Modelling Issues in Three-state Progressive Processes

by

Karen Arlene Kopciuk

A thesis
presented to the University of Waterloo
in fulfilment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Statistics

Waterloo, Ontario, Canada, 2001

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

This dissertation focuses on several issues pertaining to three-state progressive stochastic processes. Casting survival data within a three-state framework is an effective way to incorporate intermediate events into an analysis. These events can yield valuable insights into treatment interventions and the natural history of a process, especially when the right censoring is heavy. Exploiting the uni-directional nature of these processes allows for more effective modelling of the types of incomplete data commonly encountered in practice, as well as time-dependent explanatory variables and different time scales.

In Chapter 2, we extend the model developed by Frydman (1995) by incorporating explanatory variables and by permitting interval censoring for the time to the terminal event. The resulting model is quite general and combines features of the models proposed by Frydman (1995) and Kim *et al.* (1993). The decomposition theorem of Gu (1996) is used to show that all of the estimating equations arising from Frydman's log likelihood function are self-consistent. An AIDS data set analyzed by these authors is used to illustrate our regression approach.

Estimating the standard errors of our regression model parameters, by adopting a piecewise constant approach for the baseline intensity parameters, is the focus of Chapter 3. We also develop data-driven algorithms which select changepoints for the intervals of support, based on the Akaike and Schwarz Information Criteria. A sensitivity study is conducted to evaluate these algorithms. The AIDS example is considered here once more; standard errors are estimated for several piecewise constant regression models selected by the model criteria. Our results indicate that for both the example and the sensitivity study, the resulting estimated standard errors of certain model parameters can be quite large.

Chapter 4 evaluates the goodness-of-link function for the transition intensity between states 2 and 3 in the regression model we introduced in chapter 2. By embedding this hazard function in a one-parameter family of hazard functions, we can assess its dependence on the specific parametric form adopted. In a simulation study, the goodness-of-link parameter is estimated and its impact on the regression parameters is assessed. The logistic specification of the hazard function from state 2 to state 3 is appropriate for the discrete, parametric-based data sets considered, as well as for the AIDS data. We also investigate the uniqueness and consistency of the maximum likelihood estimates based on our regression model for these AIDS data.

In Chapter 5 we consider the possible efficiency gains realized in estimating the survivor function when an intermediate auxiliary variable is incorporated into a time-to-event analysis. Both Markov and hybrid time scale frameworks are adopted in the resulting progressive three-state model. We consider three cases for the amount of information available about the auxiliary variable: the observation is completely unknown, known exactly, or known to be within an interval of time. In the Markov framework, our results suggest that observing subjects at just two time points provides as much information about the survivor function as knowing the exact time of the intermediate event. There was generally a greater loss of efficiency in the hybrid time setting.

The final chapter identifies some directions for future research.

Acknowledgements

To write this thesis and fulfill a dream was possible because of the support of many people. Although I will only acknowledge a few individuals here, my sincerest thanks are extended to everyone who helped me along the long, and sometimes winding, Ph.D. road.

I am deeply indebted to my supervisor, Professor David Matthews, who gave me sound advice and unwavering support throughout my doctoral program. Over the past five years, I greatly appreciated David's unending patience, valuable insights, financial support, and encouragement to pursue my own research ideas.

I would also like to extend sincere thanks to my committee members, Professors Jerry Lawless and Richard Cook. I benefited immensely from their assistance, suggestions, and generous gifts of time throughout the course of this research.

To the members of the Department of Statistics and Actuarial Science — faculty, staff, and graduate students — a special thank-you for making my time at UW so wonderful. I cannot imagine a more supportive nor intellectually challenging place to pursue graduate studies.

I want to express my deepest appreciation to my Christian sisters and brothers. Your prayerful support, encouragement, and acts of kindness will be remembered eternally!

My life at UW was enriched immensely by the friendship and support of a few special individuals: Rhonda Rosychuk, Jacinte Jean, Ruth Malinowski, Joan Hu, Hongmei Zhu, Andreas Sashegyi, Matt Schonlau, Katrin Rohlf, George Chen, Daniel Fong, Edmund Ng, Colin Campbell, Ker-Ai Lee, and Min Zhan.

Words cannot adequately express my gratitude to my long-suffering family. My heartfelt thanks for your unwavering support, for lifting me up in prayer, for your unending encouragement, and for celebrating with me the joys along the way!

Lastly, but above all, I want to thank my God, to whom all praise is due.

Karen A. Kopciuk

He has showed you, O man, what is good.
And what does the Lord require of you?
To act justly and to love mercy
and to walk humbly with your God.
Micah 6:8

To my parents,

Audrey and Roy Kopciuk

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Understanding the changing world around us is a longstanding preoccupation for humankind. Longitudinal studies are important research strategies, since they attempt to capture the process of change. Statistical methods used in longitudinal studies can estimate relationships between measurements made at different points in time on the same study participants, and can be used for gaining insight into underlying causal mechanisms, describing individual variation in patterns of change, and predicting future values of measured variables. The motivation for this thesis was to exploit the progressive nature of many disease processes, by explicitly incorporating this feature into statistical models for the associated longitudinal data. Uncertainty about the times of progression and the utility of intermediate information for the outcome of interest are subthemes. Hence, statistical models for progressive processes, such as chronic diseases, will be developed which exploit features of the data commonly encountered in practice and which could yield greater insights into the underlying process.

1.1 Longitudinal Studies

Longitudinal studies are employed in many disciplines, including the social and behavioural sciences, biological health sciences, and economics. These studies provide invaluable methods for studying dynamic changes over specific time periods in study participants. The most basic definition of a longitudinal study is one where more than a single observation on the same response variable is obtained from each observational unit in the sample (Lindsey, 1993). Generally, the data we consider for our multi-state modelling approach are collected from individuals, but the unit of observation need not be restricted to a person. In addition to the response of interest, explanatory variables or covariates may be measured at each observation time point for every individual. We anticipate that these covariates are related to the response and these relationships will be explored in a regression context. By taking repeated measurements on the same individuals, greater precision is anticipated and hence detection of all important effects becomes more likely. If substantial unexplained variation exists after fitting a regression model, this residual heterogeneity may be modelled using random covariates or random effects.

The key defining feature of longitudinal data is that it combines certain aspects of time series data (single response measured numerous times on one individual) with multivariate data (multiple responses measured on many individuals at a single time point). For longitudinal data, typically, the number of time points is few and the observations for any individual are correlated. Both of these features need to be taken into account when applying statistical methods to a set of longitudinal observations. The consistency of a pattern occurring across many different individuals, independent of one another, is what makes robust inference possible (Diggle, Liang and Zeger, 1994). Longitudinal studies are fraught with such difficulties as unbalanced designs, missing data, loss to follow-up, and time-dependent covariates (Ware, 1985). The major advantage of longitudinal studies, when compared to the competitor, cross-sectional studies, is that the former approach can separate changes over time within individuals from differences among people in their baseline levels (Diggle, Liang and Zeger, 1994). Even when cross-sectional studies are conducted at more than one time point, we are unable to determine precisely the nature of the change we may observe.

Longitudinal data may be collected either retrospectively or prospectively. If they are obtained retrospectively, by examining existing records for individuals or by asking them to remember previous events, missing data and inaccuracies typically occur. However, if longitudinal measurements are measured prospectively, by following individuals forward in time, these problems are generally avoided, so the quality of the data is usually superior (Diggle, Liang and Zeger, 1994).

1.2 Statistical Methods for Longitudinal Data

When adopting any statistical method for analyzing longitudinal data, we assume that our statistical model embodies the evolution of some natural process (McGarty, 1974). That is, we believe the model is a useful approximation of the real process; it should allow inference about the model parameters and prediction of the underlying process (Diggle, Liang and Zeger, 1994). Depending on whether the scientific focus concerns changes to population averages or changes to individual responses, two distinct approaches are currently being used to analyze longitudinal data: marginal models and transitional models (Kosorok and Chao, 1996). Marginal models describe the marginal expectation of the response variable, $E(Y)$, while assuming some sort of correlation structure between the repeated observed responses.

Transitional models examine changes in an individual's responses over time as a function of the previous responses and covariates; these models will be the focus of this thesis. Since this approach expresses the conditional mean as a function of the previous responses and any covariates, it is often referred to as a conditional model. In contrast with marginal models, transitional models for individuals utilize all of the information in the data. Such models for longitudinal studies are particularly suited to modelling individual changes over time and the influence of covariates on those changes (Ware and Lipsitz, 1988). The underlying time scale may be either discrete or continuous. The stochastic dependence between successive responses and between each response and any associated covariates must be explicitly specified. It is common to use multi-state models in this approach, with Markov or semi-Markov assumptions.

Multi-state models consist of a finite number of states $2, \dots, M$, where each state is defined on the basis of a measurable feature of the underlying process. For example, in studies examining the progression of cancer after treatment, states can be defined as "alive" with or without metastases, and "dead". Multi-state models have proven useful for studying chronic diseases, such as cancer, AIDS, arthritis, and diabetic retinopathy, e.g. Kay, 1986; Klein, Klotz and Grever, 1984; Frydman, 1995; Kim, De Gruttola and Lagakos, 1993; Gladman *et al.*, 1998; Andersen, 1988; Marshall and Jones, 1995, and the natural development and growth of children (Goldstein, 1979). Properties and variations of multi-state models will be discussed in greater detail in the next chapter.

1.3 Thesis Outline

We will expand some topics concerning longitudinal data analysed within a progressive three-state model framework in the remainder of this thesis.

In chapter 2, we consider three-state stochastic processes within a discrete time framework. We apply the decomposition theorem of Gu (1996) to a log likelihood developed by Frydman (1995) and show self-consistency for all three estimating equations. Two extensions to the basic model of Frydman are then considered: adding other covariates, and allowing interval censoring for both random variables. The chapter concludes with an analysis of the data set considered by Frydman and several others, illustrating our regression approach.

Chapter 3 studies the use of piecewise constant baseline hazards in the regression formulation of chapter 2. A data-driven algorithm for choosing the number and location of the breakpoints is given. The algorithm uses the Akaike and Schwarz Information Criteria to guide breakpoint decisions, as well as quantiles from the nonparametrically estimated cumulative distribution functions for the sojourn times in states 1 and 2. Standard errors are obtained for the model parameters and the model-fitting process is illustrated using the example analysed in chapter 2.

Goodness-of-link assessment for the regression model of chapter 2 is considered in chapter 4. The dependence of the intensity function on explanatory variables and the form of the link function is examined by embedding the assumed link function within a family of link functions indexed by a single parameter. The family of link functions considered for the hazard function between the last two states includes the logistic and complementary log-log specifications. A simulation study looks in detail at the logistic link specification for models often encountered in practice. The AIDS data set is used to illustrate the effects of link estimation on the regression parameters. We determine whether the self-consistent estimates identified in chapter 2 for this data set are also the maximum likelihood estimates. The uniqueness and consistency of these estimates are studied.

In chapter 5, we discuss the use of auxiliary data to improve efficiency in three-state progressive models. Parametric distributions for times spent in states 1 and 2 are adopted, as are two underlying time scales. Chronological time is modelled via a simple Markov model, and duration in state one is modelled using a hybrid time scale model. The periodic examination of individuals is varied as well, to assess the effect on efficiency of different observation schemes.

Chapter 6 identifies ideas to be explored in future research, including extensions to the basic model, other aspects of model assessment, and the use of modified information criteria for the piecewise constant approach to modelling progressive processes.

Chapter 2

Three-state Regression Models

Some common regression approaches to modelling longitudinal data when the responses are categorical include generalized estimating equations, autoregressive models, and Markov chains. After a short review of the first two approaches, the focus will shift to multi-state models which assume some type of Markov property.

Generalized estimating equations (GEEs), as the name suggests, involve working directly with the estimation step instead of with the statistical model. This approach, first proposed by Liang and Zeger (1986), incorporates stochastic dependence by directly modelling some form of covariance structure. The main advantage of GEEs is that the complete probability structure of the data need not be specified, since the marginal means and the covariance structure are modelled separately. To achieve a probabilistic interpretation, many of the GEE approaches assume a normal distribution for the response of interest, since the joint, conditional and marginal distributions all belong to the normal distribution family (Lindsey, 1993).

The autoregressive approach uses a conditional model, where the probability distribution associated with the current response depends on the previous M responses and possibly other explanatory variables. The difficulties encountered with this approach have to do with the choice of the link function used to relate the current mean response to the previous values of the response and explanatory variates, the complexity of the covariance structure, and the fact that the conditional and multivariate distributions of any non-normal responses are of different forms (families of distributions).

Perhaps the most widely adopted approach to modelling longitudinal categorical data involves the use of Markov chains (MCs). The current response is modelled conditionally on the previous responses and may also be modelled conditionally on discrete explanatory variables. When explanatory variables are included, then multiway contingency tables may be used to represent the data. Typically, log-linear and logistic models are then used to analyze the data. Properties of Markov chains that affect their complexity include the order of the chain (number of previous responses the current response is dependent on), whether it is time-homogeneous, time-reversible, and whether the equilibrium (limiting) distribution exists.

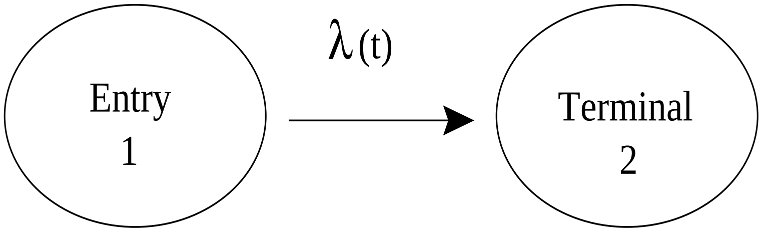
In the next section, we describe some common multi-state models for discrete-time data which adopt Markov or semi-Markov properties.

2.1 Multi-state Models

Modelling a process, such as a chronic disease, as movement through stages or states has proven to be a very useful approach. Depending on the available data and the complexity of the underlying process, multi-state models for discrete data can vary considerably in the number, types (transient or absorbing), and allowable transition directions between states. Explanatory variables can be included in most models.

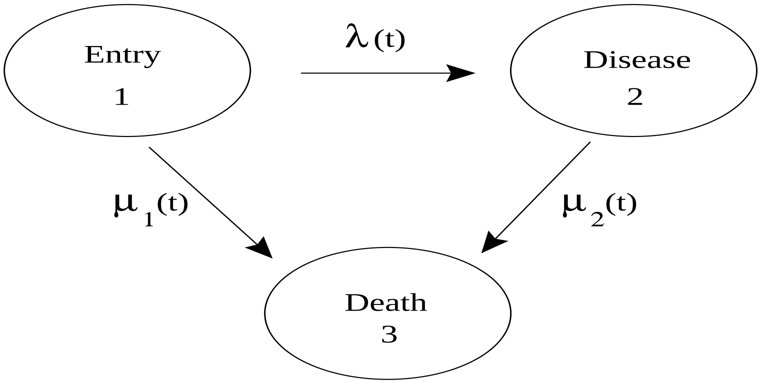
A standard model for survival data, with failure time T, is equivalent to a two-state process (Andersen, 1988). The initial state, labelled 1, corresponds to study entry and the terminal state, labelled 2, represents the event of interest. The transition from state 1 to state 2 is indicated by the directional arrow and the intensity function by $\lambda(t)$; see Figure 2.1. Casting the survival analysis problem in a counting process framework has enabled large sample properties of many estimators to be studied. Multi-state processes can also be adopted in this framework, which would allow more detailed life history data to be modelled. Properties of generalized versions of survival analysis estimators have been extensively studied as well.

Figure 2.1: Two state transition model.



A popular three-state model for life history processes is the illness-death or disability model (see Figure 2.2). Study participants would generally begin in state 1 and may be observed to make transitions to the disease state (2) or to the terminal state (3). In some settings, transitions may occur from state 2 back to state 1. For example, if states 1 and 2 were defined on the basis of levels for a serum blood marker for cancer, such as alpha-fetoprotein for hepatocelluar cancer (Kay, 1986), then transitions back to state 1 are possible when the marker level decreases.

Figure 2.2: Illness-death model.



Multi-state models may have more than one absorbing state. In a competing-risks model, a subject may move from the initial state to one of several absorbing states. They may also have more than three states. Four-state models have been used by Gladman *et al.* (1998) to model the number of damaged joints in psoriatic arthritis patients, and by Marshall and Jones (1995) to model diabetic retinopathy. Multi-state models may have any finite number of states, although the number and type will depend on the process being modelled, and the available information in the data. Two examples of complex multi-state models are given in Klein and Qian (1996), and Aalen *et al.* (1997). Klein and Qian model bone marrow transplant data in a multi-state model which includes two intermediate and two terminal states. Aalen *et al.* adopt a ten-state model to describe HIV disease progression, from HIV infection to seropositivity through to a diagnosis of AIDS.

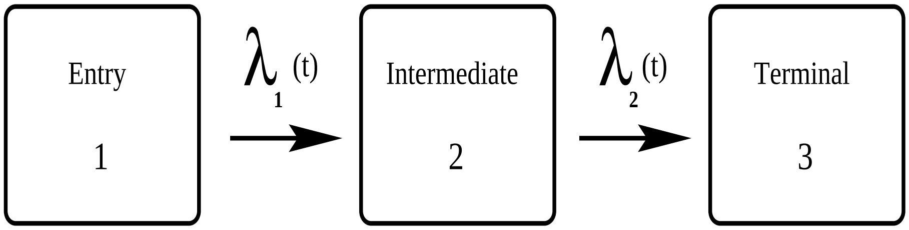
When adopting a multi-state model, the stages of the disease or process must be classified into a finite number of non-overlapping states. Usually clinical or laboratory measures determine most of the stages of the disease or process. Explanatory variables can be incorporated in several different ways, including a discrete analogue of the proportional hazards model. Terminal events, such as death or diagnosis, constitute absorbing states. Possible progression to a worse state, e.g. a higher-numbered state, and possible regression to a better, or lower-numbered state, depend on the process. The next section will describe progressive three-state models in more detail.

2.2 Progressive Three-state Models

Consider a disease or illness that becomes progressively worse in time. Examples could include cancer, where an individual may experience a recurrence of the disease and then death, or infection with the human immunodeficiency virus (HIV), where the immune system becomes more and more ineffective as the infection progresses to full-blown AIDS. In a three-state progressive model, the first two states, say 1 (entry) and 2 (intermediate), are transient states and the final state, 3 (terminal), is absorbing (see Figure 2.3). The terminal event is generally the event of interest, and the intermediate event must occur prior to the final event. A standard survival analysis may be recast in a bivariate framework and modelled as a three-state process, if an appropriate intermediate event can be identified.

Returning to the cancer example mentioned above, the entry state is defined as remission of cancer, the intermediate state is defined as the recurrence of cancer, and the final state as death from cancer. In the AIDS example, if we define the occurrence of intermediate states based on the first time the CD4 counts for individuals infected with HIV decrease to a certain level, then we can model the transition from state 1 (uninfected) through state 2 (infected) to diagnosis of AIDS (state 3) via progression through a succession of intermediate states. Statistical methods have been developed for the analysis of this special type of bivariate survival data, where the failure time of the initial response measurement represents the time origin of the second survival variable.

Figure 2.3: Three state transition model.



In their 1989 paper, De Gruttola and Lagakos proposed nonparametric methods for estimating the distribution of the time between the two events, as well as the distribution of the chronological time of the first event. They assumed the two random variables of interest were independent, and generalized the self-consistency algorithm of Turnbull (1976) in order to find the maximum likelihood estimates of the model parameters. Two generalizations of their approach that they identified for further research were to include covariates and to allow for dependency between the two survival variables. In a follow-up paper in 1993 (Kim, De Gruttola and Lagakos, 1993), the authors proposed a method for incorporating covariate information into these special, bivariate survival data. The intra-event distribution is modelled semi-parametrically, and the distribution of the first survival variable is modelled nonparametrically. The independence of the two random variables is still assumed.

Frydman (1992) also considered nonparametric estimation of bivariate survival data, but within a non-homogeneous Markov framework; hence, the independence assumption is relaxed in Frydman's work. She proposes a self-consistent algorithm for computing the estimates of the cumulative transition intensity functions. In a succeeding paper (Frydman, 1995), she extends the earlier work by modelling the second transition intensity semi-parametrically while still estimating the distribution of time to the first transition nonparametrically. However, not all of the equations involved in the likelihood-based solutions that she describes are identified as self-consistent.

We begin with a simple model of a progressive disease process that can be modelled with three states. We randomly sample n individuals from some population of interest and assume that all are observed to begin in state 1. Each person is observed at the same regular time points until he or she reaches state 3 or until the end of the study. Hence, the observation times are intended to be identical for everyone, although missing data may occur if appointments are not kept. At each visit, the information recorded is the value of the response (or the state) for each person, as well as the values of any time-dependent covariates. We will assume the response of interest has a Markov structure, and hence adopt a conditional approach to modelling the natural history of the process.

2.2.1 Notation and data

Our time origin is defined as the beginning of the study, which is a point in real time. We denote the ordered observation times for an individual as $a_0 < a_1 < a_2 <$

$\ldots < a_M < \infty$, or as the intervals

$$[a_0, a_1),\ [a_1, a_2),\ [a_2, a_3),\ \ldots\ , [a_{M-1}, a_M),\ [a_M, \infty)\ .$$

Equivalently, we may define t as a one-to-one mapping of the intervals into the non-negative integer values

$$t = 1,\ 2,\ 3, \ldots\ ,\ M,\ M+1, \qquad\qquad t = 0 = [a_0, a_1)\ .$$

Let $\{Y(t),\ t = 0, 1, 2, \ldots\ , m\}$ be a discrete-valued life history process, taking on the values corresponding to the three states (1, 2, 3). Since we are considering only progressive processes, subjects begin in state one, may only move to a higher state and, by assumption, may only make a maximum of one transition per time interval. One possible realization of this process is $\{y(0) = 1, y(1) = 1, y(2) = 2, y(3) = 2, y(4) = 2, y(5) = 3\}$, where an individual who began in state 1 at time 0 was first observed in state 2 at $t = 2$ and in state 3 at $t = 5$.

However, a more insightful way to define our process would be as the sum of two simultaneous binary processes. Thus $Y(t) = Y_1(t) + Y_2(t) + 1$ where

$$Y_1(t) = \begin{cases} 0 & \text{if in state 1,} \\ 1 & \text{otherwise,} \end{cases}$$

and

$$Y_2(t) = \begin{cases} 1 & \text{if in state 3,} \\ 0 & \text{otherwise.} \end{cases}$$

Now our realization example, $\{y(0) = 1, y(1) = 1, y(2) = 2, y(3) = 2, y(4) = 2, y(5) = 3\}$, can be written as

$$\{[y_1(0) = 0, y_2(0) = 0],\ [y_1(1) = 0, y_2(1) = 0],\ [y_1(2) = 1, y_2(2) = 0],$$
$$[y_1(3) = 1, y_2(3) = 0],\ [y_1(4) = 1, y_2(4) = 0],\ [y_1(5) = 1, y_2(5) = 1]\}\ .$$

By transforming the original process with categorical outcomes to a pair of binary processes, it is now easier to interpret it as a special case of bivariate survival data. The first process, $Y_1(t)$, can be identified with the random variable X, which is the duration in state 1 or the time to the first event. Similarly, we can associate the second process, $Y_2(t)$, with the random variable T. This random variable corresponds to the time to the second event or, equivalently, the total time on study. Another way to interpret the process $(Y_1(t), Y_2(t))$, is as a trinomial one, since there are only three possible outcomes {(0,0), (1,0), and (1,1)}.

If we let Δ and δ denote the censoring indicator functions for transition from state 1 to 2 and from state 2 to 3, respectively, then the likelihood contribution from an individual, n, can be written as

$$\begin{aligned} L_n &= P(X_n = x_n \mid X_n \geq x_n)^{\Delta_n} \, P(X_n \geq x_n) \; \times \\ &\qquad P(T_n = t_n \mid T_n \geq t_n, X_n = x_n)^{\delta_n} \, P(T_n \geq t_n \mid X_n = x_n) \, . \end{aligned}$$

Now letting $\alpha(x)$ represent the hazard function between states 1 and 2 and $\lambda(t,x)$ represent the corresponding hazard function between states 2 and 3, we can rewrite the likelihood contribution as

$$L_n = \alpha(x_n)^{\Delta_n} \prod_{0<x_u<x_n} \{1-\alpha(x_u)\} \;\; \lambda(t_n, x_n)^{\delta_n} \left[\prod_{x_n<t_r<t_n} \{1-\lambda(t_r, x_n)\} \right]^{\Delta_n} \, .$$

Since the model is unidirectional, if $\Delta_n = 0$, then we assume there is no information on T, and, hence, $\delta_n = 0$ as well. Covariates may be included by appropriately parametrizing the hazard rates. Some discrete hazard rate regression models include the logistic, proportional hazards, grouped Cox, and linear log-odds models.

2.2.2 Assumptions

To help fix ideas, let us assume that our sample of size n individuals was chosen randomly from some larger homogeneous population. We could relax this assumption by incorporating other sampling schemes that still allow independence among the selected individuals. We will further assume that each individual begins in state 1 at $t = 0$ and is subsequently followed in chronological time until either entry into state 3 occurs or the record is censored at the last observation time before the end of the study. The assumption that everyone must begin observation in state 1 at the same time could be relaxed if we consider left censoring for the time to the intermediate event or allow immigration at later time points. A maximum of one transition per individual is considered possible in each time interval. Censored or missing data are assumed to occur at the end of a given time interval.

Covariates or explanatory variables may be time-independent or fixed, such as sex, or may be time-dependent, such as age. We will assume that a value for a covariate remains fixed over each time interval but may change at the beginning of the next time interval. Time-dependent covariates are assumed to follow some deterministic function, that is, they may be reconstructed retrospectively without error. An example of this type of covariate is the age variable, which is a linear function of time. The observation scheme is noninformative in the sense of Grüger, Kay and Schumacher (1991). These noninformative observation schemes may include prestudy fixed visits (e.g., every 6 months) for every study participant, or fixed visits which may vary by participant, but are not determined by the participant.

2.3 Incomplete data

It is very common in longitudinal studies for incomplete data to arise. Such data may include observations which are censored (interval or right) or truncated (left or right). Interval censoring of transitions occurs, for example, if an individual misses scheduled appointments during which a transition occurs. Right truncation is common in studies involving patients infected with the HIV-1 virus, as there are often reporting delays for a diagnosis of AIDS. When incomplete data are observed, special methods must be used to accommodate this incompleteness. One approach would be to include only the data from individuals measured at all occasions (Fitzmaurice, Laird and Rotnitzky, 1993). Problems with this approach are inefficiency (discarding a lot of data) and the possibility of substantial bias. Another approach uses imputation to estimate the missing data from the observed data. This approach, however, is very sensitive to model misspecification. A third approach uses all of the observed data, but in order to maximize the likelihood, an expectation-maximization (EM) algorithm is adopted.

2.4 Self-consistency and the EM algorithm

Nonparametric estimation methods which involve maximizing a likelihood function are really just specializations of a more general concept, the self-consistency principle or missing-information principle (Cox and Oakes, 1984). If the observed data are incomplete in some way — censored, truncated, or grouped — then maximization in the nonparametric setting is more complicated (Laird, 1988). With these three types of incomplete data, the self-consistent estimator will be identical to the nonparametric maximum likelihood estimator (NPMLE) of a distribution function, however, in other settings, the self-consistent estimate may not be unique. The NPMLE is always self-consistent, but the converse is not always true. If the self-consistent estimator maps into the empirical distribution function of the observed data, then Tsai and Crowley (1985) show that the self-consistent estimator must also be the NPMLE.

A popular method based on the self-consistency principle, the expectation-maximization (EM) algorithm, was developed by Dempster, Laird and Rubin in 1977.

The EM algorithm is used to maximize a likelihood when the complete data has a much simpler form than the observed data. We call $\mathbf{x}$ in sample space $\mathcal{X}$ the complete data and $\mathbf{y}$ in sample space $\mathcal{Y}$ the incomplete data iff (Gu, 1996)

$$p(\mathbf{y} \mid \phi) = \int p(\mathbf{x} \mid \phi) d\mu(\mathbf{x}) \; .$$

Here $p(\mathbf{x} \mid \phi)$ is the probability density function of $\mathbf{x}$ with parameters $\phi \in \Omega$ and μ is a measure on $\mathcal{X}$ that is free of ϕ. The log likelihood of the incomplete data may be expressed as (Gu, 1996)

$$\log L(\phi') = Q(\phi' \mid \phi) - H(\phi' \mid \phi) \; ,$$

where ϕ' is any value of the parameter $\phi \in \Omega$. Let $f(\mathbf{x} \mid \mathbf{y}, \phi)$ denote the conditional probability density of $\mathbf{x}$, given $\mathbf{y}$ and ϕ; then

$$\begin{aligned} Q(\phi' \mid \phi) &= E\{\log \; p(\mathbf{x} \mid \phi') \mid \mathbf{y}, \phi\} \; , \\ H(\phi' \mid \phi) &= E\{\log \; f(\mathbf{x} \mid \mathbf{y}, \phi') \mid \mathbf{y}, \phi\} \; . \end{aligned}$$

The Q function is just the conditional expectation of the log likelihood based on the random variable X, given the observed data (Cox and Oakes, 1984). The EM algorithm hinges on this Q function, as we can show that the maximum likelihood estimator of ϕ must satisfy the self-consistency condition

$$Q(\phi' \mid \hat{\phi}) \leq Q(\hat{\phi} \mid \hat{\phi}) \; .$$

The two steps of one iteration cycle of the EM algorithm from $\phi^{(p)}$ to $\phi^{(p+1)}$ may be defined as follows:

Expectation step: Calculate $Q(\phi' \mid \phi^{(p)})$;

Maximization step: Choose a new estimate $\phi^{(p+1)}$ to be any value of $\phi' \in \Omega$ which maximizes $Q(\phi' \mid \phi^{(p)})$.

The cycle is repeated until the sequence $\phi^{(p)}$ converges to ϕ^* and $\log L(\phi^*)$ is a maximum.

2.5 Decomposing likelihoods

In the typical approach to using the EM algorithm, the complete data are constructed for each particular missing data situation. This requires an understanding of the underlying probability structure associated with the observed likelihood. However Gu (1996), in his Ph.D. thesis, showed that for likelihood functions that have a particular form, parameter estimation via the EM algorithm is possible without constructing the complete data. Hence, if the likelihood function for the observed data is decomposable as defined by Gu, a version of the EM algorithm can be developed and self-consistent parameter estimates obtained, as we outline below.

Let $L(\phi)$ be a likelihood function such that

$$\log L(\phi) = \sum_{i \in \mathbf{I}} a_i \log p_i(\phi) \ ,$$

where $a_i \geq 0$, $p_i(\phi) \geq 0$, and $\phi \in \Omega$ is a vector of unknown parameters. Suppose that $p_i(\phi)$ can be decomposed into $\{||f_j(\phi)||; j \in \mathbf{I}_i\}$ such that $p_i(\phi) = \sum_{j\in\mathbf{I}_i} ||f_j(\phi)||$, where $f_j(\phi) \geq 0$ for any j belonging to the index set $\mathbf{I_i}$. We use $||\ \ ||$ to denote a component of a decomposition of a likelihood. Hence,

$$\log L(\phi) = \sum_{i\in\mathbf{I}} a_i \log\{\sum_{j\in\mathbf{I}_i} ||f_j(\phi)||\} \ .$$

Let

$$Q(\phi'|\phi) = \sum_{i\in\mathbf{I}} a_i \sum_{j\in\mathbf{I}_i} \frac{f_j(\phi)}{\sum_{k\in\mathbf{I}_i} f_k(\phi)} \log f_j(\phi') \ ;$$

it follows that

$$\begin{aligned} H(\phi'|\phi) &= Q(\phi'|\phi) - \log L(\phi') \\ &= \sum_{i\in\mathbf{I}} a_i \sum_{j\in\mathbf{I}_i} \frac{f_j(\phi)}{\sum_{k\in\mathbf{I}_i} f_k(\phi)} \log\{f_j(\phi')/\sum_{k\in\mathbf{I}_i} f_k(\phi')\} \ . \end{aligned}$$

The monotone property of this algorithm is guaranteed by the inequality

$$H(\phi|\phi) - H(\phi'|\phi) \geq 0$$

for all $\phi, \phi' \in \Omega$.

The approach taken by Gu considers the complete data as convenient random variables, rather than as the type of information we would have preferred to record. The distribution of the complete data depends on the parameters of interest, and the log likelihood of the incomplete data is obtained from an appropriate subset.

The next example will illustrate the usefulness of this likelihood decomposition.

2.5.1 Example: Frydman's 1995 Paper

In this paper, maximum likelihood estimators are developed and a version of the EM algorithm is used to estimate the transition rates to HIV$^+$ status (state 2) and clinical symptoms of AIDS (state 3) for a group of individuals who are initially HIV$^-$ (state 1). Both transitions are assumed to follow a duration-dependent nonhomogeneous Markov process. The distribution of the time to the first event (intermediate transition) is modelled nonparametrically using the intensity function $\alpha(t)$ of the $1 \to 2$ transition; in Frydman's notation

$$p_j = \alpha(x_j) \prod_{0<u<x_j} [1 - \alpha(u)]\,, \qquad 1 \le j \le m\,,$$

and $p_{m+1} = 1 - \sum_{j=1}^{m} p_j$. The second transition intensity, $h(x,t)$, is modelled semi-parametrically via the logistic specification

$$h(x,t) = \frac{\lambda(t) e^{\beta(t-x)}}{[1 + \lambda(t) e^{\beta(t-x)}]}\,.$$

We use X_n, T_n to denote the times at which subject n enters states 2 and 3, respectively and further assume that the exact times of entry into state 2 are not observed (interval censored), but that the exact times of entry into state 3 are known or are right censored. For each individual, the observed data includes an indicator variable (Δ_n) for the transition from state 1 to state 2, the time interval in which the individual made the $1 \to 2$ transition (A_n), another indicator variable

(δ_n) for the occurrence of the transition from state 2 to state 3, and the time, t_n, of entry into state 3. If we denote the set of distinct entry times into state 3 by $T^* = \{t_k^*, 1 \leq k \leq n^*\}$, where n^* is smaller than m, the total number of observation times, and use d_k to denote the multiplicity of t_k, then Frydman identified the following three types of contributions to the likelihood:
for individuals who did not leave state 1 ($\Delta_n = 0, \delta_n = 0$), the contribution is

$$\sum_{x_j \in A_n} p_j \ .$$

Individuals who leave state 1 for state 2 ($\Delta_n = 1$) but are not observed to leave state 2 ($\delta_n = 0$) give rise to contributions of the form

$$\sum_{A_n} p_j \prod_{x_j < w \leq t_n} [1 - h(x_j, w)] \ .$$

Finally, individuals who leave state 1 for state 2 ($\Delta_n = 1$) and who are observed to occupy state 3 ($\delta_n = 1$) contribute

$$\sum_{A_n} p_j \ h(x_j, t_n) \prod_{x_j < w < t_n} [1 - h(x_j, w)] \ .$$

After considering only the set, T^*, of observed distinct times of entry into state 3 and observing that the likelihood is maximized when we set $\lambda(t) = 0$ for $t \notin T^*$ in the logistic specification for $h(x, t)$, Frydman reparametrizes the likelihood contributions. Let $h_{jk} \equiv h(x_j, t_k^*)$ for $j = 1, \ldots, m-1$, $h(x_j, t_n) = \sum_{k=1}^{n^*} h_{jk} I_{[t_n = t_k^*]}$, $\prod_{jn} = \prod_{t_k^* \in (x_j, t_n]} (1 - h_{jk})$, and $\prod_{jn}^* \equiv \prod_{t_k^* \in (x_j, t_n)} (1 - h_{jk})$. We further define

$p = (p_j, j = 1, \dots, m+1)$, $\lambda_k = \lambda(t_{k^*})$ and lastly $\lambda = (\lambda_k, 1 \leq k \leq n^*)$. If we take the sums, $\sum_{\delta_n=0}$ and $\sum_{\delta_n=1}$, to be over subjects censored or uncensored for entry into state 3 respectively, then the log likelihood can now be expressed as

$$\log \ L(\Theta) = \sum_{\delta_n=0} \log \sum_{A_n} p_j \left[\prod\nolimits_{jn}\right]^{\Delta_n} \ + \ \sum_{\delta_n=1} \log \sum_{A_n} \ p_j \ h(x_j, t_n) \prod\nolimits^*_{jn} \ .$$

If we define indicator variables, $d_{jn} = 1$, if $x_j \in A_n$ for $\delta_n = 0$ and 0 otherwise, and $b_{jn} = 1$, if $x_j \in A_n$ for $\delta_n = 1$ and 0 otherwise, we can replace the summation over all $x_j \in A_n$ by the summation over all study time points j, $j = 1, \dots, m, m+1$. We can now rewrite Frydman's log likelihood as

$$\log \ L(\Theta) = \sum_{\delta_n=0} \log \sum_{j=1}^{m+1} d_{jn} \ p_j \left[\prod\nolimits_{jn}\right]^{\Delta_n} \ + \ \sum_{\delta_n=1} \log \sum_{j=1}^{m+1} b_{jn} \ p_j \ h(x_j, t_n) \prod\nolimits^*_{jn} \ .$$

Subject to the constraints that $\sum_{j=1}^{m+1} p_j = 1$ $(0 \leq p_j \leq 1)$ and that $0 \leq h(x_j, t_n)$, we may easily decompose this version of the likelihood using the decomposition theorem of Gu as follows:

$$\log \ L(\Theta) = \sum_{\delta_n=0} \log \sum_{j=1}^{m+1} || \, d_{jn} \ p_j \left[\prod\nolimits_{jn}\right]^{\Delta_n} || \ + \ \sum_{\delta_n=1} \log \sum_{j=1}^{m+1} || \, b_{jn} \ p_j \ h(x_j, t_n) \prod\nolimits^*_{jn} || \ ,$$

where $\Theta = (p_1, \dots, p_{m+1}, \lambda_1, \dots, \lambda_{n^*}, \beta)$.

In order to find the maximum likelihood estimates, we need to define the expectation of the complete data likelihood, conditional on the incomplete (observed) data $\boldsymbol{y}$ and the current value of Θ. For $\delta_n = 0$, let $P(X = x_j, T > t_n | \phi) = f_{jn}(\phi) = d_{jn} \ p_j \left[\prod_{jn}\right]^{\Delta_n}$ and for $\delta_n = 1$, let $P(X = x_j, T = t_n | \phi) = f_{jn}(\phi) =$

$b_{jn}\ p_j\ h(x_j, t_n)\prod^*_{jn}$. Then the conditional probability density for $\boldsymbol{(X, T)}$ is given by

$$P(X = x_j, T = t_n \mid y_{jn}, \phi) = \frac{f_{jn}(\phi)}{\sum_{j=1}^{m+1} f_{jn}(\phi)} \quad .$$

It also follows that the $P(X = x_j, T = t_n|\phi') = f_{jn}(\phi')$. It is quite straightforward now to show that this conditional expectation is given by

$$\begin{aligned} Q(p', \lambda', \beta'|p, \lambda, \beta) \quad = \quad & \sum_{\delta_n=0} \sum_{j=1}^{m+1} \left[\frac{d_{jn} p_j \left[\prod_{jn}\right]^{\Delta_n}}{\sum_{l=1}^{m+1} d_{ln} p_l \left[\prod_{ln}\right]^{\Delta_n}} \right] \log \left(p'_j \ \left[\prod_{jn}^{\Delta_n} \right]' \right) + \\ & \sum_{\delta_n=1} \sum_{j=1}^{m+1} \left[\frac{b_{jn} p_j h(x_j, t_n) \prod^*_{jn}}{\sum_{l=1}^{m+1} b_{ln} p_l h(x_l, t_n) \prod^*_{ln}} \right] \log \left(p'_j \ h'(x_j, t_n) \left[\prod_{jn}^{*} \right]' \right) \ . \end{aligned}$$

By taking partial derivatives of Q with respect to p', λ', β', the parameters of interest, we derive the following expressions that will maximize Q:

$$\hat{p}_j = \left\{ \sum_{\delta_n=0} \mu_{jn}(p, \lambda, \beta) + \sum_{\delta_n=1} \mu^*_{jn}(p, \lambda, \beta) \right\} / N \qquad (1 \le j \le m+1) \ ,$$

where N is the size of the random sample. For $\hat{\lambda}_k$ we obtain the equation

$$d_k = \sum_{\Delta_n=1} \sum_{j=1}^{m+1} \left\{ \mu^{\dagger}_{jn}(p, \lambda, \beta) + \mu^*_{jn}(p, \lambda, \beta) \right\} I_{[x_j < t^*_k \le t_n]} \ h'_{jk} \qquad (1 \le k \le n^*) \ .$$

Lastly, for $\hat{\beta}$ the equation becomes

$$\begin{aligned}\sum_{\delta_n=1}\sum_{j=1}^{m+1}\mu^*_{jn}(p,\lambda,\beta)(t_n-x_j) \quad = \quad & \sum_{\Delta_n=1}\sum_{j=1}^{m+1}\left\{\mu^\dagger_{jn}(p,\lambda,\beta)+\mu^*_{jn}(p,\lambda,\beta)\right\}\times \\ & \sum_{t^*_k\in(x_j,t_n]} h'_{jk}(t^*_k-x_j)\ ,\end{aligned}$$

where

$$\mu_{jn}(p,\lambda,\beta)=\frac{d_{jn}p_j\left[\prod_{jn}\right]^{\Delta_n}}{\sum_{l=1}^{m+1}d_{ln}p_l\left[\prod_{ln}\right]^{\Delta_n}}\quad ,$$

$$\mu^*_{jn}(p,\lambda,\beta)=\frac{b_{jn}p_jh(x_j,t_n)\prod^*_{jn}}{\sum_{l=1}^{m+1}b_{ln}p_l\ h(x_l,t_n)\prod^*_{ln}}\quad ,$$

and for $\Delta_n=1,\delta_n=0$

$$\mu^\dagger_{jn}(p,\lambda,\beta)=\frac{d_{jn}p_j\prod_{jn}}{\sum_{l=1}^{m+1}d_{ln}p_l\prod_{ln}}\quad .$$

Frydman found essentially the same score equations based on a log likelihood for a discrete time, nonhomogeneous Markov chain using Lagrangian multipliers. She identifies some of her sets of equations as self-consistent, but notes that the estimating equation for β does not seem to have a self-consistent interpretation. Using the decomposition theorem of Gu, we can conclude that indeed all of these equations are self-consistent.

2.6 Extensions

If we continue with the ideas and notation developed by Frydman (1995), several extensions are natural ones to consider. In all of these extensions, the second transition intensity is modelled as still depending on both chronological time and duration in state 2. The extensions involve modelling other covariates (both time-dependent and time-independent), and allowing the second transition time to be possibly interval-censored as well as right-censored.

2.6.1 Covariates

2.6.1.1 Time-independent Covariates

In this extension, both transition intensities are modelled semi-parametrically as functions of time-independent (fixed), external covariates. Define the covariate process to be $Z' = (z_1, z_2, \ldots, z_p)$. The logistic specifications for both transition intensities are

$$\alpha(x;z) = \frac{\eta(x)\, e^{z'\boldsymbol{\gamma}}}{1+\eta(x)\, e^{z'\boldsymbol{\gamma}}} \ ,$$

and

$$h(x,t;z) = \frac{\lambda(t)\, e^{\beta(t-x)+z'\boldsymbol{\nu}}}{1+\lambda(t)\, e^{\beta(t-x)+z'\boldsymbol{\nu}}} \ .$$

If we denote the censoring time or equivalently the first possible transition time for individual n in interval A_n as x_l^n, the log likelihood of the observed data becomes

$$
\begin{aligned}
\log\ L(\Theta) &= \sum_{\Delta_n=0,\delta_n=0} \log \prod_{0<u<x_l^n} [1-\alpha(u;z_n)] \\
&+ \sum_{\Delta_n=1,\delta_n=0} \log \sum_{A_n} \alpha(x_j;z_n) \prod_{0<u<x_j} [1-\alpha(u;z_n)] \prod_{jn}(z_n) \\
&+ \sum_{\delta_n=1} \log \sum_{A_n} \alpha(x_j;z_n) \prod_{0<u<x_j} [1-\alpha(u;z_n)]\ h(x_j,t_n;z_n) \prod_{jn}^{*}(z_n)\ .
\end{aligned}
$$

We define the indicator variables d_{jn} and b_{jn} as before, redefine $\Theta = \{\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \beta\}$, and let $\tilde{\prod}_j(z_n) = \prod_{0<u<x_j}[1-\alpha(u;z_n)]$. Once again, if we only consider the set, T^*, of observed distinct times of entry into state 3, then when we decompose the log likelihood, we obtain the corresponding $Q(\Theta'|\Theta)$ function

$$
\begin{aligned}
Q(\Theta'|\Theta) &= \sum_{\Delta_n=0,\delta_n=0} \sum_{u=x_1}^{x_{l-1}^n} \log [1-\alpha'(u;z_n)] \\
&+ \sum_{\Delta_n=1,\delta_n=0} \sum_{j=1}^{m+1} \left[\frac{d_{jn}\ \alpha(x_j;z_n) \tilde{\prod}_j(z_n) \prod_{jn}(z_n)}{\sum_{l=1}^{m+1} d_{ln}\ \alpha(x_l;z_n) \tilde{\prod}_l(z_n) \prod_{ln}(z_n)} \right] \times \\
&\quad \log \left(\alpha'(x_j;z_n)\ \left[\tilde{\prod}_j(z_n) \right]' \left[\prod_{jn}(z_n) \right]' \right) \\
&+ \sum_{\delta_n=1} \sum_{j=1}^{m+1} \left[\frac{b_{jn}\ \alpha(x_j;z_n) \tilde{\prod}_j(z_n)\, h(x_j,t_n;z_n) \prod_{jn}^*(z_n)}{\sum_{l=1}^{m+1} b_{ln}\ \alpha(x_l;z_n) \tilde{\prod}_l(z_n)\, h(x_l,t_n;z_n) \prod_{ln}^*(z_n)} \right] \times \\
&\quad \log \left(\alpha'(x_j;z_n)\ \left[\tilde{\prod}_j(z_n) \right]' h'(x_j,t_n;z_n) \left[\prod_{jn}^{*}(z_n) \right]' \right)\ .
\end{aligned}
$$

In order to maximize this likelihood, we take partial derivatives of $Q(\Theta'|\Theta)$ with respect to Θ'. We first compute the partial derivative of Q with respect to the chronological time factor, $\eta(x)$, of $\alpha(x; z)$. If we define $\eta(x_j) = \eta_j$, then for $1 \leq j \leq m$

$$\begin{aligned}
\frac{\partial Q}{\partial \eta'_j} &= \sum_{\Delta_n=0,\delta_n=0} \frac{\partial}{\partial \eta'_j} \sum_{u=x_1}^{x^n_{l-1}} \log\left[1 - \alpha'(u; z_n)\right] \\
&+ \sum_{\Delta_n=1,\delta_n=0} \mu^{\dagger}_{jn}(\Theta) \frac{\frac{\partial}{\partial \eta'_j} \alpha'(x_j; z_n)}{\alpha'(x_j; z_n)} \\
&+ \sum_{\Delta_n=1,\delta_n=0} \sum_{r=j+1}^{m+1} \mu^{\dagger}_{rn}(\Theta) \frac{\frac{\partial}{\partial \eta'_j} \tilde{\prod}'_r(z_n)}{\tilde{\prod}'_r(z_n)} \\
&+ \sum_{\delta_n=1} \mu^{*}_{jn}(\Theta) \frac{\frac{\partial}{\partial \eta'_j} \alpha'(x_j; z_n)}{\alpha'(x_j; z_n)} \\
&+ \sum_{\delta_n=1} \sum_{r=j+1}^{m+1} \mu^{*}_{rn}(\Theta) \frac{\frac{\partial}{\partial \eta'_j} \tilde{\prod}'_r(z_n)}{\tilde{\prod}'_r(z_n)} \ ,
\end{aligned}$$

where we continue to use the expressions $\mu^{\dagger}_{jn}$ and μ^{*}_{jn}, which are defined now, to show their dependence on covariates, as

$$\mu^{\dagger}_{jn}(\Theta) = \frac{d_{jn}\, \alpha(x_j; z_n) \tilde{\prod}_j(z_n) \prod_{jn}(z_n)}{\sum_{l=1}^{m+1} d_{ln}\, \alpha(x_l; z_n) \tilde{\prod}_l(z_n) \prod_{ln}(z_n)} \ ,$$

$$\mu^{*}_{jn}(\Theta) = \frac{b_{jn}\, \alpha(x_j; z_n) \tilde{\prod}_j(z_n)\, h(x_j, t_n; z_n) \prod^{*}_{jn}(z_n)}{\sum_{l=1}^{m+1} b_{ln}\, \alpha(x_l; z_n) \tilde{\prod}_l(z_n)\, h(x_l, t_n; z_n) \prod^{*}_{ln}(z_n)} \ .$$

But, if we define the indicator variable, $c_{jn} = 1$, if $x_j < x^n_l$ for $\Delta_n = 0$ and 0 otherwise, then

$$\begin{aligned}
\frac{\partial}{\partial \eta'_j} \sum_{u=x_1}^{x^n_{l-1}} \log\left[1-\alpha'(u;z_n)\right] &= -\, c_{jn}\, \alpha'(x_j;z_n)/\eta'_j \,, \\
\frac{\partial}{\partial \eta'_j} \alpha'(x_j;z_n) &= \alpha'(x_j;z_n)\left[1-\alpha'(x_j;z_n)\right]/\eta'_j \,, \\
\frac{\partial}{\partial \eta'_j} \tilde{\prod}'_r(z_n) &= -\tilde{\prod}'_r(z_n)\, \alpha'(x_j;z_n)/\eta'_j \,.
\end{aligned}$$

Hence, the partial derivative of Q with respect to η'_j becomes

$$\begin{aligned}
\frac{\partial Q}{\partial \eta'_j} = & - \sum_{\Delta_n=0,\delta_n=0} c_{jn}\, \alpha'(x_j;z_n)/\eta'_j \\
& + \sum_{\Delta_n=1,\delta_n=0} \mu^{\dagger}_{jn}(\Theta)\left[1-\alpha'(x_j;z_n)\right]/\eta'_j \\
& - \sum_{\Delta_n=1,\delta_n=0} \sum_{r=j+1}^{m+1} \mu^{\dagger}_{rn}(\Theta)\alpha'(x_j;z_n)/\eta'_j \\
& + \sum_{\delta_n=1} \mu^{*}_{jn}(\Theta)\left[1-\alpha'(x_j;z_n)\right]/\eta'_j \\
& - \sum_{\delta_n=1} \sum_{r=j+1}^{m+1} \mu^{*}_{rn}(\Theta)\alpha'(x_j;z_n)/\eta'_j \,.
\end{aligned}$$

Now if we set these equations to zero, split the sums involving $[1-\alpha'(x_j;z_n)]$ and collect like terms, the equation which yields the MLE for η_j reduces to

$$\begin{aligned}
& \sum_{\Delta_n=1} \left\{ \mu^{\dagger}_{jn}(\Theta) + \mu^{*}_{jn}(\Theta) \right\} = \sum_{\Delta_n=0,\delta_n=0} c_{jn}\, \alpha'(x_j;z_n) + \\
& \qquad \sum_{\Delta_n=1,\delta_n=0} \alpha'(x_j;z_n) \sum_{r=j}^{m+1} \mu^{\dagger}_{rn}(\Theta) + \sum_{\delta_n=1} \alpha'(x_j;z_n) \sum_{r=j}^{m+1} \mu^{*}_{rn}(\Theta) \qquad (1 \le j \le m)\,.
\end{aligned}$$

We next differentiate Q with respect to the unknown coefficients, $\boldsymbol{\gamma}$, of the covariate process, Z, which appear in the linear predictor of the first transition intensity $\alpha(x; z)$. For element s of the vector, $\boldsymbol{\gamma}$ $(1 \leq s \leq p)$, the partial derivative of Q with respect to γ'_s is given by

$$\begin{aligned}
\frac{\partial Q}{\partial \gamma'_s} &= \sum_{\Delta_n=0, \delta_n=0} \sum_{u=x_1}^{x^n_{l-1}} \frac{\partial}{\partial \gamma'_s} \log[1 - \alpha'(u; z_n)] \\
&+ \sum_{\Delta_n=1, \delta_n=0} \sum_{j=1}^{m+1} \mu^{\dagger}_{jn}(\Theta) \frac{\frac{\partial}{\partial \gamma'_s} \alpha'(x_j; z_n)}{\alpha'(x_j; z_n)} \\
&+ \sum_{\Delta_n=1, \delta_n=0} \sum_{j=1}^{m+1} \mu^{\dagger}_{jn}(\Theta) \frac{\frac{\partial}{\partial \gamma'_s} \tilde{\prod}'_j(z_n)}{\tilde{\prod}'_j(z_n)} \\
&+ \sum_{\delta_n=1} \sum_{j=1}^{m+1} \mu^{*}_{jn}(\Theta) \frac{\frac{\partial}{\partial \gamma'_s} \alpha'(x_j; z_n)}{\alpha'(x_j; z_n)} \\
&+ \sum_{\delta_n=1} \sum_{j=1}^{m+1} \mu^{*}_{jn}(\Theta) \frac{\frac{\partial}{\partial \gamma'_s} \tilde{\prod}'_j(z_n)}{\tilde{\prod}'_j(z_n)} \, .
\end{aligned}$$

However,

$$\begin{aligned}
\frac{\partial}{\partial \gamma'_s} \log(1 - \alpha'(u; z_n)) &= -\, z_{ns}\, \alpha'(u; z_n)\,, \\
\frac{\partial}{\partial \gamma'_s} \alpha'(x_j; z_n) &= \quad z_{ns}\, \alpha'(x_j; z_n)\ [1 - \alpha'(x_j; z_n)]\ , \\
\frac{\partial}{\partial \gamma'_s} \tilde{\prod}'_j(z_n) &= -\, z_{ns} \tilde{\prod}'_j(z_n) \sum_{u=x_1}^{x_{j-1}} \alpha'(u; z_n)\,,
\end{aligned}$$

where z_{ns} corresponds to element s in the covariate vector, z_n, for individual n.

