Statistical Methods for Life History Analysis Involving Latent Processes

by

Hua Shen

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

Waterloo, Ontario, Canada, 2014

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

Incomplete data often arise in the study of life history processes. Examples include missing responses, missing covariates, and unobservable latent processes in addition to right censoring. This thesis is on the development of statistical models and methods to address these problems as they arise in oncology and chronic disease. Methods of estimation and inference in parametric, weakly parametric and semiparametric settings are investigated. The specific problems are discussed as follows.

Studies of chronic diseases routinely sample individuals subject to conditions on an event time of interest. In epidemiology, for example, prevalent cohort studies aiming to evaluate risk factors for survival following onset of dementia require subjects to have survived to the point of screening. In clinical trials designed to assess the effect of experimental cancer treatments on survival, patients are required to survive from the time of cancer diagnosis to recruitment. Such conditions yield samples featuring left-truncated event time distributions. Incomplete covariate data often arise in such settings, but standard methods do not deal with the fact that the covariate distribution is also affected by left truncation. In Chapter 2 we develop a likelihood and algorithm for estimation for dealing with incomplete covariate data in such settings. An expectation-maximization algorithm deals with the left truncation by using the covariate distribution conditional on the selection criterion. An extension to deal with sub-group analyses in clinical trials is described for the case in which the stratification variable is incompletely observed.

In studies of affective disorder, individuals are often observed to experience recurrent symptomatic exacerbations of symptoms warranting hospitalization. Interest lies in modeling the occurrence of such exacerbations over time and identifying associated risk factors to better understand the disease process. In some patients, recurrent exacerbations are temporally clustered following disease onset, but cease to occur after a period of time. We develop a dynamic mover-stayer model in which a canonical binary variable associated with each event indicates whether the underlying disease has resolved. An individual whose disease process has not resolved will experience events following a standard point process model governed by a latent intensity. If and when the disease process resolves, the complete data intensity becomes zero and no further events will arise. In Chapter 3, an expectation-maximization algorithm is developed for parametric and semiparametric model fitting based on a discrete time dynamic mover-stayer model and a latent intensity-based model of the underlying point process. The method is applied to a motivating dataset from a cohort of individuals with affective disorder experiencing recurrent hospitalization for their mental health disorder.

Interval-censored recurrent event data arise when the event of interest is not readily observed but the cumulative event count can be recorded at periodic assessment times. Extensions on model fitting techniques for the dynamic mover-stayer model are discussed in Chapter 4 which incorporate interval censoring. The likelihood and algorithm for estimation are developed for piecewise constant baseline rate functions and are shown to yield estimators with small empirical bias in simulation studies. Data on the cumulative number of damaged joints in patients with psoriatic arthritis are analysed to provide an illustrative application.

Future research is outlined and discussed in Chapter 5.

Acknowledgements

First and most importantly, I would like to express my deepest gratitude to my advisor, Dr. Richard Cook, who has been a tremendous mentor for me. Dr. Cook guided, encouraged and supported my research with his profound knowledge and energy. He created great opportunities for me to collaborate with leading researchers and work with wonderful colleagues. He was very generous in providing computational devices and financial support to facilitate my research. Moreover Dr. Cook always has his students' best interests at heart and his advice is invaluable. His devotion to work, enthusiasm in life and kindness to me and others made him my role model. I consider myself very lucky to be both Dr. Cook's student and employee.

I thank Dr. Steve Brown, Dr. Leilei Zeng, Dr. Paul Peng and Dr. Shannon Majowicz for serving as my thesis examining committee. Their expertise from different aspects is inspiring. Their insightful comments and suggestions to better my thesis are greatly appreciated. I would also like to thank Novartis Pharmaceutical, Dr. Lars Vedel Kessing, Dr. Per Kragh Andersen and Dr. Dafna Gladman for the stimulating examples and permission to use data. My study was supported by a grant from Division of High Impact Clinical Trials of the Ontario Institute for Cancer Research which will be always remembered.

I also benefited significantly from the excellent study environment among strong faculty, supportive staff and warm classmates in Department of Statistics and Actuarial Science. The vibrant research atmosphere at University of Waterloo is essential. As a high performance computing consortium, Shared Hierarchical Academic Research Computing Network greatly accelerated my computation work.

A special thanks goes to my family and friends who have always been my source of energy. I would like to express my sincerest appreciation to my husband Wei for his love, patience and companion. Last but not least, I would like to thank my daughter who joined me at the end of this journey of discovery and fun with an easy and joyful pregnancy.

Dedication

To my parents and grandparents.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 An Overview of Life History Data

Life history data pertain to the events and conditions that individuals experience over their lifetime. Often events are transient and it is meaningful to model event occurrence, but in other settings interest lies in modeling changes of state where events are more naturally viewed as representing transitions in the status of an individual. Often it is of interest to study the effect of fixed or time-varying covariates on event occurrence or state transitions. Life history analysis is carried out by fitting models and conducting statistical inferences about particular features of the stochastic mechanisms giving rise to life history data. Such methods are relevant to diverse areas including population and clinical research, sociology, actuarial science, and engineering. A brief overview of the topics covered in this research is presented as follows.

1.1.1 Analysis of Time to Event Data

Survival analysis involves the modeling of time to event data, where the event times are clearly specified in terms of an unambiguous time origin, a consistent nonnegative scale of measurement and precisely defined event of interest. The time origin can be birth, the calendar time of randomization in a clinical trial, or the time of purchase of a product warranty settings. The survival time can be measured in real time or operational time as appropriate. The event of interest can be death, disease onset, marriage, warranty claim, and so on. In survival analysis, subjects are usually followed over a specified time period, thus incomplete data arise in the form of censored observations.

Let T denote the failure time, which can be either continuous or discrete, and let C denote the censoring time. When the study ends before an individual experiences the event of interest or if an individual drops out or becomes lost of follow-up during the study, $C < T$ and the time of interest is right-censored, then the observed time $X = \min(T, C)$ and the censoring indicator $\delta = I(T < C)$ are recorded. Interval censoring arises when the event of interest is only known to occur within a time interval (e.g. $T \in [L, R]$) as is often the case in studies involving periodic follow-ups. Current status data is a special case of interval-censored data where all the subjects' event times are either left-censored or right-censored.

Truncation is a term used to describe the effect of a selection condition in which individuals are screened for inclusion in a study. Individuals are included in the study and their event times can be observed only if events occur within the truncation region. Truncation differs from censoring in the sense that it is an inclusion criterion. Data are left-truncated when individuals are only included if they have not yet experienced the event of interest at a certain time point and they are then prospectively followed to observe right-censored event times. A common type of left truncation arises when subjects free of an event enter a study at random ages and are followed from this "delayed entry time" until the event is observed, subject to right censoring.

Let $F(t) = P(T < t)$, $\mathcal{F}(t) = P(T \geq t)$, $f(t) = \frac{d}{dt}F(t)$,

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} = \frac{f(t)}{\mathcal{F}(t)} \ ,$$

and $\Lambda(t) = \int_0^t \lambda(u)du$ denote the cumulative distribution function, survival function, density function, hazard function and cumulative hazard function for T respectively, where θ parameterizes the distribution of T. We can then construct the likelihood functions for survival data assuming independent censoring as

$$L(\theta) = \prod_{i \in \mathcal{O}} f(T_i; \theta) \prod_{i \in \mathcal{L}} F(T_i; \theta) \prod_{i \in \mathcal{R}} \mathcal{F}(T_i; \theta) \prod_{i \in \mathcal{I}} [\mathcal{F}(L_i; \theta) - \mathcal{F}(R_i; \theta)]$$

where $\mathcal{O}$, $\mathcal{L}$, $\mathcal{R}$ and $\mathcal{I}$ represent the subsets where event times are exactly observed, left-censored, right-censored and interval-censored. Only partial information about the event times is available if censoring occurs. Conditional probability is needed when data involves truncation.

Regression models are often used to study the relationship between the event time T and the vector of explanatory variables Z that might affect the distribution of T. Proportional hazards regression models offer a popular formulation where the effect of one unit increase in a covariate is assumed to result in a multiplicative effect on the hazard rate. The proportional hazards assumption states

$$h(t|Z) = h_0(t)g(Z;\beta) \ ,$$

where $h_0(t)$ is the baseline hazard function, and $g(Z)$ is a function which describes how the hazard changes according to the covariates, often the special form $g(Z;\beta) = \exp(Z'\beta)$ is used, where β reflects the effect of covariate vector Z on the event process. Note that $h_0(t)$ could be of a parametric form, or a non-parametric form. The Cox proportional-hazards regression model leaves the baseline hazard function unspecified and this semi-parametric model is most widely used in survival analysis (Cox, 1972). The partial likelihood function can then be constructed to facilitate estimation and inference.

1.1.2 Analysis of Recurrent Events

In some epidemiological and medical studies, an event of interest may occur multiple times for the same subject during the period of follow-up. Such processes are referred to as recurrent event processes, and the data consisting of information on the events and covariates over time are called recurrent event data. Examples include migraines, seizures, heart attacks, strokes, sporting injuries, hospitalization and so on. Researchers are often interested in characterizing the event process, identifying the sources of variation across individuals in the study population, comparing different groups of processes, and quantifying the effect of covariates on event occurrence.

There are several approaches to analyze recurrent event data. Cook and Lawless (2007) gives a recent account for the statistical research done on different frameworks. Intensity functions and counting processes are very useful in modeling and data analysis (Andersen et al., 1993), but Markov or semi-Markov processes are perhaps most often employed in modeling recurrent events.

Suppose process starts at $t = 0$, and $T_1 < T_2 < \dots$, where T_j is the time of the jth event. Let $N(t) = \sum_{j=1}^{\infty} I(T_j \leq t)$ record the number of events over $[0, t]$, $N(s, t) = N(t) - N(s)$ record the number of events over $(s, t]$, and $\{N(t), t \geq 0\}$ denote a counting process. The event process is usually censored at time C which is often assumed to be independent of the event process, and the observed data are (X_j, δ_j), where $X_j = \min(T_j, C)$ and $\delta_j = I(X_j < C)$. Assuming a continuous time framework for which two events do not occur at the same time, the event intensity function is the instantaneous probability of an event occurring at time t, conditional on the process history

$$\lambda(t|H(t)) = \lim_{\Delta t \to 0} \frac{P(\Delta N(t) = 1|H(t))}{\Delta t} \ ,$$

where $H(t) = \{N(s), 0 \leq s < t\}$ denotes the event history up to time t, and $\Delta N(t) = N(t + \Delta t^{-}) - N(t^{-})$ is the number of events over $[t, t + \Delta t)$. If an individual is observed over $[0, \tau]$, the joint probability of m events at $t_1, \dots, t_m$ is

$$\prod_{j=0}^{m} \lambda(t_j|H(t_j)) \exp\left(-\int_0^{\tau} \lambda(u|H(u))du\right) \ ,$$

where $m = N(\tau)$ is the number of events that occurred over $[0, \tau]$.

Marginal methods are developed based on rate and mean functions when interest lies in the expected number of events as a function of time since study entry (Lawless and Nadeau, 1995; Lin et al., 2000). Poisson models are often used as the canonical model for rate function analysis, with $\lambda(t|H(t)) = \rho(t)$, where $\rho(t)$ is the rate function for Poisson process and $\mu(t) = E(N(t)) = \int_0^t \rho(u)du$ is the mean function. The number of events in any time interval follows a Poisson distribution, with the number of events in disjoint intervals being statistically independent. If $\rho(t)$ is a constant, a time-homogeneous Poisson process is assumed.

When events are generated according to processes with a cyclical feature, methods based on times between events are often appropriate and these are usually based on a natural adaption of methods for survival analysis. Methods involving hazard rate functions are frequently employed. Gap times $W_j = T_j - T_{j-1}$, $j = 1, 2, \ldots$, with

$$P(W_j > w|T_{j-1} = t_{j-1}, H(t_{j-1})) = \exp\left(-\int_{t_{j-1}}^{t_{j-1}+w} \lambda(u|H(u))du\right) ,$$

are often useful. Renewal models are the canonical models for gap time analysis, with $\lambda(t|H(t)) = h(B(t))$, where $h(\cdot)$ is hazard function for W_j and $B(t) = t - t_{N(t^-)}$ denotes the time since last event. Times between successive events are often assumed to be independent and identically distributed in renewal processes. Gap times are statistically independent for the homogenous Poisson process; they are not independent in general.

Proportional hazard models representing multiplicative relationship are usually used, such as in the conditional model (Prentice et al., 1981), the marginal event-specific model (Lin et al., 2000), and the counting process formulation (Andersen et al., 1993). Andersen and Gill (1982) proposed a semiparametric regression model where the baseline rate function is not assumed to have any particular parametric form; the generalized Nelson-Aalen (Breslow) estimate can be obtained for the baseline mean function. Non-parametric estimation of $\mu(t)$ was proposed by Lawless and Nadeau (1995) for the one-sample problem. Unobserved heterogeneity between individuals can be modeled by random effects as in Lawless (1987b,a), for example. Data for recurrent event processes may provide the exact times of successive events. Sometimes only the total numbers of events occurring in specific time intervals are observed resulting interval-censored recurrent data (Thall and Lachin, 1988; Lawless and Zhan, 1998).

1.1.3 Latent Variables

Missing data is inevitable in large cohort studies. Decisions need to be made on how to deal with incomplete covariates and responses. Simply ignoring missing data may result in a loss of information and can cause bias in estimators. Three missing data mechanisms are often under discussion.

Missing Completely at Random (MCAR): If the probability that an observation is missing is independent of the value of the observation or the value of any other variables, the data are said to be missing completely at random (MCAR). In this case any particular data are just as likely to be missing as any other data. That is, if we let Y denote the data that are always observed, X denote the data that are sometimes missing, and R denote the missing status, then

$$P(R = 1|Y, X) = P(R = 1)$$

for data MCAR and the informative part of the full likelihood is proportional to $f(y|x)f(x)$, where $f(x)$ and $f(y|x)$ is the probability density function (p.d.f.) of X and conditional p.d.f. of Y given X. This nice feature of MCAR data means that the analysis remains unchanged; one may lose power but the estimators remain the usual maximum likelihood estimators.

Missing at Random (MAR): Here the probability of missing data is conditionally independent of the unobserved data given the values of the observed data. That is,

$$P(R = 1|Y, X) = P(R = 1|Y) \ ,$$

thus the likelihood is still proportional to $f(y|x)f(x)$ if the probability model for the missing data mechanism is functionally independent of the response model. When data are MAR, it can produce biased estimators of parameters in marginal (semi-parametric) models, but maximum likelihood estimators remain optimal provided the model for the missing data process does not involve any parameters in the response model (i.e. that the missing data process is non-informative).

Although the MCAR and the MAR assumptions are often realistic and particularly convenient in the sense that they lead to considerable simplification in the issues surrounding the analysis of the incomplete data, a challenging situation arises if data are neither MCAR nor MAR.

Missing Not at Random (MNAR): In this case, the probability a measurement is available depends on both observed and unobserved quantities and so $P(R|Y, X)$ cannot be simplified. The only way to obtain a consistent estimate of the parameters of interest is to model the missingness and often the associated parameters are not identifiable. Model diagnostics for the missing data model are also difficult to carry out in most cases, and so one is in the difficult position of relying heavily on model assumptions which cannot be adequately checked.

There have been some *ad hoc* approaches for dealing with missing data in analysis. By far the most common and simplest approach adopted by some statistical software packages is to exclude individuals that have missing values and to restrict analyses to the fully observed data set. This strategy, called *list-wise deletion*, or *complete case analysis*, is generally inappropriate if the researchers are interested in making inferences on the entire target population instead of the portion of it represented by available data. It normally results in a substantial loss in power and precision while consistent estimates are obtained under the MCAR assumption and bias arises when the data are not MCAR.

There are a few approaches that involve replacing missing values via imputation, including mean substitution in which the missing value is replaced with the mean of the variable estimated from available data, and regression substitution that imputes using regression analysis. These simple imputation methods are inadequate as they may reduce standard errors, inflate test statistics, give inappropriately narrow confidence intervals and invalid tests (Musil et al., 2002; Fielding et al., 2008). More modern approaches rely on maximum likelihood theory and multiple imputation (Schafer, 1999). King et al. (2001) reviewed many of the practical strengths and limitations of multiple imputation.

Little and Rubin (2002) gives a very thorough treatment of the issue of missing data. They give an extensive discussion of the theory in the context of multivariate normal models with incomplete observations; see also Anderson (1957), Afifi and Elashoff (1966), and Hocking and Smith (1968). Ibrahim (1990) examined the general problem of incomplete data for any generalized linear model (GLM) with discrete covariates and showed that the E-step of the EM algorithm can be written as a weighted complete data log-likelihood for any GLM. Horton and Laird (1999) described the method of weights in detail, illustrated its application with several examples, discussed its advantages and limitations, and reviewed extensions and applications of the method. We consider this approach in the research that follows.

Another type of latent variable is one that can never be observed, but is introduced as a way of generalizing a model. For instance, in both survival and recurrent event data, some subjects may not be observed to experience the event of interest despite the lengthy follow-up. Cure rate models (Boag, 1949; Berkson and Gage, 1952) are often used to analyze and describe survival data when long-time survivors exist. These models accommodate a sub-population of individuals who are not susceptible to the event of interest. This accommodation leads to survival curves which flatten out earlier than one would expect from a more standard distribution.

Note that here the meaning of "cure" may differ in different contexts. In chronic diseases that cannot be cured, a mixture model of this sort allows for the possibility that the disease may go into remission thereby eliminating the rise of any complications. In studies of mental health, it could be that environmental factors triggering acute episodes are eliminated. In cancer studies, cure could be said to occur when the mortality rate in the diseased group becomes the same as that of the otherwise matched control group, this could happen following a successful surgery, for example. In short, cure models are often used to model long-term survivors rather than cured patients in the general sense.

Boag (1949) first proposed a model to estimate the cure fraction in mixture model and Berkson and Gage (1952) further developed it to the standard cure rate model as

$$\mathcal{F}(t) = p + (1-p)\mathcal{F}^*(t) \ ,$$

where $\mathcal{F}(t)$ and $\mathcal{F}^*(t)$ denote the probability of being event free at time t for the mixed group and the uncured group respectively, and p is the cure rate reflecting the proportion of the population that is not susceptible. In mixture models often the probability of being cured is modeled by logistic regression and many standard models for survival data can be used for the uncured patients; the Weibull distribution and the Cox proportional hazards model are two popular choices (Farewell, 1982; Kuk and Chen, 1992; Taylor, 1995; Sy and Taylor, 2000). Many variations of mixture cure models have been proposed (Peng et al., 1998; Peng and Dear, 2000).

Mover-stayer models are more general than cure rate models, and are often discussed in the context of Markov models. They assume the study population consists of movers and stayers, where the movers make transitions following some ordinary multistage process and the stayers make no such transitions. Early references to the mover-stayer model include Blumen et al. (1955) and Goodman (1961). Further studies have lead to extensions by Spilerman (1972) and Frydman (1984). Models that incorporate dynamic mover-stayer indicators were developed by researchers including Cook et al. (2002) and Yamaguchi (2003).

1.2 Introduction to Topics

In the following sections, we describe three topics of statistical research. They involve handling missing covariates in survival data subject to left truncation (Chapter 2), dealing with right-censored recurrent event data in disease processes subject to resolution (Chapter 3), and addressing the challenges in the interval-censored recurrent event data from disease processes subject to resolution (Chapter 4).

1.2.1 Missing Covariates with Left-Truncated Event Times

Studies of chronic diseases routinely sample individuals subject to specified conditions on an event time of interest. In epidemiology, for example, prevalent cohort studies may aim to evaluate risk factors for death following onset of dementia. Such designs require subjects to have survived from the date of disease onset to the date of the screening assessment (Wolfson et al., 2001). In clinical research, randomized trials are often designed to assess the effect of experimental cancer treatments on survival and patients must survive from the time of cancer diagnosis to contact to be recruited; there may be additional conditions imposed on the times of non-fatal events related to the disease process (Hortobagyi et al., 1996). When the date of disease onset is to be used as the time origin for survival analyses, samples chosen this way feature left truncation and standard methods of survival analysis can be readily adapted to deal with this feature (Cox and Oakes, 1984; Andersen et al., 1993; Klein and Moeschberger, 1997; Kalbfleisch and Prentice, 2002; Lawless, 2002).

