested positive for both viruses. 475 patients (87.5%) were female, 62 (11.4%) were male, and 6 (1.1%) unknown. 129 patients (24.4% of known females and 21.0% of known males) were commercial sex workers (CSW). The median age at the baseline visit was 35.0 years (interquartile range: 29.0–41.0 years). Other baseline information on the patients included in our analysis dataset is presented in Table 7.1.

The visits included in the analysis dataset represent a median follow-up time of 1.83 years (IQR, 0.73–3.73 years) and 4 visits per patient (IQR, 2–8 visits). Of the 543 patients, 209 (38.5%) were in stage 1 of HIV/AIDS at the baseline visit, 193 (35.5%) in stage 2, and 141 (26.0%) in stage 3. Patients in the original dataset who were first observed in disease stage 4 are not included in the analysis dataset since they do not contribute useful longitudinal information. Five patients (0.9%) were observed in all four stages, 46 (8.5%) were observed in three stages, 235 (43.3%) in two stages, and 257 (47.3%) in one stage only (refer to Table 7.2). A summary of the observed length of time from the last observation in one state to the first observation in another state is given in Table 7.3.

Table 7.2: Distribution of observed disease stages. Note that subjects who were observed in stage 1 or stage 4 were not included in the analysis dataset, since these observations do not contribute useful longitudinal information to our model.

Number of observed states	Observed states	Number of subjects	Total number of subjects
1	2	129	257
	3	128	
2	1–2	113	235
	1–3	50	
	1–4	1	
	2–3	48	
	2–4	10	
	3–4	13	
3	1–2–3	36	46
	1–2–4	2	
	1–3–4	2	
	2–3–4	6	
4	1–2–3–4	5	5

Table 7.3: Summary of lengths of time (years) between the last observation in state i and first observation in state j, among subjects observed in states i and j and in no intermediate states (the number of such subjects is given in the second column).

Transition	Number of	Quantile		
i–j	patients	25%	50%	75%
1–2	156	0.342	0.382	0.597
2–3	95	0.331	0.381	0.682
3–4	26	0.326	0.342	0.441
1–3	52	0.345	0.478	0.604
2–4	12	0.327	0.360	0.401
1–4	1	0.441	0.441	0.441

7.3 Modeling considerations

The broad objective for this application is to characterize how the rate of disease progression depends on viral type. Since the outcome in this application is a categorical random variable, a multi-state model is a natural choice. The disease stages are progressive by nature, as we have discussed. Patients are observed at discrete timepoints and may be in any disease stage at the first visit. In spite of the sparsity of information contained in the observed data, a flexible model for the sojourn times may still be desirable. For example, a patient in stage 2 may have a greater hazard of making a transition to a more advanced stage if he has been in that stage for a long time. That is, the hazard of making a transition to another state may vary over time, and we may want to capture this variation in our model. Thus, the methods developed in this dissertation may be suitable for this application using a general progressive state model with four states. There is one caveat, however. Our approach makes the assumption that the mean total lead time before being observed in each state is relatively homogeneous, which may not be reasonable in this application. That is, among patients first observed in a given WHO stage, the time from entry into WHO stage 1 and the time of first observation may be quite variable. To mitigate this assumption, we consider a reduced general progressive model with WHO stages 2, 3, and 4. Although we continue to impose

the assumption of relative homogeneity of the lead times, the impact of this assumption is reduced, since the observed data are now contributing more information to the inference about the mean total lead time before being observed in a given state. Specifically, for the subjects who are observed in WHO stage 1, the time spent in the next observed stage before being observed now has a known upper bound. Another consequence of our reduced model is that we assume all patients visit state 2. That is, we assume that each patient visited WHO stage 2 even if he was not observed in this stage. Refer to Figure 7.1 for a graphical depiction of the way in which the observed data contribute to inference about these parameters.

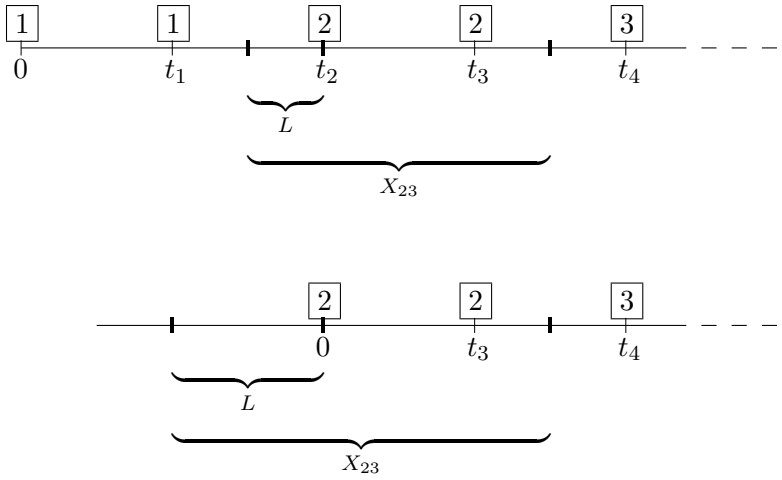


Figure 7.1: Inference about the latent state lead time L for a patient who was observed in state 1 (upper panel) and a patient not observed in state 1 (lower panel). In the first scenario, L is bounded above by $t_2 - t_1$. In the second scenario, L is unbounded above.

As we discussed in Section 5.7, we can consider a variation of our proposed approach in which we model the state lead times rather than the total lead times. Recall that for subjects who are first observed in an advanced state, the total lead time depends on latent sojourn times as well as the state lead time. Thus, modeling of state lead times does not carry the added uncertainty from sojourn times as modeling the total lead times does. In this application, we model the state lead times. We assume that the state lead times, instead

of the total lead times, are relatively homogeneous. Similarly to our choice for the total lead times, we use a truncated normal model for the state lead times, where the state lead time, defined originally in Section 5.3, is the time since entering the first observed state until this first observation occurs. We carry out inference about the mean state lead time in the i^{th} state, $\mu_{L,i}$, for $i = 2, 3$.

We have tailored the approach proposed in this dissertation in several ways to accommodate features of the dataset at hand including the sparsity of information therein and the appropriateness of assumptions for this dataset. Specifically, we model the state lead times, and include just three of the four states in the state model. A natural question to pose is whether the proposed approach continues to perform well under these modifications. We demonstrate the performance of the tailored approach via simulation study in Appendix D. Previously we noted that certain subjects were not included in the analysis dataset since they do not contribute useful longitudinal information to our model. Specifically, a subject who is observed only in state 1 contributes information about the sojourn time in state 1, but no information about the sojourn times in other states or about the trajectory through the states; since state 1 is not in the state model that we are considering in the analysis, no useful information is contributed. Similarly, a subject who is observed only in state 4 contributes information about the sojourn time in state 4, but no information about the other sojourn times or about the trajectory, and since state 4 is absorbing in our state model for the situation, this information is not useful in the model. Since the lead time parameters for each state are estimated separately in the model and no information is borrowed, the information these subjects provide about the lead times in states 1 and 4 would not be used in the model. Hence, these subjects are not included in the analysis dataset.

Finally, we note that the association between rate of disease progression, as measured by the time spent in each stage of disease, and viral type is potentially distorted by a number of patient characteristics requiring adjustment for potential confounders. In the presence of sparse information about patients' sojourn times, such as in this application, more informative priors are needed when the dimensionality grows with additional regression coefficients. Given this limitation and the absence of expert or literature information to provide justifiable prior choices, in this chapter we do not provide adjusted analysis.

Given the state model shown in Figure 7.2, we carry out inference about the sojourn time in states 2 and 3 before making a transition to the next state—X_{23}, X_{34}, and X_{24}—as well as the probability of visiting state 3. We additionally carry out inference about the mean state lead times among subjects first observed in states 2 and 3, $\mu_{L,2}$ and $\mu_{L,3}$ respectively.

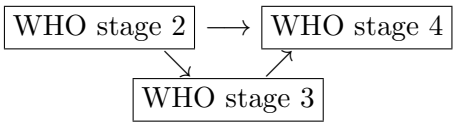


Figure 7.2: State model for data on WHO stages of HIV/AIDS.

For each of the sojourn times in the model, we want to allow for the hazard function to change over time. However, the observed data contain limited information about each subject's trajectory and sojourn time in each visited state. Thus, for our first model, we choose a Weibull model for each of the sojourn times X_{23}, X_{34}, and X_{24} in an effort to capture the potential non-constant hazard of making a transition to the next state while respecting the limitations of the data.

For simplicity we choose to model the impact of viral type on the rate of disease progression through the sojourn times only. That is, in our model the probability of visiting disease stage 3 is the same for all patients, and differences in the rate of disease progression will be captured through the sojourn times. Given the sparsity of information contained in the observed data, we assume that the impact of a covariate—in this case, viral type—on the sojourn time is the same regardless of the next state to be visited. In the notation of Chapter 6, $\boldsymbol{\beta}_{23}$ and $\boldsymbol{\beta}_{24}$ denote, respectively, the impact of covariates $\mathbf{Z}$ on the sojourn times X_{23} and X_{24}. We assume here that these coefficients are equal to a common value: $\boldsymbol{\beta}_{23} = \boldsymbol{\beta}_{24} \doteq \boldsymbol{\beta}_2$. Correspondingly we refer to the coefficients associated with X_{34} as $\boldsymbol{\beta}_{34} \doteq \boldsymbol{\beta}_3$.