Thus, the partial derivative of Q with respect to γ'_s becomes

$$\begin{aligned}
\frac{\partial Q}{\partial \gamma'_s} = & - \sum_{\Delta_n=0,\delta_n=0} z_{ns} \sum_{u=x_1}^{x^n_{l-1}} \alpha'(u;z_n) \\
& + \sum_{\Delta_n=1,\delta_n=0} \sum_{j=1}^{m+1} \mu^{\dagger}_{jn}(\Theta)\, z_{ns}\, [1-\alpha'(x_j;z_n)] \\
& - \sum_{\Delta_n=1,\delta_n=0} \sum_{j=1}^{m+1} \mu^{\dagger}_{jn}(\Theta)\, z_{ns} \sum_{u=x_1}^{x_{j-1}} \alpha'(u;z_n) \\
& + \sum_{\delta_n=1} \sum_{j=1}^{m+1} \mu^{*}_{jn}(\Theta)\, z_{ns}\, [1-\alpha'(x_j;z_n)] \\
& - \sum_{\delta_n=1} \sum_{j=1}^{m+1} \mu^{*}_{jn}(\Theta)\, z_{ns} \sum_{u=x_1}^{x_{j-1}} \alpha'(u;z_n)\ .
\end{aligned}$$

Once again, to find a turning point we set $\partial Q/\partial\gamma'_s$ equal to zero, split the sums involving $[1-\alpha'(x_j;z_n)]$, use the result that $\sum_{j=1}^{m+1}\mu^{\dagger}_{jn}(\Theta) = \sum_{j=1}^{m+1}\mu^{*}_{jn}(\Theta) = 1$, and collect like terms. The resulting equation for $\hat{\gamma}_s$ is

$$\begin{aligned}
\sum_{\Delta_n=1} z_{ns} = & \sum_{\Delta_n=0} z_{ns} \sum_{u=x_1}^{x^n_{l-1}} \alpha'(u;z_n) \\
& + \sum_{\Delta_n=1} z_{ns} \sum_{j=1}^{m+1} \left\{\mu^{\dagger}_{jn}(\Theta) + \mu^{*}_{jn}(\Theta)\right\} \sum_{u=x_1}^{x_j} \alpha'(u;z_n) \qquad (1 \le s \le p)\ .
\end{aligned}$$

Now, turning to the parameters in the second transition intensity, we begin by computing the partial derivative of Q with respect to $\lambda(t)$, the chronological time factor in $h(x,t)$. If we continue to maximize over the set, T^*, of observed distinct times of entry into state 3, we may define $\lambda(t_k^*) = \lambda_k$; consequently,

$$\begin{aligned}
\frac{\partial Q}{\partial \lambda_k'} &= \sum_{\Delta_n=1,\delta_n=0} \sum_{j=1}^{m+1} \mu_{jn}^{\dagger}(\Theta) \frac{\frac{\partial}{\partial \lambda_k'} \prod_{jn}'(z_n)}{\prod_{jn}'(z_n)} \\
&+ \sum_{\delta_n=1} \sum_{j=1}^{m+1} \mu_{jn}^{*}(\Theta) \frac{\frac{\partial}{\partial \lambda_k'} h'(x_j, t_n; z_n)}{h'(x_j, t_n; z_n)} \\
&+ \sum_{\delta_n=1} \sum_{j=1}^{m+1} \mu_{jn}^{*}(\Theta) \frac{\frac{\partial}{\partial \lambda_k'} \prod^{*\prime}{}_{jn}(z_n)}{\prod^{*\prime}{}_{jn}(z_n)} \; .
\end{aligned}$$

But,

$$\begin{aligned}
\frac{\partial}{\partial \lambda_k'} \prod_{jn}{}'(z_n) &= -\prod_{jn}{}'(z_n) I_{[x_j<t_k^*\leq t_n]} h_{jk}'(z_n)/\lambda_k' \; , \\
\frac{\partial}{\partial \lambda_k'} \prod_{jn}{}^{*\;\prime}(z_n) &= -\prod_{jn}{}^{*\;\prime}(z_n) I_{[x_j<t_k^*<t_n]} h_{jk}'(z_n)/\lambda_k' \; , \\
\frac{\partial}{\partial \lambda_k'} h'(x_j, t_n; z_n) &= I_{[t_n=t_k^*]} \left[1 - h_{jk}'(z_n)\right] h_{jk}'(z_n)/\lambda_k' \; .
\end{aligned}$$

Therefore, the partial derivative of Q with respect to λ_k' becomes

$$\begin{aligned}
\frac{\partial Q}{\partial \lambda_k'} = &- \sum_{\Delta_n=1,\delta_n=0} \sum_{j=1}^{m+1} \mu_{jn}^{\dagger}(\Theta) I_{[x_j<t_k^*\leq t_n]} h_{jk}'(z_n)/\lambda_k' \\
&+ \sum_{\delta_n=1,t_n=t_k^*} \sum_{j=1}^{m+1} \mu_{jn}^{*}(\Theta) \left[1 - h_{jk}'(z_n)\right]/\lambda_k' \\
&- \sum_{\delta_n=1} \sum_{j=1}^{m+1} \mu_{jn}^{*}(\Theta) I_{[x_j<t_k^*<t_n]} h_{jk}'(z_n)/\lambda_k' \; .
\end{aligned}$$

If we set this equation to zero, split the sums involving $\left[1 - h_{jk}'(z_n)\right]$, and use

d_k as the multiplicity of t^*_k as before, we get an estimating equation that yields the MLE of λ'_k, viz.

$$d_k = \sum_{\Delta_n=1} \sum_{j=1}^{m+1} \left\{ \mu^{\dagger}_{jn}(\Theta) + \mu^{*}_{jn}(\Theta) \right\} I_{[x_j < t^*_k \le t_n]} \, h'_{jk}(z_n) \qquad (1 \le k \le n^*) \; .$$

Likewise, if we differentiate Q with respect to β', the unknown coefficient of the covariate $(t - x)$, we find

$$\begin{aligned} \frac{\partial Q}{\partial \beta'} &= \sum_{\Delta_n=1,\delta_n=0} \sum_{j=1}^{m+1} \mu^{\dagger}_{jn}(\Theta) \frac{\frac{\partial}{\partial \beta'} \prod'_{jn}(z_n)}{\prod'_{jn}(z_n)} \\ &+ \sum_{\delta_n=1} \sum_{j=1}^{m+1} \mu^{*}_{jn}(\Theta) \frac{\frac{\partial}{\partial \beta'} h'(x_j, t_n; z_n)}{h'(x_j, t_n; z_n)} \\ &+ \sum_{\delta_n=1} \sum_{j=1}^{m+1} \mu^{*}_{jn}(\Theta) \frac{\frac{\partial}{\partial \beta'} \prod^{*\prime}_{jn}(z_n)}{\prod^{*\prime}_{jn}(z_n)} \; . \end{aligned}$$

But,

$$\begin{aligned} \frac{\partial}{\partial \beta'} \prod_{jn}{}'(z_n) &= -\prod_{jn}{}'(z_n) \sum_{t^*_k \in (x_j, t_n]} (t^*_k - x_j) h'_{jk}(z_n) \; , \\ \frac{\partial}{\partial \beta'} \prod_{jn}{}^{*\,\prime}(z_n) &= -\prod_{jn}{}^{*\,\prime}(z_n) \sum_{t^*_k \in (x_j, t_n)} (t^*_k - x_j) h'_{jk}(z_n) \; , \\ \frac{\partial}{\partial \beta'} h'(x_j, t_n; z_n) &= \sum_{k=1}^{n^*} I_{[t_n = t^*_k]} \left[1 - h'_{jk}(z_n) \right] h'_{jk}(z_n) (t^*_k - x_j) \; . \end{aligned}$$

Therefore,

$$\begin{aligned}\frac{\partial Q}{\partial \beta'} &= -\sum_{\Delta_n=1,\delta_n=0}\sum_{j=1}^{m+1}\mu^{\dagger}_{jn}(\Theta)\sum_{t^*_k\in(x_j,t_n]}(t^*_k - x_j)h'_{jk}(z_n)\\ &\quad+\sum_{\delta_n=1}\sum_{j=1}^{m+1}\mu^{*}_{jn}(\Theta)(t_n - x_j) - \sum_{\delta_n=1}\sum_{j=1}^{m+1}\mu^{*}_{jn}(\Theta)h'(x_j,t_n;z_n)(t_n - x_j)\\ &\quad-\sum_{\delta_n=1}\sum_{j=1}^{m+1}\mu^{*}_{jn}(\Theta)\sum_{t^*_k\in(x_j,t_n)}(t^*_k - x_j)h'_{jk}(z_n)\ .\end{aligned}$$

If we equate this derivative to zero and collect like terms, we get the following estimating equation for β

$$\sum_{\delta_n=1}\sum_{j=1}^{m+1}\mu^{*}_{jn}(\Theta)(t_n - x_j) = \sum_{\Delta_n=1}\sum_{j=1}^{m+1}\left\{\mu^{\dagger}_{jn}(\Theta)+\mu^{*}_{jn}(\Theta)\right\}\sum_{t^*_k\in(x_j,t_n]}h'_{jk}(z_n)(t^*_k - x_j)\ .$$

Lastly, we differentiate Q with respect to the vector of regression coefficients, $\boldsymbol{\nu}$, of the covariate process Z. For element s of $\boldsymbol{\nu}$ $(1 \leq s \leq p)$, we obtain

$$\begin{aligned}\frac{\partial Q}{\partial \nu'_s} = &\sum_{\Delta_n=1,\delta_n=0}\sum_{j=1}^{m+1}\mu^{\dagger}_{jn}(\Theta)\frac{\frac{\partial}{\partial \nu'_s}\prod'_{jn}(z_n)}{\prod'_{jn}(z_n)}\\ &+\sum_{\delta_n=1}\sum_{j=1}^{m+1}\mu^{*}_{jn}(\Theta)\frac{\frac{\partial}{\partial \nu'_s}h'(x_j,t_n;z_n)}{h'(x_j,t_n;z_n)}\\ &+\sum_{\delta_n=1}\sum_{j=1}^{m+1}\mu^{*}_{jn}(\Theta)\frac{\frac{\partial}{\partial \nu'_s}\prod^{*\prime}{}_{jn}(z_n)}{\prod^{*\prime}{}_{jn}(z_n)}\ .\end{aligned}$$

But,

$$
\begin{aligned}
\frac{\partial}{\partial \nu'_s} \prod_{jn}{}' (z_n) &= -\prod_{jn}{}' (z_n)\, z_{ns} \sum_{t_k * \in (x_j, t_n]} h'_{jk}(z_n)\,, \\
\frac{\partial}{\partial \nu'_s} \prod_{jn}{}^{*\,\prime} (z_n) &= -\prod_{jn}{}^{*\,\prime} (z_n)\, z_{ns} \sum_{t_k * \in (x_j, t_n)} h'_{jk}(z_n)\,, \\
\frac{\partial}{\partial \nu'_s} h'(x_j, t_n; z_n) &= \sum_{k=1}^{n^*} z_{ns}\, I_{[t_n = t_k^*]} \left[1 - h'_{jk}(z_n)\right] h'_{jk}(z_n)\,.
\end{aligned}
$$

Thus, $\partial Q / \partial \nu'_s$ becomes

$$
\begin{aligned}
\frac{\partial Q}{\partial \nu'_s} = \quad & - \sum_{\Delta_n = 1, \delta_n = 0} z_{ns} \sum_{j=1}^{m+1} \mu^{\dagger}_{jn}(\Theta) \sum_{t_k * \in (x_j, t_n]} h'_{jk}(z_n) \\
& + \sum_{\delta_n = 1} z_{ns} \sum_{j=1}^{m+1} \mu^{*}_{jn}(\Theta) - \sum_{\delta_n = 1} z_{ns} \sum_{j=1}^{m+1} \mu^{*}_{jn}(\Theta) h'_{jn}(z_n) \\
& - \sum_{\delta_n = 1} z_{ns} \sum_{j=1}^{m+1} \mu^{*}_{jn}(\Theta) \sum_{t_k * \in (x_j, t_n)} h'_{jk}(z_n)\,.
\end{aligned}
$$

If we equate this score function to zero and use the fact that $\sum_{j=1}^{m+1} \mu^{*}_{jn}(\Theta) = 1$, we get the estimating equation for the MLE of ν_s, viz.

$$
\sum_{\delta_n = 1} z_{ns} = \sum_{\Delta_n = 1} z_{ns} \sum_{j=1}^{m+1} \left\{ \mu^{\dagger}_{jn}(\Theta) + \mu^{*}_{jn}(\Theta) \right\} \sum_{t_k * \in (x_j, t_n]} h'_{jk}(z_n) \qquad (1 \leq s \leq p)\,.
$$

2.6.1.2 Time-dependent Covariates

In this extension, both transition intensities are modelled semi-parametrically as functions of time-dependent external covariates. The covariate process becomes $Z'(t) = \{z_1(t), z_2(t), \ldots, z_p(t)\}$. We assume that some of the covariates could be time independent and that the time-dependent covariates follow a deterministic function. The only changes to the logistic specifications for both transition intensities involve the parameters, $\boldsymbol{\gamma}$ and $\boldsymbol{\nu}$. Specifically

$$\alpha\{x; z(x)\} = \frac{\eta(x)\, e^{z'(x)\boldsymbol{\gamma}}}{1 + \eta(x)\, e^{z'(x)\boldsymbol{\gamma}}} \ ,$$

and

$$h\{x, t; z(t)\} = \frac{\lambda(t)\, e^{\beta(t-x)+z'(t)\boldsymbol{\nu}}}{1 + \lambda(t)\, e^{\beta(t-x)+z'(t)\boldsymbol{\nu}}} \ .$$

Let the indicator variables d_{jn} and b_{jn} be defined as before. We see that this log likelihood, too, may be decomposed, since

$$\begin{aligned} \log\ L(\Theta) \ &= \sum_{\Delta_n=0,\delta_n=0} \log \,||\prod_{0<u<x_l^n} [1 - \alpha\{u; z_n(u)\}]\,|| \\ &+ \sum_{\Delta_n=1,\delta_n=0} \log \sum_{j=1}^{m+1} ||\, d_{jn}\, \alpha\{x_j; z_n(x_j)\} \tilde{\prod_{0<u<x_j}} \{z_n(u)\} \prod_{x_j<w\leq t_n} \{z_n(w)\}\,|| \\ &+ \sum_{\delta_n=1} \log \sum_{j=1}^{m+1} ||\, b_{jn}\, \alpha\{x_j; z_n(x_j)\} \tilde{\prod_{0<u<x_j}} \{z_n(u)\}\, h\{x_j, t_n; z_n(t_n)\} \times \\ &\quad \prod_{x_j<w<t_n}^{*} \{z_n(w)\}\,|| \ . \end{aligned}$$

Once again we only consider the set, T^*, of observed distinct times of entry into state 3. The other expressions are now updated to show their dependence on time-dependent covariates: $\tilde{\prod}_j\{z_n(u)\} = \prod_{0<u<x_j} [1 - \alpha\{u; z_n(u)\}]$, $h_{jk}\{z_n(t_k^*)\} \equiv h\{x_j, t_k^*; z_n(t_k^*)\}$ for $(j = 1, \ldots, m-1)$, $h\{x_j, t_n; z_n(t_n)\} = \sum_{k=1}^{n^*} h_{jk}\{z_n(t_k^*)\} I_{[t_n=t_k^*]}$,

$\prod_{jn}\{z_n(t_k^*)\} = \prod_{t_k^* \in (x_j,t_n]}[1-h_{jk}\{z_n(t_k^*)\}]$, $\prod_{jn}^*\{z_n(t_k^*)\} \equiv \prod_{t_k^* \in (x_j,t_n)}[1-h_{jk}\{z_n(t_k^*)\}]$. When we decompose the log likelihood, we obtain the function

$$
\begin{aligned}
Q(\Theta'|\Theta) = & \sum_{\Delta_n=0,\delta_n=0} \sum_{u=x_1}^{x_{l-1}^n} \log\left[1-\alpha'\{u; z_n(u)\}\right] \\
& + \sum_{\Delta_n=1,\delta_n=0} \sum_{j=1}^{m+1} \left[\frac{d_{jn}\, \alpha\{x_j; z_n(x_j)\} \tilde{\prod}_j\{z_n(u)\} \prod_{jn}\{z_n(t_k^*)\}}{\sum_{l=1}^{m+1} d_{ln}\, \alpha\{x_l; z_n(x_l)\} \tilde{\prod}_l\{z_n(u)\} \prod_{ln}\{z_n(t_k^*)\}} \right] \times \\
& \log\left(\alpha'\{x_j; z_n(x_j)\} \left[\tilde{\prod}_j\{z_n(u)\} \right]' \left[\prod_{jn}\{z_n(t_k^*)\} \right]' \right) \\
& + \sum_{\delta_n=1} \sum_{j=1}^{m+1} \left[\frac{b_{jn}\, \alpha\{x_j; z_n(x_j)\} \tilde{\prod}_j\{z_n(u)\} h\{x_j, t_n; z_n(t_n)\} \prod_{jn}^*\{z_n(t_k^*)\}}{\sum_{l=1}^{m+1} b_{ln}\, \alpha\{x_l; z_n(x_l)\} \tilde{\prod}_l\{z_n(x_l)\} h\{x_l, t_n; z_n(t_n)\} \prod_{ln}^*\{z_n(t_k^*)\}} \right] \times \\
& \log\left(\alpha'\{x_j; z_n(x_j)\} \left[\tilde{\prod}_j\{z_n(u)\} \right]' h'\{x_j, t_n; z_n(t_n)\} \left[\prod_{jn}^*\{z_n(t_k^*)\} \right]' \right) .
\end{aligned}
$$

The estimators for the parameters in Θ in this time-dependent case are very similar to their time-independent covariate counterparts. The main differences appear in the terms which depend on the time-dependent covariates, specifically in the intensities (α and h) and the conditional probability functions

$$
\mu_{jn}^{\dagger}(\Theta) = \frac{d_{jn}\, \alpha\{x_j; z_n(x_j)\} \tilde{\prod}_j\{z_n(u)\} \prod_{jn}\{z_n(t_k^*)\}}{\sum_{l=1}^{m+1} d_{ln}\, \alpha\{x_l; z_n(x_l)\} \tilde{\prod}_l\{z_n(u)\} \prod_{ln}\{z_n(t_k^*)\}} \quad ,
$$

$$
\mu_{jn}^{*}(\Theta) = \frac{b_{jn}\, \alpha\{x_j; z_n(x_j)\} \tilde{\prod}_j\{z_n(u)\}\, h\{x_j, t_n; z_n(t_n)\} \prod_{jn}^*\{z_n(t_k^*)\}}{\sum_{l=1}^{m+1} b_{ln}\, \alpha\{x_l; z_n(x_l)\} \tilde{\prod}_l\{z_n(u)\}\, h\{x_l, t_n; z_n(t_n)\} \prod_{ln}^*\{z_n(t_k^*)\}} \quad .
$$

For $\hat{\eta}_j$, the estimating equation becomes

$$\sum_{\Delta_n=1} \left\{\mu^{\dagger}_{jn}(\Theta) + \mu^{*}_{jn}(\Theta)\right\} = \sum_{\Delta_n=0,\delta_n=0} c_{jn}\ \alpha'\{x_j; z_n(x_j)\}$$
$$+ \sum_{\Delta_n=1,\delta_n=0} \alpha'\{x_j; z_n(x_j)\} \sum_{r=j}^{m+1} \mu^{\dagger}_{rn}(\Theta) + \sum_{\delta_n=1} \alpha'\{x_j; z_n(x_j)\} \sum_{r=j}^{m+1} \mu^{*}_{rn}(\Theta)\ ,$$
$$(1 \le j \le m)\ .$$

For $\hat{\gamma}_s$ we obtain the equation

$$\sum_{\Delta_n=1} \sum_{j=1}^{m+1} \left\{\mu^{\dagger}_{jn}(\Theta) + \mu^{*}_{jn}(\Theta)\right\} z_{ns}(x_j) = \sum_{\Delta_n=0} \sum_{u=x_1}^{x^n_{l-1}} \alpha'\{u; z_n(u)\}\ z_{ns}(u)$$
$$+ \sum_{\Delta_n=1} \sum_{j=1}^{m+1} \left\{\mu^{\dagger}_{jn}(\Theta) + \mu^{*}_{jn}(\Theta)\right\} \sum_{u=x_1}^{x_j} \alpha'\{u; z_n(u)\}\ z_{ns}(u)\ , \qquad (1 \le s \le p)\ .$$

For $\hat{\lambda}_k$ we get the equation

$$d_k = \sum_{\Delta_n=1} \sum_{j=1}^{m+1} \left\{\mu^{\dagger}_{jn}(\Theta) + \mu^{*}_{jn}(\Theta)\right\} I_{[x_j < t^*_k \le t_n]}\ h'_{jk}\{z_n(t^*_k)\}\ , \qquad (1 \le k \le n^*)\ .$$

For $\hat{\beta}$ we have, as before,

$$\sum_{\delta_n=1} \sum_{j=1}^{m+1} \mu^{*}_{jn}(\Theta)(t_n - x_j) = \sum_{\Delta_n=1} \sum_{j=1}^{m+1} \left\{\mu^{\dagger}_{jn}(\Theta) + \mu^{*}_{jn}(\Theta)\right\} \sum_{t^*_k \in (x_j, t_n]} h'_{jk}\{z_n(t^*_k)\}(t^*_k - x_j)\ .$$

And lastly, for $\hat{\nu}_s$ we obtain the estimating equation

$$\sum_{\delta_n=1} z_{ns}(t_n) = \sum_{\Delta_n=1} \sum_{j=1}^{m+1} \left\{\mu^{\dagger}_{jn}(\Theta) + \mu^{*}_{jn}(\Theta)\right\} \sum_{t_k * \in (x_j, t_n]} z_{ns}(t^*_k)\, h'_{jk}\{z_n(t^*_k)\} \qquad (1 \leq s \leq p)\ .$$

2.6.2 T Interval Censored

Another possible extension is to allow the time to the second event, T in this setting, to be interval-censored, in addition to the time to the first event (X). This could happen, say, if the final absorbing state represents diagnosis of a disease or disorder, but would be unlikely if the event is death. The latter type of event could be affected by reporting delays, which results in truncation of the observed data. This particular complication will be the focus of another extension to Frydman's model, which will be considered in the future. For T to be interval-censored, we make the assumption that the missing observation of a possibly true transition time is not related to the underlying process (i.e., too sick to come in for an appointment). Otherwise, the censoring is informative, and our approach would not capture this dependence.

2.6.2.1 Frydman's basic model

Beginning with the basic notation and model of Frydman (1995), we assume the true times of entry (X, T) into states 2 and 3 are only known to be in the respective intervals $A = [X_L, X_R]$ and $B = [T_L, T_R]$. Assume the same discrete time scale $(0, 1, \ldots, m)$ is used to measure both the chronological time and the sojourn duration in state 2. Here m represents the last observation time in the study, which is finite. Denote the possible range of time values for X as $\{x_j;\ j = 1, \dots, J\}$ and the possible range of time values for T as $\{t_k;\ k = 1, \dots, K\}$. For all observations censored in state 1 (i.e. $\Delta = 0$), we assume there is no information available about T. For all remaining observations (i.e. $\Delta = 1$), we denote the support set of T by $supp(T)$. Define $h(x_j, t_r) = h_{jr}$ for $x_j < t_r$ $(1 \le r < k \le K)$ and $t_r \in supp(T)$. Define

$$\prod\nolimits_{j,k} = \prod_{x_j < t_r \le t_k} (1 - h_{jr}) \ , \qquad \prod\nolimits^{*}_{j,k} = \prod_{x_j < t_r < t_k} (1 - h_{jr}) \ .$$

Along with the censoring indicators Δ_n, δ_n, the observed data for subject n are now of the form $\{A_n, \Delta_n; B_n, \delta_n\}$. As before, $\Delta_n = 1$ if subject n entered state 2 in the time interval A_n, but now $\delta_n = 1$ if the subject n entered state 3 in the time interval B_n. We can write the log likelihood for this situation as

$$\begin{aligned}
\log\ L(\Theta) \ &= \ \sum_{\delta_n = 0} \log \sum_{x_j \in A_n} p_j \left[\sum_{t_k \in B_n} \prod_{x_j < t_r \le t_k} \{1 - h(x_j, t_r)\} \right]^{\Delta_n} \\
&+ \ \sum_{\delta_n = 1} \log \sum_{x_j \in A_n} p_j \sum_{t_k \in B_n} h(x_j, t_k) \prod_{x_j < t_r < t_k} \{1 - h(x_j, t_r)\} \ .
\end{aligned}$$

Define the indicator functions for individual n to be

$$d_{njk} = \begin{cases} 1 & \text{if } x_j \in A_n,\ t_k \in B_n, \text{ and } x_j < t_k \text{ for } \delta_n = 0, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$b_{njk} = \begin{cases} 1 & \text{if } x_j \in A_n,\ t_k \in B_n, \text{ and } x_j < t_k \text{ for } \delta_n = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Now the log likelihood may be written as

$$\begin{aligned} \log\ L(\Theta) &= \sum_{\delta_n=0} \log \sum_{j=1}^{J} \sum_{k=1}^{K} d_{njk}\, p_j \left[\prod_{x_j < t_r \leq t_k} \{1 - h(x_j, t_r)\} \right]^{\Delta_n} \\ &+ \sum_{\delta_n=1} \log \sum_{j=1}^{J} \sum_{k=1}^{K} b_{njk}\, p_j\, h(x_j, t_k) \prod_{x_j < t_r < t_k} \{1 - h(x_j, t_r)\} \\ &= \sum_{\delta_n=0} \log \sum_{j,k} \,||\, d_{njk}\, p_j \left[\prod_{x_j < t_r \leq t_k} \{1 - h(x_j, t_r)\} \right]^{\Delta_n} || \\ &+ \sum_{\delta_n=1} \log \sum_{j,k} \,||\, b_{njk}\, p_j\, h(x_j, t_k) \prod_{x_j < t_r < t_k} \{1 - h(x_j, t_r)\} \,||\ . \end{aligned}$$

The conditional expectation of the log of the complete data, given the observed data, corresponds to the function

$$\begin{aligned} Q(p', \lambda', \beta' | p, \lambda, \beta) &= \sum_{\delta_n=0} \sum_{j,k} \left[\frac{d_{njk}\, p_j \left[\prod_{jk}\right]^{\Delta_n}}{\sum_{l,m} d_{nlm}\, p_l\, [\prod_{lm}]^{\Delta_n}} \right] \log\ \left(p'_j\ \left[\prod{}_{jk}{}^{\Delta_n}\right]' \right) \\ &+ \sum_{\delta_n=1} \sum_{j,k} \left[\frac{b_{njk}\, p_j\, h(x_j, t_k) \prod^*_{jk}}{\sum_{l,m} b_{nlm}\, p_l\, h(x_l, t_m) \prod^*_{lm}} \right] \log\ \left(p'_j\ h'(x_j, t_k) \left[\prod{}_{jk}{}^{*}\right]' \right)\ . \end{aligned}$$

Once again, we maximize the log likelihood by taking partial derivatives of Q

with respect to the parameters of interest (p', λ', β'). For p_j we find

$$\hat{p}_j = \left\{ \sum_{\delta_n=0} \sum_{k=1}^{K} \mu_{njk}(p, \lambda, \beta) + \sum_{\delta_n=1} \sum_{k=1}^{K} \mu^*_{njk}(p, \lambda, \beta) \right\} / N \qquad (1 \leq j \leq J) \, .$$

For λ_k, the estimating equation becomes

$$\begin{aligned} \sum_{\delta_n=1} \sum_{j=1}^{J} \mu^*_{njk}(\Theta) &= \sum_{\Delta_n=1, \delta_n=0} \sum_{j=1}^{J} \mu^\dagger_{njk}(\Theta) h'(x_j, t_k) \\ &+ \sum_{\delta_n=1} \sum_{j=1}^{J} h'(x_j, t_k) \sum_{r=k}^{K} \mu^*_{njr}(\Theta) \qquad (1 \leq k \leq K) \, . \end{aligned}$$

And lastly, for β we obtain the equation

$$\begin{aligned} \sum_{\delta_n=1} \sum_{j} \sum_{k} \mu^*_{njk}(\Theta)(t_k - x_j) = & \sum_{\Delta_n=1, \delta_n=0} \sum_{j} \sum_{k} \mu^\dagger_{njk}(\Theta) \sum_{x_j < t_r \leq t_k} h'(x_j, t_r) \, (t_r - x_j) \\ & + \sum_{\delta_n=1} \sum_{j} \sum_{k} \mu^*_{njk}(\Theta) \sum_{x_j < t_r \leq t_k} h'(x_j, t_r) \, (t_r - x_j) \, , \end{aligned}$$

where

$$\mu_{njk}(p, \lambda, \beta) = \frac{d_{njk} \; p_j \left[\prod_{jk} \right]^{\Delta_n}}{\sum_{l,m} d_{nlm} \; p_l \left[\prod_{lm} \right]^{\Delta_n}} \quad ,$$

$$\mu^*_{njk}(p, \lambda, \beta) = \frac{b_{njk} \; p_j \; h(x_j, t_k) \prod^*_{jk}}{\sum_{l,m} b_{nlm} \; p_l \; h(x_l, t_m) \prod^*_{lm}} \quad ,$$

and for $\Delta_n = 1, \delta_n = 0$

$$\mu^{\dagger}_{njk}(p, \lambda, \beta) = \frac{d_{njk}\; p_j\; \prod_{jk}}{\sum_{l,m} d_{nlm}\; p_l\; \prod_{lm}} \quad .$$

2.6.2.2 Time-independent Covariate Model

The transition intensities are modelled semi-parametrically as functions of time-independent (fixed), external covariates in addition to T being possibly interval censored. The covariate process is defined as $Z' = \{z_1, z_2, \dots, z_p\}$. The logistic specifications for both transition intensities are unchanged from the case when T was observed exactly or possibly right censored, viz.,

$$\alpha(x; z) = \frac{\eta(x)\; e^{z'\boldsymbol{\gamma}}}{1 + \eta(x)\; e^{z'\boldsymbol{\gamma}}} \; ,$$

$$h(x, t; z) = \frac{\lambda(t)\; e^{\beta(t-x)+z'\boldsymbol{\nu}}}{1 + \lambda(t)\; e^{\beta(t-x)+z'\boldsymbol{\nu}}} \; .$$

The log likelihood of the observed data now becomes

$$\begin{aligned} \log\; L(\Theta) = & \sum_{\Delta_n=0, \delta_n=0} \log \prod_{0<u<x_l^n} [1 - \alpha(u; z_n)] \\ + & \sum_{\Delta_n=1, \delta_n=0} \log \sum_{j=1}^{J} \sum_{k=1}^{K} d_{njk}\; \alpha(x_j; z_n) \prod_{0<u<x_j} [1 - \alpha(u; z_n)] \prod_{jk}(z_n) \\ + & \sum_{\delta_n=1} \log \sum_{j=1}^{J} \sum_{k=1}^{K} b_{njk}\; \alpha(x_j; z_n) \prod_{0<u<x_j} [1 - \alpha(u; z_n)]\;\; h(x_j, t_k; z_n) \prod_{jk}^{*}(z_n) \; . \end{aligned}$$

We continue to define the indicator variables d_{njk} and b_{njk} as before; however, in this setting $\Theta = \{\boldsymbol{\eta}, \boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \beta\}$, and $\tilde{\prod}_j(z_n) = \prod_{0<u<x_j}[1-\alpha(u;z_n)]$. When we decompose the log likelihood, we obtain

$$\begin{aligned} Q(\Theta'|\Theta) \; = \; & \sum_{\Delta_n=0,\delta_n=0} \sum_{u=x_1}^{x_{l-1}^n} \log\left[1-\alpha'(u;z_n)\right] \\ + \; & \sum_{\Delta_n=1,\delta_n=0} \sum_{j,k} \left[\frac{d_{njk}\ \alpha(x_j;z_n)\tilde{\prod}_j(z_n)\prod_{jk}(z_n)}{\sum_{l,m} d_{nlm}\ \alpha(x_l;z_n)\tilde{\prod}_l(z_n)\prod_{lm}(z_n)}\right] \times \\ & \log\left(\alpha'(x_j;z_n)\ \left[\tilde{\prod}_j(z_n)\right]'\left[\prod_{jk}(z_n)\right]'\right) \\ + \; & \sum_{\delta_n=1} \sum_{j,k} \left[\frac{b_{njk}\ \ \alpha(x_j;z_n)\tilde{\prod}_j(z_n)\,h(x_j,t_k;z_n)\prod^*_{jk}(z_n)}{\sum_{l,m} b_{nlm}\ \alpha(x_l;z_n)\tilde{\prod}_l(z_n)\,h(x_l,t_m;z_n)\prod^*_{lm}(z_n)}\right] \times \\ & \log\left(\alpha'(x_j;z_n)\ \left[\tilde{\prod}_j(z_n)\right]'\ h'(x_j,t_k;z_n)\left[\prod^*_{jk}(z_n)\right]'\right) \ . \end{aligned}$$

By differentiating $Q(\Theta'|\Theta)$ with respect to the various parameters, we obtain the score equations which will yield the MLEs for the various parameters. Beginning with η_j, and still defining the indicator variable $c_{nj} = 1$, if $x_j < x_l^n$ for $\Delta_n = 0$ and 0 otherwise, the equation is

$$\begin{aligned} \sum_{\Delta_n=1} \sum_{k=1}^{K} \left\{\mu^{\dagger}_{njk}(\Theta) + \mu^{*}_{njk}(\Theta)\right\} = & \sum_{\Delta_n=0,\delta_n=0} c_{jn}\ \alpha'(x_j;z_n) + \\ & \sum_{\Delta_n=1,\delta_n=0} \alpha'(x_j;z_n) \sum_{r=j}^{J} \sum_{k=1}^{K} \mu^{\dagger}_{nrk}(\Theta) + \sum_{\delta_n=1} \alpha'(x_j;z_n) \sum_{r=j}^{J} \sum_{k=1}^{K} \mu^{*}_{rnk}(\Theta) \ , \\ & \qquad (1 \leq j \leq J) \ , \end{aligned}$$

where we redefine

$$\mu^{\dagger}_{njk}(\Theta) = \frac{d_{njk}\ \alpha(x_j; z_n)\ \tilde{\prod}_j(z_n)\prod_{jk}(z_n)}{\sum_{l,m} d_{nlm}\ \alpha(x_l; z_n)\ \tilde{\prod}_l(z_n)\prod_{lm}(z_n)} \quad ,$$

$$\mu^{*}_{njk}(\Theta) = \frac{b_{njk}\ \ \alpha(x_j; z_n)\tilde{\prod}_j(z_n)\, h(x_j, t_k; z_n)\prod^{*}_{jk}(z_n)}{\sum_{l,m} b_{nlm}\ \alpha(x_l; z_n)\tilde{\prod}_l(z_n)\, h(x_l, t_m; z_n)\prod^{*}_{lm}(z_n)} \quad .$$

Now the equation for $\hat{\gamma}_s$ is

$$\begin{aligned} \sum_{\Delta_n=1} z_{ns} &= \sum_{\Delta_n=0} z_{ns} \sum_{u=x_1}^{x^n_{l-1}} \alpha'(u; z_n) \\ &+ \sum_{\Delta_n=1} z_{ns} \sum_j \sum_k \left\{ \mu^{\dagger}_{njk}(\Theta) + \mu^{*}_{jn}(\Theta) \right\} \sum_{u=x_1}^{x_j} \alpha'(u; z_n) \qquad (1 \le s \le p) \ . \end{aligned}$$

Turning now to the parameters of the second transition intensity, the equation for λ_k is given by

$$\begin{aligned} \sum_{\delta_n=1} \sum_{j=1}^{J} \mu^{*}_{njk}(\Theta) &= \sum_{\Delta_n=1, \delta_n=0} \sum_{j=1}^{J} \mu^{\dagger}_{njk}(\Theta)\, h'(x_j, t_k; z_n) \\ &\quad + \sum_{\delta_n=1} \sum_{j=1}^{J} h'(x_j, t_k; z_n) \sum_{s=k}^{K} \mu^{*}_{njs}(\Theta) \ , \qquad (1 \le k \le K) \ , \end{aligned}$$

and for β we obtain

$$\begin{aligned} \sum_{\delta_n=1} \sum_j \sum_k \mu^{*}_{njk}(\Theta)\, (t_k - x_j) &= \sum_{\Delta_n=1, \delta_n=0} \sum_j \sum_k \mu^{\dagger}_{njk}(\Theta) \sum_{x_j < t_r \le t_k} h'(x_j, t_r; z_n)\, (t_r - x_j) \\ &\quad + \sum_{\delta_n=1} \sum_j \sum_k \mu^{*}_{njk}(\Theta) \sum_{x_j < t_r \le t_k} h'(x_j, t_r; z_n)\, (t_r - x_j) \ . \end{aligned}$$

Lastly the equation that we use to find the MLE for ν_s is given by

$$\begin{aligned}\sum_{\delta_n=1} z_{ns} \quad &= \quad \sum_{\Delta_n=1,\delta_n=0} z_{ns} \sum_j \sum_k \mu^{\dagger}_{njk}(\Theta) \sum_{x_j<t_r\le t_k} h'(x_j,t_r;z_n) \\ &+ \quad \sum_{\delta_n=1} z_{ns} \sum_j \sum_k \mu^{*}_{njk}(\Theta) \sum_{x_j<t_r\le t_k} h'(x_j,t_r;z_n) \qquad (1\le s\le p)\ .\end{aligned}$$

2.6.2.3 Time-dependent Covariate Model

As before, in this extension, both transition intensities are modelled semi-parametrically as functions of time-dependent external covariates. The covariate process becomes $Z'(t) = \{z_1(t), z_2(t), \ldots, z_p(t)\}$. We assume that some of the covariates could be time independent and that the time-dependent covariates follow a deterministic function. We define all the indicator variables and product variables as for the corresponding fixed covariate models, with the only changes occurring to the terms in the logistic specifications for both transition intensities. The resulting log likelihood is

$$\begin{aligned}\log\ L(\Theta) = &\sum_{\Delta_n=0,\delta_n=0} \log \prod_{0<u<x_l^n} [1-\alpha\{u;z_n(u)\}] \\ + &\sum_{\Delta_n=1,\delta_n=0} \log \sum_{j=1}^{J}\sum_{k=1}^{K} d_{njk}\ \alpha\{x_j;z_n(x_j)\} \prod_{0<u<x_j} [1-\alpha\{u;z_n(u)\}] \prod_{jk}\{z_n(t_r)\} \\ + &\sum_{\delta_n=1} \log \sum_{j=1}^{J}\sum_{k=1}^{K} b_{njk}\ \alpha\{x_j;z_n(x_j)\} \prod_{0<u<x_j} [1-\alpha\{u;z_n(u)\}]\ h\{x_j,t_k;z_n(t_k)\} \times \\ &\prod_{jk}{}^{*}\{z_n(t_r)\}\ .\end{aligned}$$

The estimating equations in this case are similar to the corresponding ones derived for the time-independent covariate model, with differences only appearing in the expressions for the intensities and for the covariates, which may be time dependent. Once again, beginning with η_j, the estimating equation is

$$\begin{aligned}\sum_{\Delta_n=1}\sum_{k=1}^{K}\left\{\mu^{\dagger}_{njk}(\Theta)+\mu^{*}_{njk}(\Theta)\right\} &= \sum_{\Delta_n=0,\delta_n=0} c_{jn}\;\alpha'\{x_j;z_n(x_j)\} + \\ \sum_{\Delta_n=1,\delta_n=0}\alpha'\{x_j;z_n(x_j)\}\sum_{r=j}^{J}\sum_{k=1}^{K}\mu^{\dagger}_{nrk}(\Theta) &+ \sum_{\delta_n=1}\alpha'\{x_j;z_n(x_j)\}\sum_{r=j}^{J}\sum_{k=1}^{K}\mu^{*}_{rnk}(\Theta) \\ &\qquad (1\le j\le J)\,.\end{aligned}$$

The expression for γ_s becomes

$$\begin{aligned}\sum_{\Delta_n=1}\sum_{j}\sum_{k}\left\{\mu^{\dagger}_{njk}(\Theta)+\mu^{*}_{njk}(\Theta)\right\}\; z_{ns}(x_j) = &\sum_{\Delta_n=0}\sum_{u=x_1}^{x^n_{l-1}}\alpha'\{u;z_n(u)\}\; z_{ns}(u) \\ + \sum_{\Delta_n=1}\sum_{j}\sum_{k}\left\{\mu^{\dagger}_{njk}(\Theta)+\mu^{*}_{njk}(\Theta)\right\}\sum_{u=x_1}^{x_j}\alpha'\{u;z_n(u)\}\; z_{ns}(u)\;, &\qquad (1\le s\le p)\,.\end{aligned}$$

The equations for the parameters in the second transition intensity are also similar to their time-independent counterparts, except that the covariates in the linear predictor may now depend on time. For λ_k, the equation becomes

$$\begin{aligned}\sum_{\delta_n=1}\sum_{j=1}^{J}\mu^{*}_{njk}(\Theta) \quad &= \quad \sum_{\Delta_n=1,\delta_n=0}\;\sum_{j=1}^{J}\mu^{\dagger}_{njk}(\Theta)\; h'\{x_j,t_k;z_n(t_k)\} \\ &\quad +\sum_{\delta_n=1}\sum_{j=1}^{J}h'\{x_j,t_k;z_n(t_k)\}\sum_{s=k}^{K}\mu^{*}_{njs}(\Theta) \qquad (1\le k\le K)\,.\end{aligned}$$

For β we obtain the equation

$$\begin{aligned}
&\sum_{\delta_n=1}\sum_{j}\sum_{k}\mu^{*}_{njk}(\Theta)\,(t_k - x_j) = \\
&\qquad\sum_{\Delta_n=1,\delta_n=0}\sum_{j}\sum_{k}\mu^{\dagger}_{njk}(\Theta)\sum_{x_j<t_r\leq t_k} h'\{x_j, t_r; z_n(t_r)\}\,(t_r - x_j) \\
&\qquad+\sum_{\delta_n=1}\sum_{j}\sum_{k}\mu^{*}_{njk}(\Theta)\sum_{x_j<t_r\leq t_k} h'\{x_j, t_r; z_n(t_r)\}\,(t_r - x_j)\,.
\end{aligned}$$

And lastly, for ν_s we get the estimating equation

$$\begin{aligned}
&\sum_{\delta_n=1}\sum_{j}\sum_{k}\mu^{*}_{njk}(\Theta) z_{n_s}(t_k) = \\
&\qquad\sum_{\Delta_n=1,\delta_n=0}\sum_{j}\sum_{k}\mu^{\dagger}_{njk}(\Theta)\sum_{x_j<t_r\leq t_k} h'\{x_j, t_r; z_n(t_r)\} z_{n_s}(t_r) \\
&\qquad\sum_{\delta_n=1}\sum_{j}\sum_{k}\mu_{njk}(\Theta)\sum_{x_j<t_r\leq t_k} h'\{x_j, t_r; z_n(t_r)\} z_{n_s}(t_r)\,.
\end{aligned}$$

2.7 Example: AIDS in Hemophilia Patients

To illustrate the advantages of our regression approach, using a data set that incorporates time-dependent covariates and interval-censoring of both random variables, X and T, we consider the AIDS data presented and analysed in several published papers (De Gruttola and Lagakos, 1989; Kim, De Gruttola and Lagakos, 1993; Frydman, 1992; Frydman, 1995). Individuals with Type A or B hemophilia who had received treatment since 1978 at two hospitals in France (Hôpital Kremlin Bicêtre and Hôpital Coeur des Yvelines) were at risk for infection with HIV through the use of contaminated blood products. All HIV infections were assumed to have resulted from the use of such infectious products, with retrospective determination of HIV status found through stored blood samples. Progression to state 2 was defined as testing positive for HIV, whereas progression to state 3 was defined as developing clinical AIDS symptoms (AIDS, lymphadenopathy, or leukopenia).

The data set we consider is the updated version appearing in Kim *et al.* (1993). This data set is slightly different than the versions appearing in the other three papers in terms of the study population size, the number of individuals who are believed to have progressed to both HIV infection (state 2) and AIDS symptoms (state 3), and the length of follow-up. The differences appear to be minor and the increased follow-up, additional covariate information, and the time intervals for the transition to AIDS symptoms all suggest this version of the data is the one to use.

The study population consists of 257 individuals, 188 of whom were found to be infected with HIV at the time of the analysis. There were 41 individuals who subsequently developed AIDS-related symptoms. The study began on January 1st, 1978 and concluded in August 1988. This time interval is discretized into six-month intervals, with $t = 1$ denoting the time period from January 1st, 1978 to June 30th, 1978, and so on. Therefore, the total number of time intervals, m, is 23. Two binary covariates were used in the analysis: treatment group and estimated age at time of infection. The treatment groups were defined as heavy/high (received at least 1000 μg/kg of blood factor for at least one year between 1982 and 1985) and light/low (received less than 1000 μg/kg of blood factor in each year). The value of the age variable was calculated by taking the expected value of age over the interval of infection, using the estimated probability distribution for the time to infection.