Incomplete covariate data often arise in studies with time to event outcomes (Little and

Rubin, 2002). This may be a consequence of the study protocol if resources are limited and a particular subset of individuals are identified for detailed examination of biomarkers, for example. In other cases it may be due to chance (e.g. noncompliance of study investigators or participants). There is a large literature on the various frameworks and methods for fitting regression models to survival data with incomplete covariate information. Methods based on the EM algorithm are developed by Lipsitz and Ibrahim (1996), Chen and Little (1999) and Herring et al. (2004) among others. Estimating function approaches incorporating inverse probability weights are given by Lipsitz and Ibrahim (1998), and Wang and Chen (2001) develop augmented estimating equations yielding more efficient estimation. Bayesian approaches for this same problem are developed by Ibrahim et al. (2008) and Bradshaw et al. (2010), and Chen and Little (2001) consider an interesting alternative approach for dealing with missing covariates in the context of linear transformation models. These methods do not deal with the setting where individuals are only sampled if they satisfy some response-dependent selection criterion (e.g. truncation). In this setting the sample covariate distributions are different from the population covariate distribution due to selection effects and in fact different individuals will have different sample covariate distributions if they have different selection criteria (Begg and Gray, 1987; Bergeron et al., 2008; Cook and Bergeron, 2011).

1.2.2 A Dynamic Mover-Stayer Model for Recurrent Events

Recurrent data arise frequently in studies of chronic disease, actuarial science, industrial research and sociology. In health research, examples include exacerbations of symptoms in patients with respiratory disease (Grossman et al., 1998), seizures in individuals with epilepsy (Pledger et al., 1994), and recurrent episodes of bleeding in patients with thrombocytopenia (Heddle et al., 2003; Webert et al., 2006). There has been considerable statistical research in the last twenty years on methods for the analysis of recurrent event data (Cook and Lawless, 2007). Models and methods can be broadly classified as intensity-based (Andersen et al., 1993), based on marginal mean or rate functions (Lawless and Nadeau, 1995), or based on random effect models (Lawless, 1987a).

Frequently the recurrent event process ends upon on the occurrence of a terminal event. Graft rejection episodes in transplant recipients, for example, cease to occur upon total graft rejection (Cole et al., 1995), skeletal complications in patients with bone metastases end when a patient dies (Hortobagyi et al., 1998), and recurrent hospitalizations for cardiovascular events end upon death (Bourassa et al., 1993). There has been considerable recent work on the development of statistical methods for the analysis of recurrent events in the presence of a terminal event. This phenomenon is naturally handled with intensity-based models (Andersen et al., 1993), but robust marginal methods have been developed (Cook and Lawless, 1997; Ghosh and Lin, 2000, 2002), as have models and methods incorporating random effects (Liu et al., 2004; Ye et al., 2007).

We consider the setting in which recurrent events arise in a chronic disease processes but where some individuals have particularly long periods of time from their last event to a right censoring time. This is motivated by the need to model recurrent event processes in which the recurrent events arise because of a transient underlying condition which can resolve. Unlike the case of a terminal event such as death, in this setting it is not known if and when the underlying condition has resolved. We handle this complication through use of a dynamic mover-stayer model. The model is comprised of an intensity function for event occurrence among individuals still experiencing the underling condition generating the events and a series of conditional probabilities for modeling the resolution of the underlying process.

Mixture models have been used extensively to model the presence of a so-called "cured fraction" in cancer studies featuring long-term survivors. Farewell (1982, 1986) proposed a parametric mixture model incorporating a logistic regression model for the latent cure status and a Weibull model for the survival times of those in the uncured group. Peng et al. (1998) extended this approach to incorporate the generalized F failure time distribution and Taylor (1995) extended this further to enable nonparametric estimation of the survival distribution among susceptible individuals through a Kaplan-Meier type estimate. Kuk and Chen (1992) extended the cure rate model to accommodate a semiparametric proportional hazard model for the survival time and proposed estimation via an EM algorithm. Peng and Dear (2000) further studied the semiparametric approach by allowing covariate effects on the cure rate. A zero-tail constraint was introduced by Sy and Taylor (2000) to deal with identifiability issues. Yamaguchi (1992) described a further interesting generalization of the notion of a cured fraction by introducing a latent failure time at which subjects became nonsusceptible to the event of interest. Asymptotic properties of maximum likelihood estimates from the cure rate model, including the existence, strong consistency and asymptotic normality, were studied by Fang et al. (2005); asymptotic variances were also derived to facilitate inferences using Wald-based pivotals.

Cure rate survival models are a special case of a more general class of mover-stayer models. In mover-stayer models the population is comprised of two sub-populations. In one sub-population, the so-called "mover" group, transitions among states are made according to a general multi-state process. In the other sub-population individuals have a zero probability of moving from the initial state, and these individuals are called "stayers". Often Markov models are adopted for the multi-state process for movers, but any multi-state model can be specified in principle. Goodman (1961) proposed methods for consistent parameter estimation to address inconsistency of estimators developed by Blumen et al. (1955) in the discrete-time setting. Spilerman (1972) further generalized the mover-stayer model to allow the individual mobility rate to follow a continuous distribution. Frydman (1984) described how to obtain maximum likelihood estimates based on the observed likelihood, while Fuchs and Greenhouse (1988) used the EM algorithm with extensions to handle incomplete follow-up in the panel studies. Models incorporating dynamic mover-stayer indicators have received some attention including the multistate models by Heckman and Walker (1987), Yamaguchi (1994, 1998, 2003) and Cook et al. (2002).

1.2.3 Interval-Censored Recurrent Event Data from Disease Processes Subject to Resolution

There are many chronic disease processes for which affected individuals experience recurrent adverse events. In some settings the events are apparent when they occur, as is the case in individuals with respiratory disease experiencing recurrent exacerbations (Grossman et al., 1998), epilepsy where the events may be recurrent seizures (Pledger et al., 1994), neurology when the events are recurrent migraine headaches among those with migraineur (Pascual et al., 2000), and angina where the events may be recurrent acute episodes (Peters et al., 2003). Statistical methods for recurrent event analysis in such settings include those reliant on intensity-based models (Andersen et al., 1993), random effect models (Lawless, 1987a), and marginal methods (Lawless and Nadeau, 1995; Lin et al., 2000). Cook and Lawless (2007) give a comprehensive account of the frameworks for analysis.

In some settings the occurrence of events is not evident, but rather can only be determined upon a radiographic examination, when blood tests are carried out, or by detailed clinical examination. Examples include the development of new tumours in bladder cancer patients (Byar et al., 1986), the occurrence of asymptomatic fractures in patients with osteoporosis (Riggs et al., 1981), and the development of new skeletal metastases in patients with cancer metastatic to bone (Hortobagyi et al., 1996).

A nonparametric approach to compare the recurrence rate of two treatment groups based on panel count data was proposed by Thall and Lachin (1988), and the nonparametric tests are further studied by researchers including Sun and Fang (2003), Zhang (2006), Park et al. (2007) and Balakrishnan and Zhao (2009). Mean function estimation was developed by Sun and Kalbfleisch (1995), which was later shown by Wellner and Zhang (2000) to be seen as a pseudo-maximum likelihood estimator under a non-homogeneous Poisson model. They proved its consistency, along with their proposed maximum likelihood estimator not relying on the Poisson assumption. Some procedures to conduct semiparametric regression analysis for interval-censored recurrent events are developed by Sun and Wei (2000), Cheng and Wei (2000), Zhang (2002) and Wellner and Zhang (2007). Regression on panel count data with informative observation times are also investigated by Huang et al. (2006), Sun et al. (2007) and Zhao and Tong (2011).

Lawless and Zhan (1998) consider multiplicative recurrent event models with piecewise constant baseline rate functions fitted using semiparametric methods via estimating functions as well as fully specified random effect models fitted using maximum likelihood. Such piecewise constant models share the advantages of parametric models and yet provide some robustness to misspecification of the parametric form of rate functions. Chen et al. (2005) extend these methods to deal with multi-type recurrent events. Sun and Zhao (2013) give an excellent account of the recent developments on methods for recurrent event analysis when data are subject to interval censoring.

In some settings the chronic condition generating the events can resolve and from the point of resolution individuals will no longer be at risk of events. Establishment of suitable medications, removal of stressors in mental health studies (Kessing et al., 2004a), or other lifestyle changes may minimize risk of future events, but it can be difficult to determine if and when such changes have taken place. In other settings the disease process resolves naturally. Polymyalgia rheumatica (Salvarani et al., 2002), for example, is a disease with different stages, and in the most active phase patients experience acute episodes of pain in the shoulder and pelvic joints. This active phase is of variable length (Healey, 1984) and upon completion of this phase the acute episodes cease to arise.

Many patients with systemic lupus erythematosus experience flares due to lupus nephritis. This condition, however, can go into remission and when this happens patients cease to experience acute flares in lupus nephritis (Barber et al., 2006). Syndesmophytes are bony growths that arise in patients with psoriatic arthritis, ankylosing spondylitis and other arthritic conditions and they are of scientific interest because they reflect a consequence of the underlying condition. Their development, however, is only detectable by radiographic examination. These rheumatic conditions can go into remission (Gladman et al., 2001; Zochling and Braun, 2006), and hence in this setting one is faced with both the challenge of interval-censored recurrent event times and the need to accommodate the possibility that the underlying condition has resolved.

1.3 Motivating Studies

In this section, we will look at the corresponding studies that motivated the methodological developments of this research.

1.3.1 Breast Cancer Patients with Skeletal Metastases

Here we consider data from a trial of 285 breast cancer patients with skeletal metastases diagnosed within three years of randomization (Hortobagyi et al., 1996). The primary purpose of this trial was to examine the effect of an experimental bisphosphonate therapy (n=133) compared to the control (standard care) therapy (n=152) on the reduction in skeletal complications arising because of these bone metastases. Secondary interest lies in the the effect of therapy on the time to death; the survival times of 42 (14.7%) of the patients were censored for death. We consider an analysis in which separate estimates of the treatment effect are desired for patients that are estrogen receptor (ER) positive and those that are ER negative, while controlling for whether the patient was 50 years of age or older at the time of diagnosis. The ER status is missing for 14.3% of patients in the experimental arm and 17.1% of patients in the control arm, but age of diagnosis was completely observed. Among the 114 individuals in the experimental arm with ER status available, 94 (82%) were ER positive, and among the 126 individuals in the control arm with available ER status, 97 (77%) were ER positive. The model in Section 2.2 is therefore suitable to address this question.

1.3.2 Danish Study of Individuals with Affective Disorder

A study of individuals with affective disorder was carried out in Denmark based on a registry of hospitalizations. For this study, a patient entered the cohort at the onset of affective disorder, defined by the first hospitalization for any mental disorder of inorganic etiology between 1994 and 1999. A total of 10523 individuals satisfied this selection condition. Over the course of the study period there was an average of 1.618 re-admissions (S.D.=1.720), with a minimum of 1 and a maximum of 90.

Kvist et al. (2007) examined the impact of misspecification of the frailty distribution, using a non-parametric estimator for the joint gap times and a marginalized estimator for marginal gap times. Cook and Lawless (2013) investigated trends in this recurrent event process and discussed the tests for trends in detail. We are now interested in extending analyses to accommodate the patients with long observed event-free periods (stayers) and ones with shorter durations (movers) in the recurrent event setting.

The present goal is to describe a model for the pattern of event occurrence where the events are the acute exacerbations of affective disorder and data feature individuals with unusually long periods of time without recurrence at the end of follow-up; see Figure 1.1. This pattern prompted the development (Winokur, 1975) and examination (Kessing et al., 2004b) of a theory that the disease process may "burn-out" for some affected individuals. This theory, in part, motivated the development of the dynamic mover-stayer model to be described in the Chapter 3 that follows. The data summary is given in Table 1.1 and we focus on the 9228 patients who remained unipolar over the entire course of study.

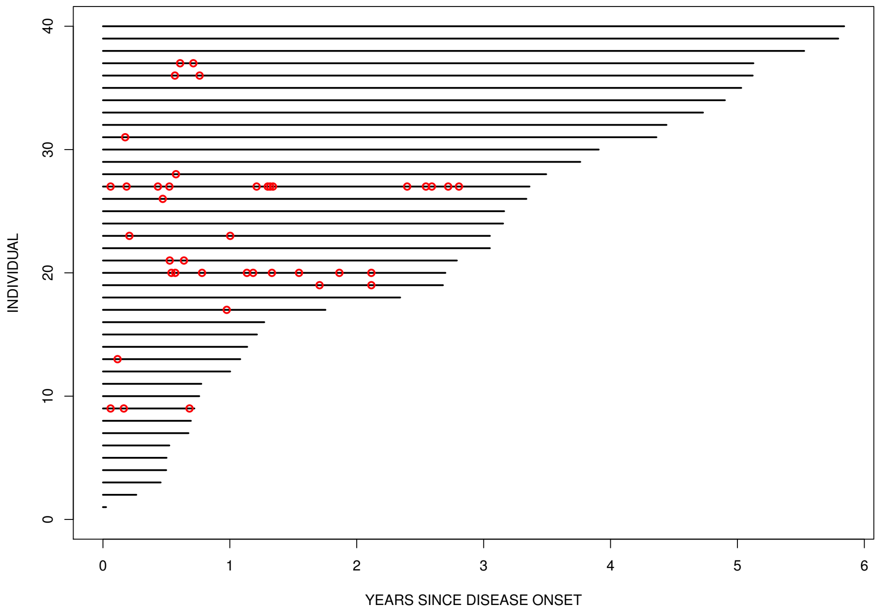

Figure 1.1: Timeline plots of recurrent acute episodes of affective disorder from time of disease onset for a selected sample of individuals from the Danish registry between 1994 and 1999

Here the recurrent event data are right-censored either due to lost of follow-up or death, whichever occurs first. Though suicide could be associated with recurrence of affective disorder, it usually happens shortly after the disease onset, and Kessing et al. (1998) found no significant association between suicides and event reoccurrence, and only a small percentage patients died by the end of study in this cohort.

Table 1.1: Data summary of the Danish registry dataset from 1994 to 1999

10523 patients in the entire course of study.				
		Female	Male	Total
	n(%)	6721(63.9%)	3802(36.1%)	10523
No. of visits	total	11132	5889	17021
	mean(std)	1.656(1.935)	1.549(1.247)	1.618(1.247)
	range	1-90	1-21	1-90
Death	n(%)	300(4.5%)	227(6.0%)	527(5.0%)
Bipolar at entry	n(%)	602(9.0%)	504(13.3%)	1106(10.5%)
Bipolar at end	n(%)	737(11.0%)	558(14.7%)	1295(12.3%)

9228 patients who have been unipolar over the entire course of study.				
		Female	Male	Total
	n(%)	5984(64.8%)	3244(35.1%)	9228
No. of visits	total	9397	4769	14166
	mean(std)	1.570(1.348)	1.470(1.101)	1.535(1.268)
	range	1-26	1-16	1-26
Death	n(%)	271(4.5%)	204(6.3%)	475(5.1%)

1.3.3 Joint Damage in Patients with Psoriatic Arthritis

Psoriatic arthritis is an inflammatory arthritis and an autoimmune disease that commonly occurs among patients with psoriasis. Patients with psoriatic arthritis may experience swelling, pain and inflammation in the affected joints. The University of Toronto Psoriatic Arthritis Clinic is the largest center in the world for specialized care and comprehensive research in this disease. The clinic, started in 1978, has been recruiting and following patients continuously since then. Data collected at clinic entry and regular follow-up clinic visits arise from a complete history, physical examination, blood and urine tests, and radiographic examination. Over 1100 patients have been closely followed over the years.

The development of joint damage is of primary interest to clinicians since this damage impairs quality of life and functional ability. Understanding the risks of rapid onset and accumulation of damage is therefore the basis of much of the scientific research in this condition (Gladman et al., 1995). Factors studied include information on family history of psoriatic arthritis and genetic information based on human leukocyte antigen (HLA) markers, for example. Radiological examinations of the hands, feet and spine are scheduled every two years, but the actual assessment times vary considerably. Moreover there are some patients who experience no joint damage over the entire course of follow-up, and others who develop damaged joints for some time but then experience long periods in which no further damage is observed. One possible explanation for the latter scenario is that these patients experience remission and hence are no longer at risk for further damage. A key point is that individuals transition from the mover (susceptible) to stayer (resolved) sub-group as time passes. Figure 1.2 displays the timing of the assessments and the number of additional damaged joints detected over the respective intervals for a sample of 15 individuals; here we restrict attention to patient data over the first 30 years from disease onset. The variability in the frequency of visits is apparent, as is the variation in the event counts both between patients.

1.4 Outline of the Thesis

This thesis aims to study and develop appropriate statistical methods to address several kinds of incompleteness problems in lifetime data: missing covariates with left truncation in

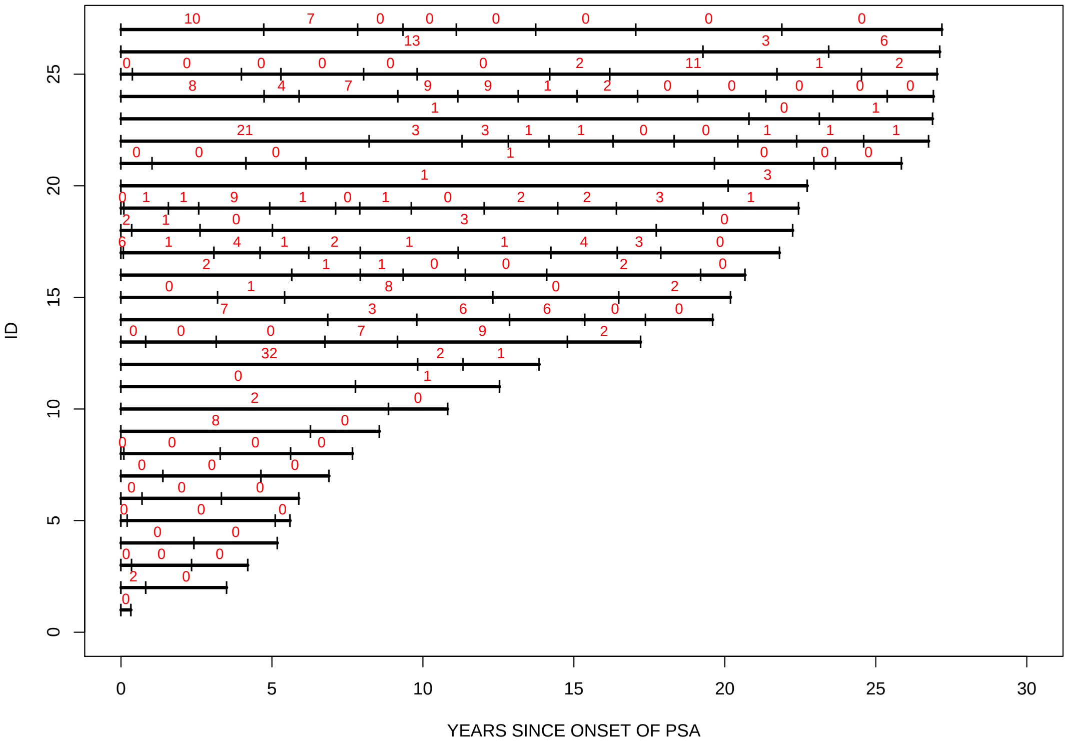

Figure 1.2: Plot of assessment times and number of additional radiological damaged joints detected between assessments (red numbers) from onset of psoriatic arthritis for a selected sample of patients from University of Toronto Psoriatic Arthritis Clinic recruited between 1978 and 2013; follow-up restricted to within 30 years of disease onset

survival analysis, unobserved latent indicator in disease process that is subject to resolution with right censoring and interval censoring in recurrent event data. The remainder of this thesis is organized as follows.