We let $\mathbf{Z} = (Z^1, Z^2)$ denote the vector of covariates for a single subject, where Z^1 indicates that the subject is infected with HIV-1 only at baseline, Z^2 indicates that he

is dually infected. In our model, the hazard of making a transition from state i to j for $(i,j) \in \{(2,3),(3,4),(2,4)\}$ at time $s \geq 0$ for a subject with covariates $\mathbf{Z}$ is given by

$$h_{ij}(s;\mathbf{Z}) \quad = \quad \exp(\boldsymbol{\beta}_i'\mathbf{Z}) \cdot h_{ij,0}(s),$$

where $h_{ij,0}(\cdot)$ is the baseline hazard function corresponding to X_{ij}. For $i = 2,3$, the vector of coefficients $\boldsymbol{\beta}_i = (\beta_i^1, \beta_i^2)$ gives the additive effect of the covariates $\mathbf{Z}$ on the log of the hazard function $h_{ij}(\cdot)$ for *any* j. That is, $\exp(\beta_i^1)$ represents the hazard ratio of making a transition from state i to another state j among patients infected with HIV-1 only relative to those infected with HIV-2 only at baseline. The interpretation of $\exp(\beta_i^2)$ is analogous.

Estimates of the model parameters can be used to obtain estimates of any derived quantities that may be of interest. For example, it may be important to estimate the mean length of time spent in stage 2, or the probability of being in state 3 five years after being infected, among patients infected with HIV-1 versus HIV-2. We can obtain point estimates and measures of uncertainty for such quantities using the results of our analysis.

Finally, we note the relationship of our model to a standard Markov model: if for every pair (i,j), the sojourn time X_{ij} were constrained to be $X_i \sim \exp(\theta_i)$ for some $\theta_i > 0$ regardless of j, then our model would be Markov. What distinguishes our approach is that we may choose a flexible model for the sojourn time in each state that can differ depending on the next state to be visited, but much more importantly, can be non-exponential.

7.4 Results

In this section we present the results of applying our proposed approach to the WHO HIV/AIDS staging data. We use the state model shown in Figure 7.2 and use our approach to carry out inference about the model parameters. Interpretations of each of the coefficients are given in Table 7.4. Results for Model 1, in which we use a Weibull model for each of the sojourn times, are presented in Table 7.5.

We used a noninformative uniform prior for each of the parameters governing the sojourn times in Model 1. Specifically, for the transition probability p_{23} we used a $Unif(0,1)$ prior. For each of the positive-valued sojourn time parameters k_{ij} and θ_{ij} we used a $Unif(0,10)$

prior; the upper limit was selected as a value that would not restrict the exploration of the parameter space. For each coefficient of the indicators of viral type, we used a normal prior with standard deviation 0.5 units centered at the null value, 0.0. For the state lead time distributions, we chose a truncated normal model and carried out inference about the means of the original normal random variables, $\mu_{L,2}$ and $\mu_{L,3}$. We set the corresponding standard deviations to be 0.2. Informative normal priors, centered at 0.0 years and with standard deviation 0.5 years, were used for the mean state lead times. These values were chosen based on examination of the observed data (refer to Table 7.3) and acknowledgment that this summary represents only patients who are observed in both of these states. We note that the parameters $\mu_{L,2}$ and $\mu_{L,3}$ may be any real number, since each of these parameters is the mean of the original normal random variable. The distribution of the state lead times is truncated normal, and the mean—which may be derived from these parameters—is necessarily positive.

The results from Model 1 are shown in Table 7.5 and suggest that, relative to HIV-2, HIV-1 is each associated with a shorter duration in states 2 and 3. The results suggest that dual infection is also associated with a shorter duration in states 2 and 3 relative to HIV-2, but there is more uncertainty associated with these estimates. The increased uncertainty is most likely due to the small number of dually-infected patients present in the dataset. The results suggested that subjects visit state 3 with very high probability (96.8%). The estimated mean lead time parameters for states 2 and 3 suggest that patients who are first observed in disease stages 2 and 3 spend a mean of 3.26 and 0.85 months in these stages before being observed for the first time. Based on sensitivity analyses in which the priors for the mean state lead times were each centered at 0.5 years rather than 0.0 years, we conclude that the results for the parameters of interest were not sensitive to this choice.

There was evidence of an increasing hazard of making a transition to state 4 among subjects in state 3, as the estimate of the shape parameter k_{34} was greater than one. However, based on Model 1, there was no indication that the time spent in state 2 had a distribution that was different from exponential. Further, since the estimated probability of skipping state 3 is quite low, inference about the parameters corresponding to X_{24} is based on few latent transitions. We may be able to learn more about the time patients

spend in state 2 by assuming that the distribution of the sojourn time in state 2 does not depend on the next state to be visited, that is, that X_{23} and X_{24} are identical; we refer to this common sojourn time as X_2. Hence, we consider a second model, Model 2, in which we model the sojourn time in state 2 as exponential with scale parameter θ_2, and the sojourn time in state 3 as Weibull, as before, with shape and scale parameters k_{34} and θ_{34}. We continue to estimate p_{23} and $\mu_{L,3}$ as before, but as a consequence of our model for X_2, we note that $\mu_{L,2}$ is no longer needed given the memoryless property under the exponential model for state 2. We used the same priors for the parameters as we did for the first model. We regard Model 2 as primary since it is a refinement of Model 1.

Results for Model 2 are presented in Table 7.6. The results for the coefficients corroborate what we found in Model 1: relative to HIV-2, HIV-1 and dual infection are associated with a shorter duration in stages 2 and 3. Specifically, we estimate that the hazard of progressing to a more severe disease stage from stage 2 is $\exp(0.260) - 1 = 29.7\%$ greater for patients infected with HIV-1 compared to those infected with HIV-2, and $\exp(0.210) - 1 = 23.4\%$ greater for dually-infected patients compared to those infected with HIV-2 only. Similarly, we estimate that the hazard of progressing to disease stage 4 from stage 3 is $\exp(0.202) - 1 = 22.4\%$ greater for patients with HIV-1, and $\exp(0.385) - 1 = 47.0\%$ greater for dually-infected patients, compared to patients infected with HIV-2.

Similarly to Model 1, the results from Model 2 suggest that patients visit disease stage 3 with high probability (97.2%). The estimate of the baseline scale parameter for the sojourn time in stage 2 is similar to the estimate of θ_{23} that we obtained from Model 1. This makes sense since the estimated shape parameter k_{23} from Model 1 was close to one, and in the present model it is assumed to be one. Inference regarding the sojourn time in stage 3 was similar to what we obtained in Model 1. Results for the parameters of interest were insensitive to the choice of prior for the mean state lead times.

We carried out convergence diagnostics for the primary model. Specifically, we carried out the stationary test and the halfwidth test of Heidelberger and Welch for the parameters of interest. Each of the parameters passed the stationarity test and the halfwidth test with precision 0.1 (see Table 7.7).

Results from Models 1 and 2 are presented graphically in Figures 7.3 and 7.4 respectively.

Specifically, Figure 7.3 has three panels, which correspond to the estimated survival curves for the three sojourn times under consideration. Each panel summarizes the posterior distribution of the survival curve corresponding to the three viral types, where for each viral type the median is given as a thick line and the endpoints of the 95% credible interval are given as thin lines. The figures portray the result that HIV-2 corresponds to slower disease progression than the other two viral types.

Table 7.4: Coefficient interpretations.

Coefficient	Interpretation
β_2^1	log HR corresponding to X_{2j} for HIV-1 v. HIV-2
β_2^2	log HR corresponding to X_{2j} for dual infection v. HIV-2
β_3^1	log HR corresponding to X_{34} for HIV-1 v. HIV-2
β_3^2	log HR corresponding to X_{34} for dual infection v. HIV-2

Table 7.1: Baseline characteristics of patients in analysis dataset, stratified by viral type. For each characteristic, statistics are given among those with non-missing values.

Baseline characteristic	Viral type		
	HIV-1 ($n = 356$)	HIV-2 ($n = 136$)	Dual infection ($n = 51$)
Demographic characteristics			
Male sex, %	11.9	9.7	14.3
Age (years)			
Median	33.0	37.0	35.0
25^{th} percentile	27.0	32.0	32.0
75^{th} percentile	39.2	44.0	44.8
Education			
None	51.8	65.6	64.6
Primary	29.5	25.8	27.1
Secondary	17.3	8.6	6.2
University	1.5	0.0	2.1
Marital status			
Single	11.8	7.7	14.6
Married, monogamous	28.8	22.3	16.7
Married, polygamous	11.2	15.4	8.3
Divorced	15.6	19.2	18.8
Separated	14.7	19.2	27.1
Widowed	17.3	16.2	14.6
Concubine	0.6	0.0	0.0
Commercial sex worker, %	21.1	26.5	35.3
Behavioral characteristics			
Contraception			
None	67.9	54.7	54.8
Condoms	17.5	30.8	35.7
Pill	6.8	6.8	4.8
Injection	4.9	2.6	0.0
Douching	0.0	0.0	0.0
Traditional	0.6	0.0	0.0
Other	2.3	5.1	4.8
Age at first sex (years)			
Median	17.0	16.0	17.0
25^{th} percentile	15.0	14.0	14.0
75^{th} percentile	20.0	18.0	20.0
Number of sex partners			
1	30.7	24.6	17.8
2–5	30.1	29.4	24.4
6–10	5.4	2.4	0.0
>10	33.9	43.7	57.8
Alcohol use, %	13.6	21.5	22.4
Smoking, %	22.8	30.8	30.0
Clinical characteristics			
CD4+ T cell count			
Median	395.0	557.5	351.0
25^{th} percentile	209.0	348.0	220.0
75^{th} percentile	577.5	793.2	538.0

Table 7.5: Results for Model 1: Weibull model for each of the sojourn times.