The split point for the binary indicator was 20 years of age.

The analysis presented in the Kim *et al.* (1993) paper assumed that the times to both events were independent, and adopted a semi-parametric modelling approach for the distribution of time between the two events, using the discrete analogue of the proportional hazards model. Obvious limitations of this approach are the independence and proportional hazards assumptions. In addition, the authors only examined the effect of covariates on the induction time, not on the time to HIV infection as well. Their findings indicated that only the treatment variable was statistically significant. The results from the modelling approach taken by Frydman (1995) suggest that incorporating the duration dependence in state 2 (HIV infection) is very important when one estimates the distribution of time from HIV infection to AIDS symptoms. Frydman does not include covariates other than the duration in state 2; indeed, separate analyses are carried out for the two treatment groups.

In our regression modelling approach, we can relax the independence assumption of Kim *et al.* (1993), include in the second transition intensity a single covariate for duration in state 2 considered by Frydman (1995) as well as the two covariates (treatment group, estimated age at infection) considered by Kim *et al.* (1993), and allow the time to the development of AIDS symptoms to be interval-censored.

The results of our analysis are summarized in Table 2.1. Under the column heading Model Description, Markov refers to the discrete, time non-homogeneous Markov chain model (chronological time scale), Hybrid refers to the basic Markov model which also includes a covariate for duration in state 2, Tx represents inclusion of the treatment group indicator variable, and Age represents inclusion of the estimated age at the time of infection indicator variable. The comparisons between models are obtained via likelihood ratio statistics (LRS). Under the null hypotheses discussed in this section, these likelihood ratio statistics are assumed to have an asymptotic χ^2 distribution, with one degree of freedom. Unless otherwise stated, a single time to infection (X) distribution function was estimated in the given model. Our results will be compared first to the results obtained by Frydman (1995) and then to those published by Kim *et al.* (1993).

Frydman found that duration in state 2 was important, but only for the heavily treated group. The LRSs comparing the simple Markov model to one in which duration in state 2 was incorporated for both treatment groups were 15.1 ($p = 0.0001$) and 0.37 ($p = 0.5430$), for the heavy and light groups, respectively. The importance of the duration in state 2 is confirmed in our analysis. The simple effect of duration in state 2 is revealed by comparing the Markov and hybrid time scale models when no other explanatory covariates are included in the models. Twice the difference between the log likelihood for the Markov model (model A; $\log L(\hat{\theta}) = -491.412$) and the log likelihood for the hybrid model (model D; $\log L(\hat{\theta}) = -489.692$) yields a LRS of 3.44 ($p = 0.0637$). Thus, there is weak evidence against the null hypothesis that $\beta = 0$. Even after adjusting for a possible difference in the distribution of infection times by estimating parameters for the treatment groups separately, the importance of duration in state 2 is revealed. By comparing the Markov model in this setting (model G; $\log L(\hat{\theta}) = -491.412$) to the hybrid model in the same setting (model I; $\log L(\hat{\theta}) = -489.692$), we find the LRS to be 3.44, which has a

Table 2.1: Log-likelihood values and regression parameter estimates (point and interval) from models based on the hemophilia patients with AIDS data.

Model			Regression Parameter Estimates		
Label	Description	log L($\hat{\theta}$)	β	ν_1	ν_2
A	Markov (Null)	-491.412			
B	Markov & Tx	-488.667		0.744 [0.122,1.471]	
C	Markov, Tx & Age	-488.649		0.755 [0.123,1.501]	0.051 [-0.671,0.755]
D	Hybrid	-489.692	0.130 [-0.009,0.279]		
E	Hybrid & Tx	-487.033	0.126 [-0.012,0.278]	0.753 [0.112,1.478]	
F	Hybrid, Tx & Age	-486.974	0.129 [-0.009,0.282]	0.775 [0.122,1.517]	0.098 (-0.624,0.822)
G	Markov (two CDFs for X)	-491.412			
H	Markov & Tx (two CDFs for X)	-488.666		0.744 [0.122,1.473]	
I	Hybrid (two CDFs for X)	-489.692	0.130 [-0.009,0.279]		
J	Hybrid & Tx (two CDFs for X)	-487.034	0.127 [-0.012,0.279]	0.753 [0.113,1.477]	

significance level of 0.0636.

The importance of the duration in state 2 may be evaluated when explanatory variates are included in the models being compared. When only the treatment information is included in the second transition intensity, twice the difference between the Markov model with this single explanatory variate (model B; $\log L(\hat{\theta}) = -488.667$) and the corresponding hybrid model (model E; $\log L(\hat{\theta}) = -487.033$) is 3.27 ($p = 0.0707$). When both explanatory variates (treatment group and estimated age at infection) are included in the second transition intensity, the LRS comparing the Markov model (model C; $\log L(\hat{\theta}) = -488.649$) and the hybrid model (model F; $\log L(\hat{\theta}) = -486.974$) is 3.35, which has a significance level of 0.0672.

When the time to infection is estimated nonparametrically using separate distribution functions for the heavily and lightly treated groups, and the effect of the treatment variable to the second transition intensity is included in the model (model H; $\log L(\hat{\theta}) = -488.666$), the addition of the duration in state 2 variable to the second transition intensity (model J; $\log L(\hat{\theta}) = -487.034$) is still important. The resulting LRS has an observed value of 3.27, with a corresponding p-value of 0.0708.

Thus, whether covariates are added to the model or not, or whether the time to infection is estimated separately or jointly for the treatment groups, the effect of the duration in state 2 is weakly important. The estimated value of the regression coefficient (β) for the duration in state 2 varies only slightly between most of the models, although it is somewhat larger when no covariates are included in the model (models D and I).

We now turn to the results from the 1993 paper by Kim *et al.* and compare our results with their findings. Three different models were fit in their paper, so each of these models will be discussed in turn here.

Their model 1 included the treatment explanatory variate in the second transition intensity, and they incorporated a single distribution for the time to infection. Using a Wald statistic which, like the LRS, has an asymptotic χ^2 distribution, they reported a p-value of 0.04 for testing the hypothesis that the regression coefficient (β in their paper) for the effect of treatment was zero. In our analysis, when a single distribution function for the first transition time was also adopted, the addition of the treatment level information (model B; $\log L(\hat{\theta}) = -488.667$) to an underlying Markov model (model A; $\log L(\hat{\theta}) = -491.412$) results in a significance level of 0.0191 for the LRS (${\chi^2}_1 = 5.49$). In the hybrid time scale model, when this covariate is added (model E; $\log L(\hat{\theta}) = -487.033$) to the underlying model (model D; $\log L(\hat{\theta}) = -489.692$), the observed p-value is very similar (0.0211) to the value obtained in the Markov time scale model. Thus, as in Kim *et al.* (1993), the models we fitted revealed strong evidence against the hypothesis that treatment level information (ν_1 in our model) is unimportant.

Model 2 in Kim *et al.* evaluated the effect on estimation of the regression coefficients for treatment when fitting separate distributions for the time to infection. Again, using a Wald statistic, the authors reported a p-value of 0.06 for testing the hypothesis that this regression coefficient was zero. The point estimate for this regression coefficient (0.65) was slightly attenuated from the value obtained in model 1 (0.69), but the estimated standard error was unchanged (0.34). In the Markov time scale framework, and using a likelihood ratio statistic, we found that adding treatment level information (model H; $\log L(\hat{\theta}) = -488.666$) to the underlying model (model G; $\log L(\hat{\theta}) = -491.412$) gave rise to a LRS value of 5.49. The corresponding significance level of 0.0191 was equivalent to the value noted above ($p = 0.0191$) when only a single time to infection distribution is used. Not surprisingly, in the hybrid time scale framework, we obtained a similar result. We calculated a LRS value of 5.32 ($p = 0.0211$) when treatment information (model J; $\log L(\hat{\theta}) = -487.034$) is added to the underlying model (model I; $\log L(\hat{\theta}) = -489.692$).

The reason for the strong similarity between the models with one and two separate distributions for the time to infection is that the estimating equations are essentially the same. Following the approach of Kim *et al.*, we define a treatment group indicator function by letting $\epsilon_n = 1$ if individual n was in the low or light treatment group, and 0 otherwise. Let p_{1j} and p_{0j} denote the nonparametric estimators for the distributions of time to infection for the light and heavy treatment groups respectively. We can write the log likelihood function corresponding to their model 2 in our setting as

$$\begin{aligned}\log\ L(\Theta) \;&=\; \sum_{\delta_n=0} \log \sum_{j=1}^{J} \sum_{k=1}^{K} d_{njk}\, {p_{1j}}^{\epsilon_n}\, {p_{0j}}^{1-\epsilon_n} \left[\prod_{x_j < t_r \le t_k} \{1 - h(x_j, t_r)\} \right]^{\Delta_n} \\ &+\; \sum_{\delta_n=1} \log \sum_{j=1}^{J} \sum_{k=1}^{K} b_{njk}\, {p_{1j}}^{\epsilon_n}\, {p_{0j}}^{1-\epsilon_n}\, h(x_j, t_k) \prod_{x_j < t_r < t_k} \{1 - h(x_j, t_r)\}\end{aligned}$$

$$
\begin{aligned}
= & \sum_{\delta_n=0,\epsilon_n=1} \log \sum_{j=1}^{J} \sum_{k=1}^{K} d_{njk}\, {p_{1j}}^{\epsilon_n} \left[\prod_{x_j<t_r\leq t_k} \{1-h(x_j,t_r)\} \right]^{\Delta_n} \\
+ & \sum_{\delta_n=1,\epsilon_n=1} \log \sum_{j=1}^{J} \sum_{k=1}^{K} b_{njk}\, {p_{1j}}^{\epsilon_n}\, h(x_j,t_k) \prod_{x_j<t_r<t_k} \{1-h(x_j,t_r)\} \\
+ & \sum_{\delta_n=0,\epsilon_n=0} \log \sum_{j=1}^{J} \sum_{k=1}^{K} d_{njk}\, {p_{0j}}^{1-\epsilon_n} \left[\prod_{x_j<t_r\leq t_k} \{1-h(x_j,t_r)\} \right]^{\Delta_n} \\
+ & \sum_{\delta_n=1,\epsilon_n=0} \log \sum_{j=1}^{J} \sum_{k=1}^{K} b_{njk}\, {p_{0j}}^{1-\epsilon_n}\, h(x_j,t_k) \prod_{x_j<t_r<t_k} \{1-h(x_j,t_r)\} \ .
\end{aligned}
$$

The resulting estimating equations derived from the expectation of the complete data likelihood, conditional on the observed data, for the parameters β, ν_1 in the second transition intensity are equivalent to the corresponding estimating equations based on a single distribution function, but which have been split into double the number of pieces. Hence, the parameter estimates for β, ν_1 are nearly identical, except for a small amount of rounding error. Although this approach does not offer any advantages for estimating the parameters in the second transition intensity, it does provide separate nonparametric estimates of the time to infection distribution for each treatment group.

The estimates of the regression coefficients associated with treatment changed very little in our analyses between the models with one or two time to infection distributions. In the hybrid time scale model, e.g. model J, the estimated coefficient was only slightly attenuated from that obtained in model E (0.7531 vs. 0.7532). In the corresponding Markov time scale model, the estimated regression coefficient for treatment was essentially the same to four decimal places between Model H and Model B (0.7444 vs. 0.7444). Hence, we can conclude that the treatment is important with respect to the progression to AIDS, irrespective of whether the time to infection is estimated separately or jointly for the two treatment groups.

The third and final model considered by Kim *et al.* (1993) included both treatment and the estimated age at infection in a model with a single infection time distribution. Using a LRS, they tested the hypothesis that the regression coefficient for the estimated age at infection was zero, and found no contradictory evidence (${\chi^2}_1 = 0.0122, p = 0.90$). They concluded that the estimated age at infection was not an important predictor of progression to clinical symptoms of AIDS. To test this hypothesis in our approach, we compared models with both explanatory covariates to models without the estimated age at infection variable. In the Markov setting, the model with both covariates (model C; $\log L(\hat{\theta}) = -488.649$) was compared to the model with only treatment information (model B; $\log L(\hat{\theta}) = -488.667$). The resulting LRS of 0.04 has a rather large p-value of 0.8512. In the hybrid time scale setting, the model with both covariates (model F; $\log L(\hat{\theta}) = -486.974$) was also compared to the model involving only treatment level (model E; $\log L(\hat{\theta}) = -487.033$). The resulting LRS of 0.12 also has a large p-value of 0.7303. Therefore, we concluded as well that the estimated age at infection was not important with respect to the progression to AIDS, after an individual tested positive for HIV.

For this AIDS example, interval estimates of the regression parameters were obtained using a profile likelihood approach. In this procedure, a single regression parameter was fixed in each model and the remaining parameter estimates were found by maximizing the constrained log likelihood. The values reported in Table 2.1 are the 95% confidence interval estimates, assuming a χ^2 distribution with one degree of freedom. Like the point estimates, the interval estimates for ν_1, ν_2 are very similar between the Markov and corresponding hybrid models. The hybrid models tend to have slightly wider intervals, primarily because the lower endpoint is further away from the point estimate than in the Markov setting. In all the hybrid models, the interval estimates for β barely include zero — a finding which is consistent with the approximately 0.07% significance level for the point estimates. The interval estimates in both frameworks become wider as the number of explanatory variables in a model increases.

In summary, our results were in very close agreement with the findings of Frydman (1995) and Kim *et al.* (1993). Like Frydman, we found that sojourn duration in state 2 was important. Like Kim *et al.*, we found that (a) the treatment level is important with respect to progression to AIDS, and is barely affected by any difference in the distribution of infection times, and (b) the estimated age at infection is not important with respect to progression to AIDS. Unlike Frydman and Kim *et al.*, in our approach we were able to demonstrate these effects in various models where regression coefficients for these explanatory variables were estimated *simultaneously.*

Since we were able to extend the original model by Frydman (1995) to incorporate covariates in the first transition intensity, we wanted to fit models that assessed the importance of covariates in both transition intensities. Unfortunately, the estimation of these new models was not possible with this data set. Neither explanatory variable seems to be closely related to time of infection, but rather only to the time of developing clinical symptoms of AIDS. The treatment variable was defined on the basis of the amount of blood factor received between 1982 and 1985, yet this time range comprises only the latter half of the support for the time to infection distribution. The age variable, as we previously explained, is calculated as the expected age over the interval of infection and then defined as an indicator variable for the split point of 20 years of age. Thus, it appears to be a more appropriate explanatory variable for the second transition time.

The semi-parametric model formulation would be useful for studying the effect of relevant covariate effects on the time to the intermediate event and would only require the additional estimation of the associated regression coefficients. This approach, which estimates the baseline intensities directly, that is, $\eta(x_j)$ in $\alpha(x_j) = \frac{\eta(x_j)}{1+\eta(x_j)}$, instead of nonparametrically estimating the transition probabilities $p(x_j)$, where $p(x_j) = \alpha(x_j)\prod_{0<u<x_j}[1-\alpha(u)]$, seems to be as numerically stable as the nonparametric approach. For example, in the Kim *et al.* data set, almost identical estimates of the cumulative distribution function (CDF) of the infection time were obtained using these two estimation methods.

2.8 Conclusions

In conclusion, our approach incorporates model features proposed by the previously-mentioned authors. Our analysis results for the AIDS data example are consistent with the findings of both Frydman (1995) and Kim *et al.* (1993). However, our methodology is more general and flexible. It allows us to incorporate time-dependent explanatory covariates, including the duration in state 2. The times of transitions to both the intermediate and final states may be interval or right censored, which is a scenario frequently encountered in practice.

One obvious drawback to our approach is that estimates of the variability in the data, except for individual regression parameters, are not available. In contrast to Kim *et al.* (1993), who used the Newton-Raphson algorithm, standard errors for the estimated regression parameters are not easily obtained nor are they jointly estimated using a profile likelihood method. Possible remedies to this lack of standard errors could include using some type of bootstrap technique, using the SEM algorithm by constructing the complete data for this application (Meng and Rubin, 1991), or perhaps using either a weakly or fully parametric model.

The next chapter will discuss a selection approach when piecewise constant intensity functions are adopted, based on the initial estimates provided by the non- and semi-parametric approaches used in this chapter. Once the number of pieces and the associated changepoints have been estimated, standard errors for the parameters can be calculated. Thus, estimated standard errors can be obtained directly from the full likelihood for all of the model parameters.

Chapter 3

Piecewise Constant Hazard Function Model

Adopting a piecewise constant form for the intensity function between states 2 and 3 avoids many difficulties encountered with a nonparametric treatment of the censored data, including that of obtaining standard errors. This modelling approach also avoids the stronger assumptions associated with fully parametric models. It offers extensive flexibility, as the number of pieces may vary considerably. However, this flexibility can make it difficult to chose an appropriate number of pieces and the corresponding times at which the hazard changes. If the appropriate number of pieces should be intermediate between a single piece for a fully parametric model, and the entire support set for a completely nonparametric model, more than ad-hoc reasons should guide that decision. The purpose of this chapter is to explore some practical guidelines for choosing the number of piecewise constant intervals and their concomitant breakpoints in a three-state progressive process.

3.1 Introduction

A recent paper by Lindsey and Ryan (1998) provides an overview of available methods for interval-censored survival data. The possibilities they consider — fully parametric, piecewise exponential, Turnbull's nonparametric and logspline models — all exist in current software or are simple to program for each application. The performances of these four methods were compared via two well-known data sets. The authors recommend the use of piecewise constant hazard models, citing their flexibility and weaker parametric assumptions. They derive an EM algorithm for the piecewise exponential model in this interval-censored context, where covariate effects are included in the fitted model via a proportional hazards assumption.

A difficulty with this approach, they note, is that the number of changes in the hazard function and breakpoints for those changes need to be selected. They fit three different piecewise exponential models to the same AIDS data set, with two or three intervals of support and different breakpoints for the resulting two-piece model. They found the EM algorithm took much longer to converge if the expected number of events in each interval was very unbalanced. However, the conclusions for the covariate effects did not vary between the three models they considered. Lindsey and Ryan concluded that for this AIDS data set, the piecewise exponential method is quite robust to the number of intervals. They also commented on the need for better guidelines in choosing the number of pieces used in piecewise constant modelling of interval-censored survival data.

In another 1998 paper, Lawless and Zhan adopted piecewise-constant rate functions in the analysis of interval-grouped recurrent-event data. They develop methods of estimation for such processes which incorporate piecewise-constant baseline rate functions in log-linear regression models. In their simulation study, one of the goals was to assess the performance of the piecewise-constant rate functions in estimating a continuous, smooth mean rate function. They concluded that the piecewise-constant rate functions did indeed provide excellent estimation of the mean rate function, as well as of the regression coefficients. They used eight rate function pieces in their simulation studies, and from their experience recommended that the number of pieces used in this setting range from four to ten.

Lindsey and Ryan (1993) used piecewise constant baseline transition rates in a three-state illness-death model for rodent tumourgenicity experiments. Their approach is particularly suitable for this type of problem, since the number of intervals is not tied to the number of sacrifices, which is the case for many non-parametric methods. In the two examples considered in their paper, a hazard with two breakpoints, i.e., three pieces, was used to model the baseline transition rates, with each piece including at least one set of sacrifices. They, too, discussed the decision-making process for determining the number of pieces and corresponding breakpoints in the fitted model. There is a trade-off between having as few pieces as possible, thereby making stronger parametric assumptions, and having as many pieces as possible, which corresponds to a nonparametric model. They noted it would be better to have a data-driven algorithm to determine the breakpoints.

Both the 1993 and 1998 articles by Lindsey and Ryan refer to an earlier paper by Friedman (1982), who examined the use of piecewise exponential models for right-censored survival data. By discovering a similarity between the likelihood functions for the piecewise exponential model and for a log-linear model of frequency data, he exploits the known results concerning MLEs for log-linear models arising from a contingency table. The existence of the MLEs from a log-linear model for survival data is determined, as well as the properties of asymptotic convergence and asymptotic normality of the MLEs. In order to show existence of the MLEs in the log-linear model, his Condition B requires that the lengths of the support intervals all go to zero and the expected number of events in each interval should be of the same magnitude. He recommends that a moderate number of intervals (5 - 7) be chosen at the beginning of an analysis which adopts a piecewise exponential model. From an initial model fit, the parameter estimates and standard errors of the baseline rates need to be examined for any monotone trends; if discovered, then Friedman recommends transforming the time scale.

In a three-state progressive process, with two possible events per subject and potential interval as well as right censoring, the decision concerning the choice of breakpoints to use in a piecewise constant hazard rate model is even more complex. Similar difficulties arise in mixture analyses and fitting splines, and the literature for these methods can offer possible solutions for our modelling problem. Several recent papers have used the Akaike Information Criterion (AIC) to help determine the mixture structure and the knot locations for fitting splines. The use of this criterion seems appropriate to our situation too. After a brief overview of the AIC and a competing alternative, the Schwarz Information Criterion, we summarize their use in other situations before exploring how either criterion might help to resolve the problem of selecting breakpoints in a piecewise constant progressive three-state model.

The Akaike Information Criterion (1973) or AIC, estimates the Kullback-Leibler information between the model generating the data and a fitted candidate model. It is an approximately unbiased estimator if the sample size is large and the dimension of the candidate model is small. A competitor to the AIC is the Schwarz (1978) Information Criterion or SIC, which was derived as an asymptotic approximation to a transformation of the Bayesian posterior probability of a candidate model. Since the SIC does not require the specification of priors, it has been used in frequentist applications as well. The SIC, which is often called BIC because of its Bayesian interpretation, has the attractive property that if the true model for the data is one of the candidate models considered, the asymptotic probability that the SIC will select the true model is unity. Hence, the SIC is often preferred in practice, as it does not tend to overfit, i.e., to select models involving excessive numbers of parameters.

The forms we adopt for the AIC and SIC select the model with the minimal value of a criterion. If we denote the observed data by $\boldsymbol{Y}$, the incomplete-data observed likelihood by $L(\Theta \mid \boldsymbol{Y})$, and the sample size by N, then expressions for each criterion are

$$\begin{aligned} AIC &= -2 \log L(\hat{\Theta} \mid \boldsymbol{Y}) + d\, 2 \ , \\ SIC &= -2 \log L(\hat{\Theta} \mid \boldsymbol{Y}) + d \log N \ ; \end{aligned}$$

the log likelihood is evaluated at the MLE, $\hat{\Theta}$, of Θ, and the number of free parameters to be estimated, that is, the dimension of Θ, is equal to d.

Solka *et al.* (1998) used the AIC to prune back an overdetermined model generated by a nonparametric adaptive mixture density estimation procedure. Their approach produces a distribution of estimated model complexities, and in a simulation study could determine the true model complexity for some mixture problems. In the mixture model setting, the AIC was a useful tool for choosing between competing models, that is, for finding an "optimal" model, and for providing a distribution of pruned models. This distribution of pruned models can give insight into the underlying true mixture as well.

The AIC and SIC were used in the context of a finite mixture problem applied to image classification or segmentation problems by Liang, Jaszczak, and Coleman (1992). If an image, such as a PET (positron emission tomography) scan, is modelled as a mixture, then each mixture component corresponds to an image class. The parameters of that component can be estimated, including the mean and standard deviation, and the number of classes in an image can be determined by applying an information criterion. The results from real medical images and small computer simulations found that the information criteria generally selected the correct number of classes, although the AIC was prone to overfitting.

Rosenberg (1995) recently used the AIC for selecting the number and location of knots when the hazard function in a survival analysis is modelled as a linear combination of splines. He favours use of the AIC rather than a trial-and-error approach or a method that involves selecting the knot locations conditional on the number of intervals of support or pieces. In his knot-selection procedure, he uses quantiles of the empirical distribution function (EDF) of the observed time to failure as the locations for the knots. This ensures that each segment of support will have a comparable amount of information since the number of observed failures will be about the same. The number of knots selected is fixed at four for both the simulation study which he describes and the illustrative analysis of an AIDS data set.

In this chapter, we will examine use of the AIC and SIC for determining the number and location of breakpoints for fitting piecewise constant baseline hazard functions in a three-state progressive process. Initial choice of breakpoints will rely on the median or tertiles from the nonparametric estimates of the Markov-based CDFs obtained under the model assumptions of chapter 2. From this starting point, the breakpoints will be systematically increased and decreased in order to find a better-fitting model. The “best” model chosen in this iterative approach will be compared to the model initially selected. Hence, we will examine the performance of this data-driven algorithm for selecting the breakpoints when adopting a piecewise constant hazard function model in a progressive three-state setting. Estimated standard errors for all model parameters will be obtained, but will only be reported for the regression parameters.

We consider four issues in our simulation study. Since the AIC is prone to overfitting, i.e., to choosing models which are unnecessarily complex, a goal of this study will be to evaluate the performance of the AIC with respect to the SIC. We expect the SIC will tend to select less complex models than the AIC. A second aim is to examine the effects of increasing the interval censoring on this model-selection process, when the true underlying number of pieces is one. Data sets with more uncertainty should result in less discrimination amongst possible models. Another purpose is to evaluate whether the strict use of quantiles from the nonparametrically estimated CDFs leads to better-fitting models, based on either model criterion, than allowing greater flexibility in the choice of breakpoints. The final goal will be to assess the estimated standard errors for the regression parameter, β, in the piecewise constant approach, by comparing results between the two estimation methods (nonparametric and piecewise constant).

3.2 Likelihood Function and Simulation Study

In the standard survival analysis where the event times are subject to interval censoring, the observed data log-likelihood for the piecewise exponential model includes sums of integrals, which do not have closed-form expressions. The complete data log-likelihood, on the other hand, turns out to involve sums of closed-form expressions. Thus, the EM algorithm is a natural method of choice for maximization of the log likelihood. In the three-state progressive setting, the complete data, with no interval or right censoring, is still very complex, so there is no apparent advantage in adopting the EM algorithm to maximize the observed data log-likelihood. The complexity is due in part to the bivariate form of the data and in part to the logistic specification for the hazard functions between all states. Hence, we chose to maximize the log likelihood of the incomplete data directly, using a quasi-Newton algorithm.

Both transition intensities are modelled semi-parametrically, although possible covariates are included only in the second transition intensity. Let the covariate process be $Z' = (z_1, z_2, \ldots, z_p)$, and take the baseline rate functions, $\eta(x)$ and $\lambda(t)$, to be piecewise constant. That is,

$$\eta(x) = \eta_h, \quad x \in E_h = (e_{h-1}, e_h],$$

and

$$\lambda(t) = \lambda_i, \quad t \in L_i = (l_{i-1}, l_i],$$

where $0 = e_0 < e_1 < \ldots < e_H < \infty$ and $0 = l_0 < l_1 < \ldots < l_I < \infty$ are the discrete time scales for the time-to-event variables, X and $T - X$, respectively. The logistic specifications for both transition intensities are now given by

$$\alpha(x) = \frac{\eta(x)}{1 + \eta(x)} = \frac{\eta_h}{1 + \eta_h}, \qquad x \in E_h,\ h = 1, \ldots, H\ ,$$

and

$$h(x, t; z) = \frac{\lambda(t)\ e^{\beta(t-x)+z'\boldsymbol{\nu}}}{1 + \lambda(t)\ e^{\beta(t-x)+z'\boldsymbol{\nu}}} = \frac{\lambda_i\ e^{\beta(t-x)+z'\boldsymbol{\nu}}}{1 + \lambda_i\ e^{\beta(t-x)+z'\boldsymbol{\nu}}}, \quad t \in L_i,\ i = 1, \ldots, I\ .$$

As we previously defined in chapter 2, the observed data for each individual includes an indicator variable (Δ_n) for the occurrence of the transition from state 1 to state 2, the time interval, $A_n = [x_l^n, x_u^n]$, during which this transition occurred, another variable (δ_n) indicating the occurrence of the transition from state 2 to state 3, and the time interval, $B_n = [t_l^n, t_u^n]$, during which this second transition occurred. If $\Delta_n = 0$, then we only know that $X_n \geq x_l^n$ and assume no knowledge is available about T_n, while if $(\Delta_n = 1, \delta_n = 0)$, then $T_n \geq t_l^n$. We use the additional indicator variables, $\epsilon_{hj} = 1$, to denote the event $x_j \in E_h$ and 0 otherwise, and $\rho_{ik} = 1$, for the event $t_k \in L_i$ and 0 otherwise. Support sets for the times of entry (X, T) into states 2 and 3 are given by $supp(X) = \{x_u, 1 \leq u \leq J\}$ and $supp(T) = \{t_r, 1 \leq r \leq K\}$, where J and K represent the respective sizes of the support sets.

Three types of contributions to the likelihood are possible. A subject who did not leave state 1 $(\Delta_n = 0, \delta_n = 0)$ contributes

$$\exp\left\{ - \sum_{x_u = x_1}^{x_{l-1}^n} \sum_{h=1}^{H} \epsilon_{hu} \log(1 + \eta_h) \right\} , \qquad x_u \in supp(X)$$

to the log likelihood function for the parameters of the model. Likewise, a subject labelled n who leaves state 1 for state 2 $(\Delta_n = 1)$ but who is never observed to occupy state 3 $(\delta_n = 0)$ gives rise to a contribution of the form

$$\sum_{x_j \in A_n} \exp\left\{ \sum_{h=1}^{H} \epsilon_{hj} \log\left(\frac{\eta_h}{1 + \eta_h} \right) \right\} \exp\left\{ - \sum_{x_u = x_1}^{x_{j-1}} \sum_{h=1}^{H} \epsilon_{hu} \log(1 + \eta_h) \right\} \times$$
$$\exp\left\{ - \sum_{t_r = x_{j+1}}^{t_{l-1}^n} \sum_{i=1}^{I} \rho_{ir} \log(1 + \lambda_i e^{\beta(t_r - x_j) + z' \boldsymbol{\nu}}) \right\} , \; x_u \in supp(X), \; t_r \in supp(T) \, .$$

Finally, an individual who leaves state 1 for state 2 $(\Delta_n = 1)$ and who is subsequently observed to occupy state 3 ($\delta_n = 1$) contributes

$$\begin{aligned}
&\sum_{x_j\in A_n}\sum_{t_k\in B_n}\exp\left\{\sum_{h=1}^{H}\epsilon_{hj}\log\left(\frac{\eta_h}{1+\eta_h}\right)\right\}\exp\left\{-\sum_{x_u=x_1}^{x_{j-1}}\sum_{h=1}^{H}\epsilon_{hu}\log(1+\eta_h)\right\}\times\\
&\quad\exp\left\{-\sum_{t_r=x_{j+1}}^{t_{k-1}}\sum_{i=1}^{I}\rho_{ir}\log(1+\lambda_i e^{\beta(t_r-x_j)+z'\boldsymbol{\nu}})\right\}\times\\
&\quad\exp\left\{\sum_{i=1}^{I}\rho_{ik}\log\left(\frac{\lambda_i e^{\beta(t_k-x_j)+z'\boldsymbol{\nu}}}{1+\lambda_i e^{\beta(t_k-x_j)+z'\boldsymbol{\nu}}}\right)\right\}, \qquad x_u\in supp(X),\ t_r\in supp(T)
\end{aligned}$$

to the log likelihood function.

The observed data log-likelihood can now be written as the sum of these contributions for the three distinct types of observations, viz.

$$\begin{aligned}
&\log L(\Theta) =\\
&\quad\sum_{\Delta_n=0,\delta_n=0}\left\{-\sum_{x_u=x_1}^{x^n_{l-1}}\sum_{h=1}^{H}\epsilon_{hu}\log(1+\eta_h)\right\}\\
&\quad+\sum_{\Delta_n=1,\delta_n=0}\log\sum_{x_j\in A_n}\exp\left\{\sum_{h=1}^{H}\epsilon_{hj}\log\left(\frac{\eta_h}{1+\eta_h}\right)\right\}\times\\
&\quad\exp\left\{-\sum_{x_u=x_1}^{x_{j-1}}\sum_{h=1}^{H}\epsilon_{hu}\log(1+\eta_h)\right\}\exp\left\{-\sum_{t_r=x_{j+1}}^{t^n_{l-1}}\sum_{i=1}^{I}\rho_{ir}\log(1+\lambda_i e^{\beta(t_r-x_j)+z'\boldsymbol{\nu}})\right\}\\
&\quad+\sum_{\Delta_n=1,\delta_n=1}\log\sum_{x_j\in A_n}\sum_{t_k\in B_n}\exp\left\{\sum_{h=1}^{H}\epsilon_{hj}\log\left(\frac{\eta_h}{1+\eta_h}\right)\right\}\times\\
&\quad\exp\left\{-\sum_{x_u=x_1}^{x_{j-1}}\sum_{h=1}^{H}\epsilon_{hu}\log(1+\eta_h)\right\}\exp\left\{-\sum_{t_r=x_{j+1}}^{t_{k-1}}\sum_{i=1}^{I}\rho_{ir}\log(1+\lambda_i e^{\beta(t_r-x_j)+z'\boldsymbol{\nu}})\right\}\\
&\quad\times\ \exp\left\{\sum_{i=1}^{I}\rho_{ik}\log\left(\frac{\lambda_i e^{\beta(t_k-x_j)+z'\boldsymbol{\nu}}}{1+\lambda_i e^{\beta(t_k-x_j)+z'\boldsymbol{\nu}}}\right)\right\}\ .
\end{aligned}$$

The simulation study involved five data sets, with the number of true underlying intervals of support in the simulated data being one, two or three for both transition intensity functions. An interval of support, or piece, was required to include at least two consecutive time points. The score or gradient function for the observed data log-likelihood was supplied in the quasi-Newton maximization routine but the matrix of second partial derivatives was estimated by finite-differencing. Only single realizations from each parameter configuration were studied due to the considerable time required for maximizing all of the models within a simulation run. For instance, in the data set specifying three true underlying pieces to model the first transition intensity, it took five days on a dedicated PC to fit the total of 1637 models.

Each data set consisted of approximately 100 subjects, whose event times were subject to right and interval censoring. To create a data set, random values were generated from different exponential distributions for each interval of support. So, if there were two pieces specified for the transition from state 1 to state 2 in the true data, about 100 random event times were created from two exponential distributions with different rate parameters, $\mu_{1_j}, j = 1, 2$; if a single interval of support or piece was adopted for the last transition, approximately another 100 random event times were created from yet a third exponential distribution. The exponential rate parameters varied from 1/4 to 2 in our study. Next, the data were discretized using a data-dependent partition. Dividing the time to the largest observed data value by 24, the number of observation times, the remaining data values were distributed so that about the same number of observations were allotted to each interval of time. Corresponding integer values were then assigned to the observations within each of the 24 intervals.

The next step in the data creation process was to impose right censoring. First, all the data values were subject to being randomly right censored in state 1. Then, from the uncensored subset, data values were now randomly subjected to being right censored in state 2. Right censoring of the event times resulted in 25% of the subjects ending the study in states 1 and 3, and the remaining 50% in state 2. To combine the right-censored data generated for each piece into a single, final data set, splitpoints were employed. Returning to the example with two intervals of support for the intensity between states 1 and 2, values less than time point six from one distribution were selected for this data set, while values greater than or equal to this splitpoint from the second distribution made up the rest of this data set. The values of the rate parameters used in this study were chosen, in part, to ensure that each piece or interval of support had the same amount of data in the final data set.

The last step taken in the creation of a data set was to impose interval censoring. For all subjects whose duration times were right censored in state 1, the left or lower interval endpoint was the discrete, true event time. For the remaining subjects who made a transition out of state 1, a random number generated from a discrete uniform distribution was added on to the discrete form of the original time value. An integer from zero to a specified maximum interval width could be appended, with the resulting new value representing the right or upper interval endpoint. Similarly, a new random uniform number from zero to the same specified width was subtracted from the original discrete duration time, forming the left or lower interval endpoint. This process ensured that the true value was contained in the constructed interval, and allowed for varying amounts of overall interval censoring. For duration times right censored in state 2, the left or lower interval endpoint was set to the last observation time point (24) for the study. The process of adding and subtracting a random uniform number from zero to a possibly different maximum interval width was repeated for the remaining uncensored data values. These remaining observations represented those subjects known to have made a transition to state 3.

The amount of interval censoring for transitions between states 1 and 2 was about two time points on average, with one exception. For the data set with single intervals of support for both transition intensities, the level of interval censoring for the transition intensity between states 1 and 2 was first set to about two time units on average, and then doubled in a second data set. The differential effect on estimation that resulted from increasing the interval censoring was of interest, but was only studied in the simplest case. The amount of interval censoring for the transition intensity between states 2 and 3 was about two time points on average in all five data sets.

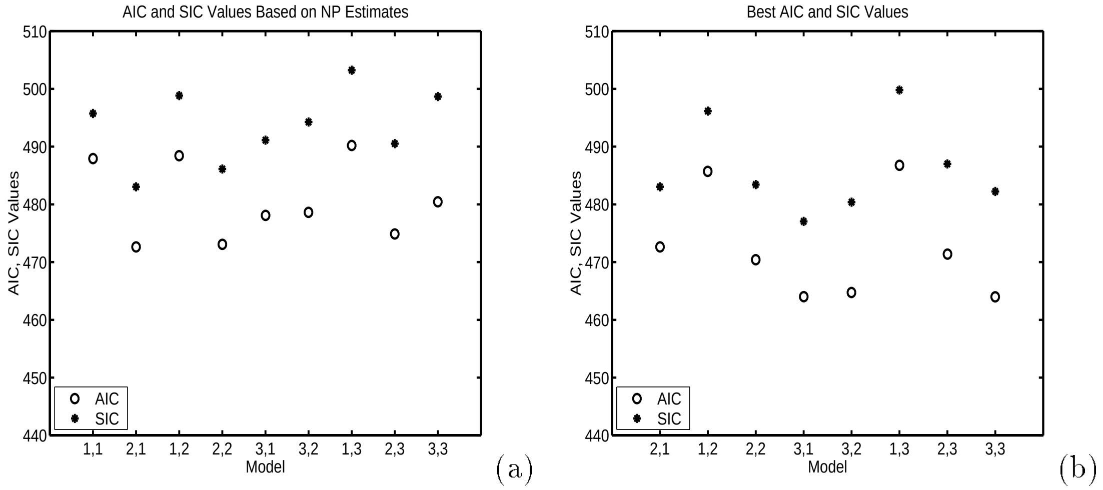
We now describe how the best-fitting piecewise-constant models in this progressive three-state process, as adjudged by the SIC and AIC values, were determined. First, using the final version of the simulated data set, nonparametric estimates of the CDFs were obtained using the basic model proposed by Frydman (1995), and described in §2.5.1. Second, a piecewise constant model was fitted, assuming single intervals of support (H = 1, I = 1) for both transition intensities. The parameter and related standard error estimates were found, as well as the SIC and AIC values. Next, assuming there were two intervals of support (H = 2) for the transition intensity between states 1 and 2, but only one interval of support (I = 1) for the intensity between the last two states, we fit a new piecewise constant model. The cutpoint for the two pieces was based on the quantiles of the nonparametrically estimated CDF for the duration in state 1 variable. New estimates and criterion values were also recorded.

The cutpoint was now systematically moved up to three time points above and below the initial changepoint, with a model fitted at each new time point. For each new fitted model, the SIC and AIC values obtained were compared to the smallest values recorded thus far. If either model criterion was smaller than any corresponding previous values, a new best model was noted. The parameter estimates, estimated standard errors, criterion values, and quantiles of the estimated CDF for the overall best fitting model were recorded. After determining the best fitting model assuming two and one intervals of support for the respective transition intensities, we now assumed that there were three intervals of support (H = 3) for the first transition intensity, and still only a single interval of support (I = 1) for the second. The same series of model fitting procedures ensued, with the same types of results recorded. We continued this method of fixing the number of intervals of support, fitting a model using the quantiles of the estimated CDF, and then finding a best-fitting model, for a total of nine piecewise constant settings.

The simplest type of data set considered consisted of a single true underlying distribution for each transition intensity and only minimal amounts of interval censoring. On average, the interval censoring in this data set for the first transition intensity was 1.82 units, and for the transition intensity between states 2 and 3, it was 1.85 units. The rate parameter, μ_1, for the first transition intensity was set at two, while the rate parameter, μ_2, for the second transition intensity was fixed at one. Using the quantiles from the nonparametrically estimated CDFs to select the changepoints initially, both the AIC and SIC were minimized in a model which included two pieces for the intensity of transitions between states 1 and 2 and a single piece for the intensity between the last two states. In the left panel of Figure 3.1, we see that this model is clearly the smallest for the SIC values (designated by closed circles) when the interval width for the first time to event is "narrow". In this plot, and in all other figures comparing AIC and SIC values, the horizontal scale identifies the model complexity associated with each model criterion value. At the first x-axis tick location, for example, (2,1) represents a model with two intervals of support ($H = 2$) for the transition intensity between states 1 and 2 and a single interval of support ($I = 1$) for the transition intensity out of state 2. For the AIC values (indicated by open circles) in the left panel of Figure 3.1, the model with two pieces in both hazard functions has a similar value to the minimal model. When the Best models are found by iterating the changepoints above and below the initial values, both the AIC and SIC are now minimized in a model with three pieces for the first hazard function and still a single piece for the second hazard function (see the right panel in Figure 3.1).

To graphically compare an estimated piecewise constant model with its non-

Figure 3.1: AIC and SIC Values for the data set with a single interval of support for each transition intensity, narrow amount of interval censoring.



parametrically estimated counterpart, we defined the cumulative hazard functions for X and $T - X|X$ in the piecewise constant model as

$$A(x) = \sum_{j:x_j<x} \sum_{h=1}^{H} \epsilon_{hj} \, \frac{\eta_h}{1+\eta_h} \,, \quad H(x,t) = \sum_{j:x_j<x} \sum_{k:x_j<t_k<t} \sum_{i=1}^{I} \rho_{ik} \, \frac{\lambda_i \; e^{\beta(t_k - x_j)}}{1+\lambda_i \; e^{\beta(t_k - x_j)}} \,.$$

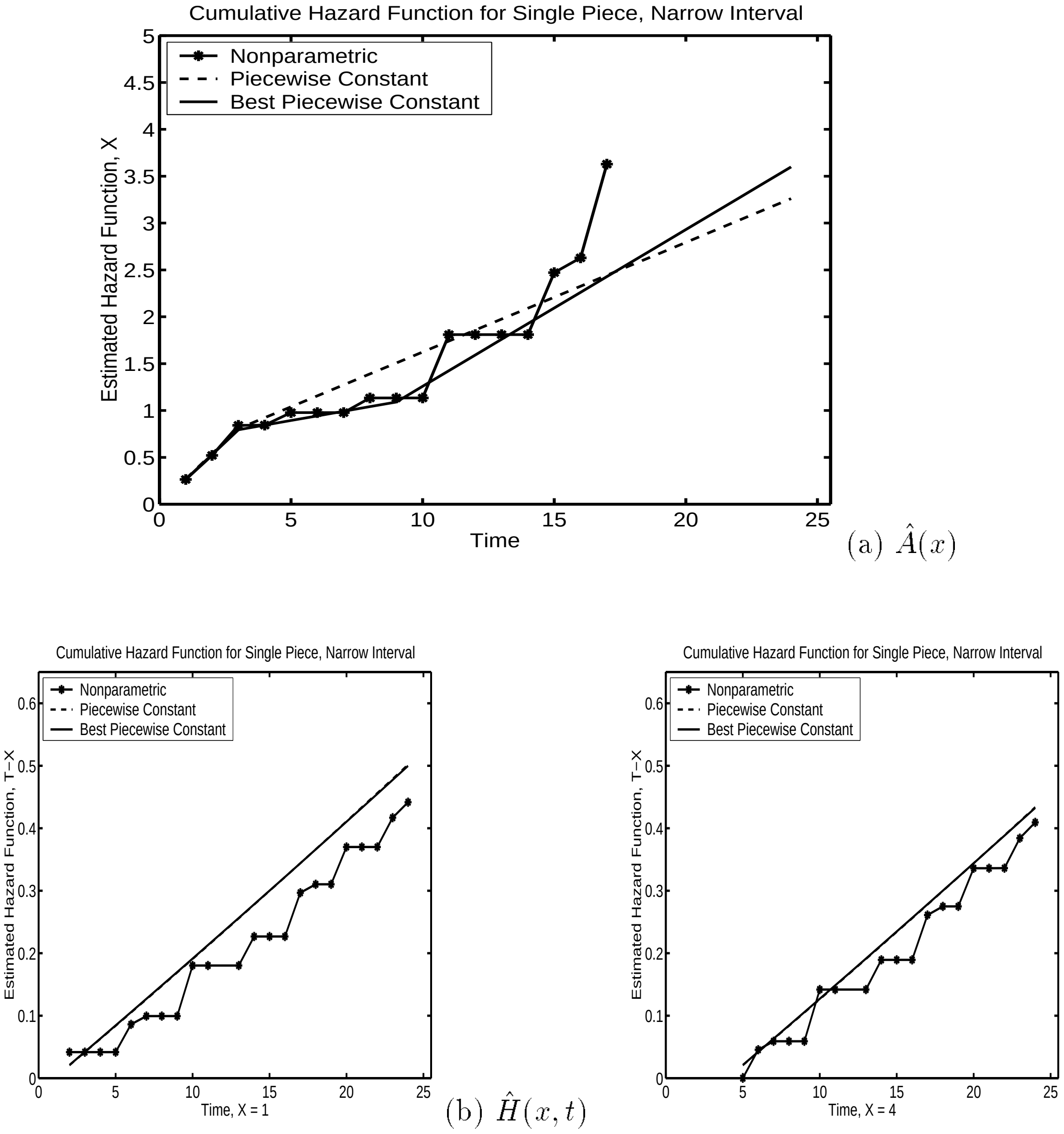
The indicator functions, ϵ_{hj} and ρ_{ik} are defined above. Since $H(x,t)$ depends on the values of T and X, we fixed X at two different time points in order to study the behaviour of this cumulative hazard function at different cross sections of duration time in state 1. Although the value of X could vary for each value of T, we chose to use the smallest possible value of X for each data set, as well as one of moderate size. This moderate value corresponded to about the 25th percentile from the support set for X; this choice helped ensure that there were still enough time points in the support set for the transition out of state 2 to compare functions, since by assumption, $T > X$. In the nonparametric model, we fixed the value of X at the same two time points and defined the cumulative hazard functions for X and $T - X|X$ as

$$A(x) = \sum_{j:x_j<x} \frac{\eta(x_j)}{1+\eta(x_j)} \ , \quad H(x,t) = \sum_{j:x_j<x} \sum_{k:x_j<t_k<t} \frac{\lambda(t_k)\, e^{\beta(t_k-x_j)}}{1+\lambda(t_k)\, e^{\beta(t_k-x_j)}} \ .$$

The nonparametric estimated cumulative hazard functions should be specified as step functions in the various plots; however, to facilitate visual comparison with the piecewise constant model estimates, we chose to join the points as De Gruttola and Lagakos (1989) and Kim *et al.* (1993) did.

Looking at the plots of the cumulative hazard functions of the nonparametrically estimated model, and the Initial and Best piecewise constant models in the top panel, (a), of Figure 3.2, we see that both piecewise constant models closely follow the nonparametric curve from the first time point until the fourth. The Best piecewise constant model appears to follow the nonparametric curve very closely until time point 10, while the Initial piecewise constant model tends to overestimate the nonparametric cumulative hazard function in this interval. Neither piecewise constant model follows the estimated nonparametric curve very closely beyond 10 time units. The sharp rise at time point 17 in the nonparametrically estimated curve is likely due to two possible failures in a risk set of size two; the nonparametric estimates will be sensitive to events in the small risk set, while the weakly parametric ones will not. In the lower left panel of Figure 3.2, when X is fixed at time point one, both piecewise constant curves coincide with each other and tend

Figure 3.2: Nonparametric, piecewise constant, and Best piecewise constant estimated hazard functions for the data set with a single interval of support for each transition intensity, narrow amount of interval censoring.



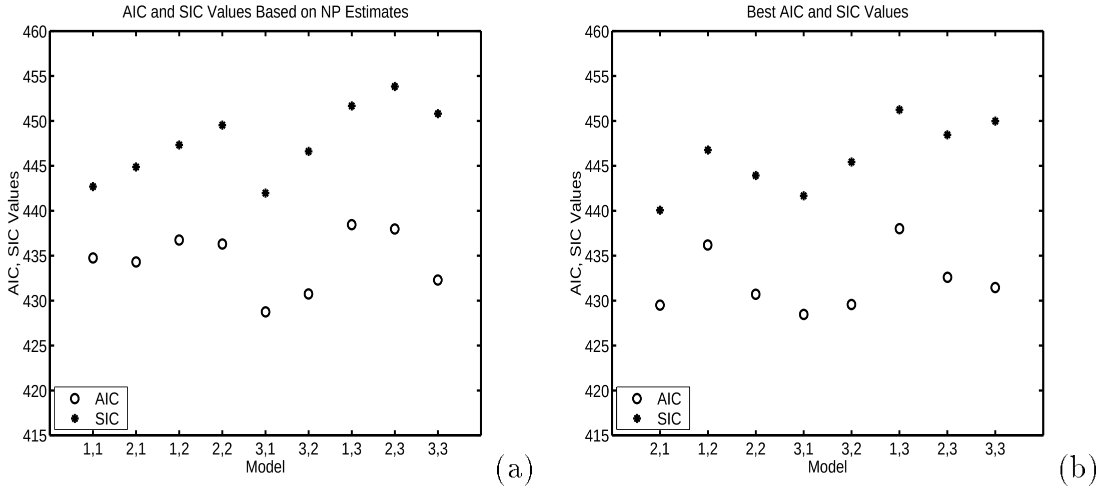
(a) $\hat{A}(x)$

(b) $\hat{H}(x, t)$

to lie above the nonparametric estimate of $H(x,t)$. The nonparametric curve has more but smaller jumps in this second transition intensity. Similar observations are made when X is set to four in the lower right panel of this same figure. Overall, the Best model does seem to follow the nonparametrically estimated cumulative hazard functions more closely than the initial estimates that rely on the quantiles of the nonparametrically estimated CDFs.