Chapter 2 focuses on the missing covariate problem in survival data with left truncation. In Section 2.1 we define notation, give the complete data likelihood, and describe how to carry out the maximization step of the EM algorithm using standard software. Additional technical details on EM algorithm including the realization of the E-step and estimation of the information-based variance are given in Appendix 2A. We then assess the empirical performance of estimators arising from a complete case analysis, a misspecified likelihood which uses the population rather than the appropriate sample covariate distribution, and the proposed method. Extensions to facilitate robust estimation using piecewise-constant baseline hazards are described in Appendix 2B. The extension dealing with the case of a missing stratification variable to be used in a secondary sub-group analysis is developed in Section 2.2 and the illustrative application is given in Section 2.3. Concluding remarks and a recap of the contributions are given in Section 2.4.

In Chapter 3 we consider the situation in which events arise in a chronic disease processes but where individuals under observation tend to have a long period from their last event to a censoring time. We handle this using a dynamic mixture model formulation. We consider a point process model augmented to include a dynamic mover-stayer indicator which is generated each time an event occurs. In Section 3.1 we introduce the notation and model formulation for general case. In Section 3.2 we first give the complete data likelihood for a general model, then describe how to implement the EM algorithm, and give specific details on how to fit a semiparametric latent Markov model. The performance of the proposed algorithm for parametric and semiparametric models is examined empirically in Section 3.3. Several models are fitted to the motivating Danish study of affective disorder in Section 3.4 and concluding remarks are given in Section 3.5.

Chapter 4 describes methods which aim to handle interval-censored recurrent events arising from disease processes subject to resolution. The dynamic mover-stayer model of Shen and Cook (2013a) is reviewed in Section 4.1, the detailed EM algorithm for fitting a dynamic mover-stayer model to interval-censored recurrent event data under a piecewise constant baseline rate function is described in Section 4.2. We empirically exam the performance of the proposed approach in Section 4.3. Data from a psoriatic arthritis cohort is analysed in Section 4.4 and general remarks are given in Section 4.5.

Further comments regarding proposed methods and topics warranting future research are discussed in Chapter 5.

Chapter 2

Incomplete Covariates and Left-Truncated Survival Data

The aim of this chapter is to consider the missing covariate problem in survival data with left truncation and propose a simple strategy for dealing with it. We describe an EM algorithm (Dempster et al., 1977) for dealing with incomplete discrete covariate data. The algorithm involves the conceptualization of a complete data set which includes information on both the missing covariates and the number of unsampled individuals in the population who did not satisfy the truncation condition (Turnbull, 1976). The maximization step is shown to be easily implemented using standard survival analysis software provided it can accommodate left-censored data. A generalization of this algorithm is then developed for sub-group analyses in clinical trials where information on the stratification variables is missing. An application to data from a recently completed trial of patients with metastatic cancer is used for illustration.

2.1 Notation and Statement of the Problem

2.1.1 The Observed Data Likelihood

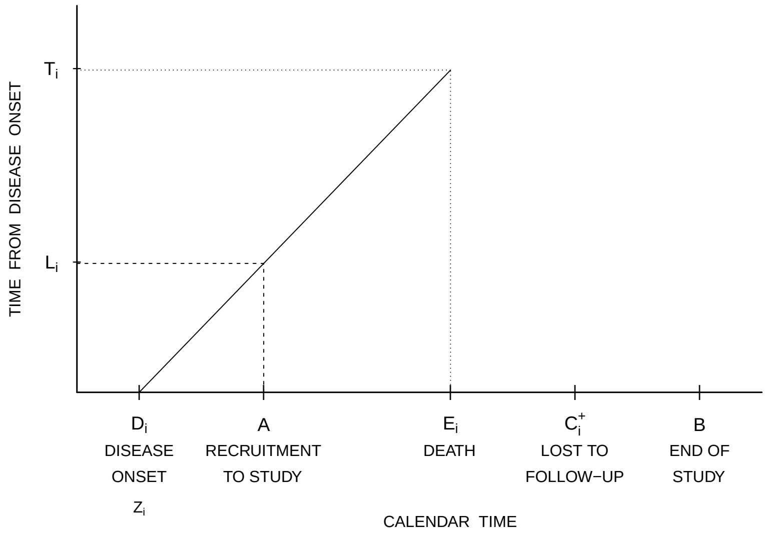

Figure 2.1: Lexis diagram of calendar event times and left-truncated failure time data

We consider first a cohort study in which a sample of m individuals is obtained by randomly sampling from a population of diseased individuals. As shown in Figure 2.1, let A denote the calendar time at which subjects are accrued, and B denote the calendar time of the end of the study; the duration of the study is then $C = B - A$. Let D_i denote the calendar time of disease onset and E_i denote the calendar time of the event, say death, for individual i; then $T_i = E_i - D_i$ is the corresponding survival time from disease onset. To be included in the study it is necessary that $T_i > L_i = A - D_i$, and so the survival time of a recruited individual is left-truncated at L_i. If $C_i^\dagger$ $(A < C_i^\dagger < B)$ is a random calendar time at which an individual is lost to follow-up, let $C_i = \min(B, C_i^\dagger) - D_i$ denote the censoring time measured from disease onset, $X_i = \min(T_i, C_i)$ denote the observation time, and $\delta_i = I(X_i = T_i)$ indicate whether individual i is observed to die. Consider a proportional hazards model

$$h(s|Z_i; \theta) = h_0(s; \alpha) \exp(Z_i' \beta)$$

specified to assess the effect of a covariate vector Z_i on the survival time, where $h_0(s; \alpha)$ is the baseline hazard function indexed by α, β is a vector of regression coefficients, and $\theta = (\alpha', \beta')'$. Let $H_0(s,t;\alpha) = \int_s^t h_0(u;\alpha)du$, $H(s,t|Z_i;\theta) = \int_s^t h(u|Z_i;\theta)du$ and we denote $H_0(0,t;\alpha)$ and $H(0,t|Z_i;\theta)$ by $H_0(t;\alpha)$ and $H(t|Z_i;\theta)$ respectively. We assume $Z_i \perp D_i$ so that the composition of the population with respect to the risk factors is stable over time, as is the effect of these risk factors on disease occurrence. We also assume $T_i \perp (D_i, C_i^\dagger)|Z_i$ so that the distribution of the event time does not depend on the calendar time of disease onset and censoring is conditionally independent of the event time.

Suppose a sample of m individuals is recruited at the start of the study. For illustration we suppose that the covariate vector is of the form $Z_i = (Z_{i1}, Z_{i2})'$ and contains risk factors for event at the time of diagnosis, where Z_{i1} is a binary covariate which is not observed for all individuals and Z_{i2} is another binary covariate which is always observed, $i = 1, \ldots, m$; extensions to handle other types of categorical covariates are straightforward. Let $R_i = I(Z_{i1} \text{ is observed})$, $\mathcal{R} = \{i : R_i = 1\}$, and $\bar{\mathcal{R}} = \{i : R_i = 0\}$. The conditional probability mass function for Z_{i1} given Z_{i2} is $P(Z_{i1}|Z_{i2};\eta)$ where

$$\text{logit} P(Z_{i1} = 1|Z_{i2}) = \eta_0 + \eta_1 Z_{i2} \ ,$$

with $\eta = (\eta_0, \eta_1)'$ and $\psi = (\theta', \eta')'$. We assume that Z_{i1} is missing at random according to $P(R_i = 1|D_i, Z_i, T_i, C_i) = P(R_i = 1|Z_{i2})$, where this model does not share any parameters with ψ and hence missingness is non-informative.

In the absence of left truncation (i.e. if $L_i = 0$, $i = 1, \ldots, m$), the observed data likelihood is

$$\begin{aligned} L(\psi) &= \prod_{i \in \mathcal{R}} \left\{ h^{\delta_i}(X_i|Z_i;\theta) \exp\left(-H(X_i|Z_i;\theta)\right) P(Z_{i1}|Z_{i2};\eta) \right\} \\ &\times \prod_{i \in \bar{\mathcal{R}}} \left\{ \sum_{Z_{i1}} h^{\delta_i}(X_i|Z_i;\theta) \exp\left(-H(X_i|Z_i;\theta)\right) P(Z_{i1}|Z_{i2};\eta) \right\} . \end{aligned} \tag{2.1}$$

When a sample features left truncation, the correct probability mass function for the covariate vector of individual i is $P(Z_i|T_i > L_i;\psi)$, so the likelihood in this setting is

$$\begin{aligned} L(\psi) &= \prod_{i \in \mathcal{R}} \left\{ h^{\delta_i}(X_i|Z_i;\theta) \exp\left(-H(L_i, X_i|Z_i;\theta)\right) P(Z_{i1}|Z_{i2}, T_i > L_i;\psi) \right\} \\ &\times \prod_{i \in \bar{\mathcal{R}}} \left\{ \sum_{Z_{i1}} h^{\delta_i}(X_i|Z_i;\theta) \exp\left(-H(L_i, X_i|Z_i;\theta)\right) P(Z_{i1}|Z_{i2}, T_i > L_i;\psi) \right\} , \end{aligned} \tag{2.2}$$

where

$$P(Z_{i1}|Z_{i2}, T_i > L_i;\psi) = \frac{P(Z_{i1}|Z_{i2};\eta) \exp\left(-H(L_i|Z_i;\theta)\right)}{\sum_{Z_{i1}} P(Z_{i1}|Z_{i2};\eta) \exp\left(-H(L_i|Z_i;\theta)\right)} . \tag{2.3}$$

The likelihood (2.2) can be maximized directly, but this can be challenging if the dimension of ψ is high. An expectation-maximization (EM) algorithm can alternatively be used with a complete data likelihood analogous to (2.2) where missing covariate values are part of the complete data. The maximization step of such an algorithm, however, would require optimizing a complicated function of ψ since one cannot factor the complete data likelihood to isolate the components θ and η; see (2.3). We propose a computationally more appealing complete data likelihood by incorporating contributions associated with individuals not selected for inclusion in the sample.

2.1.2 A Turnbull-type Complete Data Likelihood

Corresponding to individual i in the sample with left truncation time L_i, one can conceptualize J_i individuals who are identical in all respects (i.e. with the same covariate vector and disease onset time as individual i), except they did not remain event-free (alive) long enough to qualify for inclusion in the sample. Turnbull (1976) used the evocative term "ghosts" to refer to such individuals and we consider a complete data likelihood which includes those individuals. All that is known about these individuals, however, is that their respective survival times are less than L_i, and hence their survival times are left-censored at L_i. The complete data likelihood incorporating these ghosts can be written as

$$L_C(\psi) = L_{C1}(\theta) \cdot L_{C2}(\eta) \ ,$$

where

$$\begin{aligned} L_{C1}(\theta) \ \propto \ & \prod_{i \in \mathcal{R}} \left\{ h^{\delta_i}(X_i|Z_i)\, \mathcal{F}(X_i|Z_i) \ [F(L_i|Z_i)]^{J_i} \right\} \cdot \\ & \prod_{i \in \bar{\mathcal{R}}} \left\{ \prod_{z_1=0}^{1} \left\{ h^{\delta_i}(X_i|(z_1, Z_{i2}))\, \mathcal{F}(X_i|(z_1, Z_{i2})) \ [F(L_i|(z_1, Z_{i2}))]^{J_i} \right\}^{I(Z_{i1}=z_1)} \right\} \ , \end{aligned}$$

and

$$L_{C2}(\eta) \propto \prod_{i \in \mathcal{R}} P(Z_{i1}|Z_{i2})^{J_i+1} \prod_{i \in \bar{\mathcal{R}}} \left\{ \prod_{z_1=0}^{1} P(Z_{i1} = z_1|Z_{i2})^{I(Z_{i1}=z_1)} \right\}^{J_i+1} ,$$

where $\mathcal{F}(t|Z_i) = \exp(-H(t|Z_i))$, $F(t|Z_i) = 1 - \mathcal{F}(t|Z_i)$, and we suppress the dependence on parameters on the right-hand sides for convenience. The primary appeal of this complete data likelihood is that it does not involve probabilities incorporating truncation, as is the case in (2.3), and as a consequence one can factor the complete data likelihood and carry out the maximization step much more easily.

Let the observed data for individual i be denoted by $Y_i = \{(Z_i, R_i, L_i, X_i, \delta_i)\}$ if $R_i = 1$ or $\{(Z_{i2}, R_i, L_i, X_i, \delta_i)\}$ if $R_i = 0$, and let $Y = (Y_1', \ldots, Y_m')'$. We let $\ell_C(\psi) = \log L_C(\psi)$ and define $Q(\psi; \psi^r) = E(\ell_C(\psi)|Y; \psi^r)$ as the conditional expectation of the complete data log-likelihood given the observed data, where the expectation is taken using the estimate ψ^r from the rth iteration of the EM algorithm. We can then write

$$Q(\psi; \psi^r) = Q_1(\theta; \psi^r) + Q_2(\eta; \psi^r) \tag{2.4}$$

with $Q_1(\theta; \psi^r) = E(\ell_{C1}(\theta)|Y; \psi^r)$ given by

$$\begin{aligned} & \sum_{i \in \mathcal{R}} [\delta_i \log h(X_i|Z_i) + \log \mathcal{F}(X_i|Z_i) + \mathcal{J}_i^r \log F(L_i|Z_i)] \qquad (2.5) \\ + & \sum_{i \in \bar{\mathcal{R}}} \zeta_i^r [\delta_i \log h(X_i|(1, Z_{i2})) + \log \mathcal{F}(X_i|(1, Z_{i2})) + \mathcal{J}_i^{1r} \log F(L_i|(1, Z_{i2}))] \\ + & \sum_{i \in \bar{\mathcal{R}}} (1 - \zeta_i^r)[\delta_i \log h(X_i|(0, Z_{i2})) + \log \mathcal{F}(X_i|(0, Z_{i2})) + \mathcal{J}_i^{0r} \log F(L_i|(0, Z_{i2}))] \end{aligned}$$

with

$$\mathcal{J}_i^r = E(J_i | Z_i, R_i = 1, T_i > L_i, X_i, \delta_i; \psi^r)\,,$$

$$\mathcal{J}_i^{zr} = E(J_i | (z, Z_{i2}), R_i = 0, T_i > L_i, X_i, \delta_i; \psi^r)\ ,$$

and

$$\zeta_i^r = E(Z_{i1} | Z_{i2}, R_i = 0, T_i > L_i, X_i, \delta_i; \psi^r)\ .$$

Expressions for these conditional expectations are provided in the Appendix 2A.

Existing software for parametric survival analysis can be used to maximize $Q_1(\theta; \psi^r)$, provided it can handle left-censored observations. This can be achieved by creating pseudo-datasets in which for each $i \in \mathcal{R}$ two lines are generated. One line corresponds to the observed or right-censored observation depending on whether $\delta_i = 1$ or $\delta_i = 0$ respectively. The second line is introduced to correspond to the left-censored failure time of the "ghosts", and has weight $\mathcal{J}_i^r$. For each $i \in \bar{\mathcal{R}}$ four lines are required. First, a contribution for the observed or right-censored failure time is required with the value $Z_{i1} = 1$ and weight ζ_i^r; a second line corresponding to the left-censored observation time with $Z_{i1} = 1$ will have weight $\zeta_i^r \mathcal{J}_i^{1r}$. A second pair of analogous lines is required to reflect the case in which $Z_{i1} = 0$, where the first will have weight $1 - \zeta_i^r$ and correspond to the sampled individual, and the second with weight $(1 - \zeta_i^r)\mathcal{J}_i^{0r}$ corresponding to the left-censored failure time of the "ghosts". Weibull regression models, for example, can be fitted with right and left-censored data, using standard packages for parametric regression including R (survreg), S-PLUS (survReg or censorReg) and SAS (PROC LIFEREG). Alternatively a more flexible piecewise constant baseline hazard function can be adopted, in which case the M-step can be carried out using software for fitting generalized linear regression models. The details on how to construct the data frame for this algorithm are described in Appendix 2B.

The function $Q_2(\eta;\psi^r) = E(\ell_{C2}(\eta)|Y;\psi^r)$ in (2.4) is

$$\begin{aligned}
&\sum_{i\in\mathcal{R}} \left[(\mathcal{J}_i^r + 1)\log P(Z_{i1}|Z_{i2})\right] \\
+ &\sum_{i\in\bar{\mathcal{R}}}\sum_{z_1=0}^{1} [\zeta_i^r]^{z_1}\left[1-\zeta_i^r\right]^{1-z_1}(\mathcal{J}_i^{z_1 r}+1)\log P(Z_{i1}=z_1|Z_{i2})
\end{aligned} \tag{2.6}$$

and can also be maximized using software for logistic regression by creating a pseudo-dataset with one line for each individual $i \in \mathcal{R}$ with weight $\mathcal{J}_i^r + 1$ and observed value of Z_{i1}. For each $i \in \bar{\mathcal{R}}$ two lines are required: one with weight $\zeta_i^r(\mathcal{J}_i^{1r}+1)$ and $Z_{i1} = 1$, and one with weight $(1-\zeta_i^r)(\mathcal{J}_i^{0r}+1)$ and $Z_{i1} = 0$. Specification of a quasi-likelihood model with a logit link function and variance function $V(\mu) = \mu(1-\mu)$ will yield the updated estimate η^{r+1}.

2.1.3 Empirical Performance of the Proposed Method

Here we evaluate the frequency properties of estimators obtained by the proposed algorithm, and we begin by a description of the method of data generation. We let $P(Z_{ik} = 1) = 0.5$, $k = 1, 2$ and the odds ratio for the association between Z_{i1} and Z_{i2} be 2, so $\eta_0 = -0.347$ and $\eta_1 = \log 2$. Suppose the survival time is Weibull distributed with hazard

$$h(s|Z_i;\theta) = h_0(s;\alpha)\exp(Z_i'\beta) \ ,$$

where $h_0(s;\alpha) = \rho\kappa(\rho s)^{\kappa-1}$, $\alpha_1 = \log\rho$, $\alpha_2 = \log\kappa$ and $\alpha = (\alpha_1, \alpha_2)'$; we set $\rho = 1$ and $\kappa = 1.5$. We consider a calendar time origin of zero, and suppose disease onset happens according to a stationary process in the population giving $D_i \sim \text{Unif}(0, A)$ where $D_i \perp Z_i$. The desired degree of left truncation is obtained by choosing A to satisfy

$$T\% = 100 \cdot (1 - P(E_i > A)) = 100 \cdot (1 - E_{Z_i}\left[E_{D_i|Z_i} P(T_i > A - D_i|D_i, Z_i)\right])$$

where $T\%$ is the truncation percentage; we consider $T\%$=25 and 50.

To generate covariate data compatible with the sampling requirement, given D_i, we generate Z_i according to $P(Z_i|T_i > L_i)$. We then generate $U_i \sim \text{Unif}(0, 1)$, and solve for the failure time T_i in $U_i = \exp(-H(L_i, T_i|Z_i))$. The probability that an individual included in the study is administratively censored given the disease onset time D_i and covariates Z_i, is

$$P(E_i > B|E_i > A, D_i, Z_i) = P(T_i > B - D_i|D_i, Z_i)/P(T_i > L_i|D_i, Z_i) \; .$$

We obtain the administrative censoring rate given Z_i by

$$P(E_i > B|E_i > A, Z_i) = E_{D_i|E_i>A,Z_i}\left[P(E_i > B|E_i > A, D_i, Z_i)\right] \; ,$$

and solve for B in

$$100 \cdot P(E_i > B|E_i > A) = 100 \cdot E_{Z|E_i>A} P(E_i > B|E_i > A, Z) \; ,$$

to obtain the desired rate, where

$$P(Z_i|E_i > A) = P(E_i > A|Z_i)\, P(Z_i) / \sum_{Z_i} P(E_i > A|Z_i) P(Z_i)\ .$$

Additional random censoring is incorporated by generating an exponential withdrawal time to give a net censoring rate of 25%.

To simulate incomplete data for Z_1, we assume a missing at random mechanism with

$$P(R_i = 1|Z_i, D_i, E_i > A, X_i, \delta_i) = P(R_i = 1|Z_{i2})$$

and let

$$\text{logit} P(R_i = 1|Z_{i2}) = \gamma_0 + \gamma_1 z_{i2}\ .$$

The net frequency of complete data in the sample is then

$$P(R_i = 1) = E_{Z_{i2}|E_i>A}(P(R_i = 1|Z_{i2}))$$

If we fix $\gamma_1 = \log 4$, and the percentage of missing covariate values at M%, one can solve for γ_0 correspondingly; we set M% = 25, 50 (i.e, $P(R_i = 1) = 0.75, 0.50$). Five hundred datasets (nsim = 500) of $m = 500$ individuals were simulated.