Parameter	Mean	Posterior SD
k_{23}	0.976	0.050
k_{34}	1.405	0.136
k_{24}	1.043	0.251
θ_{23}	4.384	0.497
θ_{34}	6.931	0.925
θ_{24}	3.649	1.856
p_{23}	0.968	0.014
$\mu_{L,2}$	0.222	0.031
$\mu_{L,3}$	-0.438	0.273
β_2^1	0.260	0.138
β_2^2	0.210	0.213
β_3^1	0.202	0.198
β_3^2	0.385	0.310

Table 7.6: Results for Model 2: exponential model for sojourn time in state 2, Weibull model for sojourn time in state 3.

Parameter	Mean	Posterior SD
θ_2	4.633	0.593
k_{34}	1.320	0.105
θ_{34}	7.487	0.842
p_{23}	0.972	0.013
$\mu_{L,3}$	-0.610	0.356
β_2^1	0.285	0.138
β_2^2	0.129	0.274
β_3^1	0.088	0.213
β_3^2	0.351	0.306

Table 7.7: Convergence diagnostic for Model 2: stationarity and interval halfwidth tests of Heidelberger and Welch.

	Stationarity		Halfwidth		
Parameter	Test	p-value	Test	Mean	Halfwidth
θ_2	passed	0.186	passed	4.415	0.08112
k_{34}	passed	0.421	passed	1.278	0.01327
θ_{34}	passed	0.337	passed	7.294	0.20291
p_{23}	passed	0.085	passed	0.975	0.00196

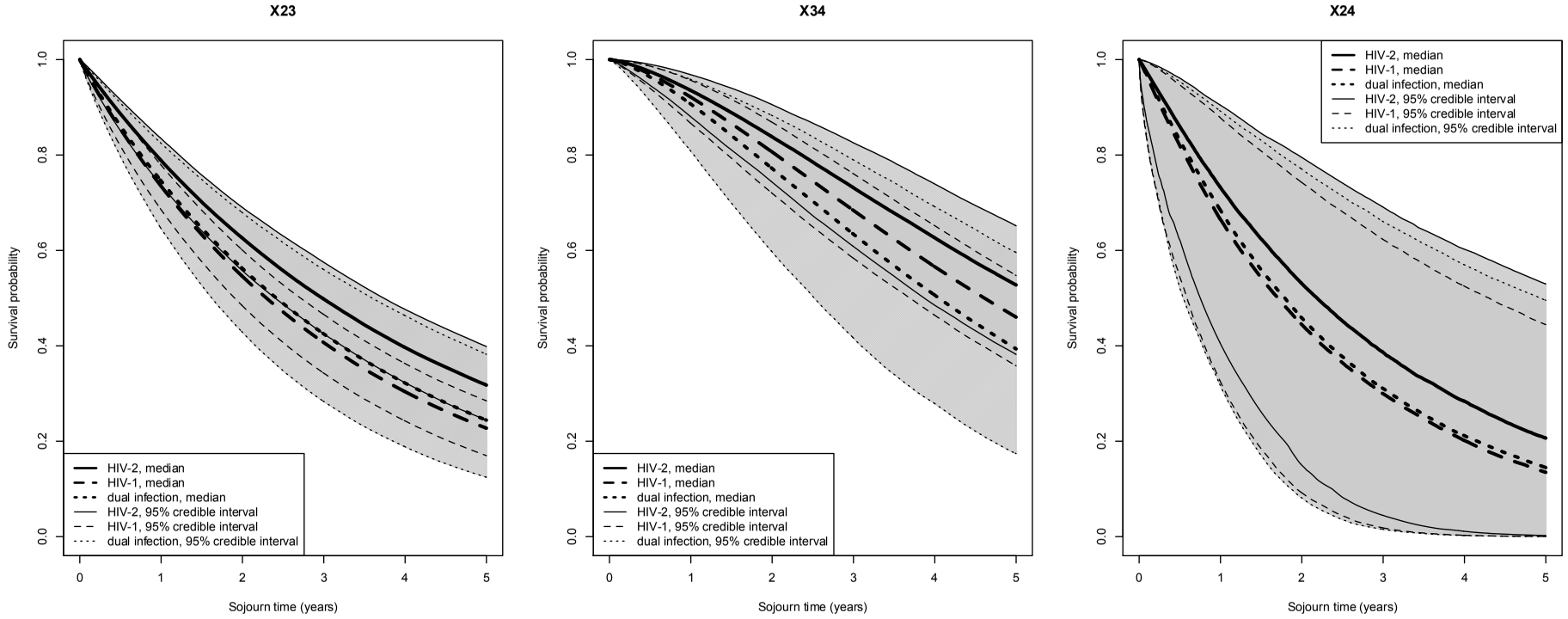


Figure 7.3: Model 1: posterior distribution of the survival probability in each state for each viral type. The left (right) panel shows a summary of the distribution of the survival probability of being in disease stage 2 at each point in time since entering this stage, among patients who go on to stage 3 (stage 4). The middle panel shows a summary of the distribution of the survival probability of being in disease stage 3 at each point in time since entering this stage. Each panel shows estimates for the three viral types, where for each time point, the posterior median of the survival probability is shown as a thick line and the 95% credible interval is shaded in grey.

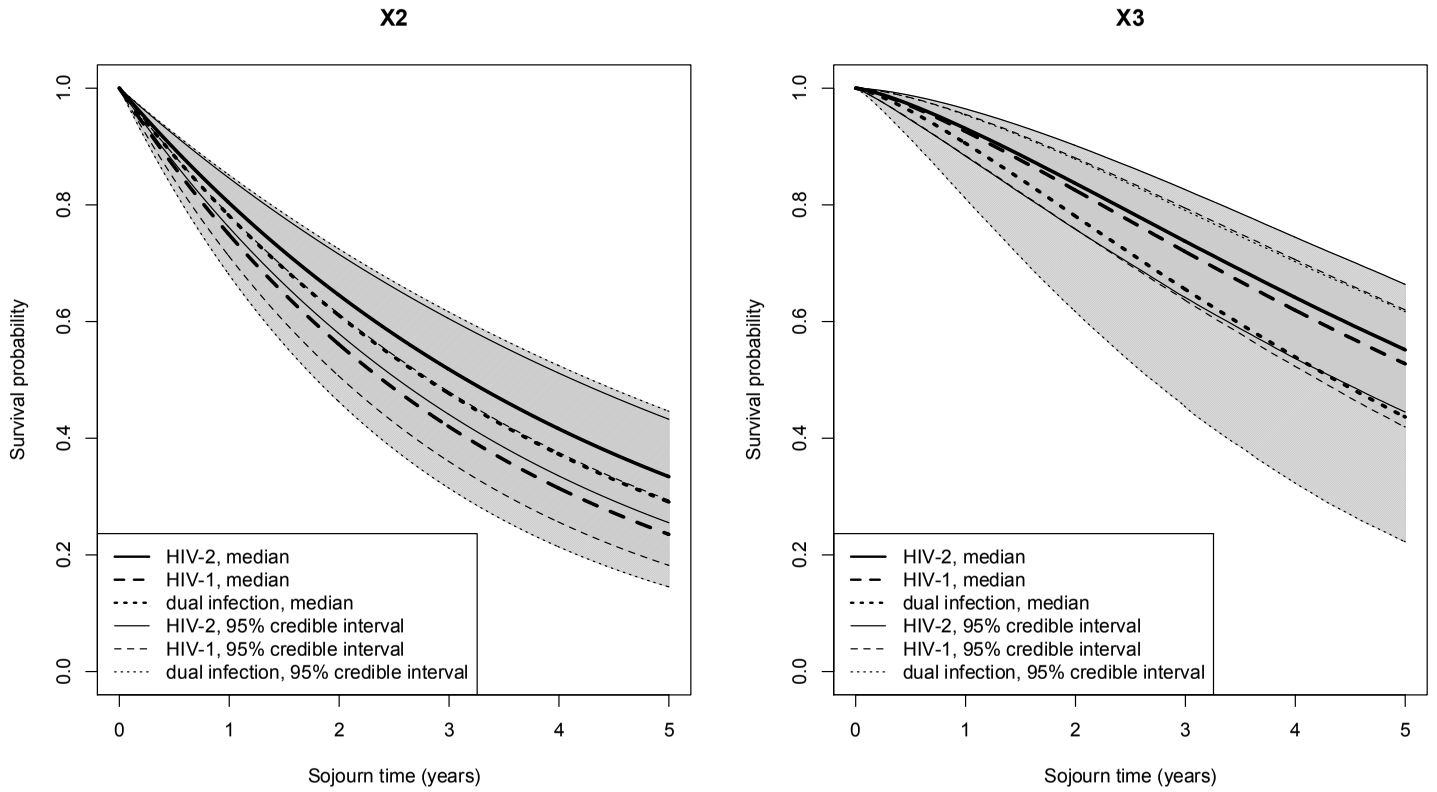


Figure 7.4: Model 2: posterior distribution of the survival probability in each state for a given viral type. The left (right) panel shows a summary of the distribution of the survival probability of being in stage 2 (stage 3) at each point in time since entering this stage. Each panel shows estimates for the three viral types, where for each time point, the posterior median of the survival probability is shown as a thick line and the 95% credible interval is shaded in grey.

7.5 Discussion

We have applied our proposed approach to longitudinal panel observations of HIV/AIDS stage on a set of untreated HIV-infected patients in Senegal. Our aim was to characterize the dependence of disease progression rate on the type of virus (HIV-1, HIV-2, or both viruses). The results of our analyses suggest that, relative to HIV-2, HIV-1 and dual infection are associated with shorter durations in disease stages 2 and 3. That is, HIV-1 and dual infection are associated with faster progression through the stages of HIV/AIDS, a finding that is consistent with existing knowledge of the natural history of infection with HIV-1 and HIV-2. The results of our analyses also suggest that it is rare for patients to progress directly from disease stage 2 to AIDS.