The same exponentially-distributed data was now subjected to a greater measure of interval censoring but the same amount of right censoring. In this "wide" case, the average width of the interval containing the true value of X was 3.62 units; the average value of the right censoring occurring in the transitions from states 2 to 3 was 1.74 units, not very different from the value of 1.85 observed in the corresponding narrow, single piece case. Using the quantiles of the nonparametrically estimated CDF, both the AIC and SIC were minimized in models with three pieces for the first intensity and a single piece for the second intensity. In the left panel of Figure 3.3, we can see that the AIC clearly favours this model (label 3,1). The SIC is minimized with this model complexity too, although the original model with one piece needed to model each hazard is a close second choice. Plotted in the right panel, (b), of Figure 3.3 are the AIC and SIC values obtained in the Best fitting models. The AIC is still minimized with a model of the same complexity as in the Initial model, but there is not a lot of separation, and hence discrimination, between this model and a model with two pieces in the second transition intensity. A third model with two pieces in the first intensity and one piece in the last intensity appears to fit well, too. The SIC values, on the other hand, clearly suggest a model

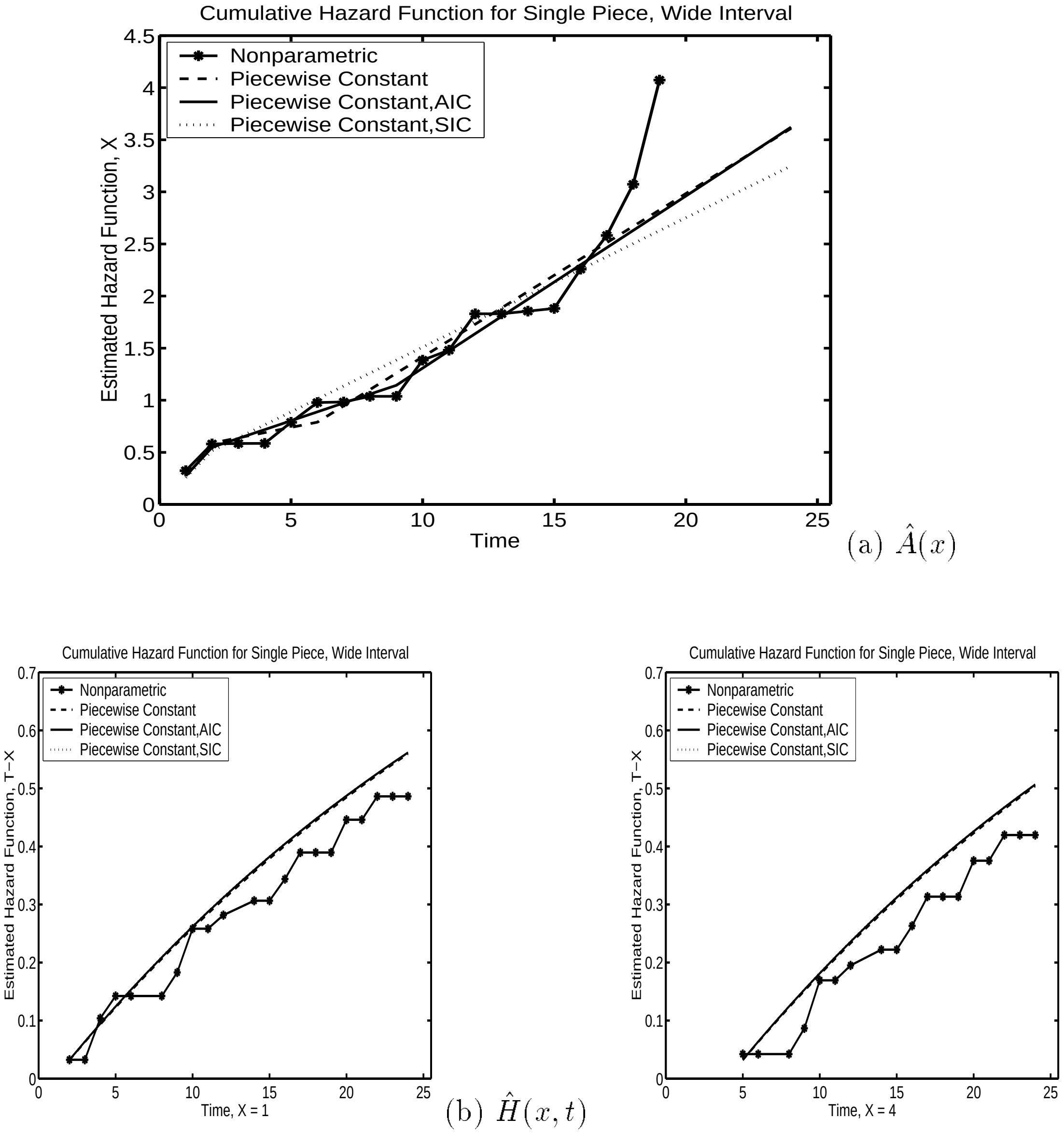
Figure 3.3: AIC and SIC values for the data set with a single interval of support for each transition intensity, wide amount of interval censoring.



with two and one pieces, respectively, for the first and second transition intensities.

Once again, the plots of the piecewise constant and nonparametric fitted models can be compared graphically. All of the fitted piecewise constant model curves seem to follow the nonparametrically estimated curve well until time point 17 for the first transition intensity function; see Figure 3.4, (a). The estimated piecewise constant models with three intervals of support for the sojourn time in state 1 generally appear to follow the changes in the slope of the nonparametric estimated cumulative hazard function a little closer. The nonparametric curve once again takes two large jumps at time points 18 and 19, likely due to possible events in very small risk sets there; the fitted piecewise constant models do not follow these jumps. The nonparametric estimate of the intensity function for transitions out of state 2 does not have any large jumps nor any abrupt changes in its slope; see both

Figure 3.4: Nonparametric, piecewise constant, and Best piecewise constant estimated hazard functions for the data set with a single interval of support for each transition intensity, wide amount of interval censoring.



(a) $\hat{A}(x)$

(b) $\hat{H}(x,t)$

lower panels in Figure 3.4. All the estimated piecewise models, which have a single interval of support for the distribution of $T - X|X$, follow the nonparametrically estimated curve quite closely for the first 10 support points, then lie above the nonparametrically estimated curve. When the value of X is increased to four, the estimated piecewise models still coincide with one another but now lie mostly above the nonparametric estimate of $H(x, t)$. Thus, compared to the data set with less interval censoring, the piecewise constant models tended be slightly more discrepant from the nonparametric estimated cumulative hazard function when the duration in state 1 variable is fixed at time point four, but similar when X is set to one. The apparent curvature in the estimated cumulative hazard functions for the piecewise constant models is likely due to the influence of a larger estimated value of β in the logistic specifications of the hazard functions.

Cases with two true underlying pieces in the second intensity and one or two true underlying pieces in the first intensity were also fit. The results obtained from fitting these situations are summarized in Table 3.1. The original number of pieces for the first transition is labelled H, while the original number of pieces for the second is labelled I. The values of the rate parameters used to generate the data for each piece are indicated in the corresponding column values for $\mu_i, i = 1, 2$. The paired values in the body of the table under the column headings for AIC and SIC refer to the number of intervals of support selected for modelling each transition intensity by that model criterion; e.g., (3,1) would be a model with three pieces used to model the first transition intensity function and one piece for the second transition intensity function.

Table 3.1: Piecewise constant models which minimize the model criteria in the data sets with two intervals of support for at least one transition intensity function.

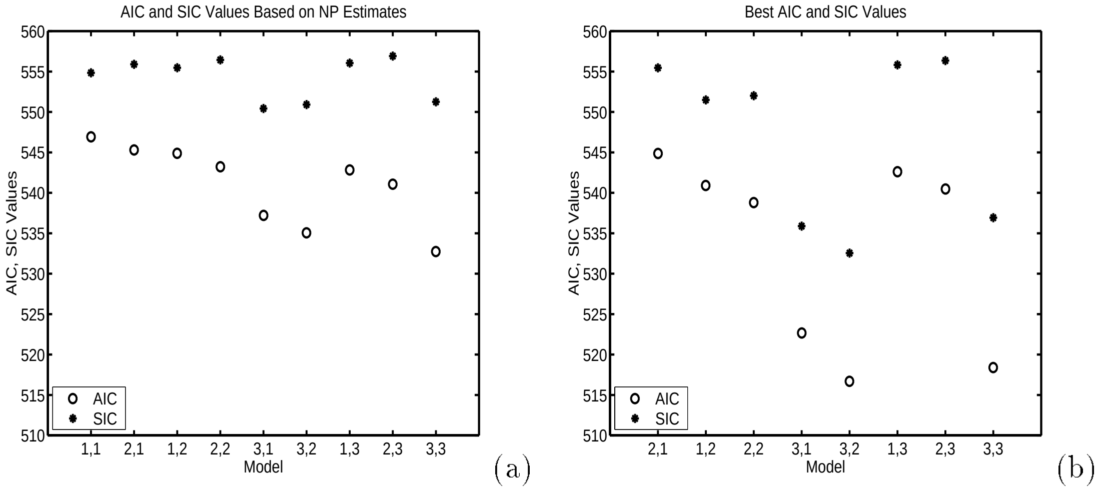
Original Number of Pieces		Rate Parameter Values		Initial Model		Best Model	
H	I	μ_1	μ_2	AIC	SIC	AIC	SIC
2	2	2, 1/2	1	(3,3)	(3,1)	(3,2)	(3,2)
1	2	1	2, 1/2	(1,1)	(1,1)	(3,2)	(3,1)

The breakpoints in the data sets were chosen to help ensure sufficient information in each piece. Nonparametric estimates of cumulative distribution functions, which were based on data generated with different exponential rate parameters, were used to help select these cutpoints. For the data set having two support intervals for the first intensity function, the breakpoint was fixed at time point five. A breakpoint in the support set for the first intensity function necessarily induces a breakpoint in the support set for the second intensity function because of the underlying model for the data. Recall that T, the chronological time on study variable, was the sum of the time to the first event variable, X, and the time to the second event variable, V, which can only occur after the first event was observed. Hence, a two-piece intensity function for the second transition hazard resulted from creating a two-piece intensity function for the first transition hazard. The true breakpoint in the second transition intensity was only known to be larger than the fifth time point. In the data set having a single interval of support for the intensity function for transitions between states 1 and 2, and two intervals of support for the transition intensity function between states 2 and 3, the breakpoint in the second intensity function was set at time point 11.

Neither the AIC nor the SIC selected the original model configuration in either of the two-piece data sets. For the data set comprised of two pieces in both transition intensities ($H = 2, I = 2$), the AIC and SIC favour models with three pieces between states 1 and 2 (see Table 3.1, first row). This level of model complexity occurred whether the quantiles for the nonparametrically estimated CDFs were used or not. When the quantiles were used (Initial model), a model with only a single interval of support for the transition between states 2 and 3 yielded the smallest SIC value. The smallest AIC value was found for a model employing three intervals of support for this transition intensity.

Looking at the plots of the AIC and SIC values in Figures 3.5 (a) and (b), it is clear that models with three pieces instead of the original number of two pieces for the first transition intensity minimize both model criteria. The plot of the estimated cumulative hazard functions in the left side of Figure 3.6 (b) reveals an apparent change in the nonparametric estimated function after time point six — one time point beyond the earliest possible change in the original underlying intensity. As with the previous two data sets, the estimates of $H(x,t)$ from the piecewise constant models tend to coincide with the nonparametric estimated version of the cumulative hazard function, and with one another for both values of X (one, six). Only the estimated cumulative hazard curve from the Initial piecewise constant model which minimized the SIC lies some distance from the other estimated curves. When X is at its smallest value (one), this estimated piecewise constant hazard function, with only a single interval of support, tended to lie above the other curves, and when X

Figure 3.5: AIC and SIC values for the data set with two intervals of support per transition intensity.

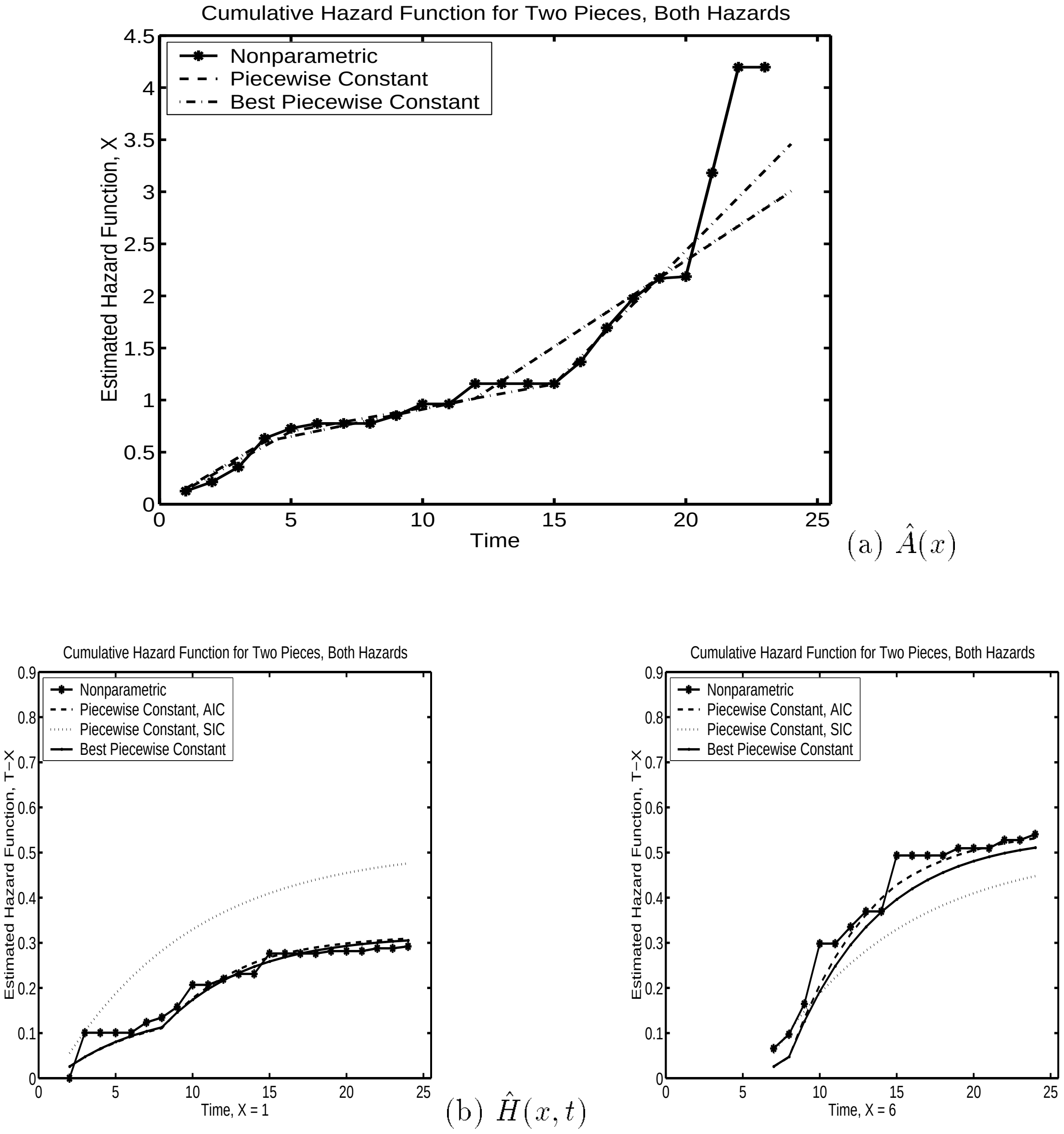


is increased to six, it tended to lie below them.

In the upper panel, (a), of Figure 3.6, the estimated nonparametric curve appears to change at three or four time points, i.e. at $x = 4, 16, 21$, and 22. The large jumps at time points 21 and 22 are likely due to possible events in the small risk sets occurring then, and so may be misleading. The Best model selected by both the AIC and SIC, with three pieces used to model the first transition intensity and two for the second transition intensity, do seem to agree graphically with the estimated nonparametric model over most of the study duration.

For the data set consisting of a single interval of support for the first intensity function and two intervals of support for the second $(H = 1, I = 2)$, both model criteria are minimized, initially, using the quantiles of the nonparametrically estimated CDFs in the simplest of all models — a single piece for each intensity (see

Figure 3.6: Nonparametric, piecewise constant, and Best piecewise constant estimated hazard functions for the data set with two intervals of support for each transition intensity.



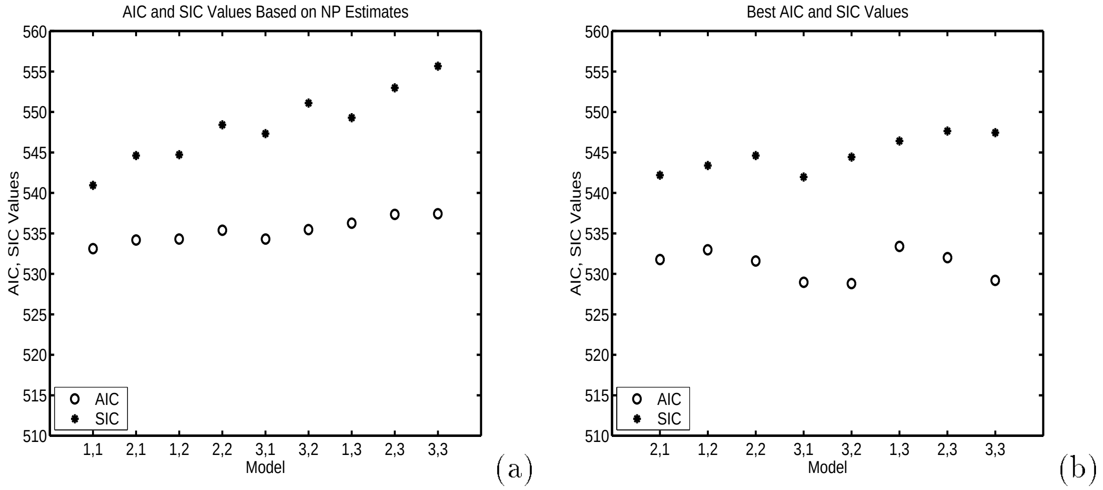
(a) $\hat{A}(x)$

(b) $\hat{H}(x, t)$

Table 3.1, second row). Moving the breakpoints above and below the initial ones resulted in the AIC and SIC selecting slightly different Best models. Both criteria are minimized in models having three intervals of support for the first intensity function, but only the AIC identifies the original model that had two intervals of support for the second intensity function. Looking at the plot of AIC and SIC values in Figure 3.7 (a), we see that for Initial models, there is very little discrimination amongst the AIC values; the points essentially form a straight line. The SIC values, on the other hand, do show more vertical variability and, hence, fairly good discrimination among competing models. The model with a single piece comprising each intensity is clearly favoured by both model criteria. When the breakpoints are chosen without relying on the quantiles from the estimated distributions, the AIC is smallest for models using three pieces in the first transition intensity, and any number of pieces used in the second transition intensity (see Figure 3.7 (b)). The model with two pieces in the second intensity was only slightly better, as judged by the AIC, than models with one or three pieces. The SIC also favoured a model with three pieces for the first transition intensity, but only a single piece for the second transition intensity.

The plots comparing the nonparametric estimated cumulative hazard function with their estimated piecewise constant counterparts provide graphical checks on the complexity of the models identified by minimal AIC and SIC values. In the upper panel of Figure 3.8, we see that the estimated piecewise constant cumulative hazard function from the Initial model follows the nonparametrically estimated cumulative hazard function fairly closely until the large jump at time point 23. It

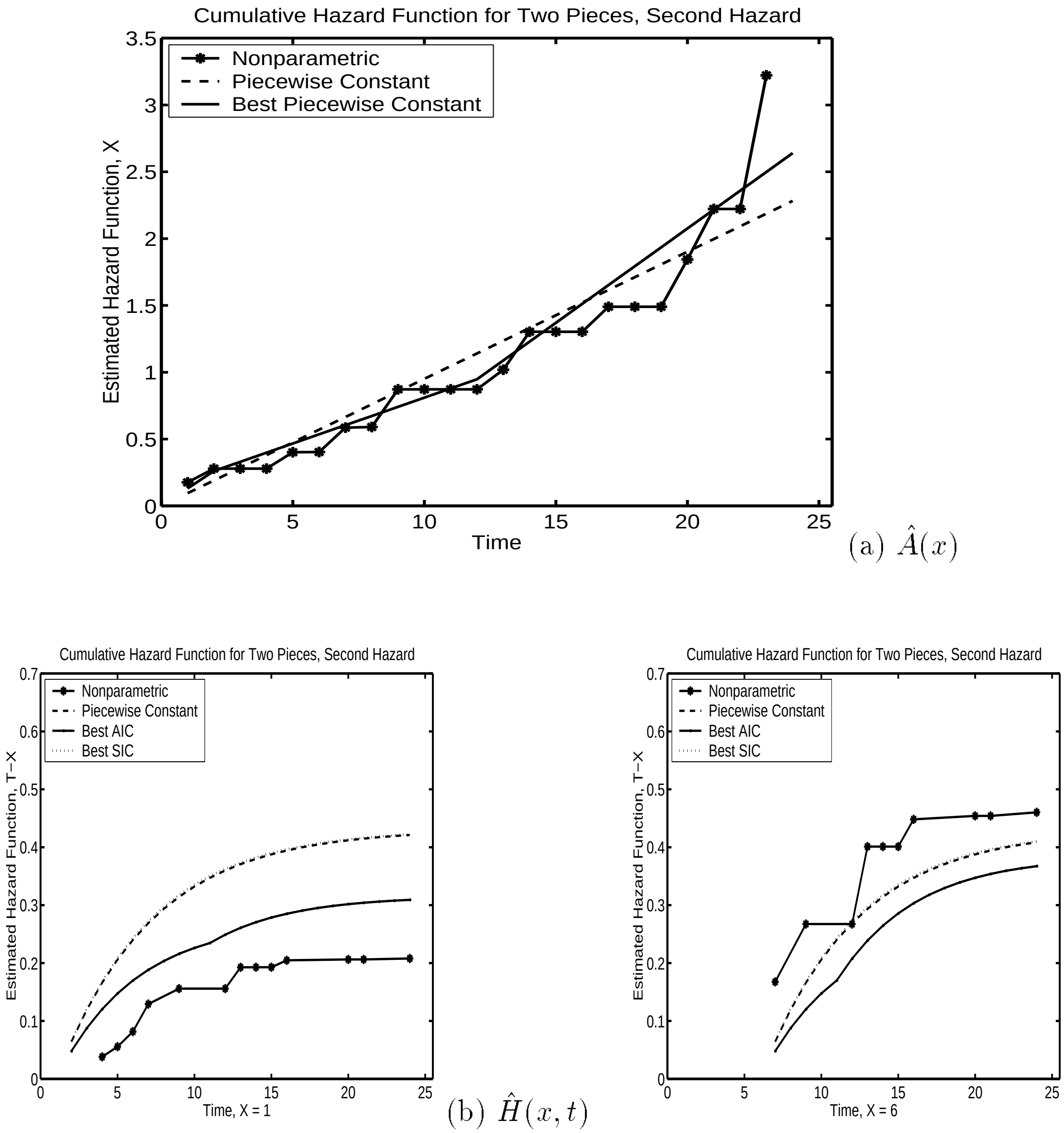
Figure 3.7: AIC and SIC values for the data set with one and two intervals of support for the transition intensity between states 1 and 2, and between states 2 and 3, respectively.



tends to overestimate the nonparametric curve, except for a couple of time points at the beginning and end of the observation period. The Best models selected by the AIC and SIC, which were found after iterating about the initial choice of breakpoints, were identical in model complexity and breakpoint locations for this first intensity function. The estimated cumulative hazard function from this Best model did not consistently over- or underestimate the nonparametric curve, and also seemed to follow the increase in the nonparametric curve past time point 17.

In the lower left panel of Figure 3.8, when $X = 1$, we see that only the Best model selected by the AIC correctly identified the breakpoint at time point 11 in the intensity function between states 2 and 3. The two single piece models do not capture the change in the true hazard function, and their estimated curves tend to lie considerably above the other two. However, when X is increased to six, the

Figure 3.8: Nonparametric, piecewise constant, and Best piecewise constant estimated hazard functions for the data set with one and two intervals of support for the transition intensity between states 1 and 2, and between states 2 and 3, respectively.



(a) $\hat{A}(x)$

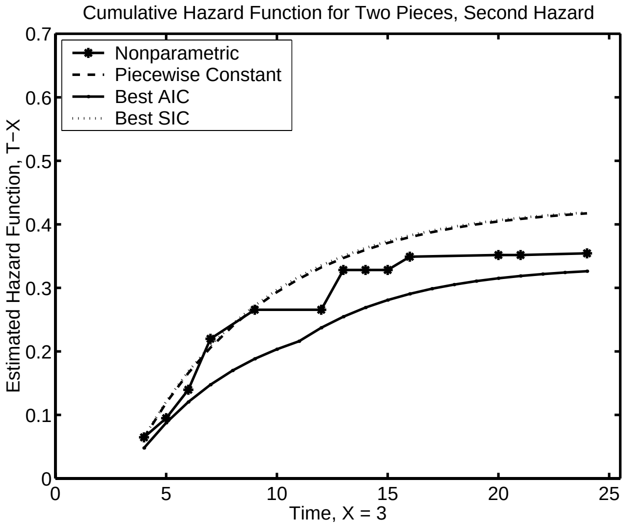
(b) $\hat{H}(x, t)$

three estimated cumulative hazard functions from the piecewise constant models tend to agree with one another and now lie beneath the nonparametric estimate of $H(x,t)$. Based on these graphical comparisons, it appears that the Best model determined by the AIC has identified a good model for these data.

The apparent discrepancy between the estimated piecewise constant and nonparametric curves may be due to the differences in the estimates of the regression parameter, β. In the nonparametric setting, the estimate of β was -0.279 (LRS = 19.06), while in the piecewise constant models, the values ranged from -0.16692 to -0.19773. Another explanation for the apparent discrepancy is the relatively large differences between the smallest value in the support sets for the transitions out of state 2. In this data set, the fourth not the second time point, was the smallest value in the support set for the nonparametric approach. Conditioning on $X = 3$ results in less separation between the estimated estimated cumulative hazard functions; see Figure 3.9. The roughness of the nonparametric estimate of $H(x,t)$ may also be a contributing factor to the observable discrepancy between the estimated curves.

The last data set we considered had three intervals of support for the transition intensity between states 1 and 2. Breakpoints for this first transition intensity were set at time points 5 and 14. The rates used to generate the data were fixed at 1/2, 2, and 1/4 for each corresponding interval of support, while the rate used to generate the data for the second intensity function was fixed at 1. The partition of the support for the first intensity function should induce a partition in the support for the second intensity function in this data set too. A model adopting three pieces

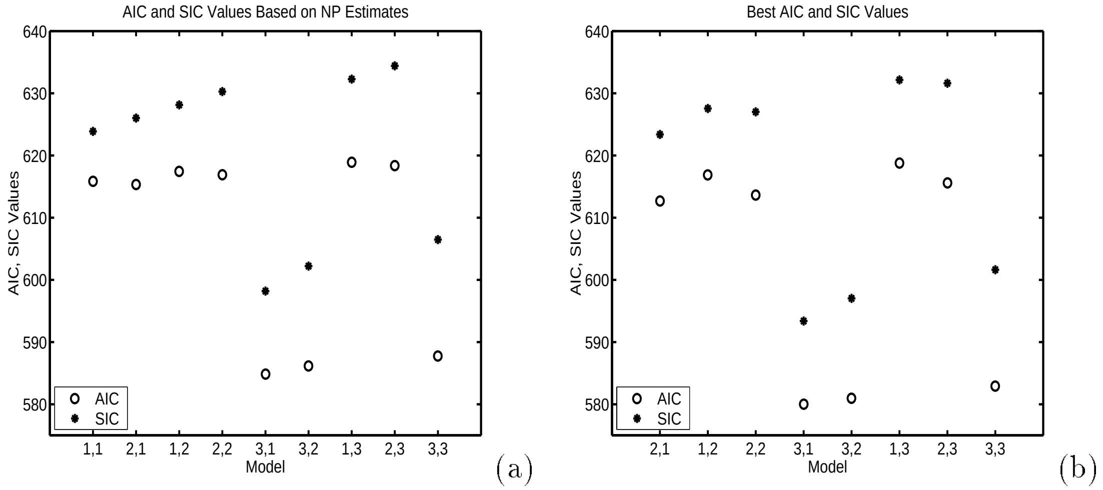
Figure 3.9: Checking on the apparent discrepancy between the nonparametric and piecewise constant estimated hazard functions for the data set with one and two intervals of support for the transition intensity between states 1 and 2, and between states 2 and 3, respectively. $X = 3$ in this plot.



for the intensity function between states 1 and 2 and a single piece for the intensity function between states 2 and 3 minimized both model criteria when the quantiles from the nonparametrically estimated CDFs were used to select the breakpoints. This Initial model identified breakpoints at time points 3 and 16. The same model complexity was chosen by both criteria when the breakpoints were systematically increased and decreased about the initial points, however, the optimal breakpoints identified in this "best" approach were 4 and 15.

The plot of the AIC and SIC values obtained using the initial breakpoints (Figure 3.10 (a)) shows that three pieces for the first intensity clearly minimize both model-selection criteria. There is considerable vertical separation between these

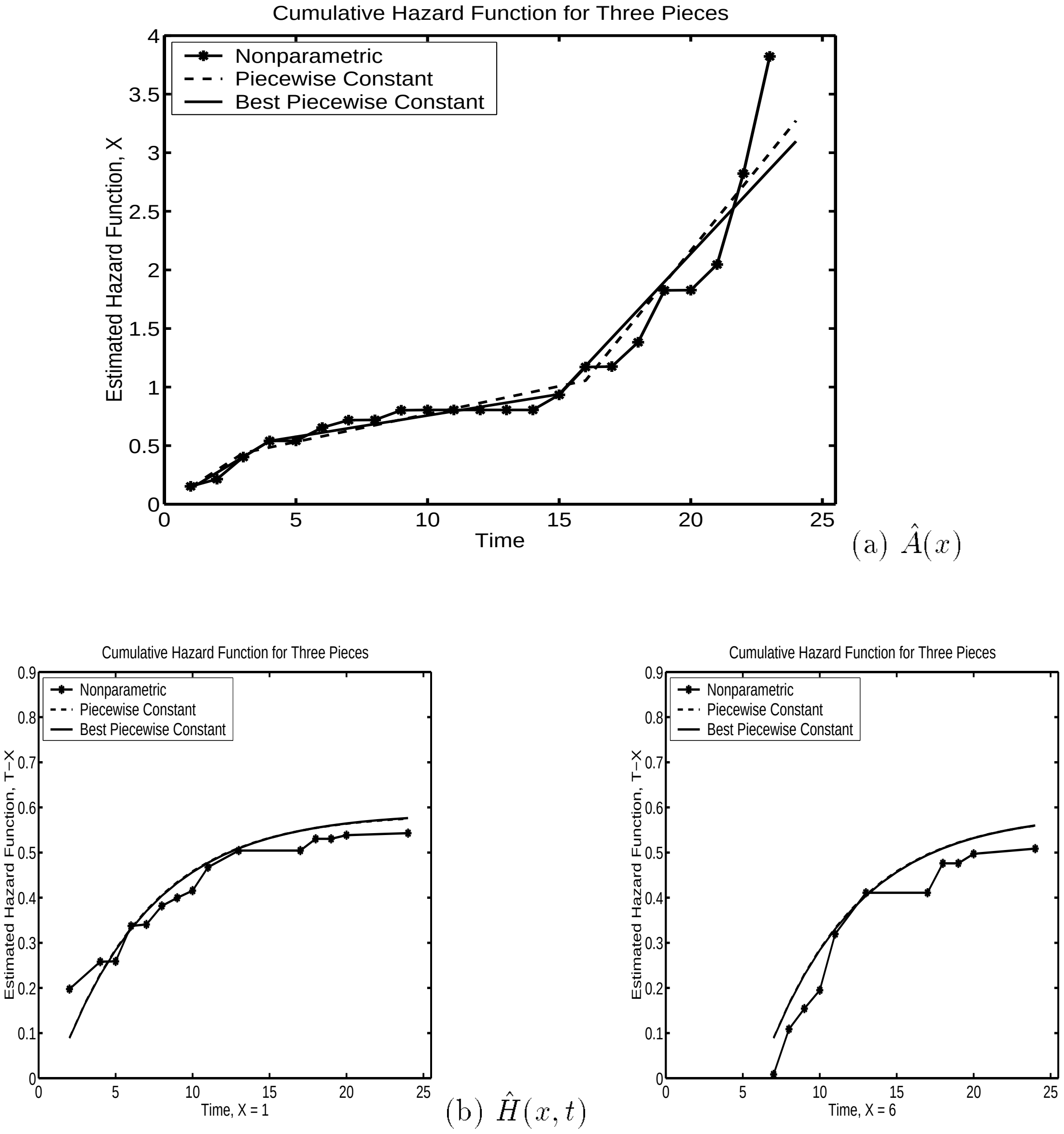
Figure 3.10: AIC and SIC values for data set with three intervals of support for each transition intensity.



three-piece models and models with one or two intervals of support. When these initial breakpoints are systematically increased and decreased to discover better fitting models, the same patterns in the model-selection criteria values emerge (see Figure 3.10 (b)). Thus, three intervals of support for the first transition intensity are correctly identified for these data. The number of pieces used to model the transition intensity between states 2 and 3 is consistently identified as one, and not three, however.

Graphical comparisons between the estimated cumulative hazard functions from the piecewise constant models and from the nonparametric model permit comparisons between the model criteria, and also provide insight into consistencies between the two modelling approaches. In the upper panel of Figure 3.11, the piecewise constant estimated cumulative hazard functions follow the nonparametric estimated

Figure 3.11: Nonparametric, piecewise constant, and Best piecewise constant estimated hazard functions for the data set with three intervals of support for each transition intensity.



(a) $\hat{A}(x)$

(b) $\hat{H}(x,t)$

curve very closely until time point 18. As with the other four data sets previously considered, there is a large jump in the nonparametric estimated curve at the last data point in the support set for this method of estimation. The risk set here includes only two individuals. The changepoints (4 and 15) determined in the Best piecewise constant model are closer to the true changepoints (5 and 14) than the changepoints (3 and 16) identified in the Initial model.

In both lower panels of Figure 3.11, the estimates of $H(x,t)$ from the Best and Initial piecewise constant models are indistinguishable from one another. They also follow the nonparametric estimate of the cumulative hazard function very closely for both values of X (one and six). Although these piecewise constant estimated curves employ only a single interval of support for this second transition intensity, additional pieces do not seem necessary.

We also investigated the standard error estimates for the regression parameter, β. To assess the adequacy of estimating the variability in this piecewise constant setting, we compared the significance levels of two different statistics evaluating the importance of the duration in state 2 covariate. These statistics, with different asymptotic distributions, should provide approximately equivalent answers for the data sets we constructed. In the piecewise constant setting, we calculated a z-statistic, where $z = \hat{\beta}/est.\,se(\hat{\beta})$, while in the nonparametric setting, we calculated a likelihood ratio statistic (LRS). The LRS was formed by comparing the values of the maximized log likelihoods for models with and without the duration in state 2 covariate. In Table 3.2, we report the results for the data set consisting of a single interval of support for both transition intensities, where the interval censoring is about two units ("narrow" case). The nonparametric estimate of β was -0.008 for this data set. The value of the LRS was 0.4723, with a corresponding p-value of 0.492. This statistic indicates that in the nonparametric framework, there is no evidence against the hypothesis $\beta = 0$. The z-values also lead to the same conclusion, although the strength of the conclusion or p-values are not the same (see Table 3.2, fifth column). The estimated values of β are near zero in the two settings, but the signs are positive in the piecewise constant models.

Table 3.2: Z statistics and associated p-values for testing $\beta = 0$ in the data set with single intervals of support per transition intensity, narrow amount of interval censoring.

Model	H	I	Z	p-value	$\hat{\beta}$
Initial	2	1	0.0083	0.993	0.004
Best	3	1	0.0068	0.995	0.003

When we consider a data set where the duration in state 2 covariate is important in the nonparametric setting, we do not find the same agreement between these statistics. Using the case where the transition intensity functions were composed of three pieces, we calculated a LRS and two z-values for the previously identified Initial and Best models (see Table 3.3). In the nonparametric setting, duration in state 2 is very important: the LRS is 15.375, with a significance level of 0.00009. The magnitude of the parameter estimate, -0.192, is larger in this case but still negative. The calculated z-values do not lead to the same conclusion about time in state 2 for the piecewise constant models. The p-values are both about 0.82, indicating no support for the hybrid model. Interestingly, the parameter estimates for β are quite similar in direction and magnitude in the two settings.

Table 3.3: Z statistics and associated p-values for testing $\beta = 0$ in the data set with three intervals of support per transition intensity.

Model	H	I	Z	p-value	$\hat{\beta}$
Initial	3	1	-0.2242	0.823	-0.174
Best	3	1	-0.2123	0.832	-0.171

To investigate this discrepancy between these two statistics, interval estimates were obtained using a profile likelihood approach. As in chapter 2, the value of β was fixed and the log likelihood was maximized to obtain estimates for the remaining model parameters. The values recorded in the last column of Table 3.4 are the 95% confidence interval estimates, assuming a χ^2 distribution with a single degree of freedom. The interval estimates of β in the nonparametric and both piecewise constant modelling approaches are very similar and do not include the null value of zero. The intervals are nearly symmetric in the piecewise constant models, but not in the nonparametric one. In this latter setting, the lower endpoint is about 1.4 times further from the point estimate than the upper endpoint.

Table 3.4: Point and interval estimates of β obtained from profile log likelihoods.

			Estimates for β	
Model	H	I	Point	Interval
Nonparametric			-0.192	(-0.332,-0.090)
Initial, Piecewise	2	1	-0.174	(-0.279,-0.087)
Best, Piecewise	3	1	-0.171	(-0.276,-0.085)

The results for the other three data sets were similar and are not reported here. In Appendix A, the parameter and standard error estimates for the Initial and Best models identified as minimizing the AIC and SIC in each of the five data sets are recorded.

To summarize our findings for the five data sets studied, we will consider the various aims of the simulation study in turn.

(a) The SIC had better discrimination than the AIC between competing models in several of the data sets. Most times, however, the two criteria selected the same models, both Initial and Best. Out of the 10 possible cases — five data sets with two modelling approaches per data set — the SIC selected less complex models three times.

(b) Increasing the interval censoring in the data set generated with a single interval of support for both transition intensities did not seem to affect the discrimination between competing models. There appeared to be good discrimination between competing models in both cases, although the range of all possible models was surprisingly smaller in the "wide" interval censoring case. When the interval censoring between state 1 and state 2 was increased, the Initial models became more complex but the Best model determined by the SIC actually became simpler. Thus, the results were not definitive.

(c) The piecewise constant models identified as Best tended to follow the nonparametric estimated cumulative hazard functions more closely than the Initial estimates that relied on the strict use of quantiles from the nonparametrically estimated CDFs. The only exception to this pattern regularly occurred at the last time point in the support set for the transition to state 2, where the risk sets are usually small.

The complexity of the Initial models selected by either model criteria was the same or smaller than the Best models in all of the data sets, with only two exceptions among the 10 possible cases. Thus, piecewise constant models which did not rely on the quantiles from the nonparametrically estimated CDFs seemed to fit the data better, although they were generally more complex than models which did use this information.

The changepoints in the Best models found by iterating above and below the initial changepoints were closer to the changepoints in the discrete version of the original data. For the first transition intensity, the Best models either identified or came closer than the Initial models to the true changepoint in two of the three data sets with changepoints in this transition intensity. The same result was observed for the second transition intensity function.

(d) If the LRS for the nonparametric model indicates that a hybrid model is appropriate, the z statistic based on the estimated parameter and standard error from the piecewise constant model do not yield the same conclusion. The parameter estimates — point and interval — in both settings were very similar, although the nonparametric estimates were always slightly larger in magnitude. These findings suggest that the estimation of the standard errors in this piecewise constant setting is not very accurate. One possible reason may be that the matrix of second partial derivatives was estimated using finite-differencing from the log-likelihood function rather than explicit evaluation.

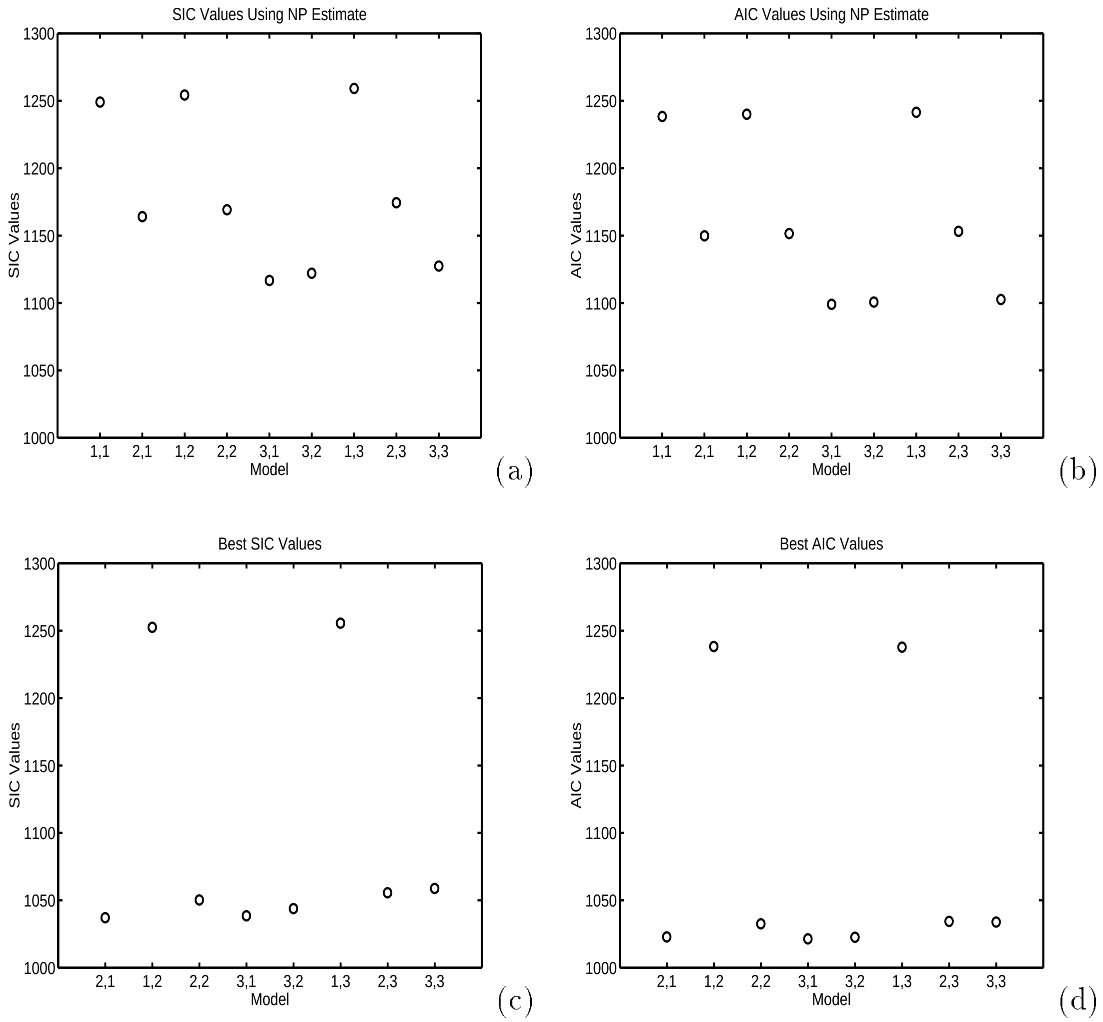
3.3 Example: AIDS in Hemophilia Patients

We analysed the example considered in chapter 2 under the assumption that the baseline intensities were piecewise constant. The hazard functions between states 1 and 2 and states 2 and 3 were modelled as having one, two, or three intervals of support each. An interval of support, or piece, was still required to include at least two consecutive time points. The results from fitting these nine combinations are presented in Table 3.5. The number of intervals of support used within each model is recorded in the first column; H refers to the number of pieces for the time to infection variable, X, whereas I refers to the number of pieces for the time to AIDS diagnosis variable, $T - X$, from the infection time. Models based on the quantiles, $\hat{F}$, of the nonparametrically estimated CDFs for X and $T - X$ are labelled Initial Models whereas models chosen by iterating above and below the changepoints selected by the Initial models are labelled Best Models. The model with the smallest value of either the AIC or SIC was selected as the Best model for the given model specification. Within all nine combinations considered, the model with the minimal value of the AIC corresponded to the model with the minimal value of the SIC. Hence, Best Models refer to the best models for both the AIC and SIC *within* that model configuration. The nonparametric estimate for the CDF between states 2 and 3 was based on a Markov model, so the explanatory variable, duration in state 2, was not included. The largest observed value of $\hat{F}_x$ was 0.9349, and for $\hat{F}_{t-x}$ the value was 0.3959. The values of $\hat{F}_x$ and $\hat{F}_{t-x}$ at the breakpoint(s) for the Initial and Best models in all nine piecewise combinations are also reported in Table 3.5.

Table 3.5: AIC and SIC values, and associated $\hat{F}_x$ and $\hat{F}_{t-x}$ values from all piecewise constant models based on the AIDS data.

		Initial Model				Best Model			
H	I	$\hat{F}_x$	$\hat{F}_{t-x}$	AIC	SIC	$\hat{F}_x$	$\hat{F}_{t-x}$	AIC	SIC
1	1	1	1	1238.4	1249.1	NA	NA	NA	NA
2	1	0.566	1	1149.8	1164.0	0.085	1	1022.8	1037.0
1	2	1	0.200	1240.1	1254.3	1	0.241	1238.3	1252.5
2	2	0.566	0.200	1151.5	1169.2	0.229	0.241	1032.5	1050.2
3	1	{0.418, 0.742}	1	1099.0	1116.7	{0.085, 0.418}	1	1021.3	1038.5
3	2	{0.418, 0.742}	0.200	1100.7	1122.0	{0.085, 0.418}	0.200	1022.5	1043.8
1	3	1	{0.200, 0.272}	1241.4	1259.2	1	{0.241, 0.322}	1237.8	1255.6
2	3	0.566	{0.200, 0.2723}	1153.1	1174.4	0.229	{0.241, 0.272}	1034.3	1055.5
3	3	{0.418, 0.742}	{0.200, 0.272}	1102.6	1127.4	{0.229, 0.812}	{0.200, 0.241}	1033.9	1058.8

Figure 3.12: Plots of AIC and SIC values from the Initial models, and AIC and SIC values from the Best models, AIDS example.



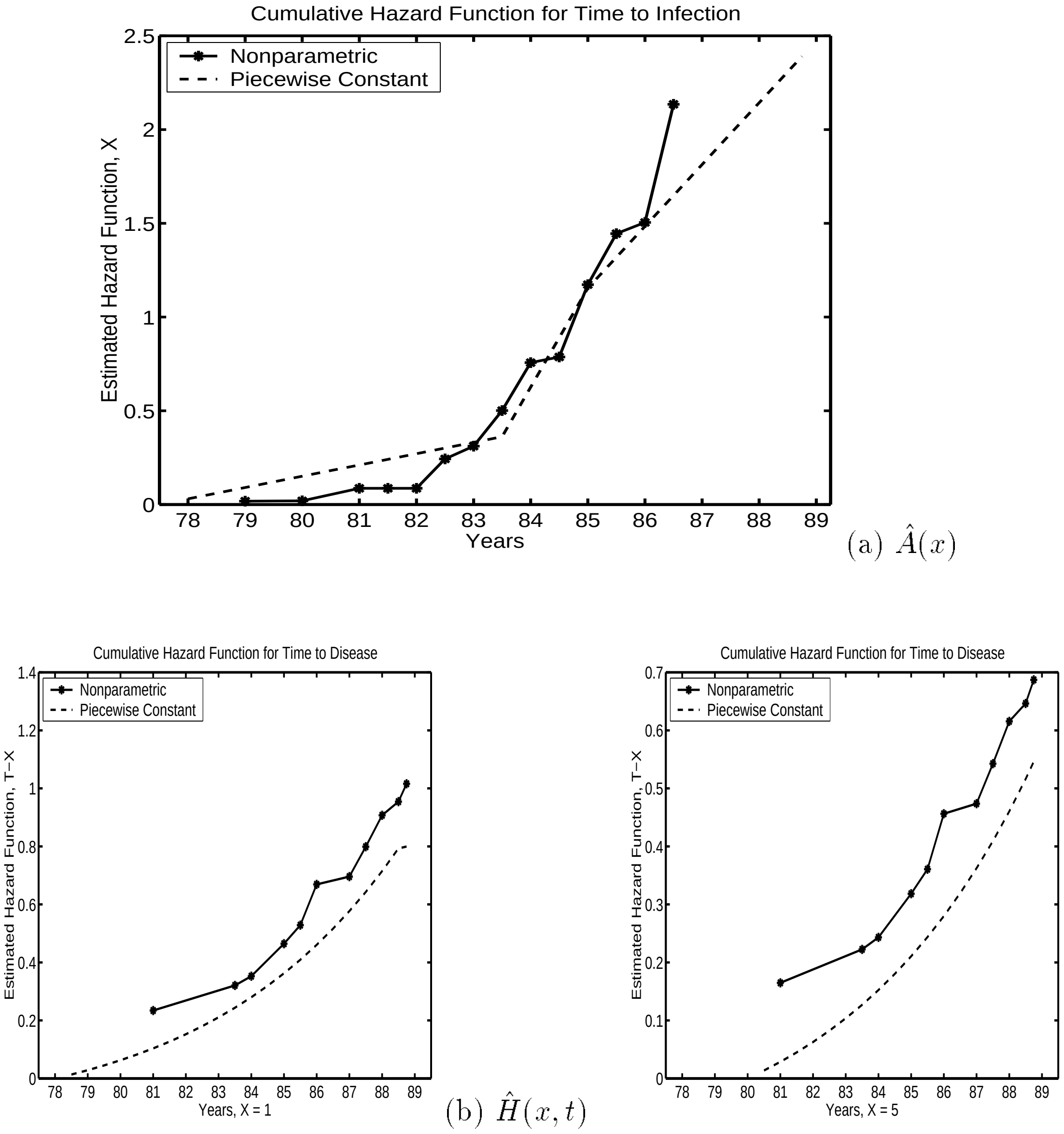
Among the Initial Models, the smallest overall AIC value (1099.0) in the nine piecewise specifications occurred for a model using three intervals of support for the first hazard function and a single interval of support for the last hazard function. The smallest all-around SIC value (1116.7) was also observed in this parameter configuration. The changepoints for the hazard function for X in this model were July 1983 and January 1985. The estimated total probability mass associated with each piece in the time to infection distribution were 0.4175, 0.3243, and 0.1931 for the first, second and third pieces, respectively.