For each simulated dataset we conducted four analyses: *i)* an analysis based on the sample including all values of the covariates (NO MISS), possible because this is a simulation study, *ii)* a complete case (CC) analysis which restricts attention to individuals in $\mathcal{R}$, *iii)* an analysis based on a misspecified likelihood (MISSPEC) with the form of (2.2) but with $P(Z_{i1}|Z_{i2};\eta)$ in place of $P(Z_{i1}|Z_{i2}, T_i > L_i;\psi)$, and *iv)* the proposed EM algorithm (EM). The analysis in *ii)* is based on a correctly specified model and yields consistent estimates of θ under this missing data mechanism, but it is inefficient since it disregards data from individuals in $\bar{\mathcal{R}}$. The analysis in *iii)* is based on the correct model for the survival time given the covariates but an incorrect model for the covariates since the population covariate distribution is used; the estimator for ψ is therefore inconsistent. For this analysis, the asymptotic theory on the behavior of maximum likelihood estimators under misspecified models could be exploited (Cox, 1961; White, 1982; Rotnitzky and Wypij, 1994), but we elect to study this through simulation. The analysis based on *iv)* is correct and so a consistent estimator of ψ is obtained, which should be more efficient than the estimator from the complete case analysis. The simulation study sheds light on the bias and efficiency trade-offs for these various approaches. Across all parameter configurations considered here, the proposed EM algorithm converged reasonably quickly with longer computing times occurring under higher rates of missing data and left truncation.

The empirical biases and empirical standard errors of the estimators from all four approaches are displayed in Table 2.1; we do not report performance of estimators of η in the first two rows of each configuration (NO MISS and CC) since the covariate distribution would not typically be modeled in these settings. The analysis based on subjects with complete data yielded estimates which had negligible empirical bias for the parameters of interest, as expected. The complete case analysis leads to estimates with negligible empirical bias but lower efficiency reflected by the greater empirical standard errors. Under the misspecified model, there were small empirical biases of estimators for θ (most appreciable for the α components), and much larger empirical biases of estimators for η, reflecting misspecification of the covariate model. As expected the estimates from the proposed EM had negligible empirical biases for the components of θ and η, and empirical standard errors which were smaller than those from the complete case analysis. Note that the efficiency gains from the correct analysis were appreciable for all elements of θ except for β_1, the regression coefficient of the partially observed covariate. Broadly similar conclusions were seen in the case $\eta_1 = 0$ (i.e. when covariates are independent) with slightly lower improvement in efficiency with the proposed EM algorithm (results not reported).

Table 2.1: Empirical biases ($\times 10^2$) and standard errors of estimators from analysis in the absence of missing data, the complete case analysis, a misspecified model, and the correct missing data model fitted via an EM algorithm of Section 2.1.2; $\alpha_1 = \log \rho = 0$, $\alpha_2 = \log \kappa = 0.405$, $\beta_1 = 0.693$, $\beta_2 = 0.405$, $\eta_0 = -0.347$, $\eta_1 = 0.693$, $\gamma_1 = 1.386$, 25% net censoring, $m = 500$, $nsim = 500$

			α_1		α_2		β_1		β_2		η_0		η_1	
T%	M%	METHOD[†]	BIAS	ESE	BIAS	ESE	BIAS	ESE	BIAS	ESE	BIAS	ESE	BIAS	ESE
50	50	NO MISS	-0.13	0.067	0.40	0.066	-0.19	0.106	0.78	0.106	- - -	- - - -	- - -	- - - -
		CC	0.31	0.107	0.70	0.096	0.35	0.149	0.46	0.155	- - -	- - - -	- - -	- - - -
		MISSPEC	1.81	0.076	-2.27	0.070	-1.41	0.144	0.16	0.110	-54.65	0.221	-10.99	0.283
		EM	-0.26	0.079	0.70	0.072	0.46	0.150	0.76	0.109	-1.86	0.313	0.54	0.394
50	25	NO MISS	-0.13	0.067	0.40	0.066	-0.19	0.106	0.78	0.106	- - -	- - - -	- - -	- - - -
		CC	-0.24	0.078	0.74	0.077	-0.51	0.123	1.17	0.121	- - -	- - - -	- - -	- - - -
		MISSPEC	0.84	0.070	-0.97	0.068	-1.23	0.119	0.67	0.109	-53.60	0.164	-11.66	0.217
		EM	-0.07	0.070	0.37	0.068	-0.56	0.121	0.97	0.108	-1.28	0.256	-0.92	0.303
25	50	NO MISS	-0.02	0.059	0.54	0.064	-0.73	0.100	0.97	0.103	- - -	- - - -	- - -	- - - -
		CC	0.40	0.098	0.96	0.092	0.12	0.146	0.31	0.156	- - -	- - - -	- - -	- - - -
		MISSPEC	0.14	0.068	0.17	0.068	-0.16	0.144	0.64	0.107	-18.56	0.216	-5.56	0.269
		EM	-0.07	0.068	0.89	0.068	0.25	0.145	0.70	0.107	-0.95	0.219	1.50	0.271
25	25	NO MISS	-0.02	0.059	0.54	0.064	-0.73	0.100	0.97	0.103	- - -	- - - -	- - -	- - - -
		CC	-0.01	0.071	0.84	0.074	-1.00	0.117	1.34	0.119	- - -	- - - -	- - -	- - - -
		MISSPEC	0.07	0.062	0.22	0.065	-1.05	0.115	1.03	0.108	-17.11	0.162	-6.81	0.220
		EM	-0.02	0.062	0.53	0.065	-0.92	0.116	1.07	0.108	0.31	0.168	0.04	0.224

[†] NO MISS is analysis in the absence of missing data, CC is complete case analysis, MISSPEC is based on a misspecified covariate model ignoring truncation, and EM is the correct algorithm described in Section 2.1.2.

2.2 Sub-group Analysis in Clinical Trials

When assessing a treatment effect on a time to event response in randomized trials it is customary to define the time origin as the date of randomization. When this time origin is adopted, one is implicitly making treatment comparisons after marginalizing over the left truncation times as well as any covariates. The time of randomization is the time at which evidence of a treatment effect could emerge and so from this standpoint it has face validity. Often however, protocols dictate that analyses be stratified according to risk factors whose effects are manifest at the time of disease onset, and hence can influence whether individuals will satisfy the entry criteria for the clinical trial. In cancer trials, for example, it may be appropriate to stratify on tumour type, or HER2 (human epidermal growth factor receptor 2) status (Gennari et al., 2008). Important secondary analyses may in fact be directed at assessing treatment effects by HER2 status and investigating whether there is evidence of differences in treatment effect between strata defined by HER2 status. The most sensible time origin for these types of analyses is the time of disease onset, and in fact this is essential to adopt to ensure valid covariate models when such data are incomplete.

We consider here the problem of conducting pre-specified subgroup analyses in which the subgroups are defined by patient characteristics and have biological rationale (Yusuf et al., 1991). We presume that the other criteria for valid sub-group analyses are satisfied and thus the trial is compliant with the CONSORT statement (Moher et al., 2001). Consider the setting of Section 2.1 with D_i, $(Z_{i1}, Z_{i2})'$ and R_i defined as in Section 2.1.1 but now suppose that at the time of accrual individuals are randomized to one of two treatment arms. To accommodate the fact that treatment does not begin until recruitment we define a time-dependent variable $Z_{i3}(s)$ such that $Z_{i3}(s) = 0$ for $0 < s < L_i$ and for $L_i \leq s$, $Z_{i3}(s) = 1$ if individual i is randomized to receive an experimental treatment, and $Z_{i3}(s) = 0$ otherwise. We then let $Z_i(s) = (Z_{i1}, Z_{i2}, Z_{i3}(s))'$ denote the full covariate vector and $Z_i^*(s) = (Z_{i2}, Z_{i3}(s))'$ denote a sub-vector containing covariates which are always observed. Next let $\bar{Z}_i(s) = \{Z_i(u), 0 \leq u \leq s\}$ and $\bar{Z}_i^*(s) = \{Z_i^*(u), 0 \leq u \leq s\}$ denote the corresponding histories at s, and $\bar{Z}_i = \bar{Z}_i(\infty)$ and $\bar{Z}_i^* = \bar{Z}_i^*(\infty)$ denote the full paths of the respective covariates.

If interest lies in estimating the effect of treatment according to subgroup defined by Z_{i1}, then a natural model is

$$h(s|Z_i(s);\theta) = h_0(s;\alpha)\exp(Z_{i1}\beta_1 + Z_{i2}\beta_2 + Z_{i3}(s)\beta_3 + Z_{i1}Z_{i3}(s)\beta_4)\ . \tag{2.7}$$

If we let $H_i(t;\theta) = H(t|\bar{Z}_i(t);\theta) = \int_0^t h(s|Z_i(s))ds$, then the complete data likelihood is

$$\begin{aligned} L_C(\psi) \ \propto \ & \prod_{i \in \mathcal{R}} \Big\{ h^{\delta_i}(X_i|Z_i(X_i)) \exp(-H_i(X_i)) \left[1 - \exp(-H_i(L_i))\right]^{J_i} \ P(Z_{i1}|\bar{Z}_i^*\})^{J_i+1} \Big\} \cdot \\ & \prod_{i \in \bar{\mathcal{R}}} \Big\{ h^{\delta_i}(X_i|Z_i(X_i)) \exp(-H_i(X_i)) \left[1 - \exp(-H_i(L_i))\right]^{J_i} \ P(Z_{i1}|\bar{Z}_i^*)^{J_i+1} \Big\}^{Z_{i1}} \cdot \\ & \prod_{i \in \bar{\mathcal{R}}} \Big\{ h^{\delta_i}(X_i|Z_i(X_i)) \exp(-H_i(X_i)) \left[1 - \exp(-H_i(L_i))\right]^{J_i} \ P(Z_{i1}|\bar{Z}_i^*)^{J_i+1} \Big\}^{(1-Z_{i1})} \ . \end{aligned}$$

Note that $E(J_i|\bar{Z}_i, T_i > L_i; \psi^r)$ and $E(J_i|Z_{i1} = z, \bar{Z}_i^*, T_i > L_i; \psi^r)$ are given by (2A.1) and (2A.2) respectively since the treatment variable is defined to be zero prior to the left truncation time. Here, however, $\zeta_i^r = E(Z_{i1}|\bar{Z}_i^*, R_i = 0, T_i > L_i, X_i, \delta_i; \psi^r)$ is

$$\frac{h^{\delta_i}(X_i|(1, Z_i^*(X_i)); \theta^r) \exp(-H(X_i|(1, \bar{Z}_i^*(X_i)); \theta^r))P(Z_{i1} = 1|Z_{i2}; \eta^r)}{\sum_{z=0}^{1} h^{\delta_i}(X_i|(z, Z_i^*(X_i)); \theta^r) \exp(-H(X_i|(z, \bar{Z}_i^*(X_i)); \theta^r))P(Z_{i1} = z|Z_{i2}; \eta^r)} .$$

Calculations like those of Section 2.1.3 can be carried out to satisfy the 25% censoring rate and particular truncation and marginal missing data rates.

Analyses based on the full sample with no missing covariates (NO MISS), a complete case analysis (CC) and the proposed EM algorithm were carried out. In Table 2.2 the empirical biases and standard errors are reported for truncation and missing data rates of 25% and 50% respectively for 500 simulated datasets of $m = 500$ individuals. The estimators of β_3 and $\beta_3 + \beta_4$, the two estimates of treatment effect for individuals with $Z_1 = 0$ and $Z_1 = 1$ respectively, are of greatest interest. As was the case in Section 2.1, we see small biases in these three analyses with the proposed algorithm giving improved efficiency over the complete case analysis for most parameters.

Table 2.2: Empirical biases ($\times 10^2$) and standard errors of estimators from analysis in the absence of missing data, the complete case analysis, and the correct missing data model fitted via the EM algorithm of Section 2.2; $\alpha_1 = \log \rho = 0$, $\alpha_2 = \log \kappa = 0.405$, $\beta_1 = 0.693$, $\beta_2 = 0.405$, $\beta_3 = 0$, $\beta_4 = -0.693$, $\eta_0 = -0.347$, $\eta_1 = 0.693$, $\gamma_1 = 1.386$, 25% net censoring, $m = 500$, $nsim = 500$

			α_1		α_2		β_1		β_2		β_3		$\beta_3 + \beta_4$	
T%	M%	METHOD[†]	BIAS	ESE	BIAS	ESE	BIAS	ESE	BIAS	ESE	BIAS	ESE	BIAS	ESE
50	50	NO MISS	-0.35	0.079	1.16	0.066	0.32	0.141	0.97	0.101	-1.23	0.123	-0.92	0.171
		CC	-1.05	0.120	2.21	0.090	1.55	0.201	1.67	0.150	-0.49	0.187	-1.60	0.227
		EM	-0.50	0.089	1.42	0.070	0.48	0.187	0.94	0.105	-1.05	0.141	-0.66	0.202
50	25	NO MISS	-0.35	0.079	1.16	0.066	0.32	0.141	0.97	0.101	-1.23	0.123	-0.92	0.171
		CC	0.26	0.095	1.21	0.076	-0.28	0.169	0.99	0.116	-2.05	0.142	-1.32	0.188
		EM	-0.26	0.083	1.18	0.068	-0.14	0.162	1.06	0.101	-1.29	0.126	-0.70	0.179
25	50	NO MISS	0.15	0.076	1.31	0.065	-0.48	0.142	0.95	0.103	-1.46	0.142	-1.03	0.152
		CC	-0.28	0.118	2.30	0.089	0.74	0.202	1.43	0.156	-0.63	0.204	-2.10	0.206
		EM	0.09	0.084	1.62	0.067	0.10	0.188	1.04	0.108	-1.06	0.166	-1.84	0.188
25	25	NO MISS	0.15	0.076	1.31	0.065	-0.48	0.142	0.95	0.103	-1.46	0.142	-1.03	0.152
		CC	0.57	0.094	1.36	0.075	-0.32	0.172	0.81	0.123	-2.24	0.164	-1.57	0.174
		EM	0.15	0.079	1.43	0.066	-0.27	0.166	0.93	0.105	-1.58	0.148	-1.06	0.165

[†] NO MISS is analysis in the absence of missing data, CC is complete case analysis, and EM is the correct algorithm described in Section 2.2.

2.3 Analysis of Data from a Metastatic Cancer Trial

Table 2.3 gives the results of fitting a model based on (2.7) under the complete case analysis and fitting a model based on the proposed EM algorithm; standard errors were obtained based on 500 bootstrap samples for the proposed EM algorithm. Note that there is no evidence of a treatment effect for any patients irrespective of estrogen receptor (ER) status. This is not surprising since this was a palliative trial in which the aim was to improve quality of life. Among individuals who are ER positive, the relative risks were close to one for both analyses, but the point estimate for ER negative patients suggests a 19.5% relative risk reduction based on the complete case analysis (p=0.491). The proposed EM algorithm, which exploits the information about the missing ER status from the left truncation time, gives a relative risk reduction estimate of 25.9% (95% CI: 0.415, 1.327; p=0.311).

Table 2.3: Relative risk estimates from complete case analysis and the proposed EM algorithm for fitting a Weibull proportional hazards model with ER status as the partially observed covariate (Z_1), age at diagnosis ($Z_2 = I(\text{age} \geq 50)$), treatment, and an ER status by treatment interaction; standard errors based on 500 bootstrap samples for proposed EM algorithm

	ER Negative			ER Positive		
Method	RR	95% CI	p-value	RR	95% CI	p-value
Complete Case	0.805	(0.433, 1.493)	0.491	1.048	(0.792, 1.387)	0.741
Proposed EM	0.741	(0.415, 1.322)	0.311	1.029	(0.775, 1.367)	0.842

2.4 Remarks

We have considered issues in the analysis of incomplete covariate data under a form of response-biased sampling which is widely encountered in epidemiologic research as well as clinical trials. This response-bias arises any time that there are conditions imposed on individuals for inclusion in a study, but in prevalent cohort studies the condition that individuals be event-free (e.g. alive) at the time of diagnosis leads to left-truncated event times. Left truncation can readily be handled using standard software when covariates are complete (Klein and Moeschberger, 1997). When covariates are incompletely observed, one strategy is to specify an observed data likelihood based on the joint distribution of the response times and the covariates. This can be challenging because the correct covariate distribution must condition on the selection criterion being satisfied and therefore involves parameters of the survival distribution. To address this we describe an EM algorithm based on a complete data likelihood including contributions from individuals who did not satisfy the truncation condition. Standard software for parametric survival analysis which handles left censoring can then be used at the maximization step. The proposed algorithm is shown to perform well for both the setting of prevalent cohort studies and clinical trials where subgroup analyses are of interest but covariates are incomplete.

Appendix 2A: Additional Details for the EM Algorithm

Appendix 2A.1: Form of Conditional Expectations

For each $i \in \mathcal{R}$, the only "missing" information is J_i, the number of "ghosts" who did not satisfy the truncation condition of the respective individual. If ψ^r denotes the parameter estimate at the rth iteration of the EM algorithm, to take the relevant expectations in (2.5) and (2.6) we note $E(J_i|Z_i, R_i = 1, T_i > L_i, X_i, \delta_i; \psi^r) = E(J_i|Z_i, T_i > L_i; \psi^r)$ and that

$$\begin{aligned} \mathcal{J}_i^r &= E(J_i|Z_i, T_i > L_i; \psi^r) \qquad (2A.1) \\ &= \frac{P(T_i < L_i|Z_i; \theta^r)}{P(T_i \geq L_i|Z_i; \theta^r)} = \frac{1 - \exp(-H(L_i|Z_i; \theta^r))}{\exp(-H(L_i|Z_i; \theta^r))} \ , \quad \text{for } i \in \mathcal{R} \ . \end{aligned}$$

For $i \in \bar{\mathcal{R}}$, in addition to the number of "ghosts", the value of Z_{i1} is missing. We note $E(J_i|(z, Z_{i2}), R_i = 0, T_i > L_i, X_i, \delta_i; \psi^r) = E(J_i|(z, Z_{i2}), T_i > L_i; \psi^r)$ and let

$$\begin{aligned} & \mathcal{J}_i^{zr} = E(J_i|(z, Z_{i2}), T_i > L_i; \psi^r) \\ = \ & \frac{P(T_i < L_i|(z, Z_{i2}); \theta^r)}{P(T_i \geq L_i|(z, Z_{i2}); \theta^r)} = \frac{1 - \exp(-H(L_i|(z, Z_{i2}); \theta^r))}{\exp(-H(L_i|(z, Z_{i2}); \theta^r))} \ , \quad \text{for } i \in \bar{\mathcal{R}}, \quad (2A.2) \end{aligned}$$

denote the expectation conditional on a particular value of $Z_i = (z, Z_{i2})'$, $z = 0, 1$. We then note $\zeta_i^r = E(Z_{i1}|Z_{i2}, R_i = 0, T_i > L_i, X_i, \delta_i; \psi^r) = E(Z_{i1}|Z_{i2}, T_i > L_i; \psi^r)$, for $i \in \bar{\mathcal{R}}$, which we obtain through

$$\zeta_i^r \quad = \quad \frac{h^{\delta_i}(X_i|(1,Z_{i2});\theta^r)\mathcal{F}(X_i|(1,Z_{i2});\theta^r)\,P(Z_{i1}=1|Z_{i2};\eta^r)}{\displaystyle\sum_{z=0}^{1} h^{\delta_i}(X_i|(z,Z_{i2});\theta^r)\mathcal{F}(X_i|(z,Z_{i2});\theta^r)\,P(Z_{i1}=z|Z_{i2};\eta^r)}\ . \tag{2A.3}$$

Appendix 2A.2: Estimation of Information-Based Variances

Standard errors can be obtained using the nonparametric bootstrap as done in the example, or using the approach of Louis (1982), implemented as follows. Let $U(\psi) = (U_1'(\theta), U_2'(\eta))'$ where $U_1(\theta) = \partial \log L_C(\psi)/\partial\theta$ and $U_2(\eta) = \partial \log L_C(\psi)/\partial\eta$, and

$$I(\psi) = -\partial U(\psi)/\partial\psi' = \begin{pmatrix} I_1(\theta) & 0 \\ 0 & I_2(\eta) \end{pmatrix} \tag{2A.4}$$

where $I_1(\theta) = -\partial U_1(\theta)/\partial\theta'$ and $I_2(\eta) = -\partial U_2(\eta)/\partial\eta'$. Then if $\mathcal{I}(\psi)$ is the information matrix from the observed data likelihood (2.2),

$$\mathcal{I}(\psi) = E_M\{I(\psi)|Y\} - E_M\{U(\psi)U'(\psi)|Y\} \tag{2A.5}$$

where M represents the missing data which is simply the number of "ghosts" J for individuals in $\mathcal{R}$, and is the number of ghosts and the covariate Z_1 for individuals in $\bar{\mathcal{R}}$. The expectations are carried out by individual, given their respective observed data. The first term in (2A.5) for example, is simply obtained by extracting the usual observed information matrices from the two analyses estimating θ and η at the final iteration of the EM algorithm and the second term is given by taking the outer product of the stacked score vectors and averaging using the weights estimated at the final iteration.