In our initial model we used a flexible model for each of the sojourn times, and the results indicated that less flexible model would have sufficed for one of these sojourn times. The results also indicated that the probability of skipping the intermediate disease stage was small. In a refinement of this initial model we addressed these findings. Our findings from the resulting primary model and the initial model had good correspondence: relative to HIV-2, HIV-1 and dual infection are associated with faster progression through the stages of disease. The results from both models suggested that HIV-1 infection was associated with greater hazard of progression from stage 2 than dual infection was, and that dual infection was associated with greater hazard of progression from stage 3 than HIV-1 alone. However, since the results are based on a small number of dually-infected patients, and on intermittent observations on all patients, we stress that these findings should not be over-interpreted.

Our application has some limitations. First, since the assumption of homogeneous total lead times may not be met for patients first observed in state 1, we considered a reduced state model. Another potential approach to handle left censoring would be to treat left-censored observations similarly to the way in which we treat right-censored observations. We will discuss this approach in the following chapter. Second, given the data sparsity and the need for informative priors for the estimation of more complex models, we did not perform covariate adjustment for potential confounders.

Examination of the observed data revealed an interesting phenomenon about subjects

who were observed in states 1 and 3 only. The times between the last observation in state 1 and the first in state 3 were quite short in general. By an assumption of our model, each of these patients visited state 2. The short intervals between observations in states 1 and 3 forced the latent sojourn times in state 2 to be correspondingly short. It is possible that many of these patients, as well as other patients in the analysis dataset, progressed directly from WHO stage 1 to stage 3 without exhibiting symptoms of the intermediate stage 2. Hence, our assumption that each patient visits state 2 may be suspect. This assumption can be eliminated by employing a state model that includes all four stages of disease.

In the model we chose for the disease process, patients were allowed to skip the third stage of disease. However, as we have just discussed, one consequence of our choice of model was that all patients were assumed to visit the second stage. There is no scientific basis for requiring a visit to one intermediate stage but not the other. Since patients may skip either or both of the intermediate stages of disease, an ideal model would allow for any progressive path through the stages.

We have several immediate plans for future analysis of this dataset. First, as we just noted, we plan to apply our proposed approach using a state model that includes all four WHO stages of HIV/AIDS. To alleviate the impact of the assumption of relative homogeneity of total lead times, we will use an exponential model for the sojourn times in disease stage 1. Additionally, we will explore other scientific questions of interest for this application. For example, considering the four-state model for the situation we can assess the impact of behavioral factors such as baseline smoking status on the rate of disease progression for a given viral type. As we consider the additional variables, we will seek to elicit informative priors from our collaborators. Additionally, recall that in this dataset, a number of patients went on antiretroviral therapy during the course of follow-up. Future analysis could assess the impact of antiretroviral therapy on rate of disease progression. This would require, however, an extension of our approach to accommodate time-varying covariates. We will discuss potential approaches for accommodating time-varying covariates in the next chapter.

Chapter 8

DISCUSSION AND FUTURE DIRECTIONS

8.1 Discussion

In this dissertation we developed methodology to model a progressive multi-state process under an intermittent observation scheme. To allow for flexibility in our model we primarily assumed that the underlying process was semi-Markov.

We began in Chapter 3 with a simple progressive process and compared the performance of our proposed approach with that of several existing methods. When the sojourn times were truly exponential, our proposed approached performed as well as the method of Kalbfleisch and Lawless (1985). For other sojourn time distributions our method outperformed that of Kalbfleisch and Lawless (1985). This is reasonable since the latter is based on the Markov assumption. The method of De Gruttola and Lagakos (1989) was able to capture deviation from a constant hazard function, but implicitly makes the assumption that subjects are observed in every visited state. Hence, although the method provides a flexible approach to modeling sojourn times, it is suitable only when subjects are frequently observed. Our approach outperforms other methods when a flexible sojourn time model is desired but subjects are observed infrequently and when it is possible that subjects are not observed in every stage of disease that they visit.

In Chapter 4 we considered a general progressive process. We focused initially on the illness-death model, for which our proposed approach performed well under a variety of scenarios, consistently estimating the parameters of both the embedded Markov chain and the sojourn times. Under the assumption of Weibull sojourn times, our estimation procedure performed very well even when subjects were observed infrequently. However, under more flexible sojourn time models, more informative priors for the sojourn time parameters were often required. We observed similar performance of our procedure under a four-state process.

Chapter 5 examined the performance of our approach for the illness-death model in the

presence of left censoring. We obtained consistent inference about the parameters of interest both when the total lead times were assumed to be common and when they were allowed to vary. However, in each of these two cases, an informative prior on the distribution of the common or mean total lead time was required. As expected, more severe left censoring led to inflated posterior standard deviations for the parameters governing the sojourn times in the first state.

In Chapter 6 we examined the performance of the approach in estimating the impact on the rate of disease progression of a single covariate in the presence of left-censored entry. We considered a binary-valued and continuous-valued covariate, in turn, and in both cases we obtained consistent inference about the covariate coefficients as well as the parameters governing the embedded Markov chain and sojourn times.

In Chapter 7 we applied our approach to the longitudinal study of WHO-defined stages of HIV/AIDS in Senegal. Subjects in the study were infected with HIV-1, HIV-2, or both viruses, and the primary scientific aim was to characterize the impact of viral type on rate of disease progression. Although there are four disease stages, we modeled progression considering only the last three disease stages to reduce the impact of the assumption of relative homogeneity of the total lead times among subjects.

8.2 Recommendations

The proposed approach is geared to the analysis of intermittent observations of a process having several stages. Using the observed stage data, we attempt to make inference about the paths that patients take through the disease stages and the lengths of time spent in each. By its very nature, this type of data contains sparse information about each patient's true trajectory. In particular, in some cases, the amount of information contained in the dataset is too low for the proposed approach to work properly. In such cases, the Markov model, with its own limitations, may be a viable alternative. Though it is difficult to assess with certainty whether the proposed approach may give sound inference for any dataset under consideration, we provide here some rough guidelines regarding the circumstances under which our proposed approach may be appropriate.

The decision of whether to apply the proposed approach to the dataset under consider-

ation depends on both the complexity of the desired model and the richness of information contained in the observed data. The complexity of the desired model depends on the number of states and allowed transitions as well as the modeling assumptions about the sojourn times in each state. It additionally depends on the number of covariates under consideration and the modeling assumptions about the impact of the covariates on the underlying process. On the other hand, the richness of the observed information depends on the number of subjects and the frequency of observation relative to the speed of the underlying process.

Table 8.1 gives a guideline for using the proposed approach for a hypothetical dataset. That is, we provide a suggestion of whether the proposed approach may be suitable under various scenarios. We first consider a 3-state general progressive process and vary the number of subjects N, the number of covariates, and the frequency of observation of the subjects. The guidelines in the table are based on simulation studies of a process in which each of the sojourn times follows a Weibull distribution. For each scenario we give a recommendation for the case in which subjects are observed less frequently than twice per year, and at least twice per year. We examined scenarios in which there were varying numbers of independent binary covariates. We also considered a 4-state general progressive process with no covariates. Each recommendation was based on the frequentist evaluation of the Bayesian estimation applied to a number of simulated datasets.

For a 3-state process, we see that the proposed approach is more suitable when the dataset is either fairly large, or is moderate in size and subjects are observed frequently. As additional covariates are added to the model, the proposed approach requires greater richness of observed information to work properly. Put another way, for a given dataset, the proposed approach is less likely to work properly as additional covariates are added to the model. This trend can be seen by inspecting the columns corresponding to 0–1 covariates and 2–3 covariates. For a 4-state process, the proposed approach is again more suitable when the dataset is fairly large. The same trend holds as covariates are added to the model (results not shown).

The results presented in the table are intended to be used merely as guidelines. That is, our recommendation for a particular dataset may depend on specific features of that dataset as well as the amount and type of expert information about the underlying process that

		3-state process				4-state process	
		0–1 covariates		2–3 covariates		0 covariates	
	annual frequency of observation	<2	≥2	<2	≥2	<2	≥2
N	50	No	No	No	No	No	No
	100	No	Yes	No	No	No	No
	200	Yes	Yes	No	Yes	Yes	Yes
	400	Yes	Yes	Yes	Yes	Yes	Yes

Table 8.1: Indicators of whether proposed approach is recommended in each case.

is available. For example, for a dataset with few subjects, our proposed approach would generally not be recommended. However, if the subjects are observed very frequently and if existing knowledge about the underlying process indicates that the times spent in each stage are not modeled well by the exponential distribution, then we may recommend our proposed approach and use the existing information to choose appropriate sojourn time models and formulate prior distributions for the corresponding parameters.

Additionally, informative priors or additional assumptions may be needed for parameters corresponding to state pairs for which there are few latent transitions. For example, in a 3-state process, if very few patients make a transition directly from state 1 to 3, then the parameters corresponding to X_{13} may be poorly estimated under noninformative priors. One option is to impose informative prior distributions on these parameters. Another option is to assume either that X_{13} has the same distribution as X_{12}, or that X_{13} has a simplified distribution, such as an exponential distribution. In such a case, the chosen assumption should be the one that is most credible given knowledge of the underlying process.

As we discussed previously, the quality of inference yielded by our proposed approach decreases as additional covariates are added to the model. Regardless of the approach that is ultimately taken, the set of covariates included in the model should be parsimonious. That is, to preserve precision of estimation, only the predictor of interest and major potential confounders should be included. If the number of covariates that must be included in the model is large and the richness of information in the observed data is modest, then the proposed approach may not be suitable.