Looking at the top half of Figure 3.12, and using the same paired values to denote the model configurations along the x-axis as in the simulation study, it is clear that all models incorporating three intervals of support for the hazard function between states 1 and 2, regardless of the number of pieces used to model the second transition intensity, had the lowest AIC and SIC values. Models with two intervals of support for the initial transition were next, followed by single piece models. The differences, vertical distances in the plots, were not as great between models of varying complexity in the second transition intensity as for the first. Both information criteria were slightly smaller for models specifying a single interval of support for the second transition intensity.

To graphically compare piecewise constant models with nonparametric models, we defined the estimated cumulative hazard functions for X and $T - X|X$ as in the simulation study. Looking at the upper panel of Figure 3.13, it seems that the piecewise constant model estimate of $A(x)$ closely matches the estimated nonparametric cumulative hazard function only for the years 1983 to 1985 inclusive. The large jump in the nonparametrically estimated hazard function from January 1985 to July 1985 was due to a second possible event time for the single person in the risk set. In the lower panels of Figure 3.13, the estimated cumulative hazard functions from the piecewise constant model for this second transition intensity function appear to follow the slope of the nonparametric estimated cumulative hazard function curve, but seem to consistently lie underneath it. The apparent difference between the curves is smaller when X is set to one, than when the value of X is set to five.

The breakpoints for the nine models were now systematically increased and decreased from the initial time points, and new SIC and AIC values calculated. For a model with a given number of pieces specified for each intensity, the reported "best" model was the one with the smallest value of either criterion, since both gave equivalent results (see Table 3.5). The smallest overall AIC value in the nine parameter configurations was once again a model employing three pieces for the first hazard function and one for the second. Now, however, the changepoints for the first hazard function are January 1982 and July 1983. The best overall model selected by the SIC is simpler, having only two intervals of support for the first hazard function and one for the second. The single changepoint used in the "best" SIC model — January 1982 — is the same as the first changepoint selected by the AIC. Both of these "best" models follow the nonparametrically estimated cumulative hazard function for the time to infection variable much more closely (see the top panel in Figure 3.14). The only time point they still do not match is July 1985 — an artifact created by a possible event in a risk set of size one in the nonparametric estimate. The estimates of $H(x, t)$ from these "best" piecewise

Figure 3.13: Plot of nonparametric and Initial piecewise constant estimated hazard functions, AIDS example.



(a) $\hat{A}(x)$

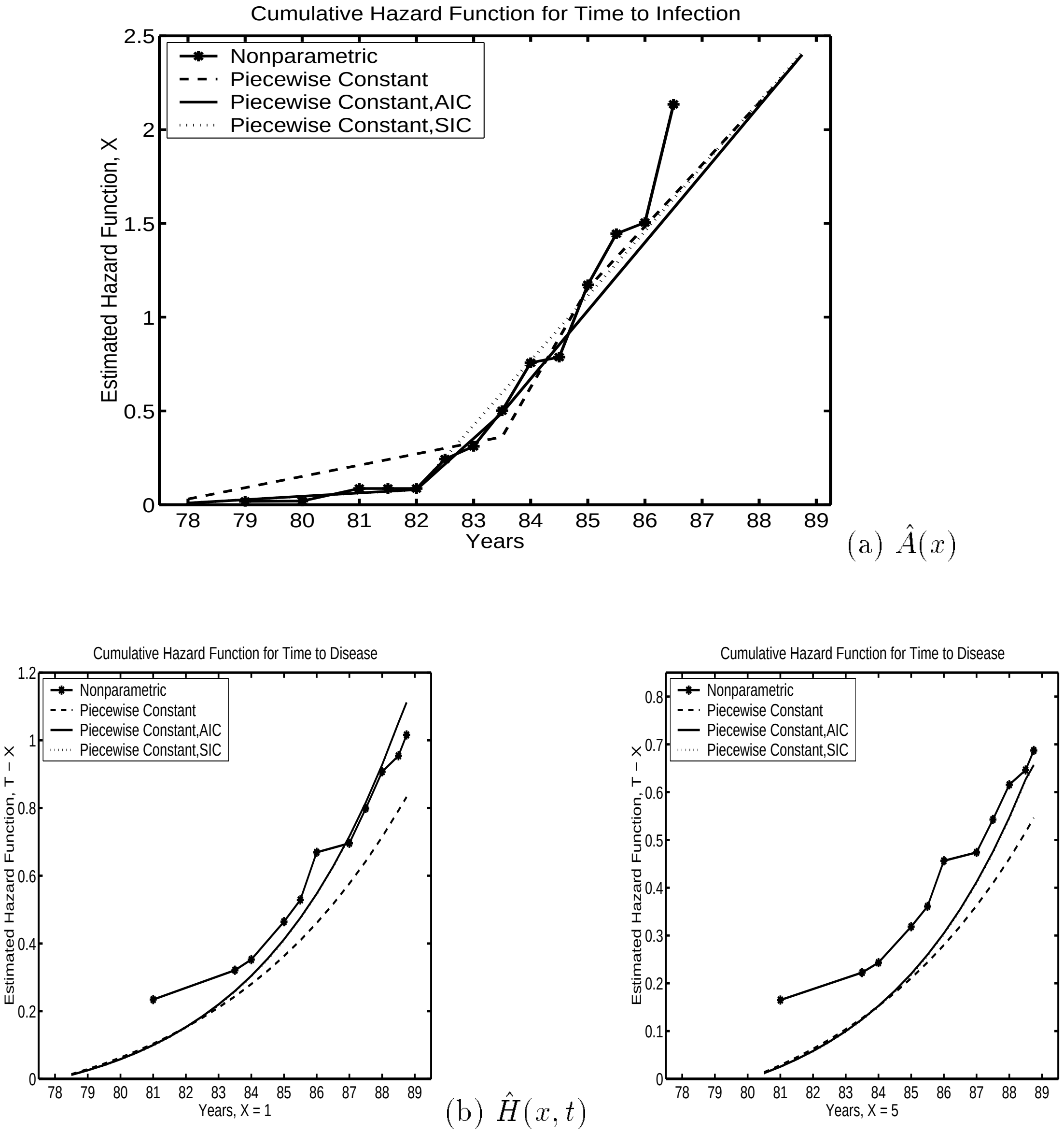
(b) $\hat{H}(x,t)$

constant models tend to follow more closely the nonparametric estimate of the cumulative hazard function between states 2 and 3, especially towards the end of the study follow-up time (see the lower panels of Figure 3.14).

Looking at the lower half of Figure 3.12, it is clear that single piece models for the time to infection variable do not fit as well as models with two or three pieces. Models using only one or two pieces for the hazard function of the time to AIDS diagnosis variable, from the time of infection, also seem to fit better than ones with three pieces. Nevertheless, six different models seem to fit roughly the same using either criterion. The extra variation in the smallest six SIC values indicates a little more discrimination among the best fitting models than in the corresponding set of six models suggested by the AIC.

Parameter estimates based on the Initial Models are summarized in Table 3.6. For models with the same number of pieces, there is a little variation among the baseline hazard estimates (η_i, $\lambda_i, i = 1, 2, 3$). As an example, if a single piecewise constant hazard is adopted for transitions between states 1 and 2, the point estimate for η_1 differs in the fourth decimal place between models which assume one, two or three piecewise constant hazard functions for the transition between states 2 and 3. Similarly, if a single piecewise constant hazard is adopted for transitions between state 2 and state 3, the point estimates for λ differ at most by 0.0003 for models which assume one, two, or three piecewise constant hazard functions for the transition between state 1 and state 2. Thus, estimation of the baseline intensity parameters for each transition intensity appears to be quite robust to the number of pieces being used in the other transition intensity.

Figure 3.14: Plot of nonparametric, Initial piecewise constant, and Best piecewise constant estimated hazard functions, AIDS example.



(a) $\hat{A}(x)$

(b) $\hat{H}(x,t)$

Table 3.6: Parameter estimates from Initial piecewise constant models, AIDS example.

Number of Pieces		Parameter Estimates						
H	I	η_1	η_2	η_3	λ_1	λ_2	λ_3	β
1	1	0.0662			0.0128			0.0722
2	1	0.0411	0.2467		0.0130			0.0804
1	2	0.0662			0.0001	0.0130		0.0704
2	2	0.0412	0.2466		0.0001	0.0132		0.0787
3	1	0.0311	0.3591	0.1974	0.0127			0.0905
3	2	0.0311	0.3589	0.1974	0.0001	0.0128		0.0890
1	3	0.0663			0.0001	0.0066	0.0142	0.0627
2	3	0.0412	0.2466		0.0001	0.0066	0.0139	0.0738
3	3	0.0311	0.3585	0.1976	0.0001	0.0089	0.0133	0.0853

Table 3.7: Parameter estimates from Best piecewise constant models, AIDS example.

Number of Pieces		Parameter Estimates						
H	I	η_1	η_2	η_3	λ_1	λ_2	λ_3	β
2	1	0.0089	0.2083		0.0106			0.1239
1	2	0.0663			0.0001	0.0141		0.0630
2	2	0.0130	0.2213		0.0001	0.0123		0.1053
3	1	0.0091	0.1577	0.2223	0.0108			0.1227
3	2	0.0091	0.1577	0.2223	0.0001	0.0109		0.1211
1	3	0.0664			0.0001	0.0049	0.0173	0.0464
2	3	0.0129	0.2211		0.0001	0.0217	0.0119	0.1091
3	3	0.0131	0.2572	0.1088	0.0001	0.0001	0.0122	0.1068

The estimated values of the regression coefficient, β, corresponding to the duration in state 2 do not vary substantially between all of these initial models either; the average value of β is 0.0781, with a standard deviation of 0.0092. All of the estimated values for β from the piecewise constant models are considerably smaller than the corresponding estimate (0.1307) from the nonparametric framework. We also found that the estimate of η_1 decreases as the number of pieces used in the first transition intensity increases. Similarly, the estimate of λ_1 decreases to its fixed lower bound of 0.0001 when two or more pieces are used in the second transition intensity. Interestingly, the estimate of η_2 increases when a third piece is added to the transition intensity between states 1 and 2, but the opposite trend is observed for the value of λ_2 when a third piece is added there.

When the optimal break points with respect to AIC and SIC values are selected, the results indicate less consistency in the parameter estimates (see Table 3.7). There is very little change in the parameter estimates for η_1 and λ_1 between models which use two or three intervals of support in the other transition intensity in these Best Models. The estimates for η_2 and λ_2 are not as congruous. For example, the values of η_2 vary from 0.2083 to 0.2213 when one to three pieces are used to model the intensity between states 2 and 3. The values of λ_2 range from 0.0109 to 0.0141 when one, two or three pieces are used to model the intensity between states 1 and 2. The values for η_3 and λ_3 also seem to depend on the complexity of the baseline hazard for the corresponding transition intensity.

The values of η_1 tend to decrease when a second piece is added to the intensity between states 1 and 2. A similar finding occurs for λ_1. The values of η_2, on the other hand, decrease when a third piece is added to the intensity between state 1 and state 2, but only if the second transition intensity has fewer than three pieces. The effect on λ_2 of adding a third piece in the second transition intensity is also mixed. The values of β vary considerably between these Best Models. The average value (0.0998) is now larger and closer to the nonparametric estimate, however, the standard deviation (0.0291) is quite a bit larger too.

The estimated standard errors for the two modelling approaches are reported in Tables 3.8 and 3.9. The second derivatives of the observed information matrix were computed using a finite-differencing approach. As we will discuss in §4.3.2, this data set appears to lack sufficient information to permit precise estimation of the model parameters in the nonparametric framework. This apparent lack of information also led to some instability in the estimation of the standard errors in this piecewise-constant approach. Despite the shortcomings of this data set, some simple observations and comparisons are appropriate.

The models with breakpoints based on the quantiles of the estimated nonparametric cumulative distribution functions demonstrated much less instability when estimating the standard errors, than did models which minimized the AIC or SIC. The estimates obtained for one intensity function still appear to be somewhat robust to the number of pieces used in the other intensity function in these Initial Models.

In the hazard function between states 1 and 2, the estimated variability increases with the addition of each new piece. In models with two pieces (H = 2), the estimated variability in the second piece is much larger than that for the first piece,

Table 3.8: Standard error estimates from Initial piecewise constant models, AIDS example.

Number of Pieces		Standard Error Estimates						
H	I	η_1	η_2	η_3	λ_1	λ_2	λ_3	β
1	1	0.1313			0.1139			0.9922
2	1	0.1068	1.1333		0.1092			0.9963
1	2	0.1439			1.0022	0.1102		0.9493
2	2	0.1111	1.4471		1.1523	0.1182		1.0239
3	1	0.0808	1.2166	5.3301	0.1130			1.1238
3	2	0.0946	1.3960	4.9168	1.0262	0.1242		1.1738
1	3	0.1330			1.0060	0.1591	0.1123	0.9546
2	3	0.0993	1.8514		1.0172	0.1841	0.1153	0.9991
3	3	0.0848	1.0949	4.3675	1.1124	0.1281	0.0937	0.9612

Table 3.9: Standard error estimates from Best piecewise constant models, AIDS example.

Number of Pieces		Standard Error Estimates						
H	I	η_1	η_2	η_3	λ_1	λ_2	λ_3	β
2	1	0.0523	0.9062		0.0841			0.9973
1	2	0.1326			1.0232	0.1255		1.0020
2	2	0.0696	6.6461		1.4742	0.0766		0.9397
3	1	0.0646	0.6426	5.3273	0.1271			1.7097
3	2	0.0821	1.1311	4.7711	1.1424	0.1037		0.9503
1	3	0.1414			1.0167	0.1308	0.1603	1.0378
2	3	0.0703	2.9882		1.0469	0.5494	0.1073	1.1709
3	3	0.0633	0.5140	0.6287	1.0222	1.1577	0.0811	0.9254

and in models with three pieces, the estimated variability is largest for the last piece and smallest for the first. When a third piece is added to a two-piece model, the standard errors for the second piece increase slightly, but only in the case where a single piecewise constant hazard is used in the second intensity function. Some opposite trends are evident in the hazard function between states 2 and 3. The estimated variability is always greatest in the first piece and smallest in the third piece. The estimated standard errors for the parameter β do not vary greatly across the nine piecewise constant settings; the range of values is from 0.9493 to 1.1738, with an average value of 1.0194 (sd = 0.0782).

In the Best Models, that is, the models selected by minimizing the AIC or SIC, the estimated standard errors were more numerically unstable and variable (see Table 3.9). The variability increased as the model complexity increased, especially in models with more than one interval of support for each intensity function. The estimated standard errors for the first piece of the intensity between states 1 and 2, η_1, were generally smaller in these models than their Initial Model counterparts. The estimated values for β are similar between the two types of models, but the variability is larger in this setting (average = 1.0916, sd = 0.2615).

We also compared the estimated standard errors for the model with the smallest AIC value in all nine parameter settings, that is, a model with three intervals of support in the first intensity and one for the second. In the Best Model setting, the values are not surprisingly smaller for all three pieces for the first intensity than in the Initial Model. The value of the baseline intensity parameter in the second transition intensity is similar in both models, while the estimated variability for β

is now larger. For the model with the smallest SIC value in all the nine parameter settings — two intervals of support for the first intensity and a single interval of support for the second — the estimated standard errors are now smaller for all the baseline intensity pieces. The estimated variability for β is essentially the same. These results are consistent with the plots in both panels of Figure 3.14. Thus, the models selected by iterating the changepoints about the initial break points have smaller standard errors for the first transition intensity, as well as AIC and SIC values, than their counterparts which use the initial breakpoints.

De Gruttola and Lagakos (1989) fitted fully parametric models to these data. They chose a uniform distribution for the time to infection variable, X, and a Weibull distribution for the induction time variable, $T-X$. Their results indicated that using a uniform distribution for the time to infection variable overestimated the probabilities in the early 1980's. The Weibull model fit reasonably well, but still did not correctly capture the change that occurred in the hazard function for subjects in the heavily treated group.

We also fitted piecewise constant models which included the covariates treatment group and estimated age at time of infection. The parameter estimates obtained in the piecewise constant models without covariates and the nonparametric estimates from a hybrid time scale model were used as the initial values for maximizing the observed data log likelihood. Tables 3.10 and 3.11 convey the findings we obtained by adding these covariates to the three different piecewise constant models selected on the basis of overall minimum AIC and SIC values.

In the Initial model, there is very little change in the parameter estimates from

Table 3.10: Parameter and standard error estimates when covariates are included in Initial piecewise constant models, AIDS example.

				Parameter Estimates				
				(Standard Error Estimates)				
H	I	η_1	η_2	η_3	λ_1	β	ν_1	ν_2
3	1	0.0312	0.3588	0.1969	0.0078	0.0970	0.7623	
		(0.0886)	(1.1057)	(1.4749)	(0.0603)	(0.9491)	(1.1358)	
3	1	0.0312	0.3587	0.1969	0.0074	0.0977	0.7879	0.1365
		(0.0364)	(0.8612)	(1.5411)	(0.0516)	(0.7655)	(3.8635)	(6.6079)

Table 3.11: Parameter and standard error estimates when covariates are included in Best piecewise constant models, AIDS example.

				Parameter Estimates				
				(Standard Error Estimates)				
H	I	η_1	η_2	η_3	λ_1	β	ν_1	ν_2
2	1	0.0089	0.2079		0.0068	0.1276	0.7321	
		(0.0533)	(0.4291)		(0.0272)	(0.6717)	(6.4667)	
2	1	0.0089	0.2079		0.0065	0.1279	0.7486	0.0956
		(0.0557)	(0.4636)		(0.0319)	(0.2219)	(3.8326)	(7.2372)
3	1	0.0091	0.1580	0.2218	0.0069	0.1262	0.7256	
		(0.0251)	(0.2496)	(1.4536)	(0.0710)	(1.4027)	(0.9695)	
3	1	0.0091	0.1581	0.2218	0.0066	0.1269	0.7460	0.1036
		(1.5986)	(0.4355)	(0.0006)	(0.6981)	(3.3962)	(6.2593)	(1.536)

the piecewise constant model when covariates are added (compare Tables 3.6 and 3.10). The estimated values of the regression coefficients did increase slightly from their corresponding values in the nonparametric framework. In a hybrid model which included the treatment group indicator variable, the nonparametric estimate of ν_1 was 0.7321 but increased to 0.7623 in this piecewise constant model. When indicator variables for treatment group and estimated age at time of infection were included, the parameter estimates also increased slightly. The nonparametric estimates for ν_1, ν_2 were 0.7751 and 0.0975, and increased to 0.7879 and 0.1365, respectively, in this piecewise constant model. Most of the estimated standard errors for the baseline intensity pieces decreased with the addition of one or both covariates (see Tables 3.8 and 3.10). This was particularly evident for the third baseline intensity parameter, η_3, between states 1 and 2. The decreases in the estimated standard errors were still not large enough, however, to statistically distinguish these parameter estimates from just noise.

In the global "best" model suggested by the SIC — a model with two pieces used for the first transition intensity and a single piece for the second transition intensity — there is very little change in either the parameter estimates or the estimated standard errors from their corresponding values in the piecewise constant or nonparametric models (see Tables 3.7, 3.9, and 3.11). When we consider the "best" overall model suggested by the AIC — a model which uses three pieces for the transition intensity between states 1 and 2 — the addition of one or both covariates does affect the estimated values. When only the treatment group covariate is added, the piecewise constant parameter estimates for the baseline intensity in the second hazard function, λ_1, and for the treatment covariate, ν_1, decrease slightly. The standard errors for all of the estimated piecewise constant parameters decrease slightly as well. When the covariate for age at infection is also added, the piecewise constant parameter estimates for λ_1 and ν_1 still decrease, but ν_2, the parameter estimate for the estimated age at infection variable increases from its nonparametric value. The estimated variability is much greater now too. This model required 489 iterations to converge, suggesting instability in the maximization process. The parameter estimates we obtained were consistent to the fixed precision level (i.e., 10^{-4}), however, for a variety of starting values. The estimated standard errors are still very large, relative to the estimated parameter values, so no significant effects are observed.

The AIC and SIC values for the three models which included these covariates were calculated as well. In Table 3.12, the first row refers to the Initial model where the changepoints were based on the quantiles of the estimated CDFs. The last two rows correspond to the Best models. Models with one covariate included only the treatment group indicator variable, while models with two covariates included the age at time of infection variable as well. The smallest AIC values were obtained for models which include only the treatment group covariate and were largest for the no covariate model. The SIC values, on the other hand, increase as the number of covariates went from none to two. Thus, the SIC would suggest adopting a simple piecewise constant model with no covariates, while the AIC would suggest adopting a piecewise constant model which included the treatment group covariate as well. The differences in the SIC values for models with and without this covariate

Table 3.12: SIC and AIC values in selected piecewise constant models with none, one, or two covariates, AIDS example.

Number of		Model					
Pieces		No Covariates		One Covariate		Two Covariates	
H	I	AIC	SIC	AIC	SIC	AIC	SIC
Initial	Model						
3	1	1099.0	1116.7	1095.7	1117.0	1097.6	1122.4
Best	Model						
2	1	1022.8	1037.0	1019.9	1037.6	1021.8	1043.1
3	1	1021.3	1038.5	1018.0	1039.3	1019.9	1044.7

are small — less than one unit. In chapter 2 we obtained a p-value of 0.0211 for the likelihood ratio statistic evaluating the inclusion of this covariate in a hybrid time scale model, using a nonparametric approach. Although the likelihood ratio statistics based on these AIC and SIC values would concur with the nonparametric results for selecting a model with just the treatment covariate, only the AIC would have identified these one-covariate models as the best-fitting models.

In summary, utilizing the quantiles from the nonparametrically estimated CDFs led to a model with three pieces for the transition intensity between states 1 and 2 and a single piece for the transition intensity between states 2 and 3. The complexity of this model seemed appropriate when compared graphically to the nonparametric estimates for both intensities. However, when the changepoints are increased and decreased from these initial points, better fitting models were found. The overall Best model selected by the AIC now follows the nonparametrically estimated cumulative hazard curves much more closely over the time on study. The

Best overall model selected by the SIC, which included only two pieces for the first transition intensity, seemed to follow the nonparametrically estimated cumulative hazard curve almost as well as a model involving three pieces.

The parameter estimates and standard errors from models whose changepoints were selected using the estimated quantiles were quite consistent across comparable models. There was considerably more variability and numerical instability in the models where the changepoints were iterated about these initial time points. Many of the resulting Best models did have smaller standard errors, suggesting a better fit. However, a lack of data did lead to some estimation instability, particularly as the number of pieces in the hazard function between states 2 and 3 increased beyond one. The addition of covariates to an underlying piecewise constant model did not affect the baseline intensity parameter estimates very much. The standard errors generally decreased when one or both covariates were added, suggesting improvement in the fit of the model. However, the estimated standard errors for all of the regression coefficients were very large relative to the estimated parameters. The goal of obtaining estimates of the variability in these regression parameters was met, but an apparent lack of information in the data led to less than satisfactory results.

The SIC and AIC performed equally well in the approach that relied on the quantiles from the nonparametrically estimated CDFs. There was very good discrimination amongst competing models for the complexity of the first transition intensity but poor discrimination amongst competing models for the complexity of the second transition intensity. The lack of discrimination is likely related to a lack of data for estimating more than a single transition intensity in this setting. The AIC tended to overfit when the quantiles from the nonparametrically estimated CDFs were ignored. However, consistent with the LRSs, the AIC correctly distinguished models with the treatment covariate as best fitting among models with this covariate, models with the additional age at infection variable, or models with no covariates.

3.4 Conclusions

The piecewise constant method of estimating the baseline intensity functions in a three-state progressive process was adopted, in part, to permit estimation of the standard errors for the regression parameters. The standard errors were easily estimated using this approach, but the results were not satisfactory. The estimated standard errors were generally very large, relative to the parameter estimates of the regression parameters, leading to some inconsistent results between this method and the nonparametric estimation approach.

Iterating above and below the initial changepoints selected by using the quantiles of the estimated CDFs yielded better models. The changepoints in these Best models were in closer agreement with the true changepoints in the simulated data, generally had smaller standard errors, and appeared to follow the nonparametrically estimated hazard curves much more closely. The Best models also seemed to be robust to the increased amount of interval censoring in the data set with single intervals of support for both transition intensities. This approach seems to have a sufficient amount of information in the data to adequately estimate each piece in the model, since the breakpoints were close to the breakpoints selected initially.

Neither the AIC nor the SIC selected the same number of pieces and breakpoints as the model from which the data were simulated. This result may not be that surprising, since the uncertainty introduced with the right and interval censoring of the data can make model identification more difficult. Liang *et al.* (1992) found that introducing Gaussian noise into a mixture of three image classes in a simulated brain image led to overestimation of the number of classes by the AIC. The SIC tended to have greater discrimination between competing models, and occasionally selected a simpler model than its competitor, the AIC.

The plots of the AIC and SIC values gave some insight into the possible choices for model complexity. For example, in the data set generated with two intervals of support in both transitions, there is very good discrimination among the possible models. The SIC and AIC values in both panels of Figure 3.5 are vertically separated from one another by varying amounts of distance. In contrast to this good discrimination situation, the SIC and AIC values in both panels of Figure 3.7 vary only a little. Many models are plausible in this setting.

The greatest drawback of this approach is the computational burden required to find best-fitting models. The number of intervals of support for the two transition intensities in this study was quite modest — one, two, or three — yet the time required to fit the various combinations of models was large. Iterating the breakpoints only increased, multiplicatively, the number of models to fit, and the resulting computing time. This time constraint limited the goals of the simulation study to the ones considered.

Given the limited scope of this simulation study, only tentative practical guidelines are suggested here. If the quantiles from the nonparametric estimates of the CDFs are used to initially select the breakpoints, a better-fitting model can often be found by iterating above and below these breakpoints. Since the SIC tended to provide greater discrimination amongst competing models, and to select simpler models in this three-state setting, it is preferred over the AIC. To reduce the computational burden of this algorithm, initial conjectures of model complexity can be based on the nonparametric estimated values. The best fitting model can then be ascertained more quickly from this starting point, instead of from a completely naïve approach.

Future research could focus on the effect of sample size and censoring levels on the estimation of the model parameters and their associated standard errors. The choice of the SIC over the AIC could be evaluated by varying these two data properties as well. Direct calculation of the matrix of second partial derivatives of the observed data log-likelihood, instead of estimating it by using a finite-differencing method, may improve the estimation of the standard errors. Other aspects of the underlying model adopted in this three-state progressive process may be influencing the estimation of the various model parameters. In the next chapter we explore some aspects of model assessment, in particular, the logistic specification of the intensity function between state 2 and state 3.

Chapter 4

Goodness-of-Link and Model Assessment

Model assessment is an important component of data analysis, since some underlying structure is assumed to hold in a given model. Although no one model is true, serious departures from the model assumptions could lead to questionable inferences being made. The primary focus of this chapter will be on the logistic specification of the hazard function for transitions from state 2 to state 3. The dependence of the hazard function upon this parametric specification for the regression models formulated in chapter two will be assessed by embedding it in a one-parameter family of hazard functions. We refer to the parametric regression specification as a link function, since it links the linear predictor to the conditional probability of an event. A simulation study will examine the impact on estimation of the regression coefficients when the link function parameter is also estimated. Likelihood ratio statistics will be used to test whether the logistic specification is consistent with the data. The AIDS in hemophilia patients example from chapter 2 will be considered in this more general context. In addition, we will verify that the derived self-consistent estimators from chapter 2 are also the maximum likelihood estimators (MLEs). The uniqueness and convergence of the MLEs will be determined.

4.1 Regression Formulation of the Hazard Function

Discrete time models for event history data are often formulated via the hazard rate. If the occurrence of an event or transition between states is thought to depend on covariates, the corresponding hazard rate may be suitably parametrized (Blossfeld, Hamerle and Mayer, 1989). For example, in the standard discrete-time survival analysis context, with T representing the time-to-event variable, the probability of an event or transition occurring in time interval t, given the covariate vector z, is

$$P(T = t \mid z) = \lambda(t \mid z) \prod_{r=1}^{t-1} \{1 - \lambda(r \mid z)\} \ .$$

Various specifications of the model are possible, including the logistic model

$$\lambda(t \mid z) = \frac{\exp(\beta_{0t} + z'\boldsymbol{\beta})}{1 + \exp(\beta_{0t} + z'\boldsymbol{\beta})} \ , \qquad t = 1, \ldots, m \ ,$$

or the extreme value model

$$\lambda(t \mid z) = 1 - \exp(-\exp(\beta_{0t} + z'\boldsymbol{\beta}))\ , \quad t = 1, \dots, m\ .$$

Equivalent specifications for the model are identified via an appropriate linearizing transformation. The logit transformation, $g(\lambda) = \log(\frac{\lambda}{1-\lambda})$, for the logistic model is linear in the predictors, whereas the complementary log-log transformation, $g(\lambda) = \log\{-\log(1-\lambda)\}$, is the appropriate transformation for the extreme value model.

The logistic specification for the intensity of the transition from states 2 to 3 chosen by Frydman (1995) was

$$h(x,t) = \frac{\lambda(t)\ e^{\beta(t-x)}}{1 + \lambda(t)\ e^{\beta(t-x)}}\ .$$

In chapter 2, we developed intensity functions which included dependence on covariates (z) in addition to the duration in state 2 variable $(t-x)$. Thus, our model specified the hazard function as

$$h(x,t;z) = \frac{\lambda(t)\ e^{\beta(t-x)+z'\boldsymbol{\nu}}}{1 + \lambda(t)\ e^{\beta(t-x)+z'\boldsymbol{\nu}}}\ .$$

This expression for the hazard function, $h(x,t;z)$, and for the conditional probability, $1 - h(x,t;z)$, of surviving the time interval t may be rewritten as

$$h(x,t;z) = \left(1 + \frac{1}{\lambda(t)\ e^{\beta(t-x)+z'\boldsymbol{\nu}}}\right)^{-1}\ , \quad 1 - h(x,t;z) = \left(1 + \lambda(t)\ e^{\beta(t-x)+z'\boldsymbol{\nu}}\right)^{-1}$$

or more generally as

$$h(x,t;z,q) = \left(1 + \frac{1}{q\,\lambda(t)\;e^{\beta(t-x)+z'\boldsymbol{\nu}}}\right)^{-q} \quad ,$$

$$1 - h(x,t;z,q) = \left(1 + \frac{\lambda(t)\;e^{\beta(t-x)+z'\boldsymbol{\nu}}}{q}\right)^{-q} \quad .$$

The parameter q may take on any positive real value, but two special cases are of interest here. If $q = 1$, then the hazard function $h(x,t;z)$ has a logistic specification. However, if $q \to \infty$, then $1 - h(x,t;z)$ has an extreme value specification.

In simple binary regression settings, embedding the assumed link function within a family of link functions indexed by one or two parameters is a useful way to assess the assumed link function and the potential impact of that choice on the model fit. Properties of the parametric link transformation family are important considerations, particularly in generalized linear models (Czado, 1997).

Common choices within a family of link functions can be assessed, graphically, by plotting the deviance values for models fit with fixed values of the link parameter (McCullagh and Nelder, 1989). The minimum deviance may be determined from such plots, as well as the link functions which are compatible with the data within specified confidence limits. Likelihood ratio tests may distinguish between two competing values of the link parameter, when the regression parameters are treated as the nuisance parameters in a profile likelihood. Alternatively, score tests can be used to assess the plausibility of a particular link parameter value of interest. An approximate version of one such test was proposed by Pregibon (1980). His goodness-of-link test is a simple technique by which one can assess the fit of a reasonable hypothesized link. Instead of specifying a parametric family of link functions, Cheng and Wu (1994) propose a test for detecting global lack of fit of a parametric model for the link function but without specifying a particular family of alternatives. Their test is based on a quasi-likelihood model and appears to have good power to discriminate between logit and probit models.

Link estimation, rather than link assessment, may be the goal in assessing model assumptions. The log likelihood of the model based on the parametrized link function can be maximized with respect to all the model parameters. The technique of Pregibon (1980) provides a one-step approximation to the link function parameters, if the hypothesized values of a link function parameter are sufficiently "close" to the true values. The remainder of this chapter will encompass both testing and estimation of the link function or the parametric specification of the hazard function. Departures from the logistic specification of the hazard function will be assessed, as will the impact on the estimation of the regression coefficients if the link parameter is not estimated jointly with them.

4.2 Simulation Study

The model we develop here is similar to the extension from the basic Frydman model of chapter 2, which permits interval censoring for the times of entry into both states 2 and 3 and inclusion of time-independent (fixed), external covariates; see §2.6.2.2. However, in this case the first transition intensity function is modelled nonparametrically and now the second transition intensity depends on the link parameter q. The parametric logistic specification for the transition intensity from state 2 to state 3 is defined as

$$h(x,t;z,q) = \left(1 + \frac{1}{q\,\lambda(t)\;e^{\beta(t-x)+z'\boldsymbol{\nu}}}\right)^{-q} \;. \tag{4.1}$$

As in previous chapters, we define the indicator variables d_{njk} and b_{njk} as before (see pp. 41-42), the covariate process as $Z' = \{z_1, z_2, \ldots, z_p\}$, and now Θ as $\{\boldsymbol{p}, \boldsymbol{\lambda}, \boldsymbol{\nu}, \beta, q\}$. We continue to use the indicator variables Δ_n and δ_n for the transitions to state 2 from state 1 and to state 3 from state 2 for individual n, respectively. Let $\prod_{jk}(z,q) \equiv \prod\limits_{x_j<t_r\le t_k} \{1 - h(x_j, t_r; z, q)\}$, and $\prod^*_{jk}(z,q) \equiv \prod\limits_{x_j<t_r<t_k} \{1 - h(x_j, t_r; z, q)\}$. Then the log likelihood of the observed data is

$$\begin{aligned}
\log\ L(\Theta) \;&=\; \sum_{\delta_n=0} \log \sum_{j=1}^{J} \sum_{k=1}^{K} d_{njk}\; p_j \left[\textstyle\prod_{jk}(z_n, q)\right]^{\Delta_n} \\
&+\; \sum_{\delta_n=1} \log \sum_{j=1}^{J} \sum_{k=1}^{K} b_{njk}\; p_j\; h(x_j, t_k; z_n, q) \textstyle\prod^*_{jk}(z_n, q) \;.
\end{aligned}$$

Let $p(\mathbf{x}|\Theta)$ denote the probability density function of the complete data, $\mathbf{X}$, with associated parameters Θ. We let $\mathbf{y}$ denote the observed data and Θ' be any value of Θ in the parameter space for $p(\mathbf{x}|\Theta)$. Then, $E\{\log\ p(\mathbf{x}|\Theta')|\mathbf{y}, \Theta\}$, the expectation of the complete data log likelihood, conditional on the observed data, is given by

$$
\begin{aligned}
Q(\Theta'|\Theta) \;=\; & \sum_{\delta_n=0}\sum_{j,k}\left[\frac{d_{njk}\; p_j\left\{\prod_{jk}(z_n,q)\right\}^{\Delta_n}}{\sum_{l,m} d_{nlm}\; p_l\left\{\prod_{lm}(z_n,q)\right\}^{\Delta_n}}\right]\times \\
& \log\left[p'_j\left\{\prod_{jk}^{\Delta_n}(z_n,q')\right\}'\right] \\
+ & \sum_{\delta_n=1}\sum_{j,k}\left[\frac{b_{njk}\;\; p_j\, h(x_j,t_k;z_n,q)\prod^*_{jk}(z_n,q)}{\sum_{l,m} b_{nlm}\; p_l\, h(x_l,t_m;z_n,q)\prod^*_{lm}(z_n,q)}\right]\times \\
& \log\left[p'_j\; h'(x_j,t_k;z_n,q')\left\{\prod_{jk}^{*}(z_n,q')\right\}'\right]\;.
\end{aligned}
$$

We differentiate $Q(\Theta'|\Theta)$ with respect to Θ', in order to maximize the likelihood. Beginning with the only parameter in the first intensity, we obtain the equations

$$
\hat{p}_j=\left\{\sum_{\delta_n=0}\sum_{k=1}^{K}\mu_{njk}(\Theta)+\sum_{\delta_n=1}\sum_{k=1}^{K}\mu^*_{njk}(\Theta)\right\}/N \qquad (1\le j\le J)\,,
$$

for p_j, where N represents the study sample size. This is essentially the same form as the estimating equation obtained previously in §2.6.2.2; however, these expressions use the new conditional probability functions

$$
\mu_{njk}(\Theta)=\frac{d_{njk}\; p_j\;\left\{\prod_{jk}(z_n,q)\right\}^{\Delta_n}}{\sum_{l,m} d_{nlm}\; p_l\;\left\{\prod_{lm}(z_n,q)\right\}^{\Delta_n}}\quad ,
$$

$$
\mu^*_{njk}(\Theta)=\frac{b_{njk}\; p_j\; h(x_j,t_k;z_n,q)\prod^*_{jk}(z_n,q)}{\sum_{l,m} b_{nlm}\; p_l\; h(x_l,t_m;z_n,q)\prod^*_{lm}(z_n,q)}\quad ,
$$

$$\mu^{\dagger}_{njk}(\Theta) = \frac{d_{njk}\; p_j\; \prod_{jk}(z_n, q)}{\sum_{l,m} d_{nlm}\; p_l\; \prod_{lm}(z_n, q)} \quad .$$

The estimating equations for the parameters in the second transition intensity now involve the additional parameter q, making the previous simplifications no longer possible. If we define $f(x,t;\Phi) = (1 + q\,\lambda(t)\; e^{\beta(t-x)+z'\boldsymbol{\nu}})^{-1}$ and $g(x,t;\Phi) = [1 + 1/\{q\,\lambda(t)\; e^{\beta(t-x)+z'\boldsymbol{\nu}}\}]^{-q}$, and let $\Phi = \{\boldsymbol{\lambda}, \boldsymbol{\nu}, \beta, q\}$, the estimating equation for λ_k becomes

$$\begin{aligned}
&\sum_{\delta_n=1}\sum_{j=1}^{J} \mu^{*}_{njk}(\Theta)\; f(x_j, t_k; \Phi) \\
&\qquad = \sum_{\Delta_n=1,\delta_n=0}\;\sum_{j=1}^{J} \mu^{\dagger}_{njk}(\Theta) \left[\frac{g(x_j,t_k;\Phi)}{1-g(x_j,t_k;\Phi)}\right] f(x_j,t_k;\Phi) \\
&\qquad + \sum_{\delta_n=1}\sum_{j=1}^{J} \left[\frac{g(x_j,t_k;\Phi)}{1-g(x_j,t_k;\Phi)}\right] f(x_j,t_k;\Phi) \sum_{r=k+1}^{K} \mu^{*}_{njr}(\Theta) \;, \\
&\qquad\qquad (1 \le k \le K) \;,
\end{aligned}$$

while for β we obtain

$$\begin{aligned}
&\sum_{\delta_n=1}\sum_{j}\sum_{k} \mu^{*}_{njk}(\Theta)\; f(x_j,t_k;\Phi)\; (t_k - x_j) = \\
&\quad \sum_{\Delta_n=1,\delta_n=0}\;\sum_{j}\sum_{k} \mu^{\dagger}_{njk}(\Theta) \sum_{x_j<t_r\le t_k} \left[\frac{g(x_j,t_r;\Phi)}{1-g(x_j,t_r;\Phi)}\right] f(x_j,t_r;\Phi)\; (t_r - x_j) \\
&\quad + \sum_{\delta_n=1}\sum_{j}\sum_{k} \mu^{*}_{njk}(\Theta) \sum_{x_j<t_r<t_k} \left[\frac{g(x_j,t_r;\Phi)}{1-g(x_j,t_r;\Phi)}\right] f(x_j,t_r;\Phi)\; (t_r - x_j) \;.
\end{aligned}$$

The equation that we use to find the MLE for ν_s corresponds to

$$\begin{aligned}
&\sum_{\delta_n=1} z_{ns} \sum_j \sum_k \mu^*_{njk}(\Theta)\ f(x_j, t_k; \Phi) = \\
&\quad \sum_{\Delta_n=1,\delta_n=0} z_{ns} \sum_j \sum_k \mu^\dagger_{njk}(\Theta) \sum_{x_j<t_r\leq t_k} \left[\frac{g(x_j, t_r; \Phi)}{1-g(x_j, t_r; \Phi)}\right] f(x_j, t_r; \Phi) \\
&\quad + \sum_{\delta_n=1} z_{ns} \sum_j \sum_k \mu^*_{njk}(\Theta) \sum_{x_j<t_r<t_k} \left[\frac{g(x_j, t_r; \Phi)}{1-g(x_j, t_r; \Phi)}\right] f(x_j, t_r; \Phi)\ , \qquad (1 \leq s \leq p).
\end{aligned}$$

Lastly, if we define $r(x, t; \Phi) = [1 + 1/\{q\ \lambda(t)\ e^{\beta(t-x)+z'\boldsymbol{\nu}}\}]^{-1}$, then the estimating equation for q, the parameter of the link function, is given by

$$\begin{aligned}
&\sum_{\delta_n=1} \sum_{j,k} \mu^*_{njk}(\Theta)\ \left[\log\{r(x_j, t_k; \Phi)\} + f(x_j, t_k; \Phi)\right] = \\
&\quad \sum_{\Delta_n=1,\delta_n=0} \sum_{j,k} \mu^\dagger_{njk}(\Theta) \sum_{x_j<t_r\leq t_k} \frac{g(x_j, t_r; \Phi)}{1-g(x_j, t_r; \Phi)} \left[\log\{r(x_j, t_r; \Phi)\} + f(x_j, t_r; \Phi)\right] \\
&\quad + \sum_{\delta_n=1} \sum_{j,k} \mu^*_{njk}(\Theta) \sum_{x_j<t_r<t_k} \frac{g(x_j, t_r; \Phi)}{1-g(x_j, t_r; \Phi)} \left[\log\{r(x_j, t_r; \Phi)\} + f(x_j, t_r; \Phi)\right]\ .
\end{aligned}$$

If the parametric complementary log-log specification for the transition intensity between states 2 and 3 is preferred, the hazard function becomes

$$h^*(x, t; z, q) = 1 - \left(1 + \frac{\lambda(t)\ e^{\beta(t-x)+z'\boldsymbol{\nu}}}{q}\right)^{-q}\ . \tag{4.2}$$

The log likelihood of the observed data, as well as the expectation of the complete data given the observed data, can be found by substituting h^* for h in the previous expressions. Estimating equations for all of the parameters in Θ can also be easily derived. Unfortunately, these equations became numerically unstable when

q is estimated from the data or became analytically unusable when q is allowed to go infinity.

When we tried to estimate the value of q, the denominators of most terms in the estimating equations for parameters in the second transition intensity went to zero for large values of q, say near 20 million. Constraining the value of q to be large, but not too large was only partially successful. After a few more iterations, the log likelihood would decrease very abruptly, suggesting the MLEs had not been found. When we took the limit as $q \to \infty$, all of the estimating equations except the one for p_j reduced to the form $0 = 0$. We then attempted to maximize the log likelihood of the observed data directly, assuming an extreme value model for the transition intensity from state 2 to state 3. A constrained nonlinear optimizer function (*fmincon*) in Matlab was not successful in finding the MLEs either. The optimizing function set many of the parameter values at their constrained nonnegative lower bounds, i.e., 10^{-6} and placed mass at only a few parameter values. Hence, we decided not to include the parametric complementary log-log specification in our simulation study.

Values used in the simulation study for the logistic specification were chosen to reflect possible hazard functions encountered in practice: hazard functions were constant (exponential), monotone increasing (Weibull), and monotone decreasing (Weibull). A left-skewed distribution, the minimum extreme value distribution, was also chosen for the second transition intensity. Using this representation of the Weibull probability density function

$$f(t; \lambda, \delta) = \lambda\, \delta\, t^{\delta-1} \exp(-\lambda t^{\delta}) \qquad t \geq 0, \quad \lambda, \delta > 0$$

the shape parameter, δ, was fixed at the values $\{1/2, 1, 3/2\}$, while the scale parameter, $\lambda_i, i = 1, 2$ was fixed at the values $\{1, 3/2\}$. The first scale parameter, λ_1, was used for generating Weibull random variables for the distribution of time in state 1, X, while the second, λ_2, was used for generating Weibull random variables for the distribution of time in state 2, V. The same value of the shape parameter was used to generate both sets of random variables within each of the twelve cases considered, i.e., both random variables followed an exponential distribution or both followed a Weibull distribution. The corresponding values of the extreme value distribution were obtained by taking the log of the random variables generated from a Weibull distribution. The representation of the probability density function for the extreme value distribution, used only for the random variable for V, is given by

$$g(t; \phi, \delta) = \delta^{-1} e^{(t-\phi)/\delta} \exp(-e^{(t-\phi)/\delta}) \qquad -\infty < t < \infty \ ,$$

where $\delta > 0$ and $-\infty < \phi < \infty$. The same three values of δ, i.e. $1/2, 1$ and $3/2$, were also used in the complementary log-log models, with λ_2 set to $3/2$. Hence, the values of $\phi = -\delta^{-1}\log(\lambda_2)$ became the combinations $\{2, 1, 2/3\} \times \log\{2/3\}$.

Data were generated according to each Weibull parameter configuration, and the logs of the data taken in the three complementary log-log settings. The data were then discretized using the following protocol. Study follow-up time was set at the largest observed failure time and the number of intervals fixed at 24. A data-dependent partition was used, so that approximately the same number of observations occurred within each interval. Transition times to state 3 from state 2 occurred at least one time point after the transition from state 1 to state 2. The sample size was approximately 100 for each simulation run, with about 25% of individuals ending the study in state 1, 50% in state 2 and 25% in state 3. One hundred simulation runs were conducted for each of the twelve Weibull and three complementary log-log settings.

For each simulation run, differences between parameter estimates from the usual hybrid model ($q = 1$) and corresponding parameter estimates obtained for the same data when q was allowed to vary were calculated. In addition to the model parameter estimate differences, the value of q was estimated and two likelihood ratio statistics (LRSs) calculated. The first LRS evaluated the importance of the duration in state 2 variable compared to a simpler underlying Markov model, i.e. a test of $\beta = 0$, while the second evaluated the logistic specification of the hazard function between states 2 and 3.

The specific goals to be evaluated in this simulation study encompass both testing and estimation issues. First, we want to evaluate whether the logistic specification of the hazard function between states 2 and 3 is appropriate for several parametric distributions that are commonly used for lifetime data. Next, the effects of estimating the link parameter on the other model parameters will be assessed. We are interested in determining whether all model parameters are affected, and how they are affected, when q is estimated.

4.2.1 Weibull Models

Beginning with the goodness-of-link parameter, q, we see in Table 4.1 that the estimated values are all larger than the null value of one. The median was chosen for reporting purposes so that occasional values that inflate the sample average would not confound comparisons between the different cases in the study. Positive skewness, and variability differences in the estimates of q are evident for all twelve Weibull cases in the box plots found in Figure 4.1. A legend explaining the labels used in this plot, and all other plots which report our study findings is given in Table 4.2. The overall average value of q, calculated from the Weibull cases A - L in Table 4.1 was 1.3122, with a standard deviation of 0.1164 units.

Table 4.1: Estimates of q and $\bigtriangledown\beta$ from all Weibull cases.