Appendix 2B: An EM Algorithm for Piecewise Exponential Models

Here we consider an extension of the algorithm of Section 2.1, i.e., Section 2.2 of Shen and Cook (2013b), to deal with more flexible weakly parametric proportional hazards models with piecewise constant baseline hazard functions. Let $0 = b_0 < b_1 < \ldots < b_{K-1} < b_K = \infty$ denote pre-specified cut points giving K sub-intervals $\mathcal{B}_k = [b_{k-1}, b_k)$, $k = 1, \ldots, K$. The baseline function has the form $h_0(t) = \alpha_k$ if $t \in B_k$, $k = 1, \ldots, K$.

Let $\mathcal{A}_i = [L_i, \infty)$ denote the truncation region for individual i, and $\mathcal{A}_i^c = [0, L_i)$. In the observational setting of Section 2.1, a complete data likelihood is given, but here we replace the term $F(L_i|Z_i)^{J_i}$ with $\prod_{j=1}^{J_i} f(t_{ij}|Z_i)$, where t_{ij} is the failure time of the jth "ghost" for individual i known to fall in $\mathcal{A}_i^c$. The reason for considering a different form is that the maximization step of the complete data likelihood becomes trivial under a piecewise constant model if the failure times are observed; this can be exploited in the algorithm that follows.

Let $I_k(t) = I(t \in \mathcal{B}_k)$ and let $w_k(t) = \int_0^t I_k(u)du$ denote the amount of time that a particular subject is at risk in $\mathcal{B}_k$ over the interval $[0, t)$. We can then write $f(t|Z_i) = h(t|Z_i) \exp(-H(t|Z_i))$ as

$$f(t|Z_i) = \left[\prod_{k=1}^{K} \left[\alpha_k \exp(Z_i'\beta) \right]^{I_k(t)} \right] \exp\Big(- \left[\textstyle\sum_{k=1}^{K} w_k(t)\alpha_k \right] \exp(Z_i'\beta) \Big). \qquad (2B.1)$$

Let $\delta_{ik} = I_k(X_i)$ indicate whether the observation time $X_i = \min(T_i, C_i)$ is in interval $\mathcal{B}_k$ for individual i, and let $S_{ik} = \int_0^{X_i} I(u \in \mathcal{B}_k)du$ denote the total time individual i was at risk of failure during the interval $\mathcal{B}_k$. By replacing $F(L_i|Z_i)^{J_i}$ with $\prod_{j=1}^{J_i} f(t_{ij}|Z_i)$ in the complete data likelihood of Section 2.1 and by taking the logarithm, we obtain

$$\begin{aligned}
\ell_C(\psi) &= \sum_{i \in \mathcal{R}} \Big\{ \sum_{k=1}^{K} \Big[\delta_i \delta_{ik} \Big(\log \alpha_k + Z_i^{'}\beta \Big) - \alpha_k S_{ik} e^{Z_i^{'}\beta} \Big] \\
&\qquad + \sum_{j=1}^{J_i} \sum_{k=1}^{K} \Big[I_k(t_{ij}) \Big(\log \alpha_k + Z_i^{'}\beta \Big) - w_k(t_{ij}) \alpha_k e^{Z_i^{'}\beta} \Big] + (J_i + 1) \log P(Z_{i1}|Z_{i2}) \Big\} \\
&+ \sum_{i \in \bar{\mathcal{R}}} \Big[Z_{i1} \Big\{ \sum_{k=1}^{K} \Big[\delta_i \delta_{ik} \Big(\log \alpha_k + Z_i^{'}\beta \Big) - \alpha_k S_{ik} e^{Z_i^{'}\beta} \Big] \\
&\qquad + \sum_{j=1}^{J_i} \sum_{k=1}^{K} \Big[I_k(t_{ij}) \Big(\log \alpha_k + Z_i^{'}\beta \Big) - w_k(t_{ij}) \alpha_k e^{Z_i^{'}\beta} \Big] + (J_i + 1) \log P(Z_{i1}|Z_{i2}) \Big\} \\
&+ (1 - Z_{i1}) \Big\{ \sum_{k=1}^{K} \Big[\delta_i \delta_{ik} \Big(\log \alpha_k + Z_i^{'}\beta \Big) - \alpha_k S_{ik} e^{Z_i^{'}\beta} \Big] \\
&\qquad + \sum_{j=1}^{J_i} \sum_{k=1}^{K} \Big[I_k(t_{ij}) \Big(\log \alpha_k + Z_i^{'}\beta \Big) - w_k(t_{ij}) \alpha_k e^{Z_i^{'}\beta} \Big] + (J_i + 1) \log P(Z_{i1}|Z_{i2}) \Big\} \Big],
\end{aligned}$$

where the event time for the jth ghost corresponding to individual i, t_{ij}, is only known to be in the interval $\mathcal{A}_i^c = [0, L_i)$. As before we can split this likelihood into two parts

$$\ell_C(\psi) = \ell_{C1}(\theta) + \ell_{C2}(\eta) \ ,$$

where $\ell_{C1}(\theta)$ is

$$\begin{aligned}
& \sum_{i\in\mathcal{R}}\sum_{k=1}^{K}\Big\{\Big[\delta_i\delta_{ik}\log(\alpha_k e^{Z_i'\beta})-\alpha_k S_{ik}e^{Z_i'\beta}\Big]+\sum_{j=1}^{J_i}\Big[I_k(t_{ij})\log(\alpha_k e^{Z_i'\beta})-w_k(t_{ij})\alpha_k e^{Z_i'\beta}\Big]\Big\} \quad \text{(2B.2)}\\
+ & \sum_{i\in\bar{\mathcal{R}}}\Big\{Z_{i1}\sum_{k=1}^{K}\Big\{\Big[\delta_i\delta_{ik}\log(\alpha_k e^{Z_i'\beta})-\alpha_k S_{ik}e^{Z_i'\beta}\Big]+\sum_{j=1}^{J_i}\Big[I_k(t_{ij})\log(\alpha_k e^{Z_i'\beta})-w_k(t_{ij})\alpha_k e^{Z_i'\beta}\Big]\Big\}\\
+ & (1-Z_{i1})\sum_{k=1}^{K}\Big\{\Big[\delta_i\delta_{ik}\log(\alpha_k e^{Z_i'\beta})-\alpha_k S_{ik}e^{Z_i'\beta}\Big]+\sum_{j=1}^{J_i}\Big[I_k(t_{ij})\log(\alpha_k e^{Z_i'\beta})-w_k(t_{ij})\alpha_k e^{Z_i'\beta}\Big]\Big\}\Big\},
\end{aligned}$$

and $\ell_{C2}(\eta)$ is given by (2.5). Thus

$$Q(\psi;\psi^r)=E(\ell_C(\psi)|Y;\psi^r)=Q_1(\theta;\psi^r)+Q_2(\eta;\psi^r)\ ,$$

where as before $Q_1(\theta;\psi^r)=E(\ell_{C_1}(\theta)|Y;\psi^r)$, and $Q_2(\eta;\psi^r)=E(\ell_{C_2}(\eta)|Y;\psi^r)$. At the rth step of the EM algorithm, we need $\mathcal{J}_i^r$, $\mathcal{J}_i^{1r}$, $\mathcal{J}_i^{0r}$ and ζ_i^r, given by (A.1), (A.2) and (A.3) respectively. The expectations regarding t_{ij} are given as follows. If the complement of the truncation interval does not intersect with $\mathcal{B}_k$ (i.e. $\mathcal{C}_{ijk}=\mathcal{A}_i^c\cap\mathcal{B}_k=\emptyset$ because $b_{k-1}>L_i$), then $E(I_k(T_{ij})|Z_i,T_{ij}<L_i,J_i)=0$. If $b_{k-1}<L_i$, $\mathcal{C}_{ijk}=\mathcal{A}_i^c\cap\mathcal{B}_k=[L_{ijk},R_{ijk})\neq\emptyset$, where $L_{ijk}=\max(b_{k-1},0)=b_{k-1}$, and $R_{ijk}=\min(b_k,L_i)$. We then take the expectation of (2B.2) at the rth step of the EM algorithm, using

$$\begin{aligned}
\iota_{ik}^r &= E(I_k(t_{ij})|Z_i,R_i=1,T_{ij}<L_i,J_i;\psi^r)=P(T_{ij}\in B_k|Z_i,T_{ij}<L_i,J_i;\psi^r)\\
&= \frac{\mathcal{F}(L_{ijk}|Z_i;\psi^r)-\mathcal{F}(R_{ijk}|Z_i;\psi^r)}{\mathcal{F}(0|Z_i;\psi^r)-\mathcal{F}(L_i|Z_i;\psi^r)}=\frac{\mathcal{F}(b_{k-1}|Z_i;\psi^r)-\mathcal{F}(\min(b_k,L_i)|Z_i;\psi^r)}{1-\mathcal{F}(L_i|Z_i;\psi^r)}\ ,
\end{aligned}$$

and

$$\begin{aligned}\iota_{ik}^{zr} &= E(I_k(t_{ij})|(z, z_{i2}), R_i = 0, T_{ij} < L_i, J_i; \psi^r)\\ &= \frac{\mathcal{F}(b_{k-1}|(z, z_{i2}); \psi^r) - \mathcal{F}(\min(b_k, L_i)|(z, z_{i2}); \psi^r)}{1 - \mathcal{F}(L_i|(z, z_{i2}); \psi^r)},\end{aligned}$$

where $z = 0, 1$.

Regarding the time at risk, $\mathcal{C}_{ijk} = \mathcal{A}_i^c \cap \mathcal{B}_k = \emptyset$, (i.e., $L_i < b_{k-1}$), each ghost j corresponding to individual i, $j = 1, \ldots, J_i$ failed before entering interval $\mathcal{B}_k$, and thus they were never at risk of failure in $\mathcal{B}_k$; in that case, $E(w_k(t_{ij})|Z_i, T_{ij} < L_i, J_i) = 0$. If $\mathcal{C}_{ijk} = \mathcal{A}_i^c \cap \mathcal{B}_k = [L_{ijk}, R_{ijk}) \neq \emptyset$, $b_{k-1} < L_i$, it is possible that they could have failed before entering $\mathcal{B}_k$, in which case there is no period at risk corresponding to the interval $[b_{k-1}, b_k)$. At the rth step of the EM algorithm, we have,

$$\omega_{ik}^r = E(w_k(t_{ij})|Z_i, R_i = 1, T_{ij} < L_i, J_i; \psi^r) = \int_{b_{k-1}}^{\min(b_k, L_i)} \frac{\mathcal{F}(u|Z_i; \psi^r) - \mathcal{F}(L_i|Z_i; \psi^r)}{1 - \mathcal{F}(L_i|Z_i; \psi^r)} du,$$

and

$$\begin{aligned}\omega_{ik}^{zr} &= E(w_k(t_{ij})|(z, z_{i2}), R_i = 0, T_{ij} < L_i, J_i; \psi^r)\\ &= \int_{b_{k-1}}^{\min(b_k, L_i)} \frac{\mathcal{F}(u|(z, z_{i2}); \psi^r) - \mathcal{F}(L_i|(z, z_{i2}); \psi^r)}{1 - \mathcal{F}(L_i|(z, z_{i2}); \psi^r)} du, \quad z = 0, 1.\end{aligned}$$

Let $K_i = \max\{k : b_{k-1} < X_i\}$ be the maximum interval over which individual i is known to have been at risk and $K_{ij} = \max\{k : b_{k-1} < L_i\}$ denote the the maximum interval over which the ghosts for individual i could have been at risk. Furthermore, let

$$Q_{i1k}(\theta;\psi^r) = \delta_i\delta_{ik}\left(\log\alpha_k + Z_i^{'}\beta\right) - \alpha_k\exp(Z_i^{'}\beta + \log S_{ik})$$

be the expectation of this kth element of the first term in the first row of (2B.2) and let

$$G_{i1k}(\theta;\psi^r) = \mathcal{J}_i^r\left[\iota_{ik}^r\left(\log\alpha_k + Z_i^{'}\beta\right) - \alpha_k\exp(Z_i^{'}\beta + \log\omega_{ik}^r)\right]$$

denote the expectation of the kth element in the second term in the first row of (2B.2). Then if $i \in \mathcal{R}$,

$$Q_{i1}(\theta;\psi^r) = \sum_{k=1}^{K_i} Q_{i1k}(\theta;\psi^r) + \sum_{k=1}^{K_{ij}} G_{i1k}(\theta;\psi^r). \tag{2B.3}$$

Similarly, for $i \in \bar{\mathcal{R}}$, let

$$Q_{i1k}^z(\theta;\psi^r) = \delta_i\delta_{ik}\left(\log\alpha_k + (z, z_{i2})^{'}\beta\right) - \alpha_k\exp(z\beta_1 + z_{i2}^{\prime}\beta_2 + \log S_{ik})\ ,$$

and

$$G_{i1k}^z(\theta;\psi^r) = \mathcal{J}_{iz}^r\left[\iota_{ik}^{zr}\left(\log\alpha_k + (z, z_{i2})^{'}\beta\right) - \alpha_k\exp(z\beta_1 + z_{i2}\beta_2 + \log\omega_{ik}^{zr})\right],$$

and then define

$$\bar{Q}_{i1}(\theta;\psi^r) = \sum_{z=0}^{1}(\zeta_i^r)^z(1-\zeta_i^r)^{1-z}\left[\sum_{k=1}^{K_i} Q_{i1k}^z(\theta;\psi^r) + \sum_{k=1}^{K_{ij}} G_{i1k}^z(\theta;\psi^r)\right]. \tag{2B.4}$$

Combining (2B.3) and (2B.4) we then obtain

$$Q_1(\theta;\psi^r) = \sum_{i\in\mathcal{R}} Q_{i1}(\theta;\psi^r) + \sum_{i\in\bar{\mathcal{R}}} \bar{Q}_{i1}(\theta;\psi^r). \tag{2B.5}$$

The function in (2B.5) can be maximized using standard software for fitting Poisson or exponential regression models. A sample section of the data frame at the rth iteration is given in Table 2.4 and 2.5 for a subject with $R_i = 1$ or 0 respectively. If one creates a factor variable based on column `K`, we could fit a Poisson model with covariates Z_1, Z_2 and factor(K) with response `int-stat` × `stat`, offset log(`len`), and weight $\texttt{weight}_z \times \texttt{weight}_J$. The updated estimate of θ is θ^{r+1} and the parameter estimates for the baseline hazard can be obtained from the coefficients of the factor variable K. The updated estimates of η are obtained as described in Section 2.1.

Table 2.4: The first part of the pseudo-data frame for maximizing $Q_1(\theta;\psi^r)$ with respect to θ for an arbitrary individual $i \in \mathcal{R}$.

R	K	Z_1	Z_2	len	int-stat	stat	weight$_Z$	weight$_J$
1	1	z_{i1}	z_{i2}	S_{i1}	δ_{i1}	δ_i	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	K_i	z_{i1}	z_{i2}	$S_{i\,K_i}$	$\delta_{i\,K_i}$	δ_i	1	1
1	1	z_{i1}	z_{i2}	ω^r_{i1}	ι^r_{i1}	1	1	$\mathcal{J}^r_i$
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	K_{ij}	z_{i1}	z_{i2}	$\omega^r_{iK_{ij}}$	$\iota^r_{iK_{ij}}$	1	1	$\mathcal{J}^r_i$

Table 2.5: Second part of the pseudo-data frame for maximizing $Q_1(\theta; \psi^r)$ with respect to θ for an arbitrary individual $i \in \bar{\mathcal{R}}$.

R	K	Z_1	Z_2	len	int-stat	stat	weight$_Z$	weight$_J$
0	1	1	z_{i2}	S_{i1}	δ_{i1}	δ_i	ζ_i^r	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
0	K_i	1	z_{i2}	$S_{i\,K_i}$	$\delta_{i\,K_i}$	δ_i	ζ_i^r	1
0	1	1	z_{i2}	ω_{i1}^{1r}	ι_{i1}^{1r}	1	ζ_i^r	$\mathcal{J}_i^{1r}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
0	K_{ij}	1	z_{i2}	$\omega_{iK_{ij}}^{1r}$	$\iota_{iK_{ij}}^{1r}$	1	ζ_i^r	$\mathcal{J}_i^{1r}$
0	1	0	z_{i2}	S_{i1}	δ_{i1}	δ_i	$1 - \zeta_i^r$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
0	K_i	0	z_{i2}	$S_{i\,K_i}$	$\delta_{i\,K_i}$	δ_i	$1 - \zeta_i^r$	1
0	1	0	z_{i2}	ω_{i1}^{0r}	ι_{i1}^{0r}	1	$1 - \zeta_i^r$	$\mathcal{J}_i^{0r}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
0	K_{ij}	0	z_{i2}	$\omega_{iK_{ij}}^{0r}$	$\iota_{iK_{ij}}^{0r}$	1	$1 - \zeta_i^r$	$\mathcal{J}_i^{0r}$

Chapter 3

A Dynamic Mover-Stayer Model for Recurrent Event Processes Subject to Resolution

In this chapter we describe a new flexible process model which involves a series of mover-stayer indicators with each one realized upon the occurrence of an event of interest. This indicator can signal the end of the event process and therefore can explain particularly long periods of time between the last observed event and a censoring time. An EM algorithm is used to carry out estimation with right-censored recurrent event data and is shown to perform well empirically for parametric and semiparametric analyses. The proposed method is then applied to data from Danish study of individuals with affective disorder.

3.1 Notation and Model Formulation

We suppose the process of interest begins with an initiating event representing the onset of disease. This could be, for example, the first seizure among individuals with epilepsy, the first acute exacerbation in persons with asthma, or the first hospitalization in individuals with affective disorder. We let $T_0 = 0$ denote the time of the initiating event and let T_j represent the time of the jth subsequent event, $j = 1, 2, \ldots$. The number of events over time period $(0, t]$ is denoted by $N(t) = \sum_{j=1}^{\infty} I(T_j \leq t)$, and $\{N(s), 0 \leq s\}$ denotes the corresponding counting process.

Information on the nature of the event, individuals' characteristics at the event time, and any fixed covariates, are recorded in a $p \times 1$ covariate vector X_j observed upon the occurrence of the jth event. We let $\bar{X}(t) = \{X_0, \ldots, X_{N(t)}\}$ denote the history of this covariate vector when viewed in continuous time; because $N(t)$ is right-continuous this history includes X_j if $t = t_j$. Likewise we let $\bar{X}_j = \{X_0, \ldots, X_j\}$ denote the covariate history as a function of event count. To accommodate the possibility that the condition of interest is resolved upon the occurrence of the jth event, we let Z_j denote a time-dependent indicator variable such that $Z_j = 1$ if the individual remains at risk for future events following the jth event, and $Z_j = 0$ otherwise, $j = 0, 1, \ldots$. The indicator Z_j is a latent variable, but we learn that $Z_j = 1$ upon the occurrence of the $(j+1)$st event, $j = 0, 1, \ldots$. As was done for the observed covariate vector, here we let $\bar{Z}(t) = \{Z_0, \ldots, Z_{N(t)}\}$ and $\bar{Z}_j = \{Z_0, \ldots, Z_j\}$.