Recall that the proposed approach requires some knowledge of the distribution of the length of time spent in the first observed state through either the observed data or prior knowledge of the process. The approach additionally requires the assumption of relative homogeneity of these times. If knowledge of these lead times is not available or if the assumption is not appropriate, then the proposed approach may not be suitable.

We have presented the proposed approach within a flexible framework with the intent that the approach may be tailored in various ways to the dataset under consideration. Any number of simplifying assumptions, whenever scientifically meaningful, could and should be made for a specific application. The model for the underlying process itself may be simplified in several ways. For example, if it is reasonable to assume that patients progress through the disease stages in a prescribed sequence, then we can assume that the underlying process is simple progressive rather than general progressive. Alternatively, disease stages may be combined. Simplifications such as these can yield greater precision in our inference. Also, any number of simplifying assumptions may be made about the covariates and their impact on the process. Additionally, once the sojourn time models are chosen, expert information about the corresponding parameters, if it is available, should be used via informative prior distributions.

If, however, the proposed approach is deemed to unsuitable for the dataset under consideration, then other exising approaches must be considered. The primary issue to consider is whether the Markov assumption—that the time spent in each state is exponentially distributed and has no dependence on the next state to be visited—may be reasonable to impose on the underlying process.

8.3 Limitations and Future Directions

Our work, while addressing a few common features in real applications, relies on some assumptions. Specifically, to accommodate left censoring we assumed that the total lead times among subjects sharing a first observed state were relatively homogeneous. This assumption may or may not be reasonable depending on the application. Further research is needed when using a less constrained assumption.

There are some potential data features that we did not address in this work. Given the

nature of longitudinal multi-state data, the possibility of sampling bias may be a concern. For example, the accrual process may be subject to length-biased sampling: patients who progress quickly through the early stages may have a smaller chance of being observed in these stages than other patients. Hence, observations in early stages of disease may be subject to truncation, and the sojourn times in these stages may be overestimated. In future work, we plan to address this source of sampling bias. We may be able to use a maximum partial likelihood approach to carry out inference in the Cox framework as in Kalbfleisch and Lawless (1991) and Wang et al. (1993).

In our extension that allows the rate of disease progression to vary across subjects, we made a number of simplifying assumptions. Considering a single explanatory covariate, we assumed that between-subject differences could be accounted for based on a single measurement of this covariate. Our approach does not address time-varying covariates. Future work could address time-varying covariates by extending methods that have been used in standard survival analysis, but we anticipate some additional challenges. First, since the value of each covariate is subject to an intermittent observation scheme, we would either need to make an assumption about the values at intermediate timepoints, or rely on self-reported values. Second, there may be a lag in the impact of the covariate on the rate of disease progression. We may assume, for example, that the value of the covariate at time t affects the state of the process at time $t + s$, or that it has a waning effect on the state of the process over the interval $[t, \infty)$. Third, in our approach we made the assumption of proportional hazards across possible values of each covariate. In some cases, this assumption may not be reasonable. For example, suppose that a treatment is effective at slowing disease progression, but takes a period of time to begin working. In this case, the true treatment effect would increase over time, and our approach would not be able to capture this increase. In such cases, one may consider modeling using time-varying coefficients. Specifically, if we suspected that the impact of a covariate on a given sojourn time violated the proportional hazards assumption, then we could model the corresponding coefficient in a piecewise manner or as a linear function of time since entry into the state to capture the deviation from constant hazard ratio. These modeling choices should depend on scientific reasoning and existing knowledge of the manner in which a covariate may affect the process

under consideration.

Finally, we assumed throughout the methodological development in this dissertation that the underlying disease process is progressive. For many disease processes, patients may exhibit reversals in disease progression. A state model that is useful in such situations is the *relapsing-remitting* state model, shown in Figure 8.1 below.

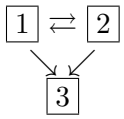


Figure 8.1: State model for relapsing-remitting disease process.

The relapsing-remitting state model is suitable for diseases for which patients may experience several bouts of illness before the disease progresses. For example, patients with Crohn's disease may have periods in which symptoms flare up, and are at higher risk of developing cancers of the bowel. The relapsing-remitting state model could be useful in this situation, with states 1–3 representing remission, the presence of symptoms, and tumor development, respectively.

When reversals of the disease process are possible, analysis of data under an intermittent observation scheme becomes more challenging since there could be, in theory, infinitely many possible paths through the states between successive observations. For example, a subject who was observed in state 1 at two successive time points could have visited state 2 any number of times between the observations. If the state model is sufficiently simple, it is possible to represent the set of all possible paths through the states between successive observations fairly easily. In the above example, the set of all possible paths between successive observations can be characterized by the number of visits to state 2, which can be 0,1,2,.... We could make a modeling assumption about the number of transitions in a given interval so that the number of possible paths is finite. Thus, we may be able to modify the methodology developed in this dissertation to accommodate a non-progressive disease process. However, for more complex state models, in which many different paths through the states may have been taken between successive observations, further research is needed.

BIBLIOGRAPHY

O. Aalen. Nonparametric inference for a family of counting processes. *Annals of Statistics*, 6(4):701–726, 1978.

AIDS Education and Training Centers National Resource Center. HIV classification: CDC and WHO staging systems. Retrieved September 6, 2012, from www.aidsetc.org/aidsetc?page=cg-205_hiv_classification.

E.A. Alvarez. Smoothed nonparametric estimation in window censored semi-Markov processes. *Journal of Statistical Planning and Inference*, 131:209–229, 2005.

T.W. Anderson and L.A. Goodman. Statistical inference about Markov chains. *Annals Of Mathematical Statistics*, 28(1):89–110, 1957.

B.J.N. Blight. Estimation from a censored sample for the exponential family. *Biometrika*, 57:(2):389–395, 1970.

B.P. Carlin and S. Chib. Bayesian model choice via Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B*, 57(3):473–484, 1995.

I.-S. Chang, Y.C. Chuang, and C.A. Hsiung. Goodness-of-fit tests for semi-Markov and Markov survival models with one intermediate state. *Scandinavian Journal of Statistics*, 2001.

H.H. Chen, S.W. Duffy, and L. Tabar. A Markov chain method to estimate the tumour progression rate from preclinical to clinical phase, sensitivity and positive predictive value for mammography in breast cancer screening. *Statistician*, 45(3):307–317, 1996.

C.L. Chiang. *An Introduction to Stochastic Processes and their Applications*. R.E. Krieger, New York, 1980.

R.E. Colvert and T.J. Boardman. Estimation in the piece-wise constant hazard rate model. *Communications in Statistical Theory and Methods*, A5(11):1013–1029, 1976.

D. Commenges, P. Joly L. Letenneur, and J.F. Dartigues. Incidence and mortality of Alzheimer's disease or dementia using an illness-death model. *Statistics in Medicine*, 23: 199–210, 2004.

C.M. Crespi, W.G. Cumberland, and S. Blower. A queueing model for chronic recurrent conditions under panel observation. *Biometrics*, 61:193–198, 2005.

V. De Gruttola and S.W. Lagakos. Analysis of doubly-censored survival data. *Biometrics*, 45:1–11, 1989.

C. Del Rio and J.W. Curran. Epidemiology and prevention of acquired immune deficiency syndrome and human immunodeficiency infection. In: Mandell, G.L., Bennett, J.E., and Dolan, R., eds. *Mandell, Douglas, and Bennetts Principles and Practice of Infectious Diseases*. 7th ed. Orlando, FL:Saunders Elsevier;2009:chap 121.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39: 1–22, 1977.

J. Esbjornsson, F. Mansson, A. Kvist, P.-E. Isberg, S. Nowroozalizadeh, A.J. Biague, Z.J. da Silva, M. Jansson, E.M. Fenyo, H. Norrgren, and P. Medstrand. Inhibition of HIV-1 disease progression by contemporaneous HIV-2 infection. *New England Journal of Medicine*, 367(3): 224–232, 2012.

Y. Foucher, M. Giral, J.-P. Soulillou, and J.-P. Daures. A semi-Markov model for multistate and interval-censored data with multiple terminal events. Application in renal transplantation. *Statistics in Medicine*, 26:5381–5393, 2007.

Y. Foucher, M. Giral, J.-P. Soulillou, and J.-P. Daures. A flexible semi-Markov model for interval-censored data and goodness-of-fit testing. *Statistical Methods in Medical Research*, 19:127–145, 2010.

H. Frydman. A nonparametric estimation procedure for a periodically observed three-state Markov process, with application to AIDS. *Journal of the Royal Statistical Society, Series B*, 54(3):853–866, 1992.

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian data analysis*. Chapman & Hall/CRC, Boca Raton, 1995.

S.J. Godsill. On the relationship between MCMC model uncertainty methods. Cambridge CB2 1PZ, UK: Signal Processing Group, Cambridge University Engineering Department; Nov. 1997. Report No.: CUED/F-INFENG/TR. 305.

G.S. Gottlieb et al. Equal plasma viral loads predict a similar rate of CD4+ T cell decline in human immunodeficiency virus (HIV) type 1- and HIV-2–infected individuals from Senegal, West Africa. *Journal of Infectious Diseases*, 185:905–914, 2002.

P.J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

U.S. Department of Health and Human Services (HHS). HIV/AIDS Basics: Prevention; Reduce your Risk; Pregnancy and Childbirth. Retrieved October 22, 2013, from aids.gov/hiv-aids-basics/prevention/reduce-your-risk/pregnancy-and-childbirth.

W.H. Hosmer, S. Lemeshow, and S. May. *Applied Survival Analysis: Regression Modeling of Time-to-Event Data.* Wiley, 2008.