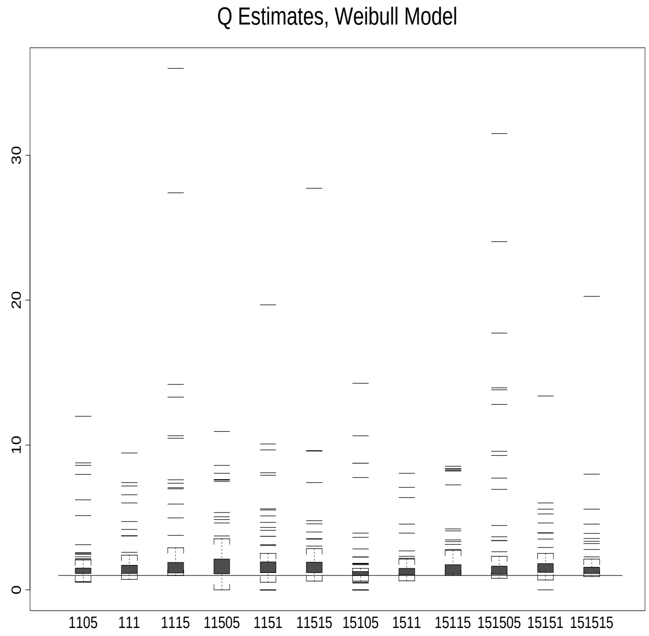
Case Label	Weibull Parameter Values			Median Parameter Estimates	
	δ	λ_1	λ_2	q	$\bigtriangledown\beta$
A	1/2	1	1	1.256	-0.033
B	1	1	1	1.345	-0.039
C	3/2	1	1	1.345	-0.045
D	1/2	1	3/2	1.380	-0.051
E	1	1	3/2	1.472	-0.056
F	3/2	1	3/2	1.460	-0.045
G	1/2	3/2	1	1.057	-0.007
H	1	3/2	1	1.214	-0.024
I	3/2	3/2	1	1.269	-0.036
J	1/2	3/2	3/2	1.238	-0.041
K	1	3/2	3/2	1.410	-0.044
L	3/2	3/2	3/2	1.301	-0.036

Although most of the estimated q values appear to be greater than one in Figure 4.1, we formally evaluated whether the logistic link function specification is plausible by calculating a LRS for each simulation run. The usual hybrid model with q fixed at one was initially fit, and the value of the log likelihood, evaluated at the MLEs,

Table 4.2: Legend for Figures 4.1, 4.3 - 4.7.

Weibull Parameter Values			Label
λ_1	λ_2	δ	
1	1	1/2	1105
1	1	1	111
1	1	3/2	1115
1	3/2	1/2	11505
1	3/2	1	1151
1	3/2	3/2	11515
3/2	1	1/2	15105
3/2	1	1	1511
3/2	1	3/2	15115
3/2	3/2	1/2	151505
3/2	3/2	1	15151
3/2	3/2	3/2	151515

Figure 4.1: Box plots of estimated q values from all Weibull cases.



was calculated. A second hybrid model, where q was now estimated along with all of the other model parameters, for the same data was now fit. This enriched log likelihood was evaluated at the new MLEs, and compared to the previous value when $q = 1$. Once again we report the median values of all 100 simulation runs for each Weibull case. In Table 4.3, the LRSs for evaluating whether q was significantly different than one are at or near zero for all twelve cases. The corresponding p-values are only less than one for cases D and G. These two cases are characterized by a decreasing hazard rate ($\delta = 1/2$) and by unequal scale parameters ($\lambda_i, i = 1, 2$) in the underlying Weibull random variables which generated the data corresponding to transition times between the three states.

Table 4.3: Likelihood ratio statistics for q, and proportions of cases where the hypothesis $q = 1$ is rejected in all Weibull models.

Case	Weibull Parameter				
Label		Values			
	δ	λ_1	λ_2	LRS	$\mathcal{P}$
A	1/2	1	1	0	0
B	1	1	1	0	0
C	3/2	1	1	0	0
D	1/2	1	3/2	0.016	0
E	1	1	3/2	0	0
F	3/2	1	3/2	0	0
G	1/2	3/2	1	0.005	0
H	1	3/2	1	0	0
I	3/2	3/2	1	0	0
J	1/2	3/2	3/2	0	0.01
K	1	3/2	3/2	0	0
L	3/2	3/2	3/2	0	0.01

In the last column of Table 4.3, we report the proportion, $\mathcal{P}$, of cases within each simulation run in which we reject the hypothesis, $q = 1$, at the 0.05 level. Assuming the large-sample properties of the LRSs hold, if q was in fact equal to one, then we would expect $\mathcal{P}$ to be near zero, whereas if q was in truth different than the null value of one, then $\mathcal{P}$ should be substantially larger than zero. We see that all of the observed proportions are at or close to zero. Only cases J and L had a single LRS, out of the 100 runs, larger than 3.841.

To evaluate whether the estimation of q was affected by the three different hazard rates — constant, increasing, or decreasing — which generated the underlying data, we constructed pseudo contrasts. Analogous with the way linear combinations of means are formally compared via contrasts in an Analysis of Variance, we informally compared averages of the q values obtained for each type of hazard. To illustrate how the calculations were made, the average value of q estimated from data generated with a constant hazard function ($\delta = 1$) between states 2 and 3 was found by summing the reported q values in Table 4.1 for cases B, E, H, K and then dividing that sum by four. Similarly, for the q estimates obtained when the underlying data were generated with nonconstant hazard functions, the respective averages for decreasing and increasing rate functions were obtained from cases A, D, G, J, and C, F, I, L. The tabulated results of these pseudo contrasts are given in the second column of Table 4.4. The average estimated value of the goodness-of-link parameter, q, is smallest (1.233) when the hazard rate used to generate the data is decreasing ($\delta = 1/2$). When the underlying hazard rate in the generated data is constant or increasing, the two corresponding pseudo contrasts (1.360 and 1.344, respectively) are larger.

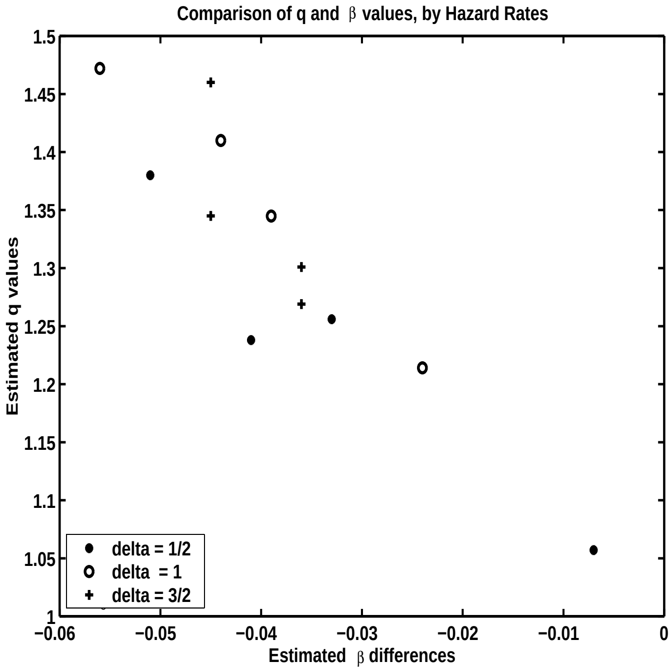
Table 4.4: Pseudo contrasts for q and $\bigtriangledown\beta$ between each hazard rate type.

Hazard Rate Value	Averages of Parameter Estimates	
δ	$\hat{q}$	$\bigtriangledown\hat{\beta}$
1/2	1.233	-0.033
1	1.360	-0.041
3/2	1.344	-0.041

Turning now to β, the parameter associated with duration in state 2, two estimates were obtained for each data set generated within a simulation run. First, estimates from the usual hybrid model ($q = 1$) were calculated, then corresponding parameter estimates were obtained in the hybrid model when q was allowed to vary. The differences between these two estimates, $\bigtriangledown\hat{\beta} = \hat{\beta}_1 - \hat{\beta}_{\hat{q}}$, became the new estimated parameter of interest. The median values of $\bigtriangledown\hat{\beta}$ from all 100 simulation runs are reported in the last column of Table 4.1. All of the values are negative, indicating these parameter estimates are amplified when q is estimated. Using all twelve Weibull cases (A - L) in our calculation, we found the average value of $\bigtriangledown\hat{\beta}$ was -0.0381, with a standard deviation of 0.0129.

It also appears that as the estimated value of q becomes larger, the estimated value of $\bigtriangledown\beta$ becomes more negative, and, hence, smaller. Figure 4.2 graphically illustrates this apparent inverse linear relationship between the estimates of these two model parameters. This result suggests that the estimated absolute value of β

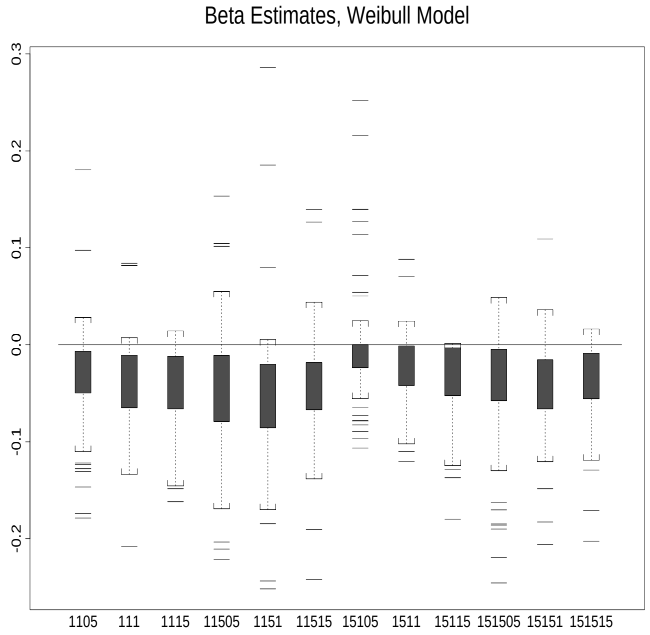
Figure 4.2: Scatterplot of estimated q and $\bigtriangledown\beta$ values.



in the model where q is free to vary increases as the estimated value of q increases.

The null value for $\bigtriangledown\beta$ is zero; however, in Figure 4.3, we see that most estimated values of the Weibull cases are less than zero. In the box plots of the data, we see variability in the length of the boxes, the length of the whiskers, the quantity and magnitude of the outliers, and the coverage of the value $\bigtriangledown\beta = 0$. To determine if the type of hazard used to generate the data might explain some of this apparent variability, we calculated pseudo contrasts for this variable too. The results are reported in the last column of Table 4.4. The average $\bigtriangledown\hat{\beta}$ values, in absolute terms, for the decreasing hazard rate, $\delta = 1/2$, is the smallest (-0.033), while for the nondecreasing hazard rates, the identical average is the largest (-0.041). However, in the plot (Figure 4.2) of the two parameter estimates, q and $\bigtriangledown\beta$, there does not

Figure 4.3: Box plots of $\bigtriangledown\beta$ estimates from all Weibull cases.



seem to be any pattern or clustering of the points from each of the three different underlying hazard types. A single outlying point near the origin for $\delta = 1/2$ may be the cause of the apparent numerical differences between these hazard types.

Likelihood ratio statistics were calculated to assess the significance of the variable representing duration in state 2, but only for the logistic model ($q = 1$). Preliminary findings indicated essentially the same results for the enriched model where q was estimated, so this additional LRS was not calculated in our study and, hence, not reported here. For the logistic model, the LRSs and associated proportions, $\mathcal{P}$, of cases where we reject the hypothesis that $\beta = 0$ are found in Table 4.5. In eleven of the twelve Weibull cases considered, the median LRSs indicate there is strong evidence to suggest that a hybrid time scale model should be adopted.

The only exception occurs for case I, where the observed value is less than 3.841. The observed proportion, 0.400, of $\mathcal{P}$ is also the smallest overall value for this case. The proportions of cases in which we reject the null hypothesis for the parameter β are generally above 0.660. Thus, in contrast with the findings for the parameter q, the results indicate that about two thirds of the 100 LRSs obtained for each of the twelve settings would reject the hypothesis $\beta = 0$. The proportions are smallest in cases H, I, and J, when λ_1 is larger than λ_2, whereas the largest proportions are observed in cases D and E, when λ_1 is smaller than λ_2. The proportions range from 0.400 to 0.883, with an average value of 0.693. We concluded that the addition of the duration in state 2 variable is often important to an underlying Markov model in the Weibull cases considered, although there were many instances where this was not the case. However, jointly estimating the goodness-of-link parameter, q, tends to increase the estimated value of β.

We now focus on the vector-valued parameters of our three-state model, beginning with the nonparametric estimator, $\boldsymbol{p}$, of the probability mass function for the time to infection variable, X. Two estimates of $\boldsymbol{p}$ were found in each simulation run: one from the standard logistic model ($q = 1$) and another for the same data set when q was free to vary. The difference between the estimates, $\bigtriangledown\hat{\boldsymbol{p}} = \hat{\boldsymbol{p}}_1 - \hat{\boldsymbol{p}}_{\hat{q}}$, was calculated for each run. To gain an understanding of the magnitude of the discrepancy between the two settings, we calculated the maximum and minimum values of this estimated difference parameter for each simulation run. We continue to use the median values to accurately compare these extreme values between the various Weibull cases. Only the median values of the maximum differences for $\boldsymbol{p}$,

Table 4.5: Likelihood ratio statistics for β, and proportions of cases where the hypothesis $\beta = 0$ is rejected in all Weibull models.

Case Label	Weibull Parameter Values				
	δ	λ_1	λ_2	LRS	$\mathcal{P}$
A	1/2	1	1	7.013	0.750
B	1	1	1	7.686	0.670
C	3/2	1	1	5.599	0.660
D	1/2	1	3/2	10.055	0.883
E	1	1	3/2	10.228	0.882
F	3/2	1	3/2	7.398	0.790
G	1/2	3/2	1	4.431	0.543
H	1	3/2	1	4.616	0.530
I	3/2	3/2	1	2.812	0.400
J	1/2	3/2	3/2	6.561	0.690
K	1	3/2	3/2	9.024	0.828
L	3/2	3/2	3/2	6.238	0.684

however, are reported in Table 4.6, due to the substantial symmetry about zero of the minimum values of $\bigtriangledown \boldsymbol{p}$.

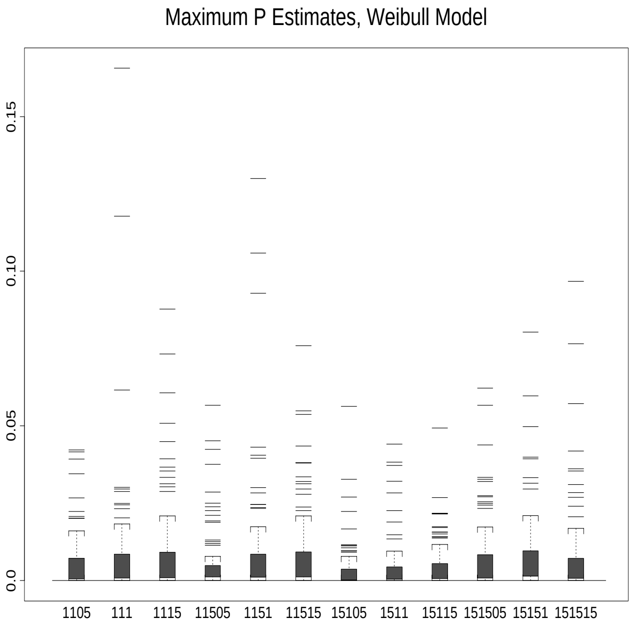
Perhaps what is most striking about these estimates is how uniformly small they are — the range of values is from 0.001 to 0.004. The average value of the twelve cases A - L is 0.003. This feature is not just evident in the summary statistics, but also in box plots of the data from the various Weibull cases. In Figure 4.4, we see that the maximal estimates for $\bigtriangledown \boldsymbol{p}$ in the data are heavily concentrated near the null value of zero, with only an occasional larger value beyond 0.05.

We also calculated pseudo contrasts for these estimates, based on the values reported in Table 4.6. Given the very small overall differences in these median values of $max(\bigtriangledown \hat{\boldsymbol{p}})$, it is not surprising that the average values for the three types of

Table 4.6: Estimates of $\bigtriangledown \boldsymbol{p}$ and $\bigtriangledown \boldsymbol{\lambda}$ for all Weibull cases.

Case Label	True Parameter Values			Median Parameter Estimates		
	δ	λ_1	λ_2	$\max(\bigtriangledown \boldsymbol{p})$	$\min(\bigtriangledown \boldsymbol{\lambda})$	$\max(\bigtriangledown \boldsymbol{\lambda})$
A	1/2	1	1	0.003	-0.071	0.005
B	1	1	1	0.003	-0.080	0.003
C	3/2	1	1	0.002	-0.076	0.013
D	1/2	1	3/2	0.003	-0.116	0.019
E	1	1	3/2	0.004	-0.110	0.056
F	3/2	1	3/2	0.003	-0.107	0.022
G	1/2	3/2	1	0.001	-0.038	0.001
H	1	3/2	1	0.002	-0.062	0.001
I	3/2	3/2	1	0.003	-0.074	0.007
J	1/2	3/2	3/2	0.003	-0.074	0.007
K	1	3/2	3/2	0.004	-0.100	0.052
L	3/2	3/2	3/2	0.002	-0.061	0.015

Figure 4.4: Box plots of estimated $\max(\bigtriangledown \boldsymbol{p})$ values from all Weibull cases.



hazard functions in the underlying Weibull data are so similar. The pseudo contrast values, found in the last column of Table 4.7, are identical to three decimal places (0.003). Thus, the extreme differences in the nonparametric estimates for $\boldsymbol{p}$, which is the probability mass function for the duration in state 1, are not substantially impacted by estimating the extra parameter, q, in the hazard function between states 2 and 3 within any of the Weibull models considered.

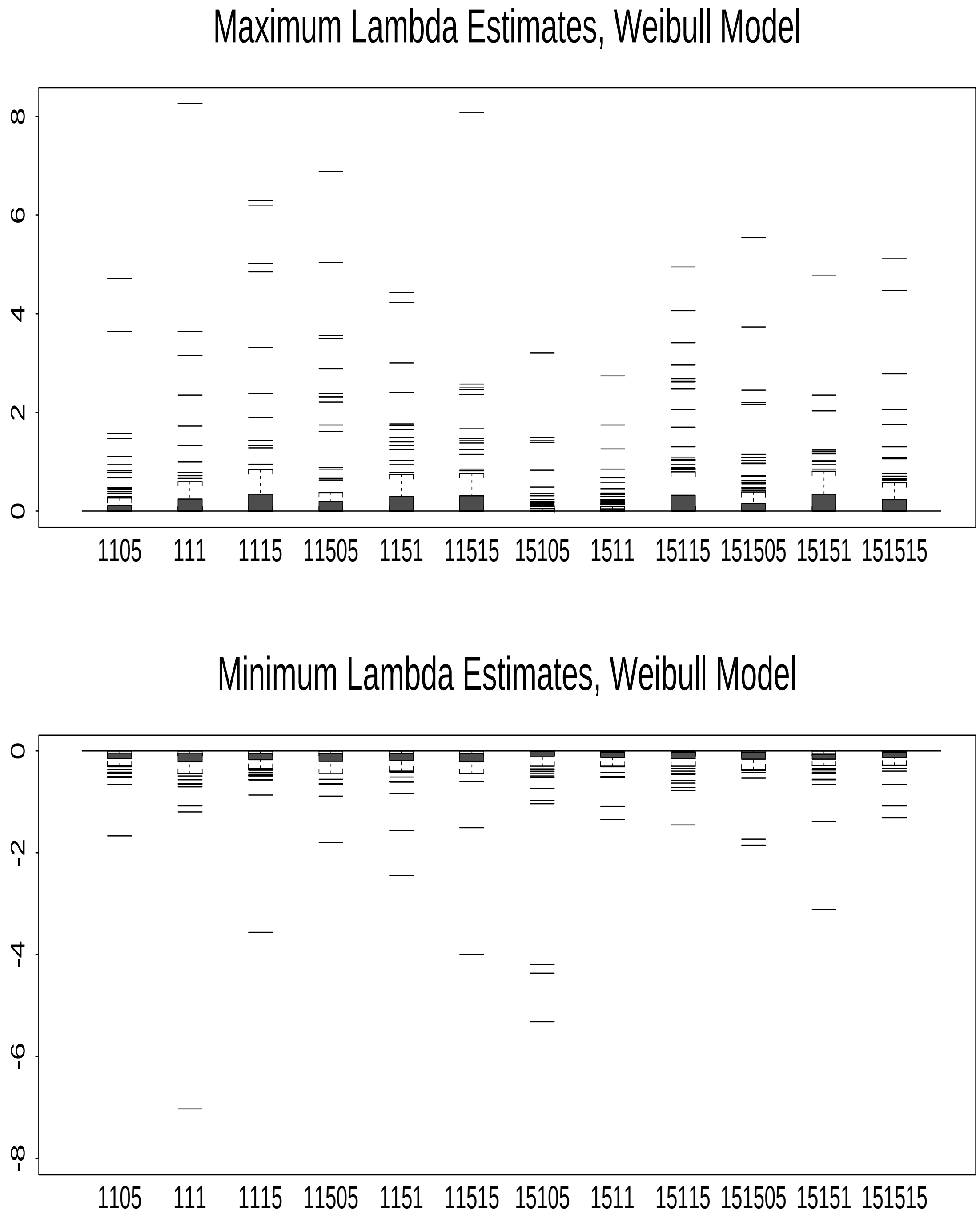
Table 4.7: Pseudo contrasts for $\bigtriangledown \boldsymbol{p}$ and $\bigtriangledown \boldsymbol{\lambda}$ between each hazard rate type.

Hazard Rate Value	Averages of Parameter Estimates		
δ	$\min(\bigtriangledown \boldsymbol{\lambda}\ )$	$\max(\bigtriangledown \boldsymbol{\lambda})$	$\max(\bigtriangledown \boldsymbol{p})$
1/2	-0.075	0.008	0.003
1	-0.088	0.028	0.003
3/2	-0.077	0.013	0.003

The last parameter to be evaluated from our three-state model with a logistic specification for the transition intensity from state 2 to state 3 is the baseline intensity vector, $\boldsymbol{\lambda}$. Once again we obtained two estimates of $\boldsymbol{\lambda}$ per simulation run, one from a model where q is fixed at unity and another where q is estimated. Like the nonparametric estimates of $\boldsymbol{p}$, we considered only the extreme estimated differences. Unlike the nonparametric estimates of $\boldsymbol{p}$, we did not detect symmetry about zero in these estimated differences. Hence, for $\bigtriangledown \hat{\boldsymbol{\lambda}} = \hat{\boldsymbol{\lambda}}_1 - \hat{\boldsymbol{\lambda}}_{\hat{q}}$, we will report the median values of the maximum and minimum differences.

Looking at the plots of the minima and maxima of the parameter estimate

Figure 4.5: Box plots of $max(\bigtriangledown\boldsymbol{\lambda}), min(\bigtriangledown\boldsymbol{\lambda})$ estimates from all Weibull cases.


Maximum Lambda Estimates, Weibull Model

Minimum Lambda Estimates, Weibull Model

differences for the twelve Weibull cases in Figure 4.5, we note more variability and outlying data points in the box plots of the maximal values. The whiskers and boxes of the boxplots for the minimal values are generally more compact, too. In spite of these graphical distinctions between the Weibull cases, the median values of the minimum of $\bigtriangledown\hat{\boldsymbol{\lambda}}$ were larger in absolute terms than those for the maximum. This result is easily seen by comparing the last two columns of Table 4.6 for the cases A - L. To determine if the underlying hazard rates generating the data for these Weibull models might be the source of some of these apparent differences, we again calculated pseudo contrasts.

Beginning with the median values of the minimal differences, we see in the second column of Table 4.7 that the average values calculated from the nonconstant hazard rates which generated the data are almost identical (-0.075, and -0.077, respectively). When the underlying hazard rate is constant, the average value of cases B, E, H, and K is comparable (-0.088), but further from zero. When we examine the median values of the maximal differences, we find similar results. The average value calculated from cases A, D, G, and J for the decreasing hazard rate is 0.008, while the average value for the increasing hazard rate, calculated from cases C, F, I, and L, is 0.013. The average value for the constant hazard case, 0.028, is again furthest from the null value of zero.

4.2.2 Extreme Value Models

From the twelve Weibull cases considered thus far, we selected a subset of three cases to compare with corresponding ones from an extreme value distribution. The same three values, $\{1/2, 1, 3/2\}$, of the shape parameter, δ, were used in both the Weibull and extreme value distributions. The first scale parameter, λ_1, used for generating Weibull random variables for the distribution of time in state 1 was fixed at 1, while the second, λ_2, used for generating Weibull random variables for the distribution of time in state 2 was fixed at 3/2. The resulting values of the extreme value parameter, $\phi = -\delta^{-1}\log(\lambda_2)$, are recorded to four decimal places in the upper half of Table 4.8. The types of estimated parameter differences calculated for these extreme value cases were identical to the ones used in the Weibull setting.

Table 4.8: Comparison of estimated parameters between three extreme value and Weibull cases.

Case Label	Extreme Value Parameter Values			Median Parameter Estimates				
	ϕ	λ_1	λ_2	q	$\bigtriangledown\beta$	$\max(\bigtriangledown\boldsymbol{p})$	$\min(\bigtriangledown\boldsymbol{\lambda})$	$\max(\bigtriangledown\boldsymbol{\lambda})$
M	-0.8109	1	3/2	2.452	-0.176	0.004	-0.182	0.181
N	-0.4055	1	3/2	2.091	-0.172	0.006	-0.159	0.310
O	-0.2703	1	3/2	2.021	-0.164	0.004	-0.152	0.468

Case Label	Weibull Parameter Values			Median Parameter Estimates				
	δ	λ_1	λ_2	q	$\bigtriangledown\beta$	$\max(\bigtriangledown\boldsymbol{p})$	$\min(\bigtriangledown\boldsymbol{\lambda})$	$\max(\bigtriangledown\boldsymbol{\lambda})$
D	1/2	1	3/2	1.380	-0.051	0.003	-0.116	0.019
E	1	1	3/2	1.472	-0.056	0.004	-0.110	0.056
F	3/2	1	3/2	1.460	-0.045	0.003	-0.107	0.022

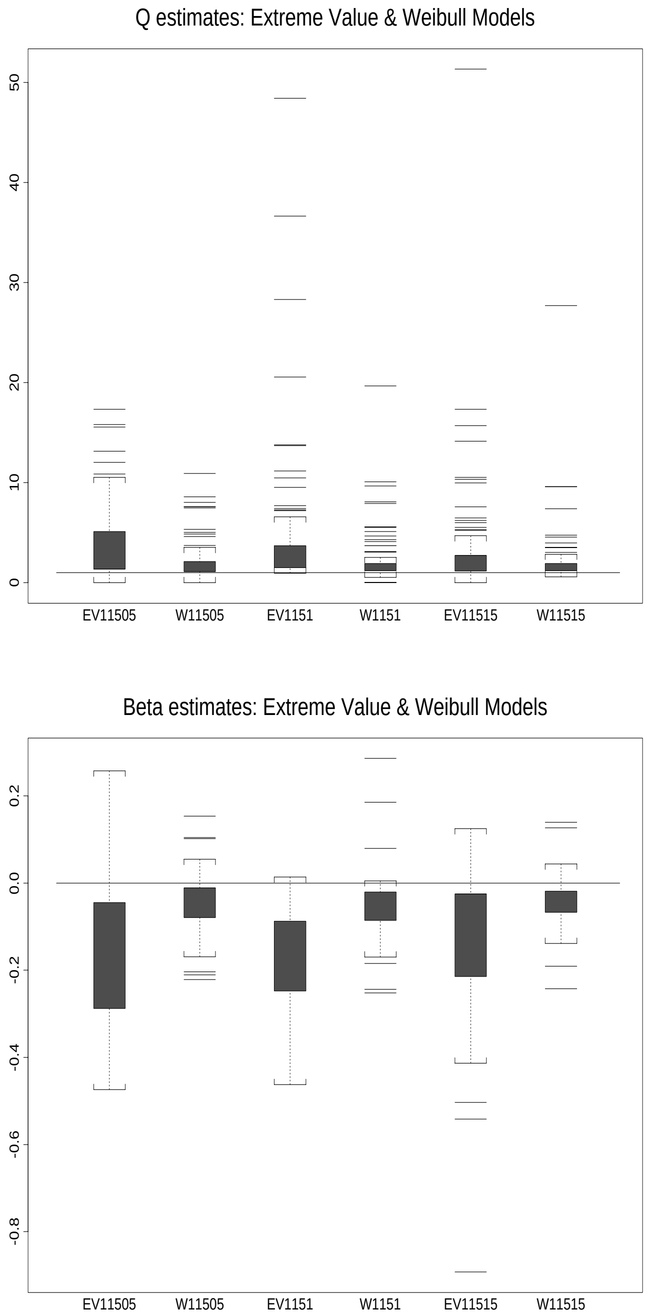
The same general results obtained in all the Weibull cases hold for the three extreme value cases M, N, and O in Table 4.8. The median values of q were all greater than one, and the range in the estimates was relatively narrow, from 2.021 to 2.452. The differences in the parameter estimates for the duration in state 2 variable, $\bigtriangledown\beta$, are also negative, indicating the estimation of q in the enriched model amplifies the regression parameter in this setting too. The median values for the maximum differences in the nonparametric estimates of $\boldsymbol{p}$ are once again uniformly small (all < 0.006). However, for these extreme value cases, the median values of the maximum of $\bigtriangledown\boldsymbol{\lambda}$ tend to be larger than the corresponding absolute minimum values. Only for case M where ϕ is the smallest (most negative) are the resulting estimates equivalent.

Comparing the parameter estimates obtained from the two distribution settings numerically, we found the median values of the parameter estimate differences for the extreme value models were all larger than their Weibull counterparts. We constructed pseudo contrasts between the two settings, in order to quantify this disparity between the cases of the two distributions studied. The pseudo contrasts for each setting were calculated by averaging over all three types of the underlying hazard function created by δ in the data. For example, the average value of q, 2.188, in the extreme value setting was found by summing the observed values of q for cases M - O; that is, $2.188 = 1/3 \times (2.452 + 2.091 + 2.021)$. We found the simple averages of q, $max(\bigtriangledown\boldsymbol{p})$ and $min(\bigtriangledown\boldsymbol{\lambda})$ were all about about one and a half times larger in the extreme value setting. The average value of $\bigtriangledown\beta$ is now 3.4 times larger, while the average value of $max(\bigtriangledown\boldsymbol{\lambda})$ is 9.9 times larger. Thus, the direction of the differences is still generally consistent with the Weibull models, but the magnitude of those differences is greater in these extreme value models. The two parameters in the logistic specification of the hazard function between states 2 and 3, β and $\boldsymbol{\lambda}$, were the most affected by estimating q.

Graphical comparisons confirm these numerical differences. In the plots of the various parameter estimates, we use the same legend as before (Table 4.2, lines 4-6) for the labels beneath the individual box plots and further distinguish between the two distributions by using "W" for the Weibull setting and "EV" for the extreme value setting. In the upper panel of Figure 4.6, we see there is greater variability in the estimates of q from the EV cases: the interquartile range is larger (shaded box area), the whiskers are longer, and the outliers are more extreme. In the lower panel of this same figure, the interquartile ranges are larger and the whiskers much longer for $\bigtriangledown\beta$ in the EV cases. Thus, when estimating q in the EV cases, the estimates of β are generally even larger than in the Weibull cases, although the range of possible values is larger as well.

In the upper panel of Figure 4.7, the box plots of the estimates of $max(\bigtriangledown\boldsymbol{\lambda})$ in the three extreme value cases have much larger interquartile ranges. The whiskers are also longer. Thus, the large numerical differences noted in the pseudo contrasts and the median values are evident in this plot as well. There is very little difference between the Weibull and EV cases in the estimates of $min(\bigtriangledown\boldsymbol{\lambda})$, which is again consistent with the numerical results. In the lower panel of Figure 4.7, the box plots of the $max(\bigtriangledown\boldsymbol{p})$ estimates from the paired cases are very similar. The interquartile range and whiskers for the EV cases tend to be slightly larger, but overall these discrepancies are quite small.

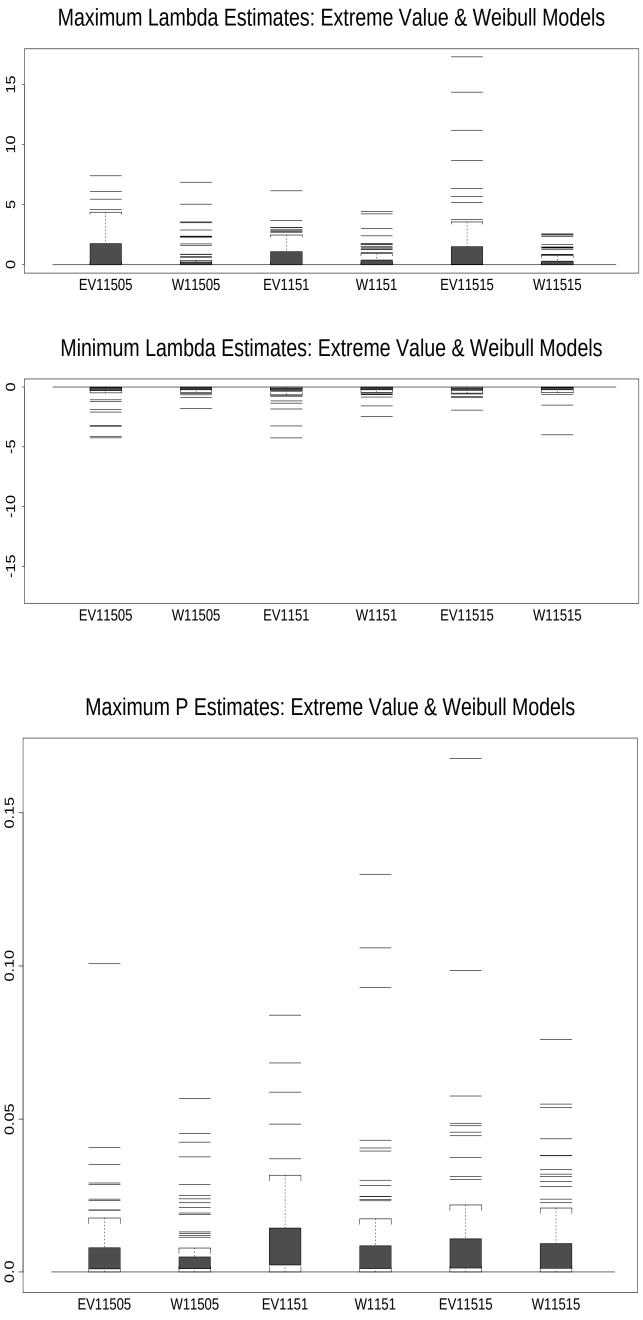
Figure 4.6: Box plot comparisons of estimated q and $\bigtriangledown\beta$ values between three extreme value and Weibull cases.



(a) q estimates

(b) $\bigtriangledown\beta$ estimates

Figure 4.7: Box plot comparisons of estimated $\bigtriangledown\boldsymbol{\lambda}$ and $\bigtriangledown\boldsymbol{p}$ values between three extreme value and Weibull cases.



(a) $max(\bigtriangledown\boldsymbol{\lambda}), min(\bigtriangledown\boldsymbol{\lambda})$ estimates

(b) $max(\bigtriangledown\boldsymbol{p})$ estimates

Table 4.9: Likelihood ratio statistics for β, q, and proportions of cases where the hypotheses $\beta = 0$, $q = 1$ are rejected in the extreme value models.

Case	Extreme Value			LRS		LRS	
	Parameters						
Label	ϕ	λ_1	λ_2	β	$\mathcal{P}$	q	$\mathcal{P}$
M	-0.8109	1	3/2	3.718	0.495	0.001	0.052
N	-0.4055	1	3/2	8.475	0.790	0	0.030
O	-0.2703	1	3/2	9.842	0.818	0	0

We also assessed the significance of the duration in state 2 variable and tested whether the goodness-of-link parameter was equal to one. The median LRSs for testing $\beta = 0$ are found in column 5 of Table 4.9. The values are all greater than 3.718, which suggests hybrid models are appropriate for these three extreme value situations too. The proportions, $\mathcal{P}$, of cases in which we reject the hypothesis $\beta = 0$ at the 0.05 level for cases N and O are comparable in magnitude to their Weibull counterparts (cases E and F). Only case M in this extreme value setting is half of the observed value in the corresponding Weibull setting.

Analogous to our findings in the Weibull cases, there is absolutely no evidence to suggest that q is different than one for these extreme value cases. The LRSs for testing whether the logistic specification is appropriate were all basically zero, and the associated p-values virtually one. The proportions of cases, $\mathcal{P}$, we report in the final column of Table 4.9 are almost identical to the corresponding proportions in the Weibull cases; see cases D, E, and F in Table 4.3. Thus, in this setting too, the proportions of cases in which we reject the hypothesis, $q = 1$, at the 0.05 level are nearly zero or are zero.

In summary, the logistic link function is an appropriate link function for all of the models considered here. The proportion of cases where $q = 1$ was rejected was generally zero or very close to zero in all the model settings. The median likelihood ratio statistics were in agreement with this finding, since they too indicated the logistic link function ($q = 1$) was not incongruous for any case, Weibull or extreme value. The value of q was slightly larger in the Weibull models where the underlying hazard rate was nondecreasing, while in the extreme value setting, the estimated value of q was slightly larger when the value of ϕ was the smallest (most negative).

The parameter estimates in the second transition intensity are the most affected by estimating q, while the parameters in the first transition intensity are not. The extreme values of the baseline intensity function, $\bigtriangledown\boldsymbol{\lambda}$, are affected by estimating the goodness-of-link parameter. In the twelve Weibull cases, the median values of the estimated minimum of $\bigtriangledown\boldsymbol{\lambda}$ are furthest away from the null value of zero, in absolute terms, but the maximum values display more variability and outlying observations. In the two extreme value cases where the underlying hazard function is nondecreasing, the opposite result is observed.

The regression parameter, β, is also affected by estimating q, and there appears to be a positive linear relationship between estimated values of q and β. As the estimated value of q becomes larger, the corresponding estimated absolute value of β also becomes larger, in all of the Weibull and EV cases we considered. The duration in state 2 variable was generally important in all of the models we studied, so it remains to be seen how the estimation of q in true Markov models might be affected. Thus, for the nonzero values of β observed in the various settings, the estimated hazard function between states 2 and 3 was magnified — either more strongly increasing or decreasing — when the link parameter was estimated than when it was not estimated. When the underlying hazard function is decreasing ($\delta = 1/2$) in the Weibull cases, the values of $\hat{q}$ are smallest and the values of $\bigtriangledown\hat{\beta}$ are generally the largest (least negative).

The estimated values of the goodness-of-link parameter, q, were generally larger than one for the Weibull cases and larger than two for the extreme value cases. The median values of the parameter estimate differences for the remaining model parameters were larger and more variable in the extreme value setting compared to the corresponding Weibull cases.

Hence, not estimating the value of q may affect the estimated values of the parameters associated with the logistic transition intensity between states 2 and 3. Depending on the purpose of an analysis, the link function parameter may be estimated to provide adjusted estimates of the model parameters, in addition to assessing the appropriateness of the logistic specification of the hazard function.

4.3 Example: AIDS in Hemophilia Patients

In §2.7, we illustrated the advantages of our regression approach using a data set presented and analysed in several articles (De Gruttola and Lagakos, 1989; Kim, De Gruttola and Lagakos, 1993; Frydman, 1992; Frydman, 1995). We consider that example once more, so that three aspects of model fit may be assessed. First, the appropriateness of the logistic specification for the hazard function between states 2 and 3 can be determined. Second, the impact on estimation of the other regression coefficients in the model when maximizing the enriched log likelihood, which includes the parameter q, can be evaluated. Lastly, under the assumption that the logistic specification is correct and, therefore, q can be fixed at one, the uniqueness and consistency of the maximum likelihood estimators are examined.

4.3.1 Link Function Assessment

The link function based on the logistic specification (eq 4.1) for the hazard function for transitions from state 2 to state 3 was used here to fit most of the models discussed in §2.7. The results of that estimation process are summarized in Table 4.10, where $\bigtriangledown$ in front of the various symbols in the column headings indicates the difference between the estimate from the usual hybrid model, and the estimate obtained from the enriched hybrid model, where q was free to vary. For example, the estimate of $\bigtriangledown\beta$ represents $\beta_1 - \beta_q$, where the subscripts 1 and q indicate corresponding parameter estimates in the logistic model, where q is fixed at 1, and in the enriched model, where q is estimated. In the following, the results for the nonparametric estimates of the time to infection distribution ($\hat{\boldsymbol{p}}$) will not be discussed, but only reported in Table 4.10. As we found in the simulation study, these estimates were quite robust to changes in the value of q and varied only slightly from those obtained in the logistic models ($q = 1$). The resulting maximum and minimum differences in the estimated values of $\boldsymbol{p}$ were negligible, and were symmetric about zero. The results of the maximum differences are summarized in the final of column of the table.

When a simple time non-homogeneous Markov model was chosen, the final

Table 4.10: Parameter estimates for various models when the logistic specification is adopted, AIDS example.

Model Description	q	$\bigtriangledown\beta$	$\bigtriangledown\nu$	$\min(\bigtriangledown\boldsymbol{\lambda})$	$\max(\bigtriangledown\boldsymbol{p})$
Markov (Null)	1.0006			-0.0002	1.0e-17
Markov & Tx	1.0277		0.0018	-0.0067	9.4e-05
Hybrid	1.0673	0.0029		-0.0105	0.0003
Hybrid & Tx	1.1304	0.0071	0.0747	-0.0211	0.0006

Table 4.11: Estimated q values, likelihood ratio statistics and associated p-values for various models employing the logistic specification, AIDS example.

Model Description	q	LRS	p-value
Markov (Null)	1.0006	0	1
Markov & Tx	1.0277	0	1
Hybrid	1.0673	0.0182	0.8927
Hybrid & Tx	1.1304	0.0372	0.8471

estimate for q was 1.0006. Given this minimal increase from one, it is not surprising that there was very little change in the chronological time factors of the second transition intensity ($\boldsymbol{\lambda}$). The greatest change in an element of this vector-lengthed parameter was -0.0002 and the observed LRS was set to zero ($p = 1$); see the initial row of Table 4.11. We report only the minimum values of $\bigtriangledown\boldsymbol{\lambda}$ in this table, as the maximum values were all zero. When the treatment variable was added to this model, the estimated value of q increased slightly to 1.0277. Again, the minimum difference between the parameters for the chronological time factors of the second transition intensity ($\boldsymbol{\lambda}$) was quite small (-0.0067). The difference, $\bigtriangledown\nu$, between the estimates of the coefficient for the treatment covariate was small as well (0.0018), and the observed value of the LRS was also set to zero.

When the hybrid time scale model, formed by adding a covariate for duration in state 2 to the Markov model, is fitted using the general logistic specification, the resulting value of $\hat{q}$ is 1.0673. The minimum change in the chronological time factors of the second transition intensity is -0.0105, and the difference between the estimates for the parameter (β) associated with the duration in state 2 covariate is 0.0029. The observed value, 0.0182, of the LRS for testing the hypothesis $\beta = 0$ has an associated p-value of 0.89. When the treatment variable is added to the underlying hybrid model, the observed value of the LRS increases to 0.0372 ($p =$ 0.8471). The differences between parameters estimated in this model when q is free to vary and one when q is fixed at one are 0.0071, 0.0747, and - 0.0211 for the regression coefficients associated with the duration in state 2, the treatment effect, and the chronological time factors, respectively.

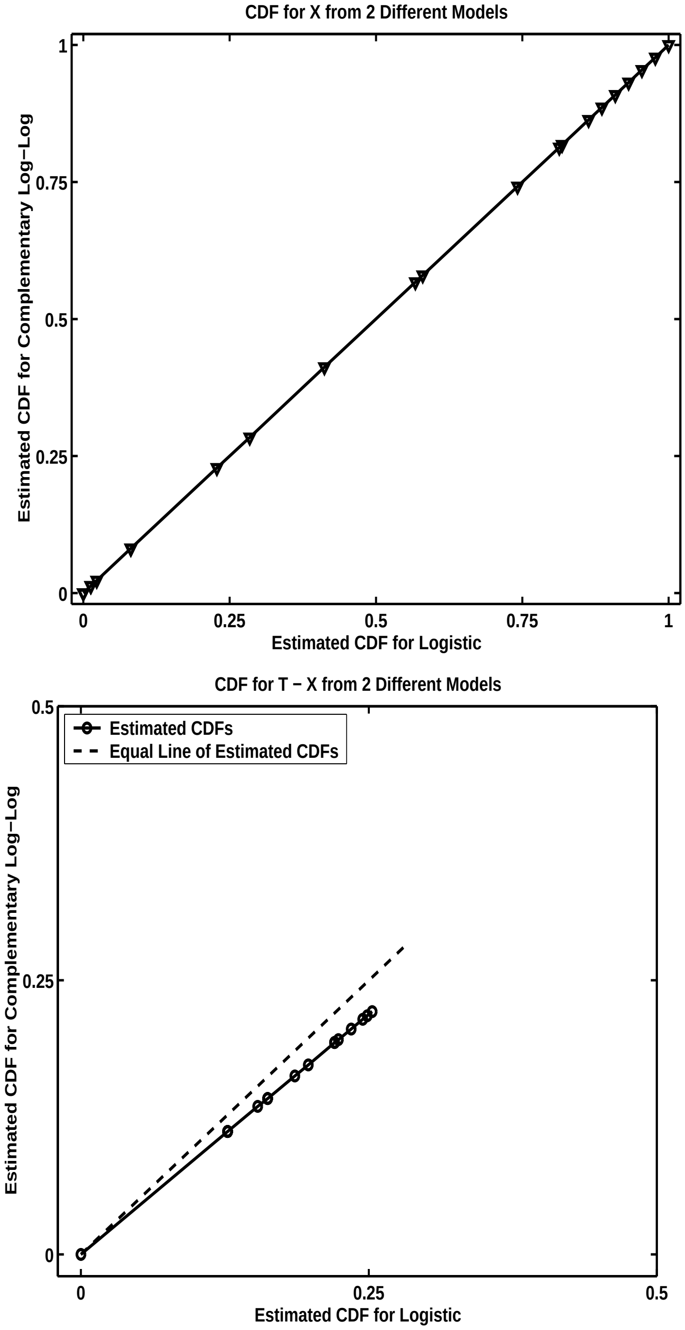
In summary, when the logistic specification for the second transition intensity was assessed, none of the conclusions described in §2.7 changed. There was virtually no change in either Markov model, and no significant changes for either of the hybrid time scale models. The regression parameter estimates (β, ν) were all slightly attenuated in the models where q was free to vary, compared to the models where it was fixed. Consistent with our findings in the simulation study, these results suggest that when fitting a model which adopts the logistic specification for the hazard function, there is a reciprocal action between the regression coefficients and the logistic specification parameter. The effect, though, is opposite for these data to what we observed for any of the Weibull or extreme value cases considered in the simulation study. For the Kim *et al.* data, the regression coefficients are attenuated, while for the simulated data, the regression coefficients are amplified. A possible reason for this discrepancy may be due to the opposite signs of the estimated coefficient, β, for the duration in state 2 variable.

The estimated value of q is largest in a model which includes both covariates, while the estimated values of the regression parameters are smallest in this setting. The first chronological time factor of the second transition intensity, (λ_1), numerically changed the most in all four models. The estimated value of q was larger in the hybrid time scale models than in either of the Markov models; it was also larger if the treatment covariate was included in either time scale model.

The hazard function, with a limiting extreme value specification (eq 4.2), for the transition between states 2 and 3 was used here to fit two of the models discussed in §2.7. Hybrid time scale models, either with or without the treatment covariate, were maximized simultaneously with the link specification parameter q. It was necessary to constrain the value of q to be 1.234e+6 in the maximization process, so the true MLEs were not found. Since we were unable to find the parameter estimates when q is at its limiting value of infinity, we cannot determine how different these parameter estimates are from those obtained when an extreme value model is fit. In spite of this limitation, we can still graphically compare functions estimated under a logistic and a complementary log-log model, to see whether one model seems to fit the data better.

Figure 4.8 provides visual comparisons between models fitted under the two specifications. In the upper panel, the estimated values of the cumulative distribution function (CDF) for a transition from state 1 to state 2 are plotted against each other. The plotted values follow the line of equal estimated cumulative probabilities, suggesting no distinction between the models. In the corresponding lower panel, the estimated cumulative distribution function for a transition to state 3 from state 2 are plotted against each other. The CDFs for $T - X|X$, under only a Markov assumption, are estimated by $\sum_t h(x,t) \prod_{k:\, t_k<t} \{1 - h(x,t_k)\}$. Thus, there is no dependence on X when the duration in state 2 variable, $T - X$, is not included in the specification of the hazard function. There is an obvious small departure from the line of equal estimated cumulative probabilities. The values of the CDF estimated under the complementary log-log model are smaller than the corresponding values estimated under the logistic model. However, with the caveat that the results from the maximization of the log likelihood for the complementary log-log formulation may not be the true MLEs, the models may not be really that

Figure 4.8: Comparison of estimated CDFs assuming different hazard specifications, AIDS example.



(a) $\hat{F}_X$

(b) $\hat{F}_{T-X}$

distinguishable.

4.3.2 MLEs - Uniqueness and Consistency

The likelihood function that Frydman (1995) used was maximized via a Lagrangian function. The estimating equations for the distribution of time to HIV infection and for the chronological time factors of the intensity function with respect to transitions from state 2 to state 3 were also shown to be self-consistent equations. She did not find a self-consistent equation for the estimating equation used to calculate the regression coefficient associated with duration in state 2. In the Kim *et al.* (1993) paper, the likelihood was maximized using a combination of estimation schemes. Estimating equations for the distribution of time in state 1 were solved using a version of the self-consistent algorithm proposed by Turnbull (1976) for singly censored data. The parameters in the distribution function for the induction time between HIV seroconversion and onset of symptoms were then estimated using a Newton-Raphson algorithm and the current estimates for the distribution of time in state 1. The extensions to Frydman's basic model which we developed in chapter 2 were shown to be self-consistent estimators. We now confirm that these estimators are also the maximum likelihood estimators.