The complete process history is denoted by $\mathcal{H}(t) = \{(N(s), X(s), Z(s)), 0 \leq s \leq t\}$, which includes the values of the latent variables realized over $[0, t]$, and the history excluding

$\bar{Z}(t)$ is denoted by $H(t) = \{(N(s), X(s)), 0 \le s \le t\}$. We let t^- denote an infinitesimal amount of time before t. Assuming two events cannot occur at the same time, the *complete data intensity function* is

$$\lambda(t|\mathcal{H}(t^-)) = \lim_{\Delta t \to 0} \frac{P(\Delta N(t) = 1|\mathcal{H}(t^-))}{\Delta t} = Z_{N(t^-)}\lambda(t|H(t^-)) \ , \tag{3.1}$$

where $\Delta N(t) = N((t+\Delta t)^-) - N(t^-)$ denotes the number of the events over the interval $[t, t+\Delta t)$ and

$$\lambda(t|H(t^-)) = \lim_{\triangle t \to 0} \frac{P(\triangle N(t) = 1|H(t^-))}{\triangle t} \tag{3.2}$$

is a *canonical event intensity function.* We use the term complete data intensity function for (3.1) because it contains the complete information over $[0, t)$ including information on the latent process; we use the term canonical intensity for (3.2) because it can be any intensity function useful for modeling recurrent event processes not subject to resolution.

It may, for example, correspond to any point process model including modulated Markov models for which

$$\lambda(t|H(t^-)) = \lambda_0(t;\alpha)\exp(X'_{N(t^-)}\beta) \ ,$$

or modulated semi-Markov models for which

$$\lambda(t|H(t^-)) = h_{N(t^-)}(B(t);\alpha)\exp(X'_{N(t^-)}\beta) \ ,$$

where $h_j(w_j;\alpha)$ is the baseline hazard for the inter-arrival time $w_j = t_j - t_{j-1}$ and $B(t) =$

$t - t_{N(t^-)}$ is the backwards recurrence time at $t > 0$ (Lawless, 1995). Mixed Markov and semi-Markov processes offer alternative frameworks (Cook and Lawless, 2007). The canonical intensity is not relevant alone for modeling the data, however, and the complete intensity is not useable since $\bar{Z}(t^-)$ is not observed. The *observed data intensity function* is obtained by marginalizing over the latent process and is of the form

$$E\{\lambda(t|\mathcal{H}(t^-))|H(t^-)\} = E(Z_{N(t^-)}|H(t^-)) \cdot \lambda(t|H(t^-)) \; . \tag{3.3}$$

As in models with fixed continuous frailty terms, here it is most convenient to adopt a latent variable approach to estimation and hence construct a complete data likelihood based on (3.1); we do so in the next section.

In general X_j can depend on the complete process history at t_j^- and the fact that an event occurred at t_j, so we denote the probability model by

$$P(X_j|\mathcal{H}(t_j^-), dN(t_j) = 1) = P(X_j|H(t_j^-), \bar{Z}_{j-1} = 1_{j-1}, dN(t_j) = 1) \tag{3.4}$$

where $\bar{Z}_{j-1} = 1_{j-1}$ is an $j \times 1$ vector of ones, and we somewhat informally let $dN(t) = \lim_{\Delta t \to 0} \Delta N(t) = 1$ if an event occurs at time t and $dN(t) = 0$ otherwise.

The probability of remaining at risk following the jth event can depend upon $\mathcal{H}(t_j^-)$ and X_j, so at t_j we write this as

$$P(Z_j = 1|\mathcal{H}(t_j^-), dN(t_j) = 1, X_j) = P(Z_j = 1|H(t_j^-), \bar{Z}_{j-1} = 1_{j-1}, dN(t_j) = 1, X_j) \, . \tag{3.5}$$

This probability may therefore depend on the times of previous events and the history of the observable covariates over $[0, t_j]$ and is only relevant if $\bar{Z}_{j-1} = 1_{j-1}$. Discrete waiting time models are suitable for the resolution of the process and we may specify them based on logistic models. If $\dot{X}_j = (1, X_j')'$, a simple model is of the form

$$\text{logit} P(Z_j = 1 | \mathcal{H}(t_j), dN(t_j) = 1, X_j) = \dot{X}_j' \eta_j \tag{3.6}$$

in which the odds the process does not resolve upon the occurrence of the jth event at t_j depends on the features X_j upon event occurrence. It is often convenient and reasonable to constrain $\eta_j = \eta$ and so there is one set of regression coefficients common across all logistic models.

Example: Suppose the canonical intensity is Markov with $\lambda(t|H(t^-)) = \lambda\alpha(\lambda t)^{\alpha-1}$, and a logistic model is used for the latent indicator with (3.6) taking the form

$$\text{logit} P(Z_j = 1 | \mathcal{H}(t_j), dN(t_j) = 1, X_j) = \eta_0 + \eta_1 j + \eta_2 X \ ,$$

where $X_j = (j, X)'$ with X being an indicator of a treatment ($X = 1$) or control ($X = 0$) condition. In this case $\exp(\eta_1)$ is the relative odds, given X, that the process remains unresolved at the jth event compared to at the previous event; the parameter η_1 therefore reflects the tendency for the process to remain unresolved upon the occurrence of each event, regardless of the times of the events. The coefficient η_2 reflects the possible effect of treatment on the odds the process remains unresolved after a given number of events.

The mean function gives the expected number of events as a function of the time and so is defined by

$$E\{N(t)|X\} = \sum_{n=0}^{\infty} n P(N(t) = n|X) \; .$$

To compute this, note that

$$\begin{aligned} P(N(t) = n|X) \;\; &= \;\; P(N(t) = n|\bar{Z}_n = 1_n, X)P(\bar{Z}_n = 1_n|X) + \\ & \quad\; P(T_n \le t|\bar{Z}_{n-1} = 1_{n-1}, X)P(\bar{Z}_{n-1} = 1_{n-1}|X) \; , \end{aligned}$$

where

$$P(N(t) = n|\bar{Z}_n = 1_n, X) = \; \Lambda(t|X)^n \, e^{-\Lambda(t|X)}/n! \, ,$$

with $\Lambda(t|X) = \int_0^t \lambda(s|X)ds$ and

$$P(T_n \le t|\bar{Z}_{n-1} = 1_{n-1}, X) = 1 - \sum_{r=0}^{n-1} P(N(t) = r|\bar{Z}_{r-1} = 1_{r-1}, X) \, ,$$

since the latent process is a Poisson process, and

$$P(\bar{Z}_n = 1_n|X) = P(Z_0 = 1|X) \prod_{j=1}^{n} P(Z_j = 1|\bar{Z}_{j-1} = 1_{j-1}, X) \, .$$

Figure 3.1 contains plots of the mean function based on the canonical intensity, and the mean functions for the marginal (observed) processes discussed here for the treatment ($X = 1$) and control ($X = 0$) groups; we set $\lambda = 36$, $\alpha = 0.50$, $\eta_1 = \log 0.95$, $\eta_2 = \log 0.75$ and determined η_0 to give $E(N(1)) = 0.75$ (left panel) or $E(N(1)) = 3$ (right panel). As expected there is a large difference in the expected number of events between the canonical and marginal models since the latter incorporate the chance that the process resolves during follow-up. The covariate effect on the mover-stayer process leads to two marginal mean functions (under the proposed model) with the difference between them reflecting magnitude of the effect of treatment on the mover-stayer indicator.

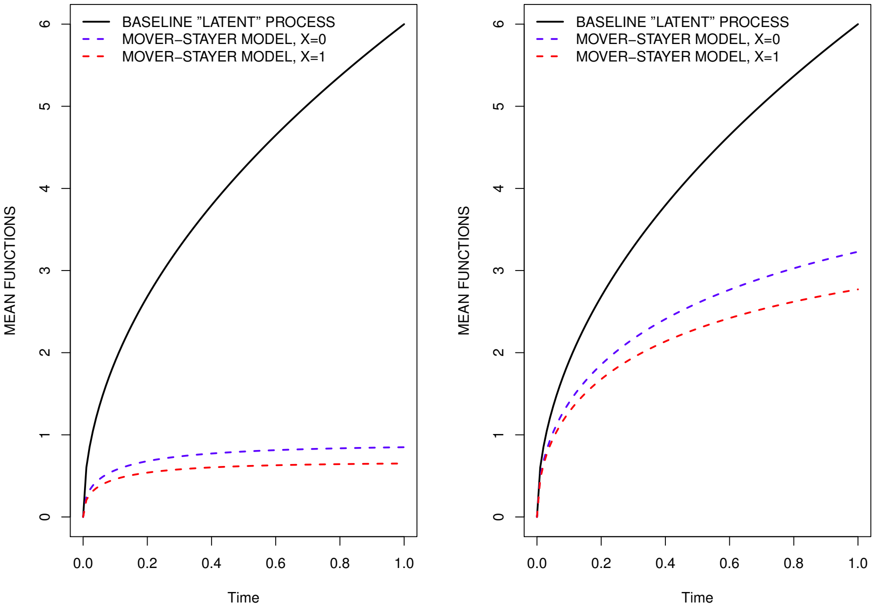

Figure 3.1: Plots of the cumulative canonical intensity $(\lambda t)^{\alpha}$ and mean functions for the treatment $(X = 1)$ and control $(X = 0)$ group in the dynamic mover-stayer model; $\lambda = 36$, $\alpha = 0.5$, $\eta_1 = \log 0.95$, $\eta_2 = \log 0.75$, η_0 is obtained to give $E(N(1)) = 0.75$ (left panel) and 3.0 (right panel)

3.2 Parameter Estimation and Statistical Inference

3.2.1 An EM Algorithm for Parametric Modeling

To describe the algorithm for estimation we return to the general case with a canonical Markov intensity of an unspecified form. Let θ_1 denote the parameter indexing the canonical intensity in (3.2), θ_2 parameterize (3.5), and θ_3 parameterize (3.4).

If the latent process were observable over an interval $[0, C]$, the complete data likelihood would be proportional to the probability of observing $\{(t_j, X_j, Z_j), j = 0, 1, \ldots, n\}$ over $[0, C]$ and is given by

$$L_C \propto L_{C1}(\theta_1) \cdot L_{C2}(\theta_2) \cdot L_{C3}(\theta_3)$$

where

$$\begin{aligned}
L_{C1}(\theta_1) &\propto \prod_{j=1}^{n} \left\{ \lambda(t_j|\mathcal{H}(t_j^-)) \exp\left(-\int_{t_{j-1}}^{t_j} \lambda(u|\mathcal{H}(u^-))du \right) \right\} \exp\left(-\int_{t_n}^{C} \lambda(u|\mathcal{H}(u^-))du \right) , \\
L_{C2}(\theta_2) &\propto P(Z_0|\mathcal{H}(0^-), dN(0) = 1, X_0) \prod_{j=1}^{n} P(Z_j|\mathcal{H}(t_j^-), dN(t_j) = 1, X_j) , \\
L_{C3}(\theta_3) &\propto P(X_0|\mathcal{H}(0^-), dN(0) = 1) \prod_{j=1}^{n} P(X_j|\mathcal{H}(t_j^-), dN(t_j) = 1)
\end{aligned}$$

and $\mathcal{H}(0^-) = \emptyset$. Terms involving the probability model for the observed covariates can be omitted if the covariate process is non-informative (i.e. the parameters indexing the distribution of the covariates are not functionally related to the parameters of the processes of interest). In this case we use the partial complete data likelihood

$$L_C(\theta) \propto L_{C1}(\theta_1) \cdot L_{C2}(\theta_2)\,, \tag{3.7}$$

where

$$L_{C1}(\theta_1) \propto \prod_{j=1}^{n} \left\{\lambda(t_j|H(t_j^-))\right\} \exp\left(-\sum_{k=0}^{n} \int_{t_k}^{t_{k+1}} Z_k^{I(k=n)} \lambda(u|H(u^-))du\right) \tag{3.8}$$

is the contribution pertaining to θ_1, with $t_0 = 0$ and $t_{n+1} = C$, and

$$\begin{aligned} L_{C2}(\theta_2) &\propto P(Z_0|H(0^-), \bar{Z}_{-1} = \emptyset, dN(0) = 1, X_0) \cdot \prod_{j=1}^{n} P(Z_j|H(t_j^-), \bar{Z}_{j-1} = 1_{j-1}, dN(t_j) = 1, X_j) \\ &\propto P(Z_0|H(0^-)) \cdot \prod_{j=1}^{n} P(Z_j|H(t_j^-), \bar{Z}_{j-1} = 1_{j-1}) \qquad (3.9) \end{aligned}$$

is the contribution related to the latent process, where $H(0^-) = \emptyset$, and $\theta = (\theta_1', \theta_2')'$. The missing variable in the above complete data likelihood is Z_n, the indicator of whether the process continues following the occurrence of the last observed event.

The expectation-maximization (EM) algorithm of Dempster et al. (1977) offers a convenient way of maximizing the observed data likelihood. To do this we define

$$Q(\theta; \widehat{\theta}) = Q_1(\theta_1; \widehat{\theta}) + Q_2(\theta_2; \widehat{\theta}) \tag{3.10}$$

where $Q_1(\theta_1; \widehat{\theta}) = E(\log L_{C1}(\theta_1)|H(C); \widehat{\theta})$ and $Q_2(\theta_2; \widehat{\theta}) = E(\log L_{C2}(\theta_2)|H(C); \widehat{\theta})$. Since

$\log L_{C1}(\theta_1)$ and $\log L_{C2}(\theta_2)$ are linear in Z_n, only

$$\zeta(\widehat{\theta}) = P(Z_n = 1 | H(C); \widehat{\theta}), \tag{3.11}$$

given by

$$\frac{P(Z_n = 1 | H(t_n^-), \bar{Z}_{n-1} = 1_{n-1}; \widehat{\theta}_2) \exp(- \int_{t_n}^{C} \lambda(u | H(u^-); \widehat{\theta}_1)\, du)}{P(Z_n = 1 | H(t_n^-), \bar{Z}_{n-1} = 1_{n-1}; \widehat{\theta}_2) \exp(- \int_{t_n}^{C} \lambda(u | H(u^-); \widehat{\theta}_1)\, du) + P(Z_n = 0 | H(t_n^-), \bar{Z}_{n-1} = 1_{n-1}; \widehat{\theta}_2)},$$

is required at the E-step to compute (3.10). The maximum likelihood estimator is obtained by iteratively maximizing (3.10) as follows. If $\widehat{\theta}^r$ denotes the estimate of θ at the rth iteration, we maximize $Q(\theta; \widehat{\theta}^r)$ with respect to θ to obtain $\widehat{\theta}^{r+1}$. This process is repeated iteratively until $\|\widehat{\theta}^{r+1} - \widehat{\theta}^r\| \leq \epsilon$ where ϵ is a pre-specified tolerance, at which point we let the final value be the maximized likelihood estimate. Variance estimation can be carried out using the method of (Louis, 1982); see Appendix 3A for details.

3.2.2 An EM Algorithm for Semiparametric Modeling of a Markov Process

The model formulation in the parametric setting is quite general. Next we consider a special model with a Markov canonical intensity with a proportional latent rate function and consider semiparametric modeling of the canonical Markov intensity. To do this we introduce subscripts to index individuals and adopt counting process notation.

Let m be the number of subjects in the study, n_i be the number of events for subject

i, $[0, C_i]$ denote the period of observation for subject i and let $Y_i(u) = I(u \leq C_i)$ indicate whether they are under observation at time u. Let $Z_i(u) = Z_{iN_i(u^-)}$ denote the latent variable expressed as a continuous time varying indicator. Under a Markov latent intensity

$$\lambda(t|H(t^-)) = \lambda_0(t)\exp(X\beta); ,$$

where $\lambda_0(t) = d\Lambda_0(t)/dt$ is the baseline latent intensity for an individual with $X = 0$. We also let $\Lambda_0(s,t) = \displaystyle\int_s^t d\Lambda_0(u)$. In counting process notation the complete data likelihood for the recurrent event process (Cook and Lawless, 2007) is

$$L_{C1}(\lambda_0(\cdot), \beta) = \prod_{i=1}^{m}\left[\prod_{j=1}^{n_i}[Y_i(u)d\Lambda(u|X_i)]^{Y_i(u)dN_i(u)}\exp\left(-\int_0^\infty Z_i(u)Y_i(u)d\Lambda(u|X_i)\right)\right]$$

and $L_{C2}(\theta_2)$ is the same as in (3.9). The complete log-likelihood is then

$$\ell_C(\theta) = \ell_{C1}(\lambda_0(\cdot), \beta) + \ell_{C2}(\theta_2)\ ,$$

where

$$\ell_{C1}(\lambda_0(\cdot), \beta) = \sum_{i=1}^{m}\left\{\int_0^\infty Y_i(u)dN_i(u)(\log d\Lambda_0(u) + X_i\beta) - \int_0^\infty Z_i(u)Y_i(u)d\Lambda_0(u)\exp(X_i\beta)\right\}\ ,$$

and

$$\ell_{C2}(\theta_2) = \sum_{i=1}^{m}\Big[\sum_{j=0}^{n_i-1}\log P(Z_{ij}|H(t_{ij}^-), \bar{Z}_{j-1} = 1_{j-1}) + \log P(Z_{in_i}|H(t_{in_i}^-), \bar{Z}_{n_i-1} = 1_{n_i-1})\Big]\ ,$$

where we define $\bar{Z}_{-1}$ as the null set. Here $\theta = (\lambda_0(\cdot), \beta', \theta_2')$ where $\lambda_0(\cdot)$ is the latent baseline rate function, β is the covariate effect on the intensity of the latent process, and for the particular model discussed in Section 3.1, for example, $\theta_2 = (\eta_0, \eta_1, \eta_2)$ is the parameter vector for the mover-stayer probability model. Then (3.10) becomes

$$Q(\theta; \widehat{\theta}) = Q_1(\lambda_0(\cdot), \beta; \widehat{\theta}) + Q_2(\theta_2; \widehat{\theta}) \ ,$$

and

$$Q_1(\lambda_0(\cdot), \beta; \widehat{\theta}) = \sum_{i=1}^{m} \left\{ \int_0^\infty Y_i(u) dN_i(u)(\log d\Lambda_0(u) + X_i\beta) - \int_0^\infty \zeta_i(u; \widehat{\theta}) Y_i(u) d\Lambda_0(u) \exp(X_i\beta) \right\}$$

where if $u < T_{in_i}$, $\zeta_i(u; \widehat{\theta}) = 1$; and if $T_{in_i} \leq u \leq C_i$, $\zeta_i(u; \widehat{\theta}) = E(Z_i(u)|H_i(C_i); \widehat{\theta})$ is given by (3.11) with $\exp(-\int_{t_n}^{C} \lambda(u|H(u^-); \widehat{\theta}_1)du)$ reduced to $\exp(-\Lambda(t_{in_i}, C_i|X_i; \widehat{\theta}_1))$, where $\Lambda(s,t|X_i; \widehat{\theta}_1) = \int_s^t \lambda(u|X_i; \widehat{\theta}_1)du$ and $\theta_1 = (\lambda_0(\cdot), \beta)'$. The argument u in $\zeta_i(u; \widehat{\theta})$ is therefore introduced to facilitate writing a general expression for this expectation.

When maximizing $Q_1(\lambda_0(\cdot), \beta; \widehat{\theta})$ with respect to $\lambda_0(\cdot)$ and β, we obtain the two equations

$$\sum_{i=1}^{m} \left[Y_i(u) dN_i(u) - \zeta_i(u; \widehat{\theta}) Y_i(u) \exp(X_i\beta) d\Lambda_0(u) \right] = 0, \ 0 < u \tag{3.12}$$

$$\sum_{i=1}^{m} \left[\int_0^\infty Y_i(u) dN_i(u) X_i - \int_0^\infty \zeta_i(u; \widehat{\theta}) Y_i(u) d\Lambda_0(u) \exp(X_i\beta) X_i \right] = 0 \, . \tag{3.13}$$

For a given β, we obtain the "profile" estimate

$$d\widehat{\Lambda}_0(u;\beta) = \frac{\sum_{i=1}^m Y_i(u) dN_i(u)}{\sum_{i=1}^m Y_i(u)\zeta_i(u;\widehat{\theta})\exp(X_i\beta)},$$

and substitute this into (3.13) to obtain the equation

$$\sum_{i=1}^m \int_0^\infty Y_i(u) dN_i(u) \left[X_i - \frac{\sum_{i=1}^m Y_i(u)\zeta_i(u;\widehat{\theta})\exp(X_i\beta)X_i}{\sum_{i=1}^m Y_i(u)\zeta_i(u;\widehat{\theta})\exp(X_i\beta)} \right].$$

This looks very much like the usual Cox partial likelihood score equation with offsets. For each subject i we can construct a pseudo-dataset with $n_i + 1$ lines: first n_i lines correspond to the period from 0 to t_{in_i} and have an offset of zero; the last line corresponds to the period from t_{in_i} to C_i and has an offset of $\log \zeta_i(u;\widehat{\theta})$. Existing software can therefore be used to obtain updated estimates of $\lambda_0(\cdot)$ and β.