R.A. Hubbard. *Modeling a non-homogeneous Markov process via time transformation*. PhD dissertation, University of Washington, 2007.

A. Iosifescu-Manu. Non-homogeneous semi-Markov processes. *Studii si Cercetari Matematice*, 24:529–33, 1972.

S. Jaffar, A.D. Grant, J. Whitworth, P.G. Smith, and H. Whittle. The natural history of HIV-1 and HIV-2 infections in adults in Africa: a literature review. *Bull World Health Organ*, 82:(6): 462–469, 2004.

J. Janssen, editor. *Semi-Markov Models: Theory and Applications*. Plenum Press, New York, 1986.

J. Janssen and R. Manca. *Applied Semi-Markov Processes*. Springer, 2006.

J.D. Kalbfleisch and J.F. Lawless. The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, 80(392):863–871, 1985.

J.D. Kalbfleisch and J.F. Lawless. Regression models for right truncated data with applications to AIDS incubation times and reporting bias. *Statistica Sinica*, 1:19–32, 1991.

M. Kang and S.W. Lagakos. Statistical methods for panel data from a semi-Markov process, with application to HPV. *Biostatistics*, 8(2):863–871, 2007.

E.L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Society*, 53:457–581, 1958.

M.Y. Kim, V. De Gruttola, and S.W. Lagakos. Analyzing doubly censored data with covariates, with application to AIDS. , 49:13–22, 1993.

J.P. Klein and M.L. Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer, 2nd edition, 1997.

S.W. Lagakos, C.J. Sommer, and M. Zelen. Semi-Markov models for partially censored data. *Biometrika*, 65(2):311–317, 1978.

C.D. Lai, M. Xie, and D.N.P. Murthy. A modified Weibull distribution. *IEEE Transactions on Reliability*, 52(1):–37, 2003.

J.F. Lawless and Y.T. Fong. State duration models in clinical and observational studies. *Statistics in Medicine*, 18:2365–2376, 1999.

J.F. Lawless and P. Yan. Some statistical methods for followup studies of disease with intermittent monitoring. In *Multiple comparisons, selection, and applications in biometry*. Marcel Dekker, 1993.

P. Lévy. Systèmes semi-Markoviens à au plus une infinité d'états possibles. *Proc. Int. Congr. Math.*, 2:294, 1954a.

P. Lévy. Processus semi-Markoviens. *Proc. Int. Congr. Math.*, 3:416–426, 1954b.

N. Limnios and G. Oprişan. *Semi-Markov Processes and Reliability*. Birkhäuser, Boston, 2001.

I.M. Longini, W.S. Clark, R.H. Byers, J.W. Ward, W.W. Darrow, G.F. Lemp, and H.W. Hethcote. Statistical analysis of the stages of HIV infection using a Markov model. *Statistics in Medicine*, 8:851–843, 1989.

G. Marshall and R.H. Jones . Multi-state Markov models and diabetic retinopathy. *Statistics in Medicine*, 14:1975–1983, 1995.

L. Meira-Machado, J. de Uña Álvarez, C. Cadarso-Suárez, and P.K. Andersen. Multi-state models for the analysis of time-to-event data. *Statistical Methods In Medical Research*, 18(2):195–222, Apr 2009.

L. Meira-Machado. Inference for non-Markov multi-state models: an overview. *Revstat Statistical Journal*, 9(1):83+, Mar 2011.

C.E. Mitchell, M.G. Hudgens, C.C. King, S. Cu-Uvin, Y. Lo, A. Rompalo, J. Sobel, and J.S. Smith. Discrete-time semi-Markov modeling of human papillomavirus persistence. *Statistics In Medicine*, 30(17):2160–2170, Jul 30 2011.

G.S. Mudholkar and D.K. Srivastava. Exponentiated Weibull family for analyzing bathtub failure-rate data. *IEEE Transactions on Reliability*, 42(2):299–302, 1993.

B. Ouhbi and N. Limnios. Nonparametric estimation for semi-Markov processes based on its hazard rate functions. *Statistical Inference for Stochastic Processes*, 2:151–73, 1999.

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

R. Pyke. Markov renewal processes: definitions and preliminary properties. *Annals of Mathematical Statistics*, 32:1231–42, 1961a.

R. Pyke. Markov renewal rprocesses with finitely many states. *Annals of Mathematical Statistics*, 32:1243–59, 1961b.

S.M. Ross. *Stochastic Processes*. Wiley & Sons, Berkeley, 2nd edition, 1996.

D.B. Rubin. *Multiple imputation for nonresponse in surveys*. Wiley, 1987.

G.A. Satten and M.R. Sternberg. Fitting semi-Markov models to interval-censored data with unknown initiation times. *Biometrics*, 55(2):507–513, 1999.

W.L. Smith. Regenerative stochastic processes. *Proceedings of the Royal Society of London, Series A*, 232:6–31, 1955.

T.R. Sterling and R.E. Chaisson. General clinical manifestations of human immunodeficiency virus infection (including the acute retroviral syndrome and oral, cutaneous, renal, ocular, metabolic, and cardiac diseases). In: Mandell, G.L., Bennett, J.E., and Dolan, R., eds. *Mandell, Douglas, and Bennetts Principles and Practice of Infectious Diseases*. 7th ed. Orlando, FL:Saunders Elsevier;2009:chap 121.

M.R. Sternberg and G.A. Satten. Discrete-time nonparametric estimation for semi-Markov models of chain-of-events data subject to interval censoring and truncation. *Biometrics*, 55(2):514–522, 1999.

M.A. Tanner. *Tools for statistical inference: observed data and data augmentation methods*. Springer-Verlag, Heidelberg, 1991.

M.A. Tanner and W.H. Wong. The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82(398):528–540, 1987.

T.M. Therneau and P.M. Grambsch. *Modeling survival data: extending the Cox model*. Springer, New York, 2000.

A.C. Titman and L.D. Sharples. Semi-Markov models with phase-type sojourn distributions. *Biometrics*, 66:742–752, 2010.

F.D. Tóth, A. Bácsi, Z. Beck, and J. Szabó. Vertical transmission of human immunodeficiency virus. *Acta Microbiol Immunol Hung*, 48:413–427, 2001.

B.W. Turnbull. The empirical distribution function with arbitrarily gropued, censored and truncated data. *Journal of the Royal Statistical Society, Series B*, 38:290–295, 1976.

J.G. Voelkel and J. Crowley. Nonparametric inference for a class of semi-Markov processes with censored observations. *Annals of Statistics*, 12(1):142–160, 1984.

M.-C. Wang, R. Brookmeyer, and N. Jewell. Statistical models for prevalent cohort data. *Biometrics*, 49:1–11, 1993.

J.H. Ware and D.L DeMets. Reanalysis of some baboon descent data. *Biometrics* 32:(2): 459–163, 1976.

G.H. Weiss and M. Zelen. A semi-Markov model for clinical trials. *Journal of Applied Probability*, 2:269–285, 1965.

H. Whittle, A. Egboga, J. Todd, T. Corrah, A. Wilkins, E. Demba, G. Morgan, M. Rolfe, N. Berry, and R. Tedder. Clinical and laboratory predictors of survival in Gambian patients with symptomatic HIV-1 or HIV-2 infection. *AIDS*, 6:685–689, 1992.

World Health Organization. WHO case definitions of HIV for surveillance and revised clinical staging and immunological classification of HIV-related disease in adults and children. *HIV/AIDS Programme: Strengthening health services to fight HIV/AIDS*. Geneva:6–16, 2007.

Appendix A

MISSINGNESS IN DATASETS

A number of existing methods make the assumption that each subject is observed in every visited state. In a variety of applications, this assumption is far from reasonable, and methods built on this assumption are not able to handle observations with unobserved states. Here we consider a simple three–state progressive process and derive the expected probability that a subject selected at random was not observed in state 2, for each of the scenarios considered in Section 3.4.2. As in that section, we assume that subjects were observed at times $0, 1, 2, \ldots$. We derive the probability that $\delta(2) = 1$ in the cases where the sojourn times in the two states are exponential and Weibull. In Table A.1 we present the expected probabilities for each of the four scenarios and report statistics on the datasets used in the simulation studies in Section 3.4.2.

To see the impact of excluding these observations on the true sojourn times that were used to generate the datasets used in the primary simulations, we present the maximum likelihood estimates of the parameters for each scenario in Table A.2. Though the restricted set of observations has a different distribution than the one from which the data were generated, we note that a substantial difference from the true value of the parameter indicates an impact of excluding the observations. From the results in the table we can see that the true sojourn times in state 2 are particularly affected.

Table A.1: Mean (SD) of proportion of subjects for whom state 2 was unobserved.