Gentleman and Geyer (1994) applied standard convex optimization techniques to the analysis of interval-censored data. Using the necessary and sufficient conditions for constrained optimization, i.e., the Kuhn-Tucker conditions, their paper provides a straightforward method of verifying that Turnbull's (1976) self-consistent estimates are also maximum-likelihood estimates, and in addition, that the MLE is unique.

Using the notation of Gentleman and Geyer, but adapting it to our special bivariate case, we denote the intervals in which unobserved transitions for subject n may occur as $A_n = [X_{Ln}, X_{Rn}]$ for the transition to state 2, and $C_n = [V_{Ln}|X_n, V_{Rn}|X_n]$ for the transition to state 3 from state 2, given a value of X from A_n. We let $\{s_j\}_{j=0}^{J}$ denote the unique ordered elements of $\{\{X_{Ln}\}_{n=1}^{N}, \{X_{Rn}\}_{n=1}^{N}\}$ and let $\{r_k\}_{k=0}^{K}$ denote the unique ordered elements of $\{\{V_{Ln}|X_n\}_{n=1}^{N}, \{V_{Rn}|X_n\}_{n=1}^{N}\}$. In addition, define ξ_{nj} to be the indicator of the event $[s_{j-1}, s_j] \subseteq A_n$ for $j = 1, \ldots, J$, $n = 1, \ldots, N$ and τ_{nk} to be the indicator of the event $[r_{k-1}, r_k] \subseteq C_n$ for $k = 1, \ldots, K$, $n = 1, \ldots, N$, at a fixed value of X. Thus, we will examine the results based on the marginal distribution of X and the conditional distribution of $T - X|X = V|X$ separately, not jointly. This univariate treatment of the bivariate data may still provide some useful insights into the properties of the estimates obtained in our regression approach.

The first two Kuhn-Tucker conditions include a linear constraint,

$$1 - \sum_{j=1}^{J} p_j = 0 \ , \tag{4.3}$$

and two nonnegativity constraints,

$$\begin{aligned} p_j &\geq 0 \qquad (j = 1, \ldots, J) \ , \\ \lambda_k &\geq 0 \qquad (k = 1, \ldots, K). \end{aligned} \tag{4.4}$$

A solution to this constrained optimization problem will in fact be the MLEs if and only if there exist Lagrange multipiers, μ_j $(j = 1, \dots , J)$, such that the additional Kuhn-Tucker conditions

$$\mu_j \, p_j = 0 \qquad (j = 1, \dots , J) \; , \tag{4.5}$$

$$\mu_j \geq 0 \qquad (j = 1, \dots , J) \; , \tag{4.6}$$

$$\frac{\partial}{\partial p_j} \left\{ \log L(p, \lambda, \beta) + \sum_{j=1}^{J} p_j(\mu_j - \mu_o) \right\} = 0 \; , \qquad (j = 1, \dots , J) \tag{4.7}$$

$$\frac{\partial \log L}{\partial \lambda_k} \leq 0 \; , \qquad (k = 1, \dots , K) \tag{4.8}$$

$$\frac{\partial \log L}{\partial \beta} \leq 0$$

hold as well. The unrestricted Lagrange multiplier, μ_0, originates from the equality constraint given in eq. 4.3 and equals N if equations 4.4-4.7 hold simultaneously. If we let $d_j = \frac{\partial \log L}{\partial p_j} = \sum_{n=1}^{N} \frac{\xi_{nj}}{\sum \xi_{nj} \, p_j}$, then equation (4.7) can be written as $d_j + \mu_j - N = 0$. This sum represents the reduced gradient, or the gradient of the variables free to vary. If $p_j = 0$, set $\mu_j = N - d_j$, and for $p_j > 0$, set $\mu_j = 0$.

To test for convergence of the self-consistent estimator to the MLE, Gentleman and Geyer proposed using the Lagrange multipliers. If the Lagrange multipliers are nonnegative at $\hat{p}$, then the self-consistent estimator is also the maximum likelihood estimator. The reduced gradient should be approximately zero. The uniqueness of the MLE can be determined by examining the Hessian matrix, H. If H is strictly negative definite, which will occur when the log likelihood is strictly concave, then the MLEs are unique. If we let A denote the matrix containing the indicator functions ξ_{nj}, and let C denote the matrix containing the indicator functions τ_{nk}, then $H_X = A'DA$, and $H_{T-X|X} = C'EC$. The diagonal matrices D and E have respective entries $-1/(\sum_j \xi_{nj}\ p_j)^2$, and $-1/(\sum_k \tau_{nk}\ p_k)^2$. The probability mass function, p_k, for the variable representing the time to diagnosis from the infection time is estimated from the hazard function assuming a Markov model framework. The MLEs will be unique if the rank$(A) = J$ and the rank$(C) = K$.

Turning now to the data from Kim *et al.* (1993), the matrix $\{\xi_{nj}\}$ was found to be of full rank, so the maximum likelihood estimate of parameters based on the variable X is unique. The matrix, $\{\tau_{nk}\}$ for the variable, $V|X$, was not found to be of full rank for any value of X. Using the possible values of X for individuals known to have made the transition to state 2 one at a time, the value of K and the rank of the matrix C were calculated for the corresponding set of transition times out of state 2. The rank of C was often about half of what it could be; e.g. when X was set to one, the rank of C was five but K was 10. The discrepancy between the value of K and the rank of C became smaller as the value of X increased. Hence, the maximum likelihood estimate of the parameters in the transition intensity between states 2 and 3 is not unique.

The Kuhn-Tucker conditions indicate there are 14 equivalence classes or disjoint intervals for the distribution of the time to infection. Equivalence classes are defined in the Turnbull algorithm as the regions between a left-hand limit of a censoring interval which is followed next by a right-hand limit of a possibly different censoring interval. The survivor function can only make jumps within the set of disjoint intervals. The equivalence classes for X are in fact all single time points: [3,3], [5,5], [7,7], [8,8], [9,9], [10,10], [11,11], [12,12], [13,13], [14,14], [15,15], [16,16], [17,17], and [18,18]. There are 11 equivalence classes or disjoint intervals for the induction-time distribution [7,8], [12,12], [13,13], [15,15], [16,16], [17,17], [19,19], [20,20], [21,21], [22,22], and [23,23]. To check on the consistency of the MLE for F_x, we need to examine the values of the Lagrange multipliers and the reduced gradient.

The results from fitting a hybrid time scale model are found in the last three columns of Table 4.12. Column one specifies the intervals where support is possible for the time to infection distribution and the fifth column gives the estimates, $\hat{p}_j$, at those support points. The estimated probabilities for the intervals $[8,8], [9,9]$, and $[17,17]$ were set to zero, as their values were less than 10^{-15}. The values of the reduced gradient, found in the sixth column, are not all close to zero. The Lagrange multipliers, however, are all nonnegative (column 7) and indicate the Kuhn-Tucker condition (eq 4.6) is satisfied. The analysis was rerun with the probabilities for intervals $[8,8], [9,9]$, and $[17,17]$ initially set to zero. No change was evident in the reduced gradient for this model nor in the Lagrange multipliers. The convergence criterion for the observed data log-likelihood, now set at 10^{-7} and not the more lax value of 10^{-4} which was used in chapter 2, was further reduced to 10^{-9}. The values of the reduced gradient, particularly for intervals $[3,3]$ and $[5,5]$, were not affected.

The results obtained from fitting a Markov model were examined, to see if the cause of the poor fit was due to estimating the more complicated hybrid model. In columns 2-4 of Table 4.12, we see there are still some large values for the reduced gradient (column 3) for intervals $[3,3], [5,5]$. However, the largest value of the reduced gradient vector now occurs at $t = 5$, not $t = 3$. The Lagrange multipliers

Table 4.12: Equivalence classes, associated probabilities, reduced gradient and Lagrange multiplier values for Markov and hybrid time scale models, AIDS example.

	Markov			Hybrid		
Interval	$\hat{p}_j$	Reduced gradient	Lagrange multiplier	$\hat{p}_j$	Reduced gradient	Lagrange multiplier
$[3, 3]$	0.0186	24.019	0	0.0125	32.909	0
$[5, 5]$	0.0004	37.821	0	0.0108	18.765	0
$[7, 7]$	0.0664	-1.038	0	0.0577	6.827	0
$[8, 8]$	0	0	68.013	0	0	61.774
$[9, 9]$	0	0	48.569	0	0	44.732
$[10, 10]$	0.1431	2.936	0	0.1464	0.347	0
$[11, 11]$	0.0529	4.797	0	0.0575	3.047	0
$[12, 12]$	0.1354	-1.968	0	0.1259	-0.327	0
$[13, 13]$	0.1520	-1.367	0	0.1581	-3.513	0
$[14, 14]$	0.0085	3.149	0	0.0088	1.038	0
$[15, 15]$	0.1655	-2.752	0	0.1647	-2.881	0
$[16, 16]$	0.0683	-1.874	0	0.0686	-1.971	0
$[17, 17]$	0	0	3.952	0	0	4.106
$[18, 18]$	0.1888	0.112	0	0.1888	-0.043	0

are still nonnegative in this analysis.

A fourth approach for investigating the lack of convergence to the MLE, $\{\hat{F}_x\}$, considered the amount of information in the data at time points 3 and 5. The average interval width for individuals whose transition time included time point 3 was 5.8 years, and for time point 5 was 5.4 years. In contrast, the average interval width for individuals whose transition time included time point 18 was 1.8 years. The average interval width for later time points is much smaller than for earlier time points, suggesting the associated parameters could be better estimated. To investigate the hypothesis that a lack of precise data was the apparent cause of the convergence problem, the interval widths for 11 individuals making a transition to state 2 from state 1 were reduced. The minimum value of a possible transition time was increased from the common value of $t = 1$ to $t = 3$ for all 11 subjects and the maximum value fixed at $t = 5$, where previously the values ranged from 5 to 15.

The results comparing Markov models fit with the original data and the test data are found in Table 4.13. The reduction in the width of the intervals has dramatically improved the fit at time points 3 and 5. The values of the reduced gradient declined from 24.019 to 3.656 and from 37.82 to 4.512 for time points 3 and 5, respectively. The estimated probability for p_5 changed the most - increasing from 0.0004 to 0.0664. The remaining estimated probabilities and reduced gradient values changed minimally. The Lagrange multipliers for time points 8 and 9 were also somewhat reduced in this analysis. Hence, we concluded that the likelihood surface is somewhat flat. When the data for the heavily and lightly treated groups are examined separately, the same lack of convergence to the MLEs is evident.

Table 4.13: Equivalence classes, associated probabilities, reduced gradient, and Lagrange multiplier values for original and test data in a Markov model, AIDS example.

	Markov, original data			Markov, test data		
Interval	$\hat{p}_j$	Reduced gradient	Lagrange multiplier	$\hat{p}_j$	Reduced gradient	Lagrange multiplier
$[3, 3]$	0.0186	24.019	0	0.0212	3.656	0
$[5, 5]$	0.0004	37.821	0	0.0664	4.512	0
$[7, 7]$	0.0664	-1.038	0	0.0319	-1.279	0
$[8, 8]$	0	0	68.013	0	0	38.638
$[9, 9]$	0	0	48.569	0	0	26.360
$[10, 10]$	0.1431	2.936	0	0.1333	2.938	0
$[11, 11]$	0.0529	4.797	0	0.0367	4.955	0
$[12, 12]$	0.1354	-1.968	0	0.1340	-1.556	0
$[13, 13]$	0.1520	-1.367	0	0.1468	-1.117	0
$[14, 14]$	0.0085	3.149	0	0.0106	3.321	0
$[15, 15]$	0.1655	-2.752	0	0.1597	-2.575	0
$[16, 16]$	0.0683	-1.874	0	0.0721	-1.986	0
$[17, 17]$	0	0	3.952	0	0	4.190
$[18, 18]$	0.1888	0.112	0	0.1873	-0.082	0

Therefore, we concluded that the self-consistent estimators had not yet converged to the MLEs. The apparent failure to converge seems to be due to a lack of information in the data. The estimated values did satisfy most of the Kuhn-Tucker conditions, though, so the fitted models could still be useful.

Chapter 5

Efficiency Gains in Survival Analyses

Judicious use of auxiliary data can lead to important gains in efficiency in the analysis of survival data, particularly if the event times are subject to heavy censoring.

Auxiliary variables are used in a variety of clinical settings including disease prevention studies, screening trials and randomized clinical trials (Fleming *et al.*, 1994). In a disease prevention trial, we could consider the immune responses created by a vaccination series as the auxiliary variable, where the endpoint of interest is prevention of HIV infection (Redfield *et al.*, 1991). For diseases such as cancer, we employ disease screening trials to provide early detection. Asymptomatic people are screened for cancer, and the presence of premalignant cellular changes may serve as auxiliary information in this setting (McPhee, 1995). A third use of auxiliary variables is in disease treatment trials. In a cancer treatment trial, for example, the auxiliary variable might be the recurrence of that form of cancer, and the endpoint is death from that form of cancer (Frank *et al.*, 1994). Thus, the effects of auxiliary data may yield valuable insights regarding treatment interventions and the natural history of disease.

It is important to make the distinction between auxiliary and surrogate variables in the context of survival data. Prentice (1989) defines an auxiliary variable to be a response variate that can provide additional information on whether an individual will have an extended or diminished length of survival. Therefore, unlike a surrogate variable, an auxiliary variate cannot substitute for a true endpoint for the purpose of hypothesis testing regarding treatment effects. We consider the case in which the auxiliary variables represent the occurrence of an intermediate event.

Lagakos (1977) adopted a parametric approach to examine the value of auxiliary variables for estimating the survival time distribution. The three event time distributions in the illness-death model are assumed to be from exponential models. The maximum likelihood estimates of the distribution function obtained by modelling only the survival data or both the survival and auxiliary data are obtained. Comparisons are made between the mean square errors (MSEs) of the two estimators of the distribution function at the three quartiles of the distribution function. Assuming no or very light (20%) right censoring, he found that incorporating auxiliary data led to considerable improvements in the estimation of the distribution of survival time. When there is no right censoring, the improvements decrease as the quartile being estimated increases. The opposite trend occurs in the light censoring scenario. Biases in the estimators were found to be small, so the observed differences in the MSEs are attributed to variability differences.

Fleming *et al.* (1994) developed augmented score and augmented likelihood methods that incorporated auxiliary data into a standard Cox regression analysis. In the augmented likelihood method, any unknown auxiliary variable information is replaced with an estimated value, then the score equations are obtained. Many forms of auxiliary variables are possible, including time-dependent ones. The methodology was illustrated with data from a colon cancer clinical trial. Using tumour recurrence as the auxiliary variable, the authors found only modest efficiency gains of 14.3 percent relative to a standard analysis using the augmented likelihood approach. In this particular trial, the auxiliary variable was nearly an intermediate variable, since only 11 of the 192 patients who died from the study total of 619 individuals did not show recurrence of the cancer.

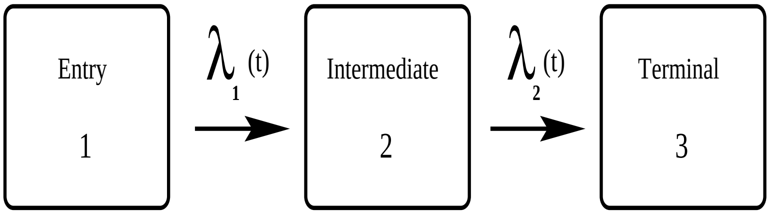
Finkelstein and Schoenfeld (1994) proposed nonparametric estimation methods within a three-state model framework. Their results indicate efficiency gains under a Markov assumption when the auxiliary event must occur before the terminal event, i.e., when the auxiliary variable is intermediate. Under a proportional hazards assumption, which related the time to progression and time from progression to death, no efficiency gains were found.

Cook and Lawless (2001) found substantial efficiency gains from using auxiliary data in a progressive three-state Markov model setting. In their simulation study, the sojourn times in states 1 and 2 were assumed to follow exponential distributions, and the amount of right censoring ranged from 0% to 90%. They calculated the empirical relative efficiency of the estimate of the survivor function for the time of entry into the third state without information on the sojourn time in state 1 (auxiliary information) to the estimate which incorporates this auxiliary information. Utilizing the auxiliary information led to substantial efficiency gains for percentiles in the left tail of the survivor function when the amount of right censoring was light, while even greater efficiency gains for percentiles in the right tail were found for heavier amounts of right censoring.

Often, more than one time scale is of interest in survival analysis. Since combining two or more time scales into one can be difficult, a common approach incorporates one of the time scales into the model as a time-dependent covariate, while the other time scale becomes the baseline measurement scale. This we term a hybrid time scale (Farewell and Cox, 1979; Oakes, 1995). In a three-state setting, the time on trial could be the baseline measurement scale, and the duration in the first state could be modelled as a time-dependent covariate. If we further allow interval censoring for the time to the intermediate state in a three-state model, then several studies have considered either a semi-Markov (e.g. De Gruttola & Lagakos, 1989) or Markov (e.g. Frydman, 1992) assumption in this setting. Frydman (1995), however, found both time scales were important in the context of the application considered by these authors.

In this chapter we examine the efficiency gains realized in estimating the survival function, when information about an interval-censored intermediate auxiliary variable is incorporated into a progressive three-state model. We adopt both the Markov and the hybrid time scale frameworks, assuming exponential distributions for the sojourn times in the initial and intermediate states. The effects of interval

Figure 5.1: Three state transition model.



censoring of the time to the intermediate event (state two), as well as possible right censoring on the time to the event of interest are examined.

5.1 Model Properties and Simulation Study

In this section we describe the features of the model (see Figure 5.1), as well as the maximum likelihood estimation (MLE) of the parameters. In a three-state model, let X represent the time from study entry until progression and V represent the time from progression to the terminal event. We specify the distribution of V conditionally on X, to accommodate a dependence in these event times. Then T, the overall time to the terminal event, is just equal to the sum of the two random variables, X and V. In this context, X represents the transition time to the intermediate state and so serves as an auxiliary variable for the purpose of making inferences about the distribution of T. We assume that each random variable follows an exponential distribution; the hazard rate for X is λ_1 and for $T - X \mid X$ is either λ_2 (Markov model) or $\lambda_2 \exp(\beta X)$ (hybrid model).

We will consider three separate cases, depending on the amount of information known about the auxiliary variable X: X is completely unknown (Case I), X is interval censored (Case II) and X is known exactly (Case III). Right censoring of T and X is possible in all three cases as well.

5.2 Markov Model

In the Markov time scale framework, the time to the terminal event only depends on the chronological time scale. Hence, the second hazard function is simply defined as λ_2, which may or may not be equal to the first hazard rate. The density functions corresponding to X and $T - X|X$ are denoted as $f_1(x;\lambda_1)$ and $f_2(t-x;\lambda_2,|\ x)$. We consider first the case when the hazard rates are unequal, then the special case when they are equal.

5.2.1 Unequal Hazard Rates

If we assume the hazard rates for the times to the intermediate and terminal events are different, that is, $\lambda_1 \neq \lambda_2$, and if we define the censoring time to be τ, and take the products, $\prod_u$ and $\prod_c$, to be over uncensored and censored subjects respectively, then the likelihood functions for the three cases are:

◇ Case I: X Completely Unknown

$$L(\theta) = \prod_u f(t_i;\theta) \prod_c \mathcal{F}(\tau;\theta) \tag{5.1}$$

where

$$\begin{aligned} f(t_i;\theta) &= \int_0^{t_i} f_1(x;\lambda_1)\ f_2(t_i - x;\lambda_2 \mid x)\,dx \\ &= \frac{\lambda_1\lambda_2}{\lambda_1-\lambda_2}\left[e^{-\lambda_2 t} - e^{-\lambda_1 t}\right]\ , \end{aligned}$$

and

$$\begin{aligned} \mathcal{F}(\tau;\theta) &= \int_\tau^\infty f(s;\theta)\ ds \\ &= \frac{1}{\lambda_1-\lambda_2}\left[\lambda_1 e^{-\lambda_2\tau} - \lambda_2 e^{-\lambda_1\tau}\right] \end{aligned}$$

is the survival function.

◇ Case II: X Interval Censored

$$L(\theta) = \prod_{j=1}^{m}\prod_{u} f^*(t_i;\theta)\ \prod_{c}\mathcal{F}^*(\tau;\theta) \tag{5.2}$$

where

$$\begin{aligned} f^*(t_i;\theta) &= \int_{I_j} f_1(x;\lambda_1) f_2(t_i - x;\lambda_2 \mid x)dx \\ &= \frac{\lambda_1\lambda_2 e^{-\lambda_2 t_i}}{\lambda_1-\lambda_2}\left[e^{-(\lambda_1-\lambda_2)L_j} - e^{-(\lambda_1-\lambda_2)U_j}\right]\ , \end{aligned}$$

$$\mathcal{F}^*(\tau;\theta) = \begin{cases} \int_\tau^\infty \int_{I_j} f_1(x;\lambda_1) f_2(\tau - x;\lambda_2 \mid x)dxds & \\ = \frac{\lambda_1 e^{-\lambda_2\tau}}{\lambda_1-\lambda_2}\left[e^{-(\lambda_1-\lambda_2)L_j} - e^{-(\lambda_1-\lambda_2)U_j}\right]\ , & X < \tau \\ e^{-\lambda_1\tau}\ , & X > \tau \end{cases}$$

and interval j is defined as $I_j = [L_j, U_j],\ j = 1, 2, \ldots, m$.

⋄ Case III: X Known

$$\begin{aligned} L(\theta) &= \prod_{u} f_1(x_i;\lambda_1)\, f_2(t_i - x_i;\lambda_2, \mid x_i) \prod_{c_2} f_1(x_i;\lambda_1)\mathcal{F}_2(\tau - x_i;\lambda_2 \mid x_i) \prod_{c_1} \mathcal{F}_1(\tau;\lambda_1) \\ &= \prod_{u} \lambda_1\lambda_2 e^{-(\lambda_1-\lambda_2)x_i-\lambda_2 t_i} \prod_{c_2} \lambda_1 e^{-(\lambda_1-\lambda_2)x_i-\lambda_2\tau} \prod_{c_1} e^{-\lambda_1\tau} \qquad (5.3) \end{aligned}$$

where the product subscripts c_2 and c_1 are taken over individuals censored in states 2 and 1, respectively.

In this Markov time scale model, the simple form of the second hazard rate resulted in closed form contributions to the likelihood for all three cases. The times to progression (X) and times from progression ($T - X$) to failure or the terminal event were generated from appropriate exponentially distributed random variables. One thousand samples of size 500 were simulated for each parameter configuration. The six parameter configurations were chosen to reflect plausible scenarios and to look for any trends. This was accomplished by keeping the second hazard rate constant, while varying the first hazard rate. In order to mimic the amount of right censoring seen in clinical trials, several censoring rates were chosen. These rates included heavy (50%), moderate (25%), and light (10%) right censoring levels.

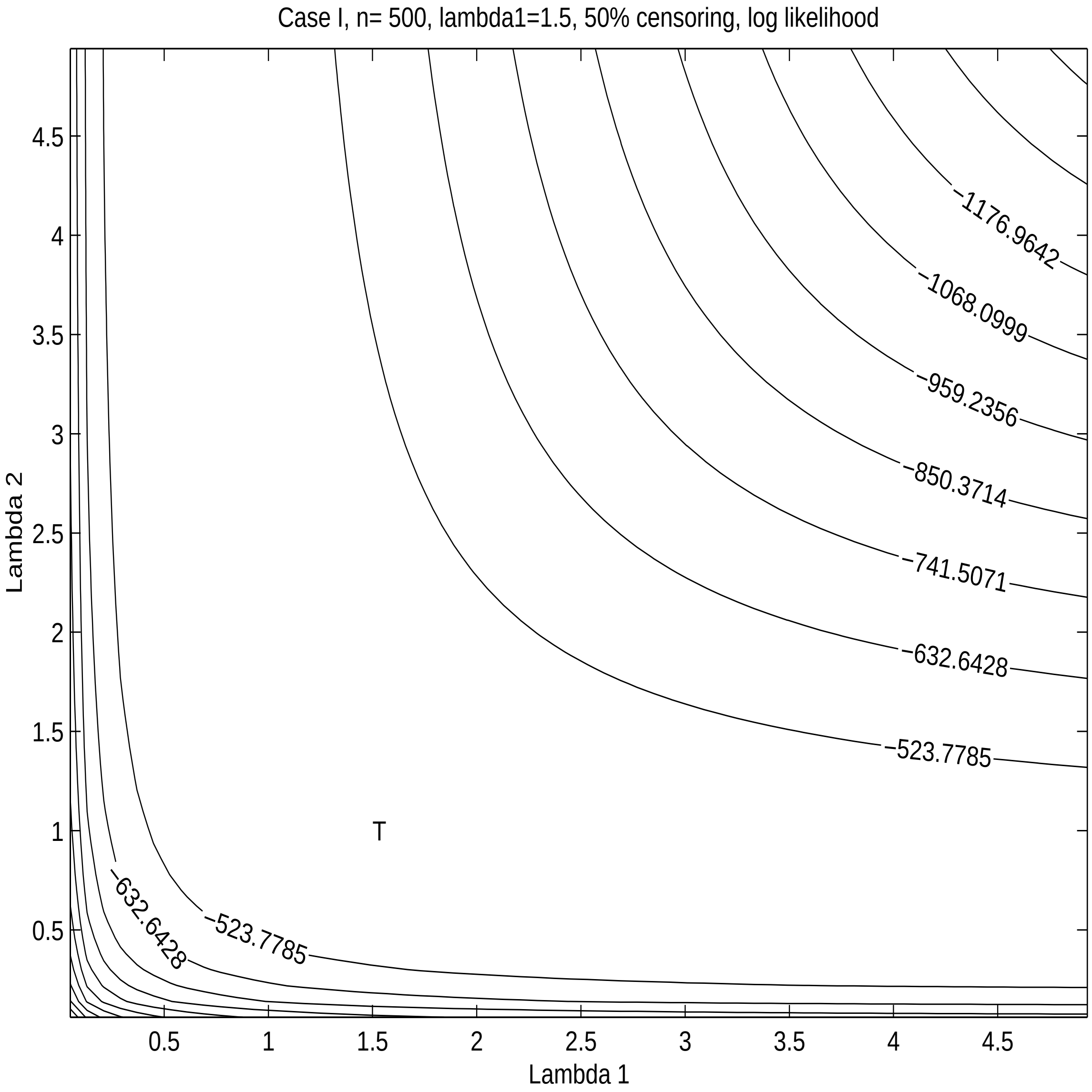
In each simulation run, the log likelihood for all three cases was maximized using the Matlab optimization function *constr*. This function employs a Sequential Quadratic Programming algorithm to find the MLEs in the nonlinear log likelihood functions, subject to the constraints that the rate parameters be positive. Examination of bias measures and contour plots revealed parameter estimation problems in the case where no information was available about the time to the first transition (Case I, equation (5.1)). In Figure 5.2, the large region of plausible values illustrates the numerical difficulties encountered when estimating λ_1 and λ_2. In the plot, 'T' represents the true value of the parameters. Hence, in the unequal hazard rates setting, the relative efficiency comparisons will only be made between cases II and III.

The results for comparing simulation variances of the estimated survivor function for T are given in Tables 5.1, 5.2, and 5.3, and Figures 5.3, 5.4, and 5.5. The estimated survivor functions were evaluated using the various MLEs and four specific percentiles of the distribution of T. These values, $\mathcal{F}(T;\theta) = 0.25, 0.50, 0.75$, and 0.95, were chosen to cover a range of possible values of the function of primary interest. Preliminary findings suggested high efficiency when only one observation on the state of the process occurred prior to the final observation. This intermediate observation was made halfway through the time on study (Table 5.1, Figure 5.3), and one-quarter (Table 5.2, Figure 5.4) and three-quarters of the way through the time on study (Table 5.3, Figure 5.5). The six parameter configurations used in the plots (top to bottom within each plot) are $(\lambda_1 = 1.75, \lambda_2 = 1), (\lambda_1 = 1.5, \lambda_2 = 1), (\lambda_1 = 1.25, \lambda_2 = 1), (\lambda_1 = 0.75, \lambda_2 = 1), (\lambda_1 = 0.5, \lambda_2 = 1), (\lambda_1 = 0.25, \lambda_2 = 1)$. The standard errors of the calculated survivor functions were derived from the sample variance, viz.

$$s.e.(\hat{\mathcal{F}}(t;\theta)) = \sqrt{\frac{1}{k-1}\sum_{i}^{k}(\hat{\mathcal{F}}_i(t;\hat{\theta}_i) - \bar{\hat{\mathcal{F}}}_{.}(t))^2} \tag{5.4}$$

where

Figure 5.2: Contour plot of the log likelihood of λ_1 and λ_2 illustrating parameter estimation problems in Case I. The symbol T marks the true value of λ_1 and λ_2.



$$\bar{\hat{\mathcal{F}}}_{\cdot}(t) = \frac{1}{k}\sum_{i}^{k} \hat{\mathcal{F}}_i(t;\hat{\theta}_i)$$

Table 5.1: Relative efficiency of Case II to Case III survivor function estimates under a Markov assumption, with an intermediate observation made at $\tau/2$.

		Censoring Rate		
True Parameter Values	Percentile of $\mathcal{F}$(T; θ)	10%	25%	50%
$\lambda_1 = 0.25, \lambda_2 = 1.0$	0.95	96.6%	98.0%	99.7%
	0.75	100.1%	99.9%	99.9%
	0.50	97.7%	99.6%	98.8%
	0.25	88.2%	95.7%	96.1%
$\lambda_1 = 0.50, \lambda_2 = 1.0$	0.95	99.8%	99.1%	100.0%
	0.75	99.7%	100.0%	99.4%
	0.50	97.2%	99.5%	98.7%
	0.25	94.0%	98.4%	98.1%
$\lambda_1 = 0.75, \lambda_2 = 1.0$	0.95	99.6%	100.0%	100.0%
	0.75	99.9%	99.7%	99.9%
	0.50	100.2%	99.0%	99.8%
	0.25	99.5%	98.3%	99.7%
$\lambda_1 = 1.25, \lambda_2 = 1.0$	0.95	98.9%	98.9%	98.5%
	0.75	99.4%	99.4%	99.2%
	0.50	99.6%	99.6%	99.5%
	0.25	99.7%	99.7%	99.6%
$\lambda_1 = 1.50, \lambda_2 = 1.0$	0.95	98.1%	99.4%	97.4%
	0.75	99.9%	100.2%	98.4%
	0.50	100.0%	100.1%	98.9%
	0.25	99.6%	99.9%	99.2%
$\lambda_1 = 1.75, \lambda_2 = 1.0$	0.95	97.5%	97.9%	96.8%
	0.75	99.9%	99.8%	98.3%
	0.50	99.6%	100.0%	99.0%
	0.25	98.4%	99.8%	99.4%

These results show there is very little loss in efficiency when assessing subjects

Figure 5.3: Histogram comparisons of the RE(case II, case III) of the survivor function estimates under a Markov assumption, when the rates are unequal, and the intermediate observation is made at $\tau/2$.

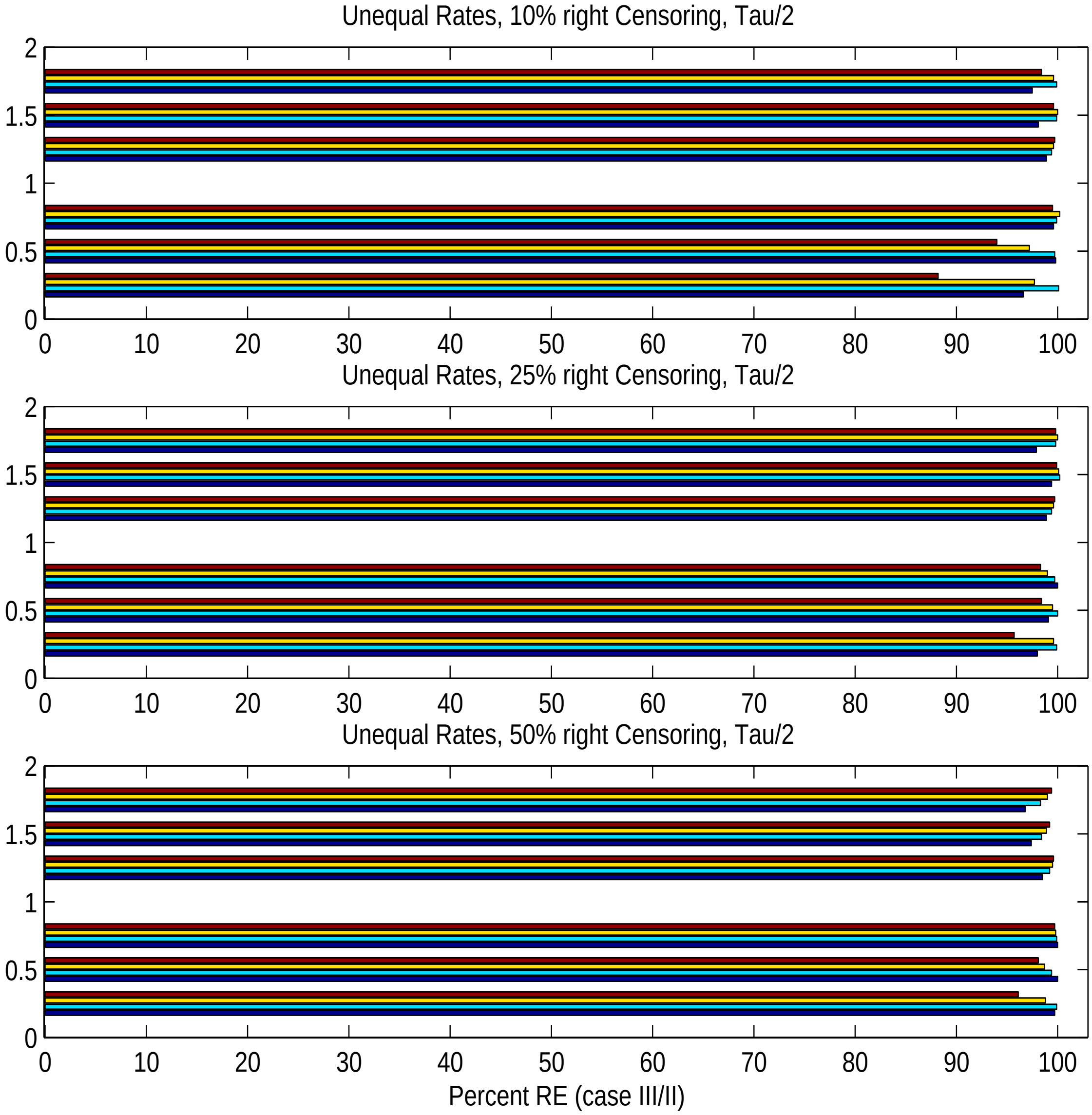


Table 5.2: Relative efficiency of Case II to Case III survivor function estimates under a Markov assumption, with an intermediate observation made at $\tau/4$.

		Censoring Rate		
True Parameter Values	Percentile of $\mathcal{F}(\mathrm{T};\,\theta)$	10%	25%	50%
$\lambda_1 = 0.25, \lambda_2 = 1.0$	0.95	97.9%	98.6%	99.4%
	0.75	100.1%	100.1%	100.0%
	0.50	97.6%	98.5%	98.7%
	0.25	89.6%	93.4%	95.0%
$\lambda_1 = 0.50, \lambda_2 = 1.0$	0.95	98.5%	99.4%	99.9%
	0.75	100.0%	99.9%	99.2%
	0.50	98.7%	98.7%	97.5%
	0.25	96.3%	96.9%	95.8%
$\lambda_1 = 0.75, \lambda_2 = 1.0$	0.95	100.1%	99.9%	99.1%
	0.75	99.6%	99.4%	98.2%
	0.50	99.0%	98.8%	97.5%
	0.25	98.4%	98.2%	96.9%
$\lambda_1 = 1.25, \lambda_2 = 1.0$	0.95	99.0%	99.0%	96.6%
	0.75	99.8%	99.6%	97.7%
	0.50	99.9%	99.8%	98.1%
	0.25	100.0%	99.9%	98.4%
$\lambda_1 = 1.50, \lambda_2 = 1.0$	0.95	100.1%	98.4%	98.2%
	0.75	100.2%	99.6%	99.2%
	0.50	99.8%	99.9%	99.5%
	0.25	99.3%	100.0%	99.7%
$\lambda_1 = 1.75, \lambda_2 = 1.0$	0.95	99.2%	98.5%	97.7%
	0.75	100.1%	99.6%	99.0%
	0.50	99.7%	100.0%	99.6%
	0.25	98.9%	100.1%	99.9%

Figure 5.4: Histogram comparisons of the RE(case II, case III) of the survivor function estimates under a Markov assumption, when the rates are unequal, and the intermediate observation is made at $\tau/4$.

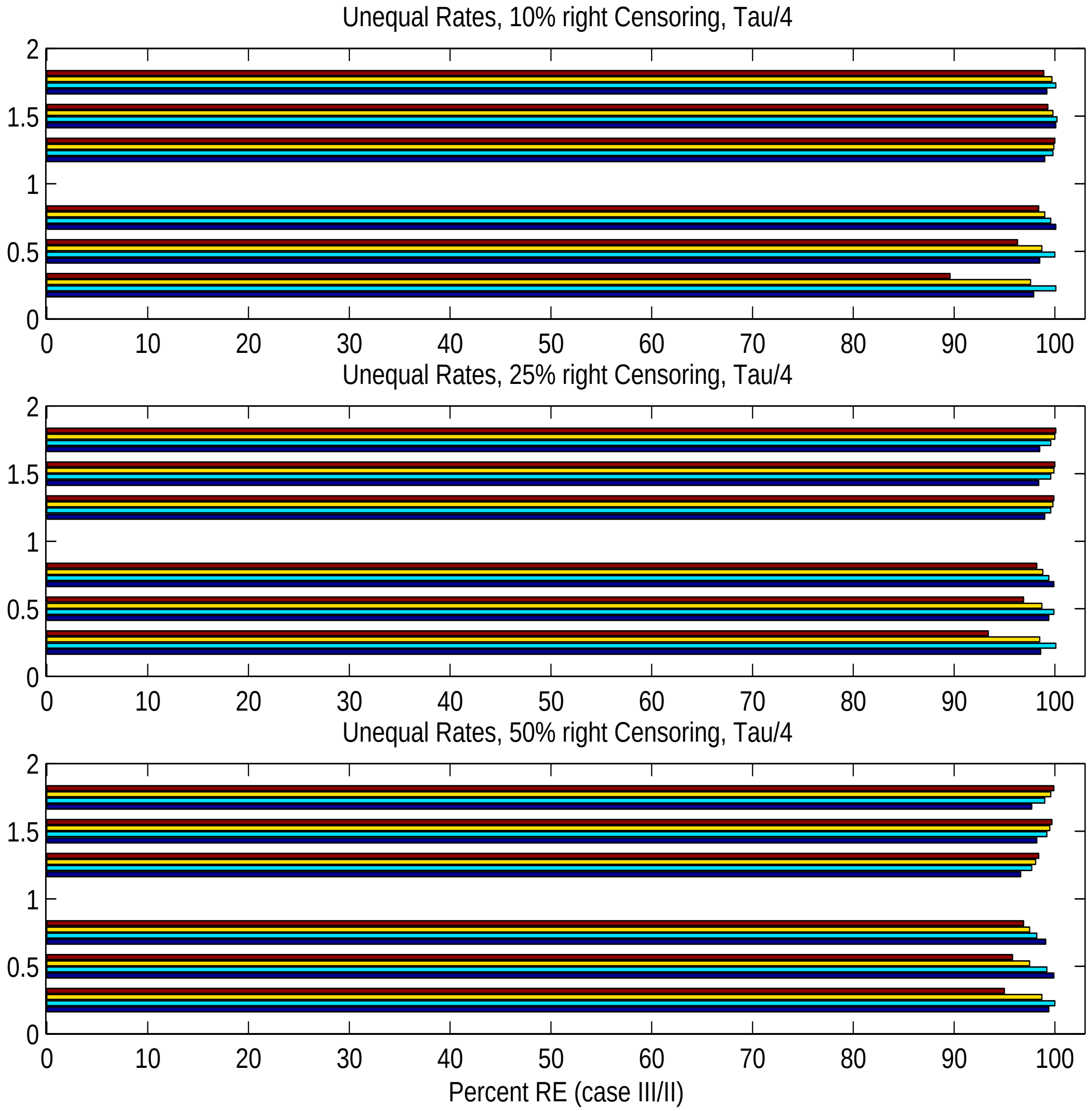
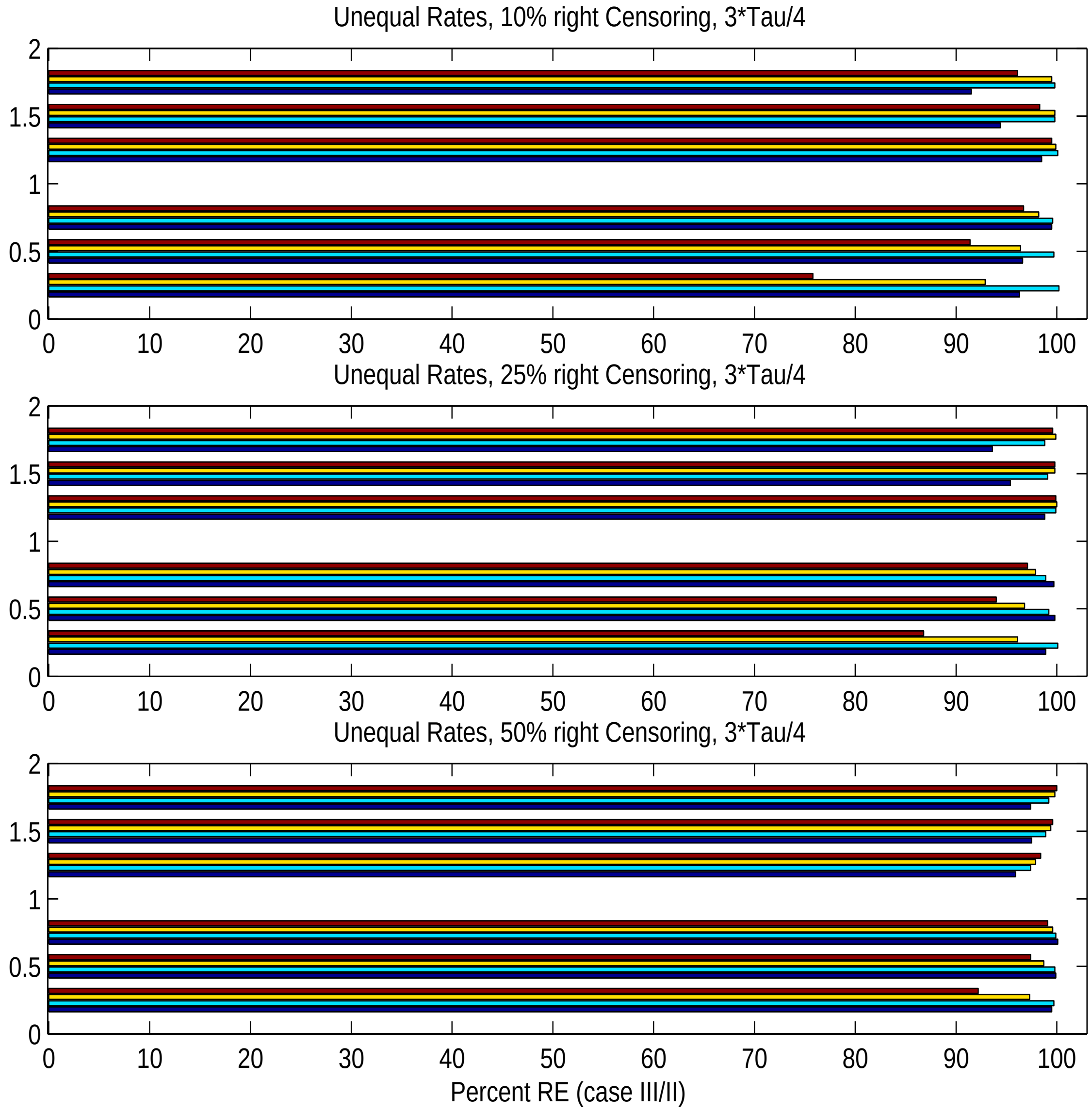


Table 5.3: Relative efficiency of Case II to Case III survivor function estimates under a Markov assumption, with an intermediate observation made at $3\tau/4$.

		Censoring Rate		
True Parameter Values	Percentile of $\mathcal{F}(\mathrm{T};\ \theta)$	10%	25%	50%
$\lambda_1 = 0.25, \lambda_2 = 1.0$	0.95	96.3%	98.9%	99.5%
	0.75	100.2%	100.1%	99.7%
	0.50	92.9%	96.1%	97.3%
	0.25	75.8%	86.8%	92.2%
$\lambda_1 = 0.50, \lambda_2 = 1.0$	0.95	96.6%	99.8%	99.9%
	0.75	99.7%	99.2%	99.8%
	0.50	96.4%	96.8%	98.7%
	0.25	91.4%	94.0%	97.4%
$\lambda_1 = 0.75, \lambda_2 = 1.0$	0.95	99.5%	99.7%	100.1%
	0.75	99.6%	98.9%	99.9%
	0.50	98.2%	97.9%	99.6%
	0.25	96.7%	97.1%	99.1%
$\lambda_1 = 1.25, \lambda_2 = 1.0$	0.95	98.5%	98.8%	95.9%
	0.75	100.1%	99.9%	97.4%
	0.50	99.9%	100.0%	97.9%
	0.25	99.5%	99.9%	98.4%
$\lambda_1 = 1.50, \lambda_2 = 1.0$	0.95	94.4%	95.4%	97.5%
	0.75	99.8%	99.1%	98.9%
	0.50	99.8%	99.8%	99.4%
	0.25	98.3%	99.8%	99.6%
$\lambda_1 = 1.75, \lambda_2 = 1.0$	0.95	91.5%	93.6%	97.4%
	0.75	99.8%	98.8%	99.2%
	0.50	99.5%	99.9%	99.8%
	0.25	96.1%	99.6%	100.0%

Figure 5.5: Histogram comparisons of the RE(case II, case III) of the survivor function estimates under a Markov assumption, when the rates are unequal, and the intermediate observation is made at $3\tau/4$.



at only two time points, rather than observing them over the entire time on study. It also does not seem to matter when the intermediate observation is made.

When the intermediate observation time is made halfway through the time on study, all of the relative efficiencies (REs) are 88.2% or larger, suggesting very high efficiency (see Table 5.1). In fact, 62 of the 72 reported REs are larger than 98%. There is a slight loss of efficiency, however, when the first hazard rate (λ_1) is the smallest (0.25) and the estimated survival function is at the lowest percentile (0.25). The RE values in the 50% right-censoring case range from 96.1% to 100.0%, in the 25% right-censoring case from 95.7% to 100.2%, and in the 10% right-censoring case from 88.2% to 100.2%. The average REs, calculated across all six parameter configurations, are 98.9%, 99.3% and 98.5%, for the 50%, 25% and 10% right-censoring cases respectively. It should be noted that the upper bound is 100% and the reported values larger than this reflect sampling variability.

When the intermediate observation time occurs earlier, at one-quarter of the time on study, similar findings result (see Table 5.2). All of the relative efficiencies (REs) are 89.6% or larger. In this scenario, 58 of the 72 reported REs are larger than 98%. As in the situation when the intermediate observation is made midway through the follow-up time, there is a slight loss of efficiency when the first hazard rate (λ_1) is the smallest (0.25) and the estimated survival function is at the lowest percentile (0.25). The RE values have similar ranges too: from 95.0% to 100.0% in the 50% right-censoring case, from 93.4% to 100.1% in the 25% right-censoring case, and from 89.6% to 100.1% in the 10% right-censoring case. The average RE values are also very similar: 98.4%, 99.0% and 98.8% for the 50%, 25% and 10%

right-censoring cases respectively.

When the intermediate observation time occurs later, at three-quarters of the time on study, the results show a slight loss of efficiency, compared to the previous two situations. All of the REs are 75.8% or larger, and now only 46 of the 72 reported REs are larger than 98% (see Table 5.3). However, 64 of the 72 reported REs are still larger than 95%, so the loss is minimal. As we observed previously, the greatest loss of efficiency occurred when λ_1 is the smallest (0.25) and the estimated survival function is at the lowest percentile (0.25). Additional losses in relative efficiency are seen for the next smallest value (0.50) of the first hazard rate, again when the estimated survival function is at the lowest percentile (0.25) and for the largest value (1.75) of the first hazard rate, but only when the estimated survival function is at the highest percentile (0.95). The RE values have wider ranges of values: 92.2% to 100.1% in the 50% right-censoring case, 86.8% to 100.1% in the 25% right-censoring case, and 75.8% to 100.2% in the 10% right-censoring case. The average RE values, although quite similar for the 50% and 25% right censoring cases (98.5%, 97.9%), are slightly lower for the 10% right censoring case (96.7%).

Several checks on the simulation study results were carried out. In case III, where complete information was available about the transition time to the intermediate state, X, closed form expressions for the MLEs of both parameters were derived. If we continue to define τ to be the censoring time, d_i to be the number of failures in state $i, i = 1, 2, 3$, and take the sums, $\sum_u$ and $\sum_{c_2}$, to be over uncensored and censored subjects in state 2 respectively, then the MLEs for case III are

$$\hat{\lambda}_1 = \frac{d_2 + d_3}{d_1\tau + \sum_{c_2,u} x_i}$$

$$\hat{\lambda}_2 = \frac{d_3}{d_2\tau + \sum_u t_i - \sum_{c_2,u} x_i} \quad .$$

These formulas can be interpreted much the same as in a parametric survival analysis when the time to failure follows an exponential distribution (Cox and Oakes, 1984). The expression for $\hat{\lambda}_1$ is just the total number of observed failures (departures from state 1) divided by the total time at risk for transition out of state 1. Similarly, the expression for $\hat{\lambda}_2$ is just the total number of failures (departures from state 2) divided by the total time at risk for transition out of state 2. The terms in the denominator adjust for the time taken to reach state 2 and, hence, are the true total time at risk in the intermediate state. The parameter estimates obtained from the maximized log likelihood differed only negligibly from these closed-form expression estimates. This close agreement indicated that the Matlab optimizing routine was working well in this case.