The second term is

$$\begin{aligned} Q_2(\theta_2;\widehat{\theta}) &= \sum_{i=1}^m \Big[\sum_{j=0}^{n_i-1} \log P(Z_{ij} = 1 | H(t_{ij}^-), \bar{Z}_{j-1} = 1_{j-1}) \\ &+ \zeta_i(\widehat{\theta}) \log P(Z_{in_i} = 1 | H(t_{in_i}^-), \bar{Z}_{n_i-1} = 1_{n_i-1}) \\ &+ (1 - \zeta_i(\widehat{\theta})) \log P(Z_{in_i} = 0 | H(t_{in_i}^-), \bar{Z}_{n_i-1} = 1_{n_i-1}) \Big]. \end{aligned}$$

where $\zeta_i(\widehat{\theta}) = \zeta_i(u;\widehat{\theta})$ for $T_{in_i} \leq u$. Maximization of $Q_2(\theta_2;\widehat{\theta})$ with respect to θ_2 can be done by fitting logistic regression to pseudo-datasets, which contains $n_i + 2$ lines for each subject i: the first n_i lines correspond to $Z_{i0} = 1, \ldots, Z_{i,n_i-1} = 1$ and have weight 1; the next line corresponds to the possibility that $Z_{in_i} = 1$ and has weight $\zeta_i(\widehat{\theta})$; the final line corresponds to the other possibility that $Z_{in_i} = 0$ and has associated weight $1 - \zeta_i(\widehat{\theta})$.

Additional details for the EM algorithm including its implementation and variance estimation are given in Appendix 3A.

3.3 Empirical Studies

Here we conduct simulation studies to evaluate the performance of the EM algorithm in fitting the dynamic mover-stayer model with a latent Markov process. We first generate a treatment indicator X as a Bernoulli random variable with $P(X = 1) = 1 - P(X = 0) = 0.5$. The Z_j are generated according to model (3.6) with a common η vector with $\eta_1 = \log 0.95$ and $\eta_2 = \log 0.75$ so that for given X, the probability of remaining a mover decreases with each event to create the scenario that is consistent with the burn-out theory and for each value of j the odds of remaining a mover are 25% lower in the treatment group with $X = 1$. For the baseline intensity of the latent Markov process of the form $\lambda\alpha(\lambda t)^{\alpha-1}$ we fix $\alpha = 1$ to correspond to a time-homogeneous latent process, and $\alpha = 0.50$ to correspond to a time-nonhomogeneous latent process; we set $\beta = \log 0.75$ to correspond to a 25% reduction in the rate of events among individuals at risk of events. For a given α and β, λ is determined so that the expected number of events over $(0, C]$ is specified at the particular value six among individuals who remain movers throughout the interval $(0, C]$.

We then solve for η_0 so that the marginal expectation satisfies $E[N(C)] = 0.75$, 1.5, or 3.0. Five hundred datasets of $m = 500$ individuals were simulated for each parameter configuration. Parametric analysis and semiparametric analysis were carried out for each simulated dataset. Standard errors were obtained using the method of Louis (1982) and the performance of the estimators was assessed in terms of empirical bias, empirical and model-based standard errors, and empirical coverage probability. The empirical bias (EBIAS), empirical standard error (ESE), average model-based standard error (ASE) computed according to Louis (1982), and empirical coverage probability expressed as a percentage (ECP) are given in Table 3.1, 3.2 and 3.3, for the parametric analyses; the empirical coverage probability is defined as the fraction of simulations for which the sample confidence interval contained the true parameter value. The empirical bias and empirical standard errors are also reported for the semiparametric analyses.

The empirical biases are generally small and decrease with increasing expected numbers of events. There is also good agreement between the empirical and average model-based standard errors and the empirical coverage probability is compatible with the nominal level of 95%. The results are roughly comparable for the parametric and semiparametric analyses and the methods perform well when there is a trend in the latent rate function.

Table 3.1: Empirical results for maximum likelihood estimates obtained by the EM algorithm for parametric and semiparametric models with $\lambda(t|H(t^-)) = \lambda\alpha(\lambda t)^{\alpha-1}\exp(\beta X)$ and $P(Z_j = 1|\bar{Z}_{j-1} = 1_{j-1}, X) = \text{expit}(\eta_0 + \eta_1 j + \eta_2 X)$; $m = 500$, $nsim = 500$, $E(N(C)) = 0.75$

		$E(N(C)) = 0.75$, $\eta_0 = -0.085$					
		Parametric				Semiparametric	
	VALUE	EBIAS	ESE	ASE	ECP	EBIAS	ESE
		Time Homogeneous Rate					
η_0	$- - -$	0.002	0.107	0.109	95.2	-0.001	0.107
η_1	-0.051	-0.007	0.073	0.072	95.6	0.020	0.082
η_2	-0.288	-0.006	0.144	0.141	94.2	0.013	0.146
λ	6.857	0.045	0.509	0.499	95.4		
β	-0.288	-0.012	0.116	0.116	96.0	-0.026	0.125
$\Lambda_0(C)$	6.857					-0.491	1.146
		Time Non-homogeneous Rate					
η_0	$- - -$	0.002	0.107	0.109	95.0	-0.001	0.107
η_1	-0.051	-0.007	0.074	0.072	95.2	0.020	0.082
η_2	-0.288	-0.006	0.144	0.142	94.4	0.013	0.146
λ	47.020	0.899	7.859	7.840	95.8		
α	0.500	0.001	0.022	0.023	97.2		
β	-0.288	-0.012	0.116	0.116	95.8	-0.026	0.125
$\Lambda_0(C)$	6.857					-0.491	1.146

Table 3.2: Empirical results for maximum likelihood estimates obtained by the EM algorithm for parametric and semiparametric models with $\lambda(t|H(t^-)) = \lambda\alpha(\lambda t)^{\alpha-1}\exp(\beta X)$ and $P(Z_j = 1|\bar{Z}_{j-1} = 1_{j-1}, X) = \text{expit}(\eta_0 + \eta_1 j + \eta_2 X)$; $m = 500$, $nsim = 500$, $E(N(C)) = 1.5$

		$E(N(C)) = 1.5$, $\eta_0 = 0.709$					
		Parametric				Semiparametric	
	VALUE	EBIAS	ESE	ASE	ECP	EBIAS	ESE
		Time Homogeneous Rate					
η_0	$---$	-0.000	0.104	0.103	94.8	-0.003	0.104
η_1	-0.051	-0.002	0.045	0.044	94.8	0.007	0.052
η_2	-0.288	-0.001	0.125	0.125	96.0	0.005	0.128
λ	6.857	0.010	0.365	0.359	94.0		
β	-0.288	0.002	0.086	0.084	94.8	-0.001	0.087
$\Lambda_0(C)$	6.857					-0.069	0.666
		Time Non-homogeneous Rate					
η_0	$---$	-0.001	0.104	0.103	95.1	-0.004	0.104
η_1	-0.051	-0.002	0.045	0.044	94.9	0.007	0.051
η_2	-0.288	0.000	0.125	0.125	95.5	0.006	0.128
λ	47.020	0.292	6.219	6.055	93.5		
α	0.500	0.001	0.017	0.017	95.5		
β	-0.288	0.001	0.086	0.084	95.1	-0.001	0.087
$\Lambda_0(C)$	6.857					-0.070	0.667

Table 3.3: Empirical results for maximum likelihood estimates obtained by the EM algorithm for parametric and semiparametric models with $\lambda(t|H(t^-)) = \lambda\alpha(\lambda t)^{\alpha-1}\exp(\beta X)$ and $P(Z_j = 1|\bar{Z}_{j-1} = 1_{j-1}, X) = \text{expit}(\eta_0 + \eta_1 j + \eta_2 X)$; $m = 500$, $nsim = 500$, $E(N(C)) = 3$

		$E(N(C)) = 3$, $\eta_0 = 1.733$					
		Parametric				Semiparametric	
	VALUE	EBIAS	ESE	ASE	ECP	EBIAS	ESE
		Time Homogeneous Rate					
η_0	$- - -$	0.003	0.115	0.117	95.4	0.000	0.116
η_1	-0.051	-0.000	0.036	0.035	94.6	0.004	0.041
η_2	-0.288	0.003	0.135	0.135	94.4	0.007	0.137
λ	6.857	0.010	0.275	0.262	92.8		
β	-0.288	0.001	0.066	0.061	92.0	0.001	0.067
$\Lambda_0(C)$	6.857					-0.016	0.335
		Time Non-homogeneous Rate					
η_0	$- - -$	0.003	0.115	0.117	95.6	0.000	0.116
η_1	-0.051	0.000	0.037	0.036	94.8	0.004	0.041
η_2	-0.288	0.004	0.135	0.135	94.4	0.007	0.137
λ	47.020	0.454	5.085	4.976	94.0		
α	0.500	-0.000	0.013	0.013	95.4		
β	-0.288	0.001	0.067	0.061	92.0	0.001	0.067
$\Lambda_0(C)$	6.857					-0.016	0.335

3.4 Application to a Cohort Study of Individuals with Affective Disorder

We consider the cohort of 10,523 individuals with a first episode of affective disorder between January 1, 1994 and December 31, 1999. Among these individuals, 3802 (36.1%) are male and 6721 (63.9%) are female. A total of 17,021 hospitalizations are made over this window of calendar time giving a mean of 1.618 visits per individual (S.D.=1.720). A total of 1106 (10.5%) of these individuals were bipolar at the time of the first admission; among the 9417 (89.5%) patients who were unipolar at the study entry, 9228 remain as unipolar, and 189 become bipolar by the end of follow-up. We consider a dataset comprised of 9417 patients who are unipolar at the first admission and who had a total of 14497 admissions (mean=1.539 and S.D.=1.272). Follow-up of these individuals is censored at the end of the observation period, upon the diagnosis of bipolar disorder, schizophrenia, or an organic disorder, or at the time of death. There are 3298 (35.0%) male individuals with total of 4860 visits (mean=1.474 and S.D.=1.105) and 6119 (65.0%) female patients with a total of 9637 visits (mean=1.575 and S.D.=1.352).

We fit parametric and semiparametric (Andersen and Gill, 1982) Poisson regression models for the recurrence of acute episodes, with a single covariate indicating gender ($X = 1$ for females, $X = 0$ for males). These results are reported in the first three columns of Table 3.4. Dynamic mover-stayer models are also fitted for which the latent variable model controls for the cumulative number of events (j) and gender; we denote the vector of covariates by $\dot{X}_j = (1, j, X)'$. A reduced dynamic mover-stayer model is also fitted with $\dot{X}_j = (1, j)'$ which simply controls for the cumulative number of acute episodes.

The canonical event intensity model in these dynamic mover-stayer models also controls for gender. Both parametric (top half) and semiparametric (bottom half) event intensity models are reported in Table 3.4.

Table 3.4: Results of fitting Poisson model and dynamic mover-stayer model† to study of affective disorder with parametric and semiparametric models; Markov model is a parametric Poisson model or Anderson-Gill (1982) semiparametric model, $m = 9417$

				Dynamic Mover-Stayer Models					
	Poisson Model			$\dot{X}_j = (1, j, X)$			$\dot{X}_j = (1, j)$		
	EST	S.E.	p-value	EST	S.E.	p-value	EST	S.E.	p-value
Parametric Models									
Mover-Stayer Model									
η_0	—	—	—	-0.6344	0.0376		-0.5219	0.0257	
η_1	—	—	—	0.5184	0.0232	< 0.0001	0.5210	0.0233	< 0.0001
η_2	—	—	—	0.1682	0.0433	0.0001			
Recurrent Event Model									
λ	0.1555	0.0058		1.2170	0.0548		1.1729	0.0548	
α	0.6970	0.0087		0.9574	0.0140		0.9570	0.0140	
β	0.1573	0.0299	< 0.0001	-0.0268	0.0515	0.6023	0.0222	0.0505	0.6600
Semiparametric Models									
Mover-Stayer Model									
η_0	—	—	—	-0.6418	0.0387		-0.5289	0.0264	
η_1	—	—	—	0.5760	0.0338	< 0.0001	0.5796	0.0340	< 0.0001
η_2	—	—	—	0.1685	0.0451	0.0002			
Recurrent Event Model									
β	0.1620	0.0299	< 0.0001	-0.0201	0.0539	0.7088	0.0338	0.0521	0.5158

† Standard errors for estimates from parametric models obtained by Louis' method (1982) and by nonparametric bootstrap (200 bootstrap samples) for fitted semiparametric models; p-values are based on Wald statistics

We focus the following discussion on the results of the analyses based on the semiparametric intensity model. The estimated regression coefficient for gender from the semiparametric Andersen-Gill model suggests women have a 17.6% increased rate of recurrence compared to men ($RR = 1.176$, 95% CI $(1.109, 1.247)$, $p < 0.001$). The estimates of the cumulative mean functions based on the fitted Andersen-Gill model are given in the left panel of Figure 3.2 and reveal a small absolute difference between genders in the cumulative expected number of episodes over time. The first semiparametric dynamic mover-stayer model reveals an insignificant association between gender on the latent intensity of recurrence ($RR = 0.980$, 95% CI $(0.882, 1.089)$, $p = 0.709$), but women have a significantly higher odds of remaining at risk of recurrence based on the mover-stayer component ($OR = 1.184$, 95% CI $(1.083, 1.293)$, $p < 0.001$). The dynamic mover-stayer model therefore suggests that the higher expected number of episodes for women may arise from a lower tendency for women to experience resolution of the disease. The right panel of Figure 3.2 gives the semiparametric estimate of the cumulative canonical event intensity for males and females. These estimates are much higher than those of the left panel since they correspond to the canonical process which does not accommodate resolution. Moreover the two estimates are very similar, reflecting the insignificant gender effect seen in this model.

Upon the removal of gender from the mover-stayer component (see the last three columns of Table 3.4) the effect of gender on the latent rate remains insignificant ($RR = 1.034$, 95% CI $(0.934, 1.146)$, $p = 0.516$). The findings from the parametric and semiparametric analyses are in broad agreement.

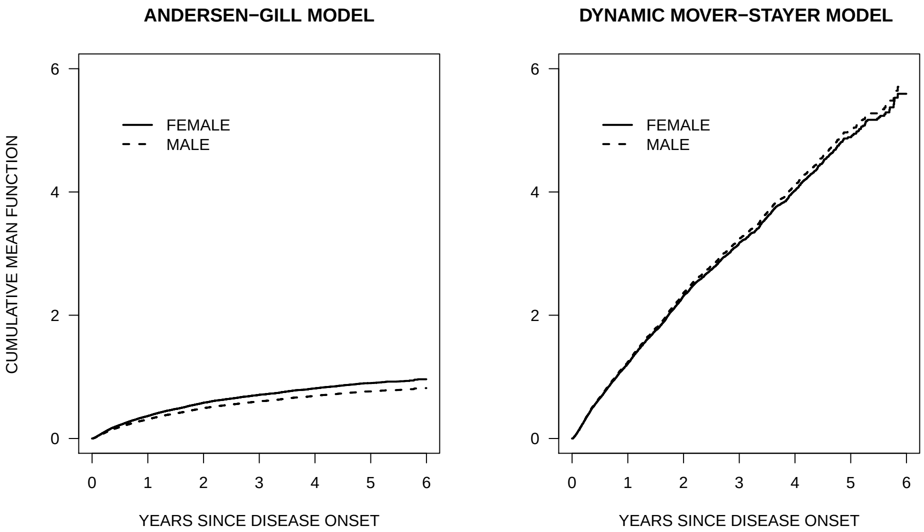

Figure 3.2: Plots of the estimated cumulative intensities for females and males with affective disorder; the left panel gives the cumulative mean function estimates based on the Andersen-Gill model and the right panel gives the cumulative canonical event intensity based on the dynamic mover-stayer model with covariate $\dot{X}_j = (1, j, X)'$ in the mover-stayer component and gender (X) in the canonical intensity model

3.5 Remarks

In this chapter we have described a dynamic mover-stayer model for the analysis of recurrent event data which is useful when there is a substantial fraction of individuals with an unduly long final gap time. This formulation is most appropriate when the underlying condition leading to the recurrent events can resolve but this resolution is not observable. There are a number of other medical conditions where this scenario can arise, and it is particularly relevant for registry studies where limited information is collected on individuals between records of events of interest. In the motivating example, the reasons for any resolution could include the identification of a suitable dose or type of medication or a change in a stressful environment leading to exacerbations of symptoms. Details on these and other possible explanations are often unavailable in the settings of registry studies but accommodation of such eventualities is often sensible in model formulation.

Appendix 3A: Additional Details for the EM Algorithm

Appendix 3A.1: Implementation of the EM Algorithm

For an individual with n events observed at times $t_1 < t_2 < \ldots < t_n < C$, the only missing quantity is Z_n. If we have a single covariate X, the dataframe used at the rth step of the EM algorithm to maximize $Q_1(\theta_1; \widehat{\theta}^r)$ has the usual counting process form with the addition of a weight which is 1 for all lines except the last one with form

ID(i)	enum(j)	estart	estop	estatus	weight	rtrunc	tstatus	X
1	0	0	t_1	1	1	NA	1	X_1
1	1	t_1	t_2	1	1	NA	2	X_1
1	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	X_1
1	$n-1$	t_{n-1}	t_n	1	1	NA	2	X_1
1	n	t_n	C	0	w^r	NA	2	X_1

In a parametric analysis with a baseline rate for the latent process of the form $\lambda_0(t;\alpha) = \alpha_2\alpha_1(\alpha_1 t)^{\alpha_1 - 1}$, $Q_1(\theta_1;\widehat{\theta}^r)$ is maximized to give $\widehat{\theta}^{r+1}$ by the R command

$$\begin{aligned}&\text{censorReg}(\text{censor}(\text{estop},\text{estatus}) \sim X, \text{truncation} = \text{censor}(\text{estart},\text{rtrunc},\text{tstatus}),\\ &\quad \text{weights} = \text{weight}, \text{distribution} = \text{“weibull”}, \text{fixed} = \text{list}(\text{scale} = 1))\ ,\end{aligned}$$

and in the semiparametric analysis by the call

$$\text{coxph}(\text{Surv}(\text{estart},\text{estop},\text{estatus}) \sim \text{X} + \text{offset}(\log(\text{weight})), \text{method} = \text{“breslow”})\ .$$

The data used to maximized $Q_2(\theta_2;\widehat{\theta}^r)$ has the form

ID(i)	enum(j)	Z	X	weight
i	0	1	X_1	1
i	1	1	X_1	1
i	2	1	X_1	1
i	$\vdots$	$\vdots$	X_1	$\vdots$
i	$n-1$	1	X_1	1
i	n	1	X_1	w^r
i	n	0	X_1	$1-w^r$

A simple logistic regression call

$$glm(Z \sim enum + X, \text{weights} = \text{weight}, \text{family} = \text{binomial}(\text{link} = \text{logit}))$$

yields $\widehat{\theta}_2^{r+1}$. New dataframes are then created with w^r replaced with w^{r+1} and the procedure is repeated until $\|\widehat{\theta}^{r+1} - \widehat{\theta}^r\| < \epsilon$ for some specified value of ϵ.