N	Scenario 1	Scenario 2	Scenario 3	Scenario 4
expected	0.229	0.687	0.077	0.159
50	0.216 (0.060)	0.689 (0.062)	0.078 (0.031)	0.156 (0.051)
100	0.232 (0.042)	0.689 (0.043)	0.078 (0.025)	0.159 (0.033)
200	0.227 (0.027)	0.684 (0.028)	0.081 (0.020)	0.161 (0.029)
400	0.229 (0.020)	0.689 (0.025)	0.077 (0.012)	0.158 (0.019)

In the general case where $X_1 \sim f_1$ and $X_2 \sim f_2$ we have

$$\begin{aligned} P(\delta(2)=1) &= \sum_{n=1}^{\infty} P(X_1 \in (n-1,n],\ X_1+X_2 \in (n-1,n]) \\ &= \sum_{n=1}^{\infty} P(X_1 = x_1 \in (n-1,n]) \cdot P(X_2 \in (n-1-x_1, n-x_1] | X_1 = x_1) \\ &= \sum_{n=1}^{\infty} \int_{n-1}^{n} \int_{0}^{n-s_1} f_1(s_1) \cdot f_2(s_2)\, ds_2\, ds_1. \end{aligned}$$

When $X_1 \sim exp(\theta_1)$ and $X_2 \sim exp(\theta_2)$ with $\theta_1, \theta_2 > 0$ and $\theta_2 \neq \theta_1$, we have

$$\begin{aligned} P(\delta(2)=1) &= \sum_{n=1}^{\infty} \int_{n-1}^{n} \int_{0}^{n-s_1} \frac{1}{\theta_1} \exp\left(-\frac{s_1}{\theta_1}\right) \cdot \frac{1}{\theta_2} \exp\left(-\frac{s_2}{\theta_2}\right) ds_2\, ds_1 \\ &= \sum_{n=1}^{\infty} \int_{n-1}^{n} \left[\int_{0}^{n-s_1} \frac{1}{\theta_2} \exp\left(-\frac{s_2}{\theta_2}\right) ds_2\right] \cdot \frac{1}{\theta_1} \exp\left(-\frac{s_1}{\theta_1}\right) ds_1 \\ &= \sum_{n=1}^{\infty} \int_{n-1}^{n} \left[1 - \exp\left(-\frac{n-s_1}{\theta_2}\right)\right] \cdot \frac{1}{\theta_1} \exp\left(-\frac{s_1}{\theta_1}\right) ds_1 \\ &= \sum_{n=1}^{\infty} \left[\int_{n-1}^{n} \frac{1}{\theta_1} \exp\left(-\frac{s_1}{\theta_1}\right) ds_1 - \int_{n-1}^{n} \frac{1}{\theta_1} \exp\left(-\frac{n}{\theta_2} + \frac{s_1}{\theta_2} - \frac{s_1}{\theta_1}\right) ds_2\right] \\ &\vdots \\ &= \sum_{n=1}^{\infty} \left[\exp\left(-\frac{n}{\theta_1}\right) \cdot \left(\exp\left(\frac{1}{\theta_1}\right) + \frac{\theta_1}{\theta_2-\theta_1} - \frac{\theta_2}{\theta_2-\theta_1} \cdot \exp\left(\frac{\theta_2-\theta_1}{\theta_1\theta_2}\right)\right)\right] \\ &= \left(\exp\left(\frac{1}{\theta_1}\right) + \frac{\theta_1}{\theta_2-\theta_1} - \frac{\theta_2}{\theta_2-\theta_1} \cdot \exp\left(\frac{\theta_2-\theta_1}{\theta_1\theta_2}\right)\right) \cdot \sum_{n=1}^{\infty} \exp\left(-\frac{n}{\theta_1}\right) \\ &= \left(\exp\left(\frac{1}{\theta_1}\right) + \frac{\theta_1}{\theta_2-\theta_1} - \frac{\theta_2}{\theta_2-\theta_1} \cdot \exp\left(\frac{\theta_2-\theta_1}{\theta_1\theta_2}\right)\right) \cdot \left[\sum_{n=0}^{\infty} \left(\exp\left(-\frac{1}{\theta_1}\right)\right)^n - 1\right] \\ &= \left(\exp\left(\frac{1}{\theta_1}\right) + \frac{\theta_1}{\theta_2-\theta_1} - \frac{\theta_2}{\theta_2-\theta_1} \cdot \exp\left(\frac{\theta_2-\theta_1}{\theta_1\theta_2}\right)\right) \cdot \left[\frac{\exp\left(-\frac{1}{\theta_1}\right)}{1-\exp\left(-\frac{1}{\theta_1}\right)}\right]. \end{aligned}$$

For $\theta_1 = \theta_2 \equiv \theta > 0$, we have

$$\begin{aligned}
P(\delta(2)=1) &= \sum_{n=1}^{\infty} \int_{n-1}^{n} \int_{0}^{n-s_1} \frac{1}{\theta} \exp\left(-\frac{s_1}{\theta}\right) \cdot \frac{1}{\theta} \exp\left(-\frac{s_2}{\theta}\right) ds_2 \, ds_1 \\
&= \sum_{n=1}^{\infty} \int_{n-1}^{n} \left[\int_{0}^{n-s_1} \frac{1}{\theta} \exp\left(-\frac{s_2}{\theta}\right) ds_2 \right] \frac{1}{\theta} \exp\left(-\frac{s_1}{\theta}\right) ds_1 \\
&= \sum_{n=1}^{\infty} \int_{n-1}^{n} \left[1 - \exp\left(-\frac{n-s_1}{\theta}\right) \right] \cdot \frac{1}{\theta} \exp\left(-\frac{s_1}{\theta}\right) ds_1 \\
&= \sum_{n=1}^{\infty} \int_{n-1}^{n} \frac{1}{\theta} \exp\left(-\frac{s_1}{\theta}\right) ds_1 - \frac{1}{\theta} \int_{n-1}^{n} \exp\left(-\frac{n}{\theta} + \frac{s_1}{\theta} - \frac{s_1}{\theta}\right) ds_1 \\
&= \sum_{n=1}^{\infty} \exp\left(-\frac{n-1}{\theta}\right) - \exp\left(-\frac{n}{\theta}\right) - \frac{1}{\theta} \exp\left(-\frac{n}{\theta}\right) \cdot \int_{n-1}^{n} ds_1 \\
&= \underbrace{\sum_{n=1}^{\infty} \left[\exp\left(-\frac{n-1}{\theta}\right) - \exp\left(-\frac{n}{\theta}\right) \right]}_{\textit{telescoping series}} - \frac{1}{\theta} \cdot \sum_{n=1}^{\infty} \exp\left(-\frac{n}{\theta}\right) \\
&= 1 - \frac{1}{\theta} \cdot \left[\frac{\exp\left(-\frac{1}{\theta}\right)}{1 - \exp\left(-\frac{1}{\theta}\right)} \right].
\end{aligned}$$

Calculations when $X_1 \sim We(k_1, \theta_1)$ and $X_2 \sim We(k_2, \theta_2)$ are not feasible with standard integration methods. We used empirical methods to find the results in the table.

Table A.2: Impact of excluding observations.

	parameter	k_1	θ_1	k_2	θ_2
Scenario 1	truth	—	2.000	—	2.000
	estimate	—	2.034	—	2.517
Scenario 2	truth	—	0.500	—	0.500
	estimate	—	0.652	—	1.011
Scenario 3	truth	2.000	2.000	2.000	2.000
	estimate	2.026	1.992	2.257	2.129
Scenario 4	truth	4.000	1.000	4.000	1.000
	estimate	3.848	0.966	4.468	1.045

Appendix B

STANDARD ERRORS FOR METHOD OF DE GRUTTOLA AND LAGAKOS (1989)

Here we present the standard errors associated with the estimates of the weights associated with the sojourn times X_1 and X_2 in the method of De Gruttola and Lagakos (1989). In the following tables we present results for Scenarios 1–4 in Table 3.3. Hence these tables of standard errors correspond to the estimates in Tables 3.9–3.10.

Table B.1: Method of De Gruttola and Lagakos (1989) (scenarios 1–2).
For each sample size n we present the standard errors corresponding to the estimated weights $\hat{w}_{1,1}, \ldots, \hat{w}_{1,r}$, which correspond to point masses $(y_{1,1}, \ldots, y_{1,r}) = (\frac{1}{2}, \ldots, \frac{2r-1}{2})$ for state 1, and those for $\hat{w}_{2,1}, \ldots, \hat{w}_{2,s}$, which correspond to point masses $(y_{2,1}, \ldots, y_{2,s}) = (1, \ldots, s)$ for state 2.

Scenario 1. True weights: $\mathbf{w}_1 = (0.393, 0.239, 0.145, 0.088, 0.053, 0.032, 0.020, 0.012, 0.007, 0.004, 0.003, 0.002, 0.001, 0.001, 0.000, 0.000, 0.000, 0.000, 0.000)$.

N	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$	$y_{1,5}$	$y_{1,6}$	$y_{1,7}$	$y_{1,8}$	$y_{1,9}$	$y_{1,10}$	$y_{1,11}$	$y_{1,12}$	$y_{1,13}$	$y_{1,14}$	$y_{1,15}$	$y_{1,16}$	$y_{1,17}$	$y_{1,18}$	$y_{1,19}$
50	0.076	0.082	0.047	0.053	0.035	0.026	0.020	0.020	0.013	0.008	0.009	0.010	0.005	0.005	0.000	0.000	0.000	0.004	0.000
100	0.058	0.047	0.040	0.033	0.024	0.018	0.015	0.014	0.009	0.010	0.006	0.005	0.004	0.004	0.000	0.000	0.001	0.000	0.000
200	0.041	0.038	0.030	0.023	0.016	0.014	0.010	0.009	0.006	0.006	0.005	0.004	0.001	0.003	0.001	0.001	0.001	0.001	0.000
400	0.026	0.024	0.021	0.015	0.010	0.010	0.008	0.006	0.004	0.004	0.002	0.002	0.002	0.002	0.001	0.001	0.001	0.001	0.000

$\mathbf{w}_2 = (0.393, 0.239, 0.145, 0.088, 0.053, 0.032, 0.020, 0.012, 0.007, 0.004, 0.003, 0.002, 0.001, 0.001, 0.000, 0.000, 0.000, 0.000, 0.000)$.