Examination of bias in the MLEs revealed no significant findings. Bias was assessed in several graphical and numerical ways. Graphical methods included histograms and box and whisker plots. Numerical methods included calculating the average bias, $\sum \frac{\hat{\theta}_i}{n_k} - \theta$, sample bias, $\sum \frac{(\hat{\theta}_i - \bar{\hat{\theta}}_.)^2}{n_k - 1}$, and one sample t statistic, $\frac{\bar{\hat{\theta}}_. - \theta}{\hat{\sigma}/\sqrt{n_k}}$. Since no evidence of bias was found in the MLEs, we concluded that the variance comparisons summarized in Tables 5.1-5.3 were appropriate.

5.2.2 Equal Hazard Rates

When the rate parameters are assumed to be equal in the Markov time scale framework, the marginal distribution for T follows a gamma distribution. This simplifying assumption also results in identical parameter information in the log likelihoods for the cases where information on X is known completely (Case III) or is interval censored (Case II). Hence, the relative efficiency of Case II to Case III is exactly one hundred percent. However, the parameter estimation problems for Case I (no information available about X) disappear, so it is now possible to examine the relative efficiency of complete or partial (interval-censored) information to no information about X (survival analysis case).

The likelihood contributions change because the estimation problem involves a single parameter. Without loss of generality, assume both transition rates are equal to λ. If we continue to define the censoring time to be τ, and take the products, $\prod_u$ and $\prod_c$, to be over uncensored and censored subjects respectively, then the likelihood functions for the three cases are:

⋄ Case I: X Completely Unknown

$$L(\theta) = \prod_u f(t_i;\theta) \prod_c \mathcal{F}(\tau;\theta) \tag{5.5}$$

where

$$f(t_i;\theta) = \int_0^{t_i} f_1(x;\lambda)\, f_2(t_i - x; \lambda, \mid x)\, dx \quad = \quad \lambda^2 t_i e^{-\lambda t_i} \ ,$$

and

$$\mathcal{F}(\tau;\theta) = \int_{\tau}^{\infty} f(s;\theta)\,ds \quad = \quad e^{-\lambda\tau}\,[\lambda\tau + 1]$$

is the survival function.

⋄ Case II: X Interval Censored

$$L(\theta) = \prod_{j=1}^{m}\prod_{u} f^{*}(t_i;\theta)\,\prod_{c}\mathcal{F}^{*}(\tau;\theta) \tag{5.6}$$

where

$$f^{*}(t_i;\theta) = \int_{I_j} f_1(x;\lambda) f_2(t_i - x;\lambda \mid x)\,dx \quad = \quad \lambda^2 e^{-\lambda t_i}\,[U_j - L_j]\ ,$$

$$\mathcal{F}^{*}(\tau;\theta) = \begin{cases} \int_{\tau}^{\infty}\int_{I_j} f_1(x;\lambda) f_2(\tau - x;\lambda \mid x)\,dx\,ds \quad = \quad \lambda e^{-\lambda\tau}\,[U_j - L_j]\ , \quad X < \tau \\ e^{-\lambda\tau}\ , \qquad X > \tau \end{cases}$$

and interval j is defined as $I_j = [L_j, U_j],\ j = 1, 2, \ldots, m$.

⋄ Case III: X Known

$$\begin{aligned} L(\theta) &= \prod_{u} f_1(x_i;\lambda)\,f_2(t_i - x_i;\lambda \mid x_i)\prod_{c_2} f_1(x_i;\lambda)\mathcal{F}_2(\tau - x_i;\lambda \mid x_i)\prod_{c_1}\mathcal{F}_1(\tau;\lambda) \\ &= \prod_{u}\lambda^2 e^{-\lambda t_i}\prod_{c_2}\lambda e^{-\lambda\tau}\prod_{c_1} e^{-\lambda\tau} \end{aligned} \tag{5.7}$$

where the product subscripts c_2 and c_1 are taken over individuals censored in states 2 and 1, respectively.

Once again, the simple form of the second hazard rate in this Markov time scale model resulted in closed-form contributions to the likelihood for all three cases. The times to progression (X) and times from progression ($T - X$) to failure or the terminal event were generated from identical exponentially distributed random variables. One thousand samples of size 500 were simulated for each parameter configuration. The eight parameter configurations included the same six values of λ_1 used in the unequal Markov case, as well two additional ones ($\lambda = 1, 2$). The same right-censoring rates (50%, 25%, and 10%) were also adopted. The Matlab optimization function *constr* was used to maximize the log likelihoods. Examination of bias measures and contour plots revealed no parameter estimation problems in Case I (see equation 5.5) this time. Hence, in the equal hazard rate setting, the relative efficiency comparisons will be made between cases I and II.

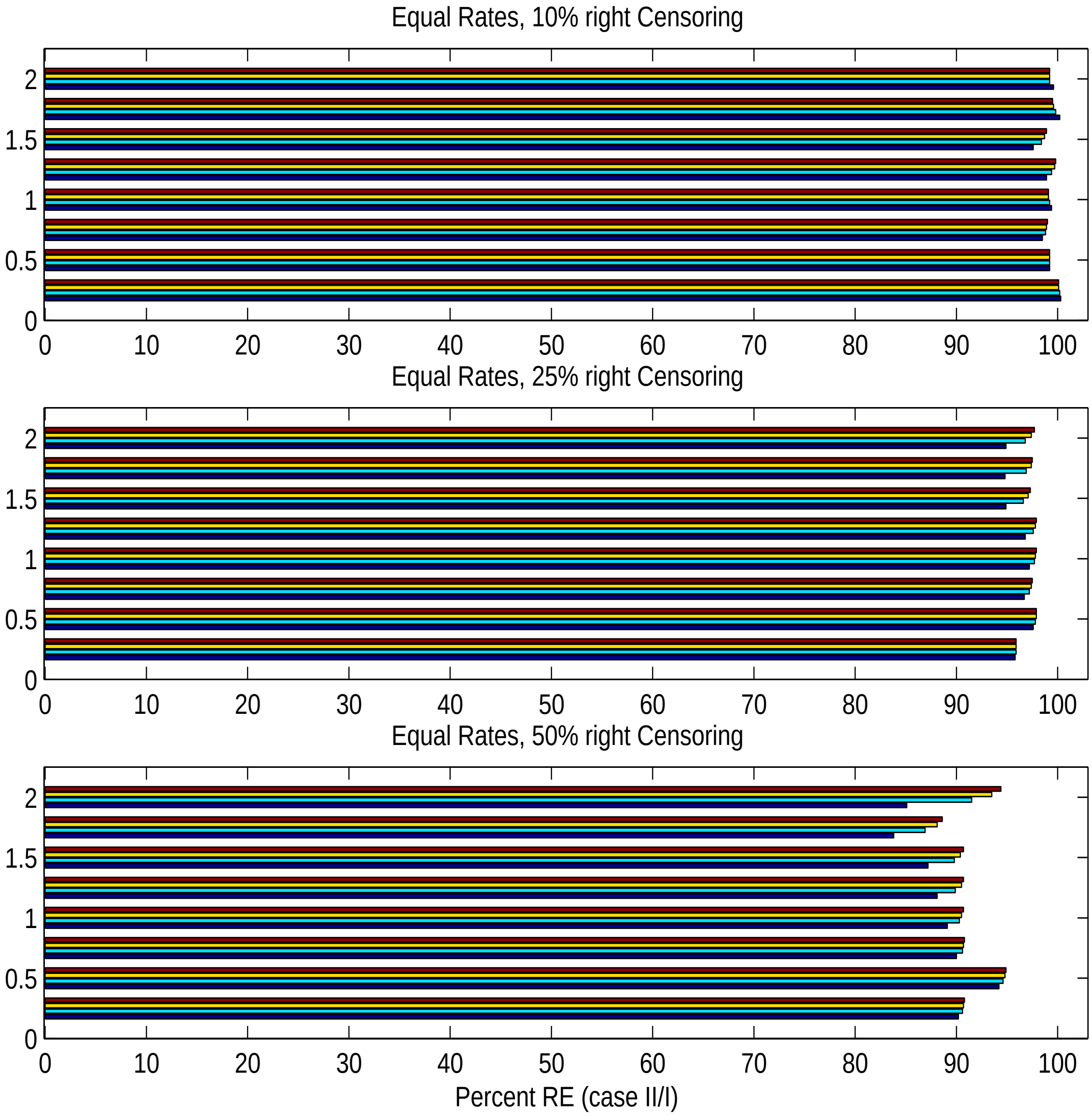
The results for comparing simulation variances of the estimated survivor function for T are given in Table 5.4 and Figure 5.6. The estimated survivor functions were evaluated using the various MLEs and the same four specific percentiles (0.25, 0.50, 0.75, and 0.95) of the distribution of T. Preliminary findings also suggested high efficiency when only one observation was made prior to the final observation. This intermediate observation was made halfway through the time on study. The eight parameter configurations used in the plot (top to bottom within each plot) are $(\lambda = 2), (\lambda = 1.75), (\lambda = 1.5), (\lambda = 1.25), (\lambda = 1), (\lambda = 0.75), (\lambda = 0.5), (\lambda = 0.25)$. The standard errors of the calculated survivor functions were derived from the sample variance, as specified in equation (5.4).

These results also show that there is generally very little loss in efficiency when the status of subjects is identified at only one time point, rather than observing

Table 5.4: Relative efficiency of Case I to Case II survivor function estimates under a Markov assumption, with an intermediate observation made at $\tau/2$.

		Censoring Rate		
True Parameter Values	Percentile of $\mathcal{F}(\mathrm{T};\,\theta)$	10%	25%	50%
$\lambda_1 = 0.25, \lambda_2 = 0.25$	0.95	100.3%	95.8%	90.2%
	0.75	100.2%	95.9%	90.6%
	0.50	100.1%	95.9%	90.7%
	0.25	100.1%	95.9%	90.8%
$\lambda_1 = 0.50, \lambda_2 = 0.50$	0.95	99.2%	97.6%	94.2%
	0.75	99.2%	97.8%	94.6%
	0.50	99.2%	97.9%	94.8%
	0.25	99.2%	97.9%	94.9%
$\lambda_1 = 0.75, \lambda_2 = 0.75$	0.95	98.5%	96.7%	90.0%
	0.75	98.8%	97.2%	90.6%
	0.50	98.9%	97.4%	90.7%
	0.25	99.0%	97.5%	90.8%
$\lambda_1 = 1.00, \lambda_2 = 1.00$	0.95	99.4%	97.2%	89.1%
	0.75	99.2%	97.7%	90.3%
	0.50	99.1%	97.8%	90.5%
	0.25	99.1%	97.9%	90.7%
$\lambda_1 = 1.25, \lambda_2 = 1.25$	0.95	98.9%	96.8%	88.1%
	0.75	99.4%	97.6%	89.9%
	0.50	99.7%	97.8%	90.5%
	0.25	99.8%	97.9%	90.7%
$\lambda_1 = 1.50, \lambda_2 = 1.50$	0.95	97.6%	94.9%	87.2%
	0.75	98.4%	96.6%	89.8%
	0.50	98.7%	97.1%	90.4%
	0.25	98.9%	97.3%	90.7%
$\lambda_1 = 1.75, \lambda_2 = 1.75$	0.95	100.2%	94.8%	83.8%
	0.75	99.8%	96.9%	86.9%
	0.50	99.6%	97.4%	88.1%
	0.25	99.5%	97.5%	88.6%
$\lambda_1 = 2.00, \lambda_2 = 2.00$	0.95	99.6%	94.9%	85.1%
	0.75	99.2%	96.8%	91.5%
	0.50	99.2%	97.4%	93.5%
	0.25	99.2%	97.7%	94.4%

Figure 5.6: Histogram comparisons of the RE(case I, case II) of the survivor function estimates under a Markov assumption, when the rates are equal, and the intermediate observation is made at $\tau/2$.



them over the entire time on study. Trends are more apparent, however, in this special case of equal hazard rates.

All of the relative efficiencies (REs) are 83.8% or larger, suggesting high efficiency (see Table 5.4). In fact, 31 of the 96 reported REs are larger than 98% and 61 of the 96 are larger than 95%. Within all eight parameter configurations, the 50% right censoring values are always smaller than the 25% right censoring values, which in turn are always smaller than the 10% right censoring values. Intuitively, this result makes sense; more complete information is available as the right-censoring rate decreases, since the follow-up time is potentially longer. This expected pattern is evident only in this special case, however.

The same trend is evident in the practically nonoverlapping ranges too: the 50% right censoring values range from 83.8% to 94.9%, the 25% right censoring values range from 94.8% to 97.9%, and the 10% right censoring values range from 97.6% to 100.3%. The average REs, calculated now across all eight parameter configurations, exhibit this trend as well. The averages are 87.9%, 96.9%, and 99.3% for the 50%, 25% and 10% right censoring cases respectively. Once again we note that the true upper bound for the RE is 100% and the reported values larger than this maximum reflect sampling variability.

Some other differences are apparent in this special case. When the right-censoring rate is high (50%), both the lowest overall efficiency (when $\lambda = 1.75$), and the greatest overall efficiency estimates (when $\lambda = 0.50$) were observed. The remaining values of λ (0.25, 0.75, 1.00, 1.25, 1.50, 2.00) have very similar estimated efficiencies and are bounded between these two extremes. There is also greater variability within the 50% right-censoring cases when the common rate parameter is large ($\lambda = 1.75, 2.00$). A slight loss of efficiency from the 75th percentile to the 95th percentile of T in the high (50%) or moderate (25%) right-censoring situations was noted.

As in the unequal hazards case, several checks on the simulation study results were carried out. A closed-form expression for the MLE of the single parameter in Case III, or equivalently Case II, was derived. If d_i is the number of failures in state $i, i = 1, 2, 3$ and the sum, $\sum_u$, is evaluated over uncensored subjects, then the MLE for λ in case III (II) is

$$\hat{\lambda} = \frac{d_2 + 2d_3}{\tau(d_1 + d_2) + \sum_u t_i} \ .$$

This expression also has a parametric survival analysis interpretation, if the time to failure follows an exponential distribution. The numerator represents the total number of observed failures (departures from state 1 and 2, and departures from state 2) while the denominator represents the total time at risk for transition to the last state.

As before, the parameter estimates obtained from the maximized log likelihood differed negligibly from the corresponding closed-form expression estimates. We concluded that the Matlab optimizing procedure was working well in this case too. Bias in the MLEs was assessed using the same graphical and numerical methods that we described previously in the unequal rate parameter setting; see §5.2.1. Since no bias was found in the MLEs, we again concluded that the variance comparisons summarized in Table 5.4 are appropriate.

5.3 Hybrid Time Scale Model

In the hybrid time scale framework, the time to the intermediate event is incorporated into the model by defining the hazard function governing transitions from state 2 to state 3 to be $\lambda_2 \exp(\beta X)$. The density functions corresponding to X and $T - X|X$ are denoted as $f_1(x; \lambda_1)$ and $f_2(t - x; \lambda_2, \beta \mid x)$.

As in the Markov time scale framework, if we define the censoring time to be τ, and take the products, $\prod_u$ and $\prod_c$, to be over uncensored and censored subjects respectively, then the likelihood functions for the three cases are:

◇ Case I: X Completely Unknown

$$L(\theta) = \prod_u f(t_i; \theta) \prod_c \mathcal{F}(\tau; \theta) \tag{5.8}$$

where

$$f(t_i; \theta) = \int_0^{t_i} f_1(x; \lambda_1)\, f_2(t_i - x; \lambda_2, \beta \mid x)\, dx \ ,$$

and

$$\mathcal{F}(\tau; \theta) = \int_\tau^\infty f(s; \theta)\, ds$$

is the survival function.

◇ Case II: X Interval Censored

$$L(\theta) = \prod_{j=1}^{m} \prod_u f^*(t_i; \theta) \prod_c \mathcal{F}^*(\tau; \theta) \tag{5.9}$$

where

$$f^*(t_i;\theta) = \int_{I_j} f_1(x;\lambda_1) f_2(t_i - x;\lambda_2,\beta \mid x)\, dx \ ,$$

$$\mathcal{F}^*(\tau;\theta) = \begin{cases} \int_\tau^\infty \int_{I_j} f_1(x;\lambda_1) f_2(\tau - x;\lambda_2,\beta \mid x)\, dx\, ds, & X < \tau \\ e^{-\lambda\tau} \ , & X > \tau \end{cases}$$

and interval j is defined as $I_j = [L_j, U_j],\ j = 1, 2, \ldots, m$.

◇ Case III: X Known

$$\begin{aligned} L(\theta) &= \prod_u f_1(x_i;\lambda_1)\, f_2(t_i - x_i;\lambda_2,\beta \mid x_i) \prod_{c_2} f_1(x_i;\lambda_1)\mathcal{F}_2(\tau - x_i;\lambda_2,\beta \mid x_i) \\ &\quad \times \prod_{c_1} \mathcal{F}_1(\tau;\lambda_1) \qquad (5.10) \\ &= \prod_u \lambda_1 e^{-\lambda_1 x_i} \lambda_2 e^{\beta x_i} e^{-(t_i - x_i)\lambda_2 e^{\beta x_i}} \prod_{c_2} \lambda_1 e^{-\lambda_1 x_i} e^{-(\tau - x_i)\lambda_2 e^{\beta x_i}} \prod_{c_1} e^{-\lambda_1 \tau} \end{aligned}$$

where the product subscripts c_2 and c_1 are taken over individuals censored in states 2 and 1, respectively.

Closed form expressions for the MLEs for cases I and II are not available, so numerical integration was employed to evaluate the likelihood contributions associated with individual subjects. The times to progression (X) and times from progression ($T - X$) to failure or the terminal event were generated from appropriate exponentially distributed random variables. Five hundred samples of size 200 were simulated for each parameter configuration. These modest values were chosen due to the computationally-intensive nature of the log-likelihood optimization procedures. The nine parameter configurations included some of the parameter values (0.5, 1, 2) used for λ_1 in the Markov time scale model, as well as the same parameter value (1) used for λ_2. Now, however, the second hazard rate varied as a function of the time spent in state one. The third parameter in this model, β, represents the effect of the duration in state one; the values of β were set at -0.5, 0, and 0.5. This approach enabled trend evaluations and also mimicked different types of disease processes. The same right-censoring rates adopted in the Markov time scale model, 50%, 25%, 10%, were chosen. A single intermediate observation occurred halfway through the time on study.

Not surprisingly, parameter estimation difficulties occurred in the case when no information was available about the time to the first transition (Case I, equation (5.8)). Therefore, only the estimated relative efficiencies for case II compared to case III are reported here. The efficiency of the estimated survivor function in case II relative to the corresponding estimate in case III was calculated via the ratio of the squared standard errors; see equation (5.4). The same four percentiles $(0.25, 0.50, 0.75, \text{and } 0.95)$ of the distribution of T were used to compare the MLEs of the estimated survivor functions.

The results reported in Table 5.5 and Figure 5.7 indicate some marked loss in efficiency when the status of subjects is identified at only two time points, rather than observing them over the entire time on study. However, this loss was generally evident only when the amount of right censoring was high (50%) or in both tails of the estimated survivor function of T $(25th, 95th$ percentiles). Trends in efficiency gains were noted as both λ_1 and $\lambda_2 \exp(\beta X)$ increased. As before, the two reported values that exceed the true upper bound of 100% reflect sampling variability.

Table 5.5: Relative efficiency of Case II to Case III survivor function estimates under a hybrid time scale assumption, with an intermediate observation made at $\tau/2$.

		Censoring Rate		
True Parameter Values	Percentile of $\mathcal{F}(\mathrm{T};\,\theta)$	10%	25%	50%
$\lambda_1 = 0.5, \lambda_2 = 1.0, \beta = -0.5$	0.95	84.7%	83.9%	63.0%
	0.75	93.4%	97.9%	68.9%
	0.50	73.6%	83.1%	91.6%
	0.25	59.9%	63.7%	80.7%
$\lambda_1 = 0.5, \lambda_2 = 1.0, \beta = \;\; 0.0$	0.95	75.9%	66.3%	63.4%
	0.75	96.2%	91.1%	76.2%
	0.50	99.9%	97.8%	94.3%
	0.25	88.1%	85.2%	93.2%
$\lambda_1 = 0.5, \lambda_2 = 1.0, \beta = \;\; 0.5$	0.95	95.9%	93.9%	80.9%
	0.75	96.0%	91.1%	89.5%
	0.50	98.3%	97.2%	95.2%
	0.25	86.0%	95.4%	95.6%
$\lambda_1 = 1.0, \lambda_2 = 1.0, \beta = -0.5$	0.95	77.1%	64.5%	61.3%
	0.75	97.3%	91.2%	77.4%
	0.50	80.7%	93.3%	97.4%
	0.25	68.1%	79.3%	86.9%
$\lambda_1 = 1.0, \lambda_2 = 1.0, \beta = \;\; 0.0$	0.95	75.9%	66.7%	71.8%
	0.75	98.6%	86.5%	83.6%
	0.50	94.3%	85.9%	98.1%
	0.25	84.1%	76.4%	94.9%

continued on next page

Table 5.5: *continued*

		Censoring Rate		
True Parameter Values	Percentile of $\mathcal{F}(\mathrm{T};\,\theta)$	10%	25%	50%
$\lambda_1 = 1.0, \lambda_2 = 1.0, \beta = 0.5$	0.95	82.5%	77.1%	61.0%
	0.75	96.3%	93.0%	83.6%
	0.50	99.2%	100.1%	98.6%
	0.25	91.1%	92.9%	96.3%
$\lambda_1 = 2.0, \lambda_2 = 1.0, \beta = -0.5$	0.95	80.4%	63.3%	64.1%
	0.75	96.2%	96.8%	80.9%
	0.50	84.2%	95.2%	97.8%
	0.25	69.0%	76.9%	93.8%
$\lambda_1 = 2.0, \lambda_2 = 1.0, \beta = 0.0$	0.95	86.7%	73.0%	68.9%
	0.75	100.4%	96.7%	84.8%
	0.50	91.0%	95.7%	98.4%
	0.25	78.0%	83.4%	92.4%
$\lambda_1 = 2.0, \lambda_2 = 1.0, \beta = 0.5$	0.95	88.6%	80.1%	65.9%
	0.75	96.9%	94.7%	83.4%
	0.50	95.9%	99.9%	97.1%
	0.25	87.3%	93.1%	99.0%

All of the relative efficiencies (REs) are 59.9% or larger, suggesting moderate to high efficiency. Of the 108 reported REs, 16 were between 59.9% and 69.9%, 13 were between 70.0% and 79.9%, 27 were between 80.0% and 89.9% and 52 were larger than 90.0%. Relative efficiences generally increased as the first hazard rate (λ_1) increased from 0.5 to 1 to 2. Similarly, REs increased as the second hazard rate, $\lambda_2 \exp(\beta X)$, increased. Since the second hazard rate varied as a function of the parameter β alone, when it increased from - 0.5 to 0 to 0.5, the REs also generally increased.

The nine parameter configurations used in Figure 5.7 (top to bottom within each subplot) are

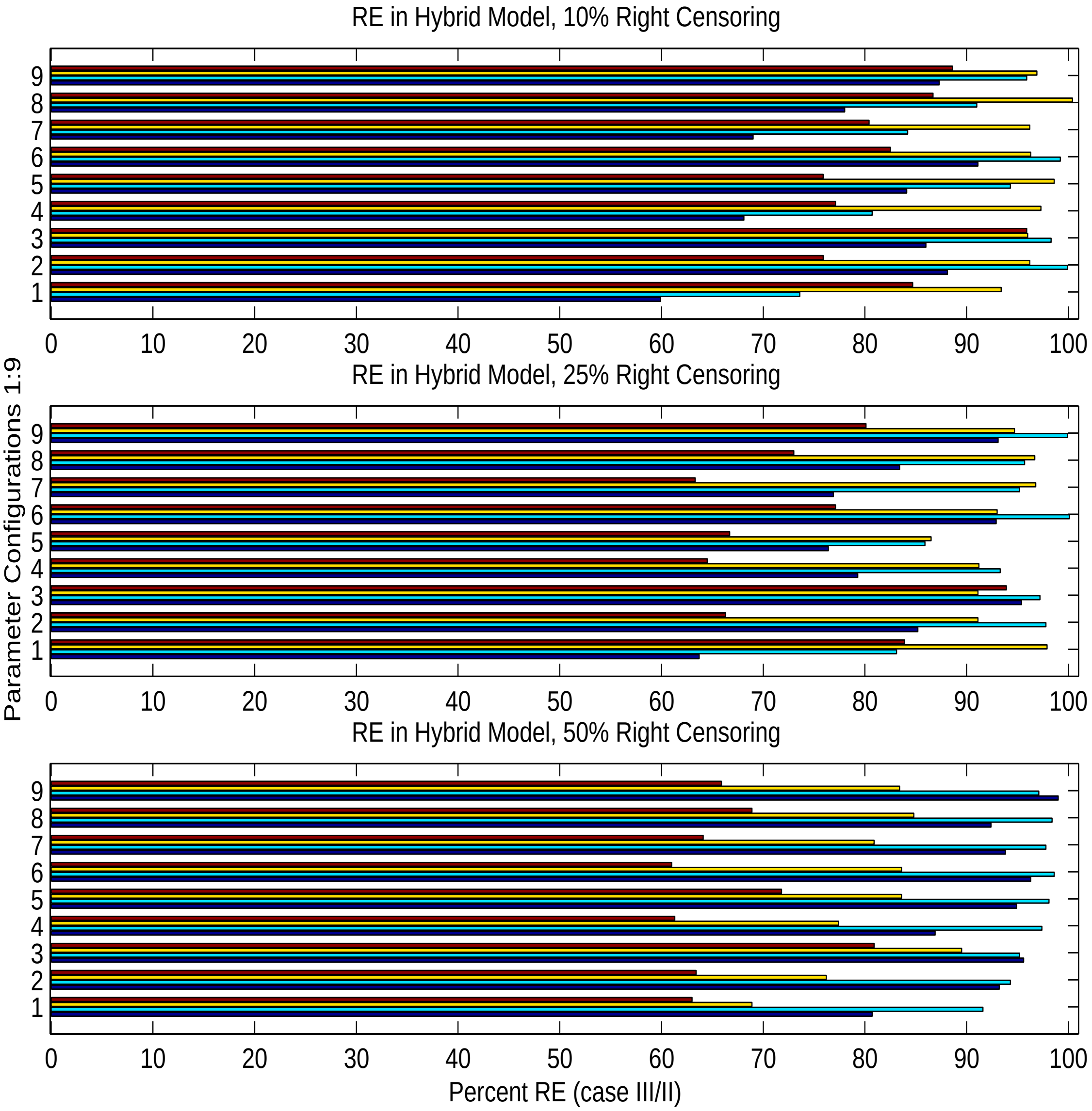
Table 5.6: Parameter configurations used in Figure 5.7.

Label	λ_1	λ_2	β
9	2.0	1	0.5
8	2.0	1	0.0
7	2.0	1	-0.5
6	1.0	1	0.5
5	1.0	1	0.0
4	1.0	1	-0.5
3	0.5	1	0.5
2	0.5	1	0.0
1	0.5	1	-0.5

When the parameter β equals zero in this hybrid time scale model, the resulting model is equivalent to the Markov time scale model considered §5.2. Comparing the results between Table 5.5 and Table 5.1 when $\lambda_1 = 0.5$, there is considerable loss of efficiency in the more complex model at the 25th and 95th percentiles of the T distribution. Similar losses of efficiency are apparent when $\lambda_1 = 1$ and $\lambda_2 = 1$, because in the equal hazards setting, the relative efficiency of Case II to Case III is 100%. The values in Table 5.5 suggest a sizable loss of efficiency, particularly in the tails of the distribution of T.

The same graphical and numerical methods, described in §5.2.1, were used to examine potential bias in the MLEs. Since no significant findings were revealed, we concluded that the variance comparisons summarized in Table 5.5 are appropriate.

Figure 5.7: Histogram comparisons of the RE(case II, case III) of the survivor function estimates under a hybrid time scale assumption, when the intermediate observation is made at $\tau/2$.



5.4 Discussion

Incorporating information about an auxiliary event can lead to more efficient estimation of the survivor function and associated parameters. In the standard survival analysis setting, which is equivalent to a two-state model, and assuming a proportional hazards model, Fleming *et al.* (1994) did not find substantial efficiency gains. Finkelstein and Schoenfeld (1994), found only modest efficiency gains in the three-state progressive model when information about an intermediate event was incorporated into the model. These gains occurred only under the Markov assumption for their nonparametric estimator of the survivor function. Lagakos (1977) found considerable gains in the illness-death model when the MSEs were compared, if an auxiliary variable was incorporated into the estimation of the survival distribution. These gains were attributed to variability differences.

Our results suggest that in this three-state model framework, subjects who are observed at just two time points provide essentially the same information about the survivor function as those for whom the time to the intermediate state is observed precisely. These observations pertain to the Markov time scale framework, and to some extent in the hybrid time scale framework. Greater loss of efficiency occurred in the hybrid time scale scenario, although not for all parameter configurations.

In the Markov model, assuming unequal hazard rates, the location of the intermediate observation does not seem to matter. Whether the observation is made at one-quarter, one-half or three-quarters of the planned length of follow up, the REs are very high. The only exception occurs for the 95th percentile of the T distribution, where efficiencies tend to decrease slightly. When equal hazard rates are assumed, REs increase as the right censoring probabilities decrease. This trend is evident across all parameter configurations. Even with heavy right censoring, most of the REs are still quite high. Hence, in the Markov model, all of the parameter configurations studied led to high efficiency gains.

In the hybrid model, more variability in the observed REs was evident. The greatest loss of efficiency occurs in the right tail of the distribution of T generally, as well as for many parameter configurations for the 25th percentile of T. Clear trends in the estimated efficiency were observed as the hazard rates increased. Relative efficiencies increased as λ_1 and $\lambda_2 \exp(\beta X)$ increased, the latter hazard rate only because β increased, not λ_2. Including the time duration in the first state is generally more efficient if the effect increases the hazard function. This effect did not seem to depend on the amount of right censoring. This may be important, for example, in clinical studies subject to heavy right censoring and where treatment interventions may increase the time spent in the initial state.

In future work, additional simulation studies could be carried out to examine a broader range of parameter values, more than one intermediate observation in the hybrid time scale model, larger sample sizes, and greater numbers of simulation runs. More flexible hazard functions, say from the gamma distribution, could be adopted for the sojourn times in both states. Tests to assess efficiency differences could also be developed.

Chapter 6

Future Work

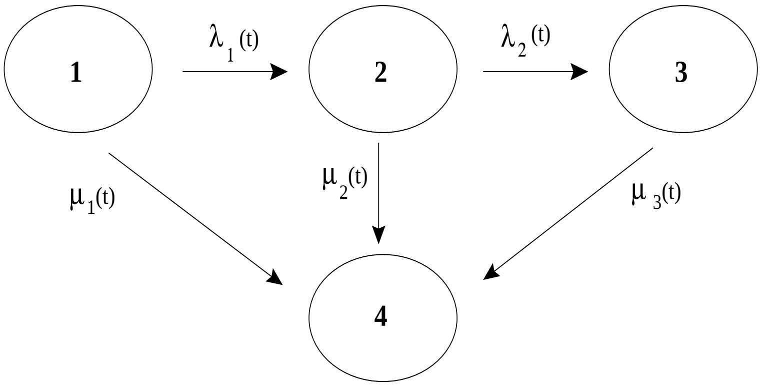
The future research ideas to be considered are organized into two general groups. In §6.1 we will describe various natural extensions, to more general models, of the work outlined in Chapter 2. We also outline some approaches to the difficult problem of model assessment that we propose to explore.

6.1 Extensions of the Basic Model

A natural extension of the three-state progressive model is to a similar framework involving $K > 3$ states. Some diseases or disorders may proceed through more than one intermediate state before progressing to the final, absorbing state. An example of a four-state progressive model is described in Gladman *et al.* (1998), where the number of joints damaged by chronic psoriatic arthritis also determines the patient's condition (state).

The simple illness-death model can be extended to a four-state model which retains the unidirectional features of the data, while allowing direct transitions to the absorbing state from all intermediate states. This type of four-state model is another extension which could be useful for modelling some disease processes where progression to some final outcome need not be through all the intermediate states.

Figure 6.1: Extended illness-death model.



The three-state model of chapter 2 can be extended to allow truncation as well as censoring. In longitudinal studies, such as those involving AIDS, reporting delays can result in right-truncated data. Left-truncated data can be the result of restricted study entry. Allowing for even more common types of missing data will generalize our current approach. Permitting different sampling schemes, other than random selection from a homogeneous population, would also increase the flexibility and utility of this multi-state model.

We also intend to explore the use of a random effect or “frailty” term in the model. If the amount of variation in the data is not satisfactorily explained by the measured covariates, a “frailty” component could capture the extra person-to-person variability. If we continue to adopt the same intensity model, the frailty term V is assumed to act multiplicatively on the "baseline" intensities

$$\alpha(x; z, v) = \frac{v\, \eta(x)\, e^{z'\gamma}}{1 + v\, \eta(x)\, e^{z'\gamma}} \ ,$$

$$h(x, t; z, v) = \frac{v\, \lambda(t) e^{\beta(t-x)+z'\nu}}{1 + v\, \lambda(t) e^{\beta(t-x)+z'\nu}} \ .$$

For simplicity we intend to make the frailty components time-invariant and subject-specific and will likely adopt a gamma distribution for each fraility component.

6.2 Model Assessment

If we adopt a modelling approach to the analysis of longitudinal data, rather than strictly nonparametric estimation, then we need to be able to evaluate several important model characteristics. Although no one model is true, there may be several competing models that seem to explain the data equally well. Thus, we need criteria with which to judge how close, in some distance metric, each model is to the data (Lindsey, 1996). Conventional approaches to model assessment focus on goodness-of-fit tests. These tools may involve tests of significance (for model parameters in a regression context), likelihood comparisons (comparing less complex nested models with more saturated versions), and diagnostic procedures to detect departures from the fitted models (identifying outliers and influential observations).

In chapter 3 we adopted a piecewise constant approach for modelling the baseline intensity functions between states 2 and 3. Although the asymptotic properties of these piecewise constant functions in a three-state model remain to be studied, our results were consistent with previous work adopting this approach. The regression parameter estimates seemed to be quite robust to the location of the breakpoints and the number of pieces used (Lawless and Zhan, 1998; Lindsey and Ryan, 1998). In a standard survival analysis, and assuming a multiplicative model for the hazard function, Friedman (1982) derived the asymptotic properties of the piecewise constant intensity functions. He recommended that the number of events be comparable within each interval of support. In our experience with the data types we considered, the parameter estimates may be consistent across models of varying complexity but the standard errors were not.

We used the Akaike and Schwarz Information Criteria to determine the complexity of the final model. Using the quantiles of the nonparametrically estimated cumulative distribution functions led to more stable parameter estimation, but did not always lead to smaller standard errors. To improve upon the choice of a final model, we would like to explore the inclusion in the SIC of terms identified as asymptotically negligible in its derivation as a transformation of the Bayesian posterior probability of a candidate model. Neath and Cavanaugh (1997) show how the inclusion of the observed data information matrix can improve upon the performance of the SIC in small to moderate data sets when applied to multiple linear regression and time series analysis applications.

They also consider the use of a nonuniform prior distribution for model selection. In our three-state model, knowledge of a change in the underlying hazard could suggest the location of a breakpoint in a piecewise constant hazard. For example, the introduction of a new treatment during the course of a study could have the effect of altering the hazard function. Including this term in small data set applications could improve upon the choice of a final model.

The derivation of standard errors for the estimated regression parameters in the transition intensity models could be found using methods other than the piecewise constant approach. Frydman (1995) suggested establishing the validity of the bootstrap for finding the properties of the estimators she derived. However, given the complexity and number of estimating equations for finding the MLE of the parameters in the extensions considered thus far, even if resampling techniques such as the jackknife or bootstrap methods are found to be valid, they may not be practical. One solution may involve extending the decomposition approach of Gu (1996) to the problem of evaluating the observed information matrix. Louis (1982) described a method for finding the observed information matrix, but his approach requires the complete-data gradient vector or the second derivative matrix. Perhaps the construction of the complete data can once again be avoided, and the observed information matrix extracted from the incomplete data, using only the mathematical form of the observed likelihood.

Assessing the logistic link function was the primary focus of chapter 4. Using a family of link functions theoretically allowed us to fit the complementary log-log model, but was not successfully implemented in practice. Although the logistic link function appears to be an appropriate choice for many common types of discrete survival data, it would still be preferable to consider other link functions besides the logit for parametrizing the hazard function. We did not parametrize the hazard function for transitions from state 1 to state 2 when we assessed the logit function. Simultaneously assessing the goodness-of-link functions would be an important aspect of model assessment when the semi-parametric form for the hazard function is used between the first two states.

We verified in chapter 4 that the self-consistent estimators for the time to infection distribution from our three-state model were also the maximum likelihood estimators for the AIDS data set. However, convergence to the MLEs for all of the parameters was not established. Upon further investigation, it seemed as though the amount of interval censoring in this data set affected this convergence. Thus, we would like to pursue additional checks for data that would assess the impact of interval-censoring characteristics on the estimation.

Lastly, we will investigate the efficiency of incorporating intermediate information into a survival analysis problem by assuming different forms of the hazard function than the characterizations considered thus far. Semi-parametric and regression models will be of interest too.

Three-state models are useful for modelling survival data, when a process is unidirectional and information about an intermediate event is available. By modelling the sojourn times associated with intermediate states, we may glean useful insights concerning the natural history of the disease process that cannot be detected by a survival analysis alone. Using a hybrid time scale model in this approach can permit estimation of the effects of a new treatment on a disease process. This is important for conditions such as AIDS, where new treatments are changing the natural history of the disease (Frydman, 1995).

Appendix A

Additional Results from Chapter 3

In chapter 3 we fit various piecewise constant models using five simulated data sets. Piecewise constant models which minimized either the AIC or SIC for each data set in the three-state progressive process were identified in two different approaches. In the Initial model approach, the quantiles from the nonparametric estimates of the CDFs were used to select the breakpoints for each specified interval of support for the transition intensity functions. In Table A.1, we report the estimated parameter and standard error values for these models. The data set labels, A through E, used in the first column of Tables A.1 and A.2, are defined as

- A: data set with single intervals of support for each transition intensity, and a "narrow" amount of interval censoring
- B: data set with single intervals of support for each transition intensity, and

a "wide" amount of interval censoring

- C: data set with two intervals of support for both transition intensities
- D: data set with a single interval of support for the transition intensity between states 1 and 2, and two intervals of support for the transition intensity between states 2 and 3
- E: data set with three intervals of support for both transition intensities

When there is only a single row associated with a data set label, then the SIC and AIC values were minimized with the same model. Otherwise, the model criterion is specified in addition to the data set label in the first column. The entries in the second column, Model Fit, refer to the number of pieces used in each transition intensity for that piecewise constant model. For example, the first row of Table A.1 records the results for data set A, the data set which had a single underlying interval of support for the both transition intensities, and the least amount of interval censoring. Both the AIC and SIC were minimized in a model with two intervals of support (H = 2) for the transition intensity out of state 1 and a single interval of support (I = 1) for the transition intensity out of state 2. The results from fitting that particular piecewise constant model using data set A are recorded in the remaining entries in the first row.

In the Best model approach, the breakpoints are selected after iterating above and below the cutpoints identified in the Initial model approach. In Table A.2 we report the parameter and standard error estimates obtained for the Best models identified for each data set. The interpretation of the entries/labels in the first two

Table A.1: Parameter and standard error estimates from all Initial models.

Data	Model	Parameter Estimates						
Set	Fit	(Standard Error Estimates)						
	H, I	η_1	η_2	η_3	λ_1	λ_2	λ_3	β
A	2, 1	0.367	0.132		0.021			0.004
		(0.953)	(0.474)		(0.117)			(0.505)
B	3, 1	0.418	0.053	0.186	0.034			-0.026
		(1.481)	(0.803)	(1.059)	(0.172)			(0.447)
C,	3, 3	0.176	0.055	0.199	0.032	0.165	0.138	-0.185
AIC		(0.531)	(0.272)	(0.816)	(0.292)	(1.305)	(1.250)	(0.869)
C,	3, 1	0.176	0.056	0.199	0.066			-0.117
SIC		(0.466)	(0.279)	(0.760)	(0.355)			(0.728)
D	1, 1	0.105			0.081			-0.167
		(0.202)			(0.467)			(0.860)
E	3, 1	0.171	0.049	0.384	0.117			-0.174
		(0.544)	(0.202)	(1.747)	(0.604)			(0.774)

Table A.2: Parameter and standard error estimates from all Best models.

Data	Model	Parameter Estimates						
Set	Fit	(Standard Error Estimates)						
	(H, I)	η_1	η_2	η_3	λ_1	λ_2	λ_3	β
A	3, 1	0.360	0.052	0.201	0.021			0.003
		(0.913)	(0.309)	(1.128)	(0.114)			(0.457)
B,	3, 1	0.378	0.093	0.198	0.034			-0.028
AIC		(0.875)	(0.466)	(1.162)	(0.185)			(0.462)
B,	2, 1	0.347	0.142		0.035			-0.028
SIC		(1.111)	(0.408)		(0.181)			(0.489)
C	3, 2	0.163	0.048	0.345	0.032	0.147		-0.181
		(0.445)	(0.229)	(2.03)	(0.276)	(0.995)		(0.829)
D,	3, 2	0.150	0.074	0.164	0.061	0.129		-0.198
AIC		(0.609)	(0.225)	(0.624)	(0.439)	(1.004)		(0.994)
D,	3, 1	0.151	0.074	0.164	0.083			-0.170
SIC		(0.608)	(0.205)	(0.694)	(0.507)			(0.999)
E	3, 1	0.156	0.038	0.316	0.115			-0.171
		(0.431)	(0.168)	(1.281)	(0.576)			(0.804)

columns of the table are the same as in Table A.1.

As we noted in chapter 3, the estimated standard errors are all larger than the corresponding parameter estimates. Hence, none of the z statistics — each statistic formed from the ratio of a parameter estimate to its associated estimated standard error — would be large enough to reject the hypothesis that the parameter was different than zero.

References

Aalen, O. O., Farewell, V. T., De Angelis, D., Day, N. E., and Gill, O. N. (1997). A Markov model for HIV disease progression including the effect of HIV diagnosis and treatment: Application to AIDS prediction in England and Wales. *Statistics in Medicine*, 16:2191–2210.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. N. and Csaki, F., editors, *Second International Symposium on Information Theory*, pages 267–281. Budapest, Hungary: Akademia Kiado.

Andersen, P. K. (1988). Multistate models in survival analysis: A study of nephropathy and mortality in diabetes. *Statistics in Medicine*, 7:661–670.

Blossfeld, H.-P., Hamerle, A., and Mayer, K. U. (1989). *Event History Analysis: Statistical Theory and Application in the Social Sciences.* Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Cheng, K. F. and Wu, J. W. (1994). Testing goodness of fit for a parametric family of link functions. *Journal of the American Statistical Association*, 89:657–664.

Cook, R. J. and Lawless, J. F. (2001). Some comments on efficiency gains from auxiliary information for right-censored data. *Journal of Statistical Planning and Inference* (to appear).

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* London: Chapman and Hall.

Czado, C. (1997). On selecting parametric link transformation families in generalized linear models. *Journal of Statistical Planning and Inference*, 61:125–139.

De Gruttola, V. and Lagakos, S. W. (1989). Analysis of doubly-censored survival data, with application to AIDS. *Biometrics*, 45:1–11.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38.

Diggle, P. J., Liang, K.-Y., and Zeger, S. L. (1994). *Analysis of Longitudinal Data.* Toronto: Oxford University Press.

Farewell, V. T. and Cox, D. R. (1979). A note on multiple time scales in life testing. *Applied Statistics*, 28:73–75.

Finkelstein, D. M. and Schoenfeld, D. A. (1994). Analyzing survival in the presence of an auxiliary variable. *Statistics in Medicine*, 13:1747–1754.

Fitzmaurice, G. M., Laird, N. M., and Rotnitzky, A. G. (1993). Regression models for discrete longitudinal responses. *Statistical Science*, 8:284–299.

Fleming, T. R., Prentice, R. L., Pepe, M. S., and Glidden, D. (1994). Surrogate and auxiliary endpoints in clinical trials with potential applications in cancer and AIDS research. *Statistics in Medicine*, 13:955–968.

Friedman, M. (1982). Piecewise exponential models for survival data with covariates. *Annals of Statistics*, 10:101–113.

Frydman, H. (1992). A nonparametric estimation procedure for a periodically observed three-state Markov process, with application to AIDS. *Journal of the Royal Statistical Society, Series B*, 54:853–866.

Frydman, H. (1995). Semiparametric estimation in a three-state duration dependent Markov model from interval-censored observations with application to AIDS data. *Biometrics*, 51:502–511.

Gentleman, R. and Geyer, C. J. (1994). Maximum likelihood for interval censored data: Consistency and computation. *Biometrika*, 81:618–623.

Gladman, D. D., Farewell, V. T., Kopciuk, K. A., and Cook, R. J. (1998). HLA markers and progression in psoriatic arthritis. *Journal of Rheumatology*, 25:730–733.

Goldstein, H. (1979). *The Design and Analysis of Longitudinal Studies: Their Role in the Measurement of Change.* London, Academic Press.

Grüger, J., Kay, R., and Schumacher, M. (1991). The validity of inferences based on incomplete observations in disease state models. *Biometrics*, 47:595–605.

Gu, X. (1996). *Statistical Analysis of Incomplete Data arising in Biomedical Studies.* PhD thesis, University of Waterloo.

Kay, R. (1986). A Markov model for analysing cancer markers and disease states in survival studies. *Biometrics*, 42:855–865.

Kim, M. Y., De Gruttola, V. G., and Lagakos, S. W. (1993). Analyzing doubly censored data with covariates, with application to AIDS. *Biometrics*, 49:13–22.

Klein, J. P. and Qian, C. (1996). Modeling multistate survival illustrated in bone marrow transplantation. In *ASA Proceedings of the Biometrics Section*, pages 93–102.

Kosorok, M. R. and Chao, W.-H. (1996). The analysis of longitudinal ordinal response data in continuous time. *Journal of the American Statistical Association*, 91:807–817.

Lagakos, S. W. (1977). Using auxiliary variables for improved estimates of survival time. *Biometrics*, 33:399–404.

Laird, N. M. (1988). Self-consistency. In *Encyclopedia of Statistical Sciences (Vol. 8)*, pages 347–351.

Lawless, J. F. and Zhan, M. (1998). Analysis of interval-grouped recurrent-event data using piecewise constant rate functions. *The Canadian Journal of Statistics*, 26:549–565.

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

Liang, Z., Jaszczak, R. J., and Coleman, R. E. (1992). Parameter estimation of finite mixtures using the EM algorithm and information criteria with application to medical imaging processing. *IEEE Transactions on Nuclear Science*, 39:1126–1133.

Lindsey, J. C. and Ryan, L. M. (1993). A three-state multiplicative model for rodent tumorigenicity experiments. *Applied Statistics*, 42:282–300.

Lindsey, J. C. and Ryan, L. M. (1998). Tutorial in biostatistics: Methods for interval-censored data. *Statistics in Medicine*, 17:219–238.

Lindsey, J. K. (1993). *Models for Repeated Measurements.* Toronto: Oxford University Press.

Lindsey, J. K. (1996). *Parametric Statistical Inference.* Toronto: Oxford University Press.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B, Methodological*, 44:226–233.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models, Second Edition.* London: Chapman and Hall.

McGarty, T. P. (1974). *Stochastic Systems and State Estimation.* Toronto: John Wiley & Sons.

McPhee, S. J. (1995). Screening for cancer: Useful despite its limitations. *Western Journal of Medicine*, 163:169–172.

Meng, X.-L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association*, 86:899–909.

Neath, A. A. and Cavanaugh, J. E. (1997). Regression and time series model selection using variants of the Schwarz information criteria. *Communications in Statistics, Part A – Theory and Methods*, 26:559–580.

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

Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Applied Statistics*, 29:15–24.

Prentice, R. L. (1989). Surrogate endpoints in clinical trials: definition and operational criteria. *Statistics in Medicine*, 8:431–440.

Rosenberg, P. S. (1995). Hazard function estimation using B-splines. *Biometrics*, 51:874–887.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6:461–464.

Solka, J. l., Wegman, E. J., Priebe, C. E., Poston, W. L., and Rogers, G. W. (1998). Mixture structure analysis using the Akaike Information Criterion and the bootstrap. *Statistics and Computing*, 8:177–188.

Tsai, W.-Y. and Crowley, J. (1985). A large sample study of generalized maximum likelihood estimators from incomplete data via self-consistency. *The Annals of Statistics*, 13:1317–1334.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B*, 38:290–295.

Ware, J. H. (1985). Linear models for the analysis of longitudinal studies. *American Statistician*, 39:95–101.

Ware, J. H. and Lipsitz, S. (1988). Issues in the analysis of repeated categorical outcomes. *Statistics in Medicine*, 7:95–107.