N	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$	$y_{2,5}$	$y_{2,6}$	$y_{2,7}$	$y_{2,8}$	$y_{2,9}$	$y_{2,10}$	$y_{2,11}$	$y_{2,12}$	$y_{2,13}$	$y_{2,14}$	$y_{2,15}$	$y_{2,16}$	$y_{2,17}$	$y_{2,18}$	$y_{2,19}$
50	0.083	0.071	0.066	0.043	0.033	0.028	0.024	0.019	0.013	0.011	0.007	0.008	0.008	0.006	0.000	0.000	0.003	0.000	0.000
100	0.059	0.045	0.033	0.028	0.026	0.020	0.014	0.013	0.009	0.007	0.005	0.003	0.005	0.003	0.003	0.001	0.000	0.002	0.000
200	0.033	0.036	0.028	0.025	0.016	0.013	0.010	0.009	0.007	0.005	0.004	0.003	0.003	0.002	0.001	0.001	0.001	0.001	0.001
400	0.026	0.028	0.023	0.017	0.013	0.011	0.008	0.005	0.005	0.003	0.003	0.002	0.002	0.001	0.001	0.001	0.001	0.001	0.001

Scenario 2.
$\mathbf{w}_1 = (0.865, 0.117, 0.016, 0.002, 0.000)$.

N	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$	$y_{1,5}$
50	0.086	0.086	0.033	0.000	0.000
100	0.063	0.059	0.027	0.009	0.000
200	0.041	0.038	0.015	0.004	0.002
400	0.027	0.026	0.013	0.004	0.001

$\mathbf{w}_2 = (0.865, 0.117, 0.016, 0.002, 0.000, 0.000, 0.000, 0.000)$.

N	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$	$y_{2,5}$	$y_{2,6}$	$y_{2,7}$	$y_{2,8}$
50	0.101	0.088	0.036	0.014	0.000	0.000	0.000	0.000
100	0.068	0.062	0.022	0.008	0.000	0.000	0.000	0.000
200	0.038	0.038	0.014	0.007	0.002	0.002	0.000	0.000
400	0.036	0.034	0.012	0.006	0.001	0.000	0.001	0.000

Table B.2: Method of De Gruttola and Lagakos (1989) (scenarios 3–4).

Scenario 3. $\mathbf{w}_1 = (0.221, 0.411, 0.262, 0.087, 0.016, 0.002, 0.000, 0.000)$.

N	$y_{1,1}$	$y_{1,2}$	$y_{1,3}$	$y_{1,4}$	$y_{1,5}$	$y_{1,6}$	$y_{1,7}$	$y_{1,8}$
50	0.072	0.061	0.066	0.041	0.018	0.006	0.000	0.000
100	0.054	0.042	0.049	0.032	0.011	0.004	0.002	0.002
200	0.034	0.033	0.035	0.022	0.009	0.003	0.001	0.000
400	0.026	0.026	0.029	0.015	0.007	0.002	0.000	0.000

$\mathbf{w}_2 = (0.221, 0.411, 0.262, 0.087, 0.016, 0.002, 0.000, 0.000)$.

N	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$	$y_{2,4}$	$y_{2,5}$	$y_{2,6}$	$y_{2,7}$	$y_{2,8}$
50	0.054	0.061	0.049	0.031	0.013	0.004	0.002	0.000
100	0.058	0.051	0.041	0.026	0.009	0.002	0.000	0.000
200	0.040	0.035	0.035	0.017	0.006	0.001	0.000	0.000
400	0.025	0.027	0.018	0.011	0.005	0.001	0.000	0.000

Scenario 4.
$\mathbf{w}_1 = (0.632, 0.368)$.

N	$y_{1,1}$	$y_{1,2}$
50	0.066	0.066
100	0.047	0.047
200	0.036	0.036
400	0.027	0.027

$\mathbf{w}_2 = (0.632, 0.368, 0.000)$.

N	$y_{2,1}$	$y_{2,2}$	$y_{2,3}$
50	0.044	0.044	0.000
100	0.031	0.031	0.000
200	0.025	0.025	0.000
400	0.017	0.017	0.000

Appendix C

BIAS AND RMSE FOR 3-STATE SIMPLE PROGRESSIVE PROCESS SUBJECT TO VARYING DEGREES OF RIGHT CENSORING

Here we present bias and root mean squared error (RMSE) corresponding to parameter estimates in selected simulation scenarios in the case of a three-state simple progressive process. We consider Scenarios 1* and 2* from Table 3.2 in Chapter 3: the exponential and Weibull cases, respectively, under correct model specification. Subjects enter state 1 at time zero and are followed up at times $0.25, 0.50, 0.75, \ldots$ years. We consider three degrees of right censoring:

- none: subjects are followed until they enter the absorbing state;
- mild: subjects are followed for a mean of ten years (standard deviation of one year) with a 10% probability of dropping out prematurely;
- severe: subjects are followed for a mean of five years (standard deviation of one year) with a 50% probability of dropping out prematurely.

Results are presented in the following tables. RMSE is computed as the square root of the sum of squared bias and model-based variance. As expected, we see that RMSE increases with the degree of right censoring, largely due to increasing variance. Also, as expected, right censoring has a greater impact on inference for the parameters corresponding to X_2 than for X_1.

Scenario 1: exponential case. Bias.

Parameter	Degree of right censoring		
	none	mild	severe
θ_1	-0.002	-0.008	-0.002
θ_2	0.023	0.024	0.023

Scenario 1: exponential case. RMSE.

Parameter	Degree of right censoring		
	none	mild	severe
θ_1	0.100	0.101	0.115
θ_2	0.103	0.109	0.145

Scenario 2: Weibull case. Bias.

Parameter	Degree of right censoring		
	none	mild	severe
k_1	0.043	0.039	0.039
θ_1	-0.003	-0.003	-0.005
k_2	-0.006	-0.007	0.002
θ_2	0.001	0.001	-0.002

Scenario 2: Weibull case. RMSE.

Parameter	Degree of right censoring		
	none	mild	severe
k_1	0.167	0.164	0.180
θ_1	0.026	0.027	0.029
k_2	0.170	0.163	0.207
θ_2	0.032	0.028	0.035

Appendix D

PERFORMANCE OF PROPOSED APPROACH MODIFIED TO ACCOMMODATE HIV/AIDS STAGING APPLICATION

We demonstrate the performance of the proposed approach when several modifications have been made to accommodate the HIV/AIDS application. Specifically, we consider the state model shown in Figure D.1 and generate simulated trajectories from the corresponding transition probability matrix:

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & p_{23} & 1-p_{23} \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

with $p_{23} \in (0,1)$. In our simulation study, we assume that each subject visits state 3 with 90% probability ($p_{23} = 0.9$). We assume that each of the baseline sojourn times in the model, X_{ij} with $(i,j) \in \{(1,2),(2,3),(3,4),(2,4)\}$, follows a Weibull distribution with common shape and scale parameters: $Weibull(k_{ij}, \theta_{ij})$ with $k_{ij} = 2.0$ and $\theta_{ij} = 4.0$ for each (i,j). Subjects are observed every three months ($\Delta = 0.25$). Each subject is followed until he enters the absorbing state. We assume noninformative uniform priors for each of the parameters of interest. Total lead times—the time from entry into state 1 until the first observation—are generated from a truncated normal distribution: $\mathcal{N}_{trunc}(\mu_T, \sigma_T^2)$ with $\mu_T = 1.0$ and $\sigma_T = 0.2$. Although we use the state lead time parameterization in the model, we generate total lead times for each subject since generating subjects' trajectories using the state lead time parameterization is quite clumsy.

We consider a categorical covariate that takes on three values to mirror the situation that we have in the HIV/AIDS staging application, where the viral type is a factor having three levels. To include this covariate in the model we choose one level as baseline and define two

binary-valued covariates to indicate the other two levels. In the simulation study we assume the three levels of this covariate are equally likely. We model the impact of the covariate on the sojourn times only, not on the probability of visiting state 3, and as in Chapter 7, we assume that the impact of the covariate on the sojourn time in state i does not depend on the next state j to be visited. So, $\exp(\beta_i^k)$ is the multiplicative increase, associated with covariate level k relative to baseline, in the hazard of making a transition from state i to the next state, for $i = 2, 3$ and $k = 1, 2$. We assume in this study that the covariate has no impact on the sojourn times in each state, so that $\beta_i^k = 0$ for each i and k. We assume a diffuse normal prior centered at zero for each of the coefficients: $\pi(\beta_i^k) \sim \mathcal{N}(0, 2)$ for each i and k.

We illustrate the performance of the approach for $N = 400$ subjects. The results, based on $M = 100$ simulated datasets, are shown in Table D.1. For each of the parameters of interest and coefficients, the posterior means and model-based as well as empirical posterior standard deviations are given. The tailored approach performs well in making inference about the parameters in the model. Though the estimate for θ_{24} is biased upward, we note that the uncertainty associated with this estimate is large, owing to the small probability of skipping state 3 and the corresponding small number of latent transitions from state 2 to 4. As we would expect, the uncertainty associated with the estimate of k_{24} is also large. The algorithm is unbiased for estimating the coefficients of the two binary covariates.

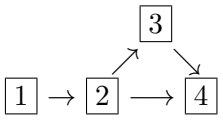


Figure D.1: State model for HIV/AIDS staging application.

Table D.1: Results for simulation study of modified approach.

Parameter	truth	mean	SD_{model}	SD_{emp}
k_{23}	2.000	1.972	0.081	0.088
k_{34}	2.000	2.001	0.084	0.082
k_{24}	2.000	2.036	0.259	0.295
θ_{23}	4.000	4.098	0.186	0.203
θ_{34}	4.000	3.980	0.181	0.190
θ_{24}	4.000	4.157	0.382	0.356
p_{23}	0.900	0.899	0.015	0.014
β_2^1	0.000	0.014	0.122	0.129
β_2^2	0.000	0.002	0.123	0.113
β_3^1	0.000	-0.015	0.129	0.134
β_3^2	0.000	-0.013	0.127	0.143