Appendix 3A.2: Variance Estimation via Louis' Method

Let $S_C(\theta) = \partial \log L_C(\theta)/\partial\theta$ and $I_C(\theta) = -\partial S_C(\theta)/\partial\theta$ where $L_C(\theta)$ is the complete data likelihood for which Z_n is treated as known, given by (3.7). If $L(\theta)$, $S(\theta) = \partial \log L(\theta)/\partial\theta$ and $I(\theta) = -\partial S(\theta)/\partial\theta$ are the observed data likelihood, score and information matrix, then according to Louis (1982), the observed information matrix is

$$I(\theta) = E_{Z_n}\left[I_C(\theta)|H(C)\right] - E_{Z_n}\left[S_C(\theta)S'_C(\theta)|H(C)\right] \tag{3A.1}$$

where $S_C(\theta) = (S'_{C1}(\theta_1), S'_{C2}(\theta_2))'$ and $S_{Ck}(\theta_k) = \partial \log L_{Ck}(\theta_k)/\partial\theta_k$, $k = 1, 2$, and

$$I_C(\theta) = \begin{bmatrix} I_{C1}(\theta_1) & 0 \\ 0 & I_{C2}(\theta_2) \end{bmatrix},$$

where $I_{Ck} = -\partial S_{Ck}(\theta_k)/\partial\theta_k$, $k = 1, 2$. We estimate $I(\widehat{\theta})$ in (3A.1) by running the EM algorithm to the point of convergence and using the expression in (3.11) evaluated at the MLE $\widehat{\theta}$ to take the required expectation. Standard software can be readily exploited to do this in both the parametric and semiparametric settings.

The first matrix on the right hand side of (3A.1) is obtained by extracting the values stored in the information matrices produced at the final M-step. Each individual contributes to the complete data likelihood and complete data score, so we can compute their contributions to $S_{C1}(\theta_1)$ and $S_{C2}(\theta_2)$, stack them and then take a weighted average to estimate the second term in (3A.1).

In the semiparametric setting, let $u_1 < \ldots < u_R$ denote the R unique event times over the entire sample, let $d\Lambda_0 = (d\Lambda_0(u_1), \cdots, d\Lambda_0(u_R))'$, and let $\theta_1 = (d\Lambda'_0, \beta)'$, then let $S_{C1}(\theta_1) = (S'_{C11}(\theta_1), S'_{C12}(\theta_1))'$, where $S_{C11}(\theta_1) = (S_{C11u_1}(\theta_1), \ldots, S_{C11u_R}(\theta_1))'$ and

$$\begin{aligned} S_{C11u}(\theta_1) = S_{C11}(\lambda_0(u)) &= Y(u)\left\{dN(u) - Z(u)d\Lambda_0(u)\exp(\beta X)\right\}, \quad 0 < u \\ S_{C12}(\theta_1) = S_{C12}(\beta) &= \int_0^\infty Y(u)X\left\{dN(u) - Z(u)d\Lambda_0(u)\exp(\beta X)\right\}\ . \end{aligned}$$

Then

$$I_{C1}(\theta_1) = \left[\begin{array}{c} -\partial S_{C11}(\theta_1)/\partial \theta_1' \\ \\ -\partial S_{C12}(\theta_1)/\partial \theta_1' \end{array} \right] = - \left[\begin{array}{cc} \partial S_{C11}(\theta_1)/\partial d\Lambda_0' & \partial S_{C11}(\theta_1)/\partial \beta \\ \\ \partial S_{C12}(\theta_1)/\partial d\Lambda_0' & \partial S_{C12}(\theta_1)/\partial \beta \ , \end{array} \right]$$

where

$$\begin{array}{rcl} \dfrac{\partial S_{C11}(\theta_1)}{\partial d\Lambda_0'(u)} & = & -Y(u)Z(u)\exp(\beta X) \\ \dfrac{\partial S_{C12}(\theta_1)}{\partial d\Lambda_0'(u)} & = & -Y(u)XZ(u)\exp(\beta X) \\ \dfrac{\partial S_{C11}(\theta_1)}{\partial \beta} & = & -Y(u)Z(u)d\Lambda_0(u)\exp(\beta X)X \\ \dfrac{\partial S_{C12}(\theta_1)}{\partial \beta} & = & -\displaystyle\int_0^\infty Y(u)Z(u)Xd\Lambda_0(u)\exp(\beta X)X \ . \end{array}$$

Then we can obtain $S_C(\theta)$ and $I_C(\theta)$ and proceed as in the parametric setting.

Chapter 4

Analysis of Interval-Censored Recurrent Events with Resolution

Shen and Cook (2013a) proposed a dynamic mover-stayer model for the analysis of right-censored recurrent event data which accommodates unusually long times from the last observed event to the censoring time. In this chapter we extend this method to deal with interval-censored recurrent event data where the underlying process is subject to resolution using a likelihood based approach, where binary mover-stayer indicators are used to indicate the status of disease resolution. An expectation-maximization algorithm is adopted to deal with the difficulty that the exact event times are unknown and the event process is coarsened so that only counts of events are known between inspection times; Lawless and Zhan (1998) refer to this as interval-grouped recurrent event data. Piecewise constant baseline intensity models are adopted for mixed-Poisson processes to provide flexibility and protection against model misspecification. This approach allows estimation of treatment effects on the event rate, baseline intensity modeling and the modeling of the mover-stayer process. The maximization-step is facilitated by making use of existing softwares. Data on the cumulative number of damaged joints in patients with psoriatic arthritis are analysed to provide an illustrative application.

4.1 Model Formulation

First we review the dynamic mover-stayer model of Chapter 3, i.e. Shen and Cook (2013a), where the notation and model formulation are similar as in the current development. Let $T_0 = 0$ denote the time of an initiating event such as the onset of a chronic disease, and let T_j represent the time of the jth subsequent event, $j = 1, 2, \ldots$. If $N(t) = \sum_{j=1}^{\infty} I(T_j \leq t)$ denotes the number of events over $(0, t]$, then $\{N(s), 0 \leq s\}$ is a counting process.

Individuals' characteristics are recorded in $p \times 1$ covariate vector X_j observed upon the occurrence of the jth event. Shen and Cook (2013a) define Z_j as a time-dependent indicator variable whereby $Z_j = 1$ provided that upon the occurrence of the jth event they remain at risk of future events, and $Z_j = 0$ otherwise, $j = 0, 1, \ldots$. The resolution of the chronic condition upon the jth event is reflected by a realization $Z_j = 0$ when $Z_{j-1} = 1$. Here Z_j is a latent variable, but we learn that $Z_j = 1$ as soon as the $(j+1)$st event occurs, $j = 0, 1, \ldots$. We let $\bar{Z}(t) = \{Z_0, \ldots, Z_{N(t)}\}$ and $\bar{Z}_j = \{Z_0, \ldots, Z_j\}$.

The complete process history is denoted by $\mathcal{H}(t) = \{(N(s), Z(s)), 0 \leq s \leq t, X_j\}$, which includes the values of the latent variables realized over $[0, t]$, and the history excluding $\bar{Z}(t)$ is denoted by $H(t) = \{N(s), 0 \leq s \leq t, X_j\}$. We let $\Delta N(t) = N((t + \Delta t)^-) - N(t^-)$

denote the number of the events over the interval $[t, t+\Delta t]$, where t^- is an infinitesimal amount of time before t. The *complete data intensity function* $\lambda(t|\mathcal{H}(t^-))$, the *canonical event intensity function* $\lambda(t|H(t^-))$ and the *intensity function* for the observable process in the absence of censoring $E\{\lambda(t|\mathcal{H}(t^-))|H(t^-)\}$ are as defined in (3.1), (3.2) and (3.3) respectively. The probability of remaining at risk following the jth event given $\mathcal{H}(t_j^-)$ is as defined in (3.5)

$$P(Z_j = 1|\mathcal{H}(t_j^-), dN(t_j) = 1, X_j) = P(Z_j = 1|H(t_j^-), \bar{Z}_{j-1} = 1_{j-1}, dN(t_j) = 1, X_j) ,$$

where $dN(t) = \lim_{\Delta t \to 0} \Delta N(t)$ indicates whether an event occurred at time t and $\bar{Z}_{j-1} = 1_{j-1}$ denotes an $j \times 1$ vector of ones as it implies $Z_0 = Z_1 = \ldots = Z_{j-1} = 1$. A simple model as in (3.6) can be adopted to model the resolution process with the form

$$\text{logit} P(Z_j = 1|\mathcal{H}(t_j^-), dN(t_j) = 1, X_j) = \dot{X}_j' \gamma_j ,$$

where $\dot{X}_j = (1, X_j')'$ and γ_j parameterizes the association between the explanatory variables and the mover-stayer indicator. Note that for simplicity we could set the covariate vector to be fixed and observed at study entry and use X to denote it, and let $\gamma_j = \gamma$ so that the sets of regression coefficients are same across all logistic models.

4.2 Estimation with Interval-Censored Data

4.2.1 The Complete Data Likelihood for Interval-Censored Data

In what follows we consider data from a single individual. Let a_0 denote the time of a baseline assessment at which a $p \times 1$ fixed covariate vector X is observed. Suppose follow-up assessments occur at times $a_1 < \cdots < a_R$ and at a_r the number of events over interval $\mathcal{A}_r = (a_{r-1}, a_r]$ is recorded, denoted by $n_r = N(a_r) - N(a_{r-1})$, $r = 1, \ldots, R$. The data for such an individual is then $D = \{(a_r, n_r), r = 1, \ldots, R, X\}$.

Let θ_1 index the canonical event intensity (3.2), and θ_2 index the mover-stayer model (3.5). A complete data log-likelihood can be constructed by considering the event times and the latent mover-stayer indicators as observed from the sample. When the covariate process is non-informative, the contribution to such a log-likelihood from an individual is then

$$\ell_C(\theta) = \ell_{C1}(\theta_1) + \ell_{C2}(\theta_2) \tag{4.1}$$

where

$$\ell_{C1}(\theta_1) = \int_0^\infty Y(u) \left[dN(u) \log \lambda(u|H(u^-)) - Z_{N(u)} \lambda(u|H(u^-)) du \right] \tag{4.2}$$

pertains to the latent event process and

$$\ell_{C2}(\theta_2) = \sum_{j=0}^{n} \log P(Z_j | H(t_j^-), \bar{Z}_{j-1} = 1_{j-1}, X) \tag{4.3}$$

pertains to the latent mover-stayer model, where $\bar{Z}_{-1} = \emptyset$ and $H(0^-) = \emptyset$.

Suppose interest lies in modeling data from individuals over the interval $[0, \tau]$ where τ is fixed. We focus here on settings with latent multiplicative Poisson processes, where

$$\lambda(t|H(t^-)) = \rho(t|X) = \rho_0(t)\exp(X^{'}\beta)$$

and $\rho_0(t)$ is the canonical baseline rate function. Given a set of cut-points $0 = b_0 < b_1 < b_2 < \cdots < b_K = \tau$, flexible piecewise constant baseline rate functions are obtained by letting $\rho_0(t) = \rho_k$ for $t \in \mathcal{B}_k = [b_{k-1}, b_k)$, $k = 1, \ldots, K$. One can then write

$$\rho(t|X;\theta_1) = \rho_0(t;\alpha)\exp(X'\beta) = \prod_{k=1}^{K}[\exp(\alpha_k + X'\beta)]^{d_k(t)} \, .$$

where $\alpha_k = \log \rho_k$, $k = 1, \ldots, K$, $\alpha = (\alpha_1, \ldots, \alpha_K)'$, $\theta_1 = (\alpha', \beta')'$ and $d_k(t) = I(t \in \mathcal{B}_k) = I(b_{k-1} \leq t < b_k)$, $k = 1, \ldots, K$.

We let $\mathcal{C}_{rk} = \mathcal{A}_r \cap \mathcal{B}_k = [C_{r,k-1}, C_{rk})$ and let $\mathcal{K}^r = \{k : \mathcal{A}_r \cap \mathcal{B}_k \neq \varnothing\}$ represent the labels for the q_r $(0 \leq q_r \leq K)$ pieces intersecting $\mathcal{A}_r$, denoted $\{k_\ell^r, \ell = 1, \ldots, q_r\}$. If we let $n_{rk} = \int I(u \in \mathcal{C}_{rk})dN(u)$ denote the number of events over $\mathcal{C}_{rk}$ and $w_k(t) = \int_0^t I(u \in \mathcal{B}_k)du$ denote the time at risk in $\mathcal{B}_k$ over $(0, t]$, then (4.2) can be rewritten as

$$\begin{aligned}\ell_{C1}(\theta_1) \;\; = \;\; & \sum_{k=1}^{K}\sum_{r=1}^{R} n_{rk}(\alpha_k + X'\beta) - \\ & \sum_{k=1}^{K}[Z_n w_k(a_R) + (1 - Z_n)w_k(t_n)]\exp(\alpha_k + X'\beta)\,.\end{aligned}$$

If we use $w_{rk}(t) = \int_0^t I(u \in \mathcal{C}_{rk})du$ to denote the time at risk in $\mathcal{C}_{rk}$ over $(0, t]$, then

$w_k(a_R) = \sum_{r=1}^{R} w_{rk}(a_R)$, $w_k(t_n) = \sum_{r=1}^{R} w_{rk}(t_n)$ and (4.2) becomes

$$\begin{aligned} \ell_{C1}(\theta_1) \;&=\; \sum_{k=1}^{K}\sum_{r=1}^{R}\Big\{ n_{rk}(\alpha_k + X'\beta) - \\ &\qquad \big[Z_n w_{rk}(a_R) + (1-Z_n) w_{rk}(t_n)\big]\exp(\alpha_k + X'\beta)\Big\} . \end{aligned} \tag{4.4}$$

4.2.2 Derivation of the Conditional Expectations

Since the actual events times and the final mover-stayer indicator are not observed, the quantities n_{rk}, $w_k(a_R)$, $w_k(t_n)$, and Z_n in (4.1) are unknown and we require expressions for their conditional expectations (Dempster et al., 1977). We focus initially on the expectations given D and Z_n, and consider first the case in which $Z_n = 1$. We let

$$\eta_{rk}^{(1)} = E(n_{rk}|D, Z_n = 1) = \frac{n_r \mu_{rk}}{\sum_{k\in\mathcal{K}^r} \mu_{rk}} \ , \tag{4.5}$$

where

$$\mu_{rk} = \int I(u \in \mathcal{C}_{rk})\rho(u|X)du = \exp(\alpha_k + X'\beta)|\mathcal{C}_{rk}| \ ,$$

denotes the cumulative intensity over $\mathcal{C}_{rk}$ and

$$|\mathcal{C}_{rk}| = \max(0, \min(b_k, a_r) - \max(b_{k-1}, a_{r-1}))$$

denotes the length of $\mathcal{C}_{rk}$, $k \in \mathcal{K}^r$, $r = 1, \ldots, R$ (Lawless and Zhan, 1998). Here we use a superscript (1) to reflect the fact that their expectation is given $Z_n = 1$. Given $Z_n = 1$ we can write

$$\omega_{rk}^{(1)} = E(w_{rk}(a_R)|D, Z_n = 1) = |\mathcal{C}_{rk}| \ .$$

Next we consider the case when $Z_n = 0$ and use a superscript (0) to reflect the fact that the conditional expectation is given $Z_n = 0$. If s denotes the index for the inspection interval containing t_n (i.e. $t_n \in \mathcal{A}_s$), then when $r < s$,

$$\eta_{rk}^{(0)} = E(n_{rk}|D, Z_n = 0) = E(n_{rk}|D, Z_n = 1)$$

as in (4.5) for any $k \in \mathcal{K}^r$. When $r > s$,

$$\eta_{rk}^{(0)} = E(n_{rk}|D, Z_n = 0) = 0$$

for any $k \in \mathcal{K}^r$ by the definition of $\mathcal{A}_s$. Note that $E(n_{sk}|D, Z_n = 0)$, $k \in \mathcal{K}^s$, can be obtained by conceptualizing a progressive time nonhomogeneous multistate Markov process with a finite number of states labelled $N(a_{s-1}), \ldots, N(a_s)$ where only $\ell \to \ell + 1$ transitions are allowed with a common "transition" intensity $\rho(u|X)$, $\ell = N(a_{s-1}), \ldots, N(a_s) - 1$ and $N(a_s) = n$ is an absorbing state. We let

$$\boldsymbol{n}_s = (n_{sk}, k = k_1^s, \ldots, k_{q_s}^s)$$

denote the counts over the sub-intervals of $\mathcal{A}_s$, let

$$\bar{n}_{sk} = n(a_{s-1}) + \sum_{j \in \mathcal{K}^s} I(j \leq k) n_{sj}$$

denote the cumulative count at C_{sk}, and let

$$\bar{\boldsymbol{n}}_s = (\bar{n}_{sk}, k = k_1^s, \ldots, k_{q_s}^s)$$

denote the vector of cumulative counts. We can then write the joint probability of the latent states over $\boldsymbol{\mathcal{A}}_s$ given the observed data when $Z_n = 0$, $P(\boldsymbol{n}_s|D, Z_n = 0)$, as

$$P(N(C_{sk}) = \bar{n}_{sk} \text{ for all } k \in \mathcal{K}^s | N(a_{s-1}) = n(a_{s-1}), N(a_s) = n(a_s), X, Z_n = 0) \ ,$$

where $n(a_{s-1}) = n - n_s$ and $n(a_s) = n$ by the definition of $\boldsymbol{\mathcal{A}}_s$. This can in turn be written as

$$P(\boldsymbol{n}_s|D, Z_n = 0) = \frac{P(N(C_{sk}) = \bar{n}_{sk} \text{ for all } k \in \mathcal{K}^s | N(a_{s-1}) = n(a_{s-1}), X, Z_n = 0)}{P(N(a_s) = n | N(a_{s-1}) = n(a_{s-1}), X, Z_n = 0)} \ , \tag{4.6}$$

where the numerator is equal to

$$\prod_{k \in \mathcal{K}^s} P(N(C_{sk}) = \bar{n}_{sk} | N(C_{s,k-1}) = \bar{n}_{s,k-1}, X, Z_n = 0) \tag{4.7}$$

by the Markov property and the denominator is

$$\sum_{\boldsymbol{n}_s \in \mathcal{N}_s} \prod_{k \in \mathcal{K}^s} P(N(C_{sk}) = \bar{n}_{sk} | N(C_{s,k-1}) = \bar{n}_{s,k-1}, X, Z_n = 0) \ , \tag{4.8}$$

where $\mathcal{N}_s = \{\boldsymbol{n}_s : n_s = \sum_{k \in \mathcal{K}^s} n_{sk}\}$ is the set of all vectors $\boldsymbol{n}_s$ compatible with observed total over $\boldsymbol{\mathcal{A}}_s$.

To determine the terms in (4.7), for a given vector $\boldsymbol{n}_s$ we further consider the specific subinterval $\mathcal{C}_{s\ell}$ containing t_n (i.e. $t_n \in \mathcal{C}_{s\ell}$). For $k \in \mathcal{K}^s$, when $k < \ell$,

$$P(N(C_{sk}) = \bar{n}_{sk} | N(C_{s,k-1}) = \bar{n}_{s,k-1}, X, Z_n = 0) = \mu_{sk}^{n_{sk}} \exp(-\mu_{sk})/n_{sk}! \; ,$$

where $\mu_{sk} = \exp(\alpha_k + X'\beta)|\mathcal{C}_{sk}|$. When $k > \ell$,

$$P(N(C_{sk}) = \bar{n}_{sk} | N(C_{s,k-1}) = \bar{n}_{s,k-1}, X, Z_n = 0) = 1 \; .$$

When $k = \ell$, a time-homogeneous Markov process governors events over $\mathcal{C}_{s\ell}$ with allowable transitions $0 \to 1 \to \ldots \to N_\ell = n_{s\ell}$ occurring with rate $\exp(\alpha_\ell + X'\beta)$. Note that the probability of making transition from state i to state j, $i, j = 0, \ldots, N_\ell$, within time t is

$$P_{ij}(t) \quad = \quad (\exp(\alpha_\ell + X'\beta)\, t)^{j-i} \exp(-\exp(\alpha_\ell + X'\beta)\, t)/(j-i)! \tag{4.9}$$

if $0 \leq i \leq j < N_\ell$, with $P_{iN_\ell}(t) = 1 - \sum_{j=i}^{N_\ell - 1} P_{ij}(t)$ if $0 \leq i \leq N_\ell - 1$. Given this we can calculate $P(N(C_{s\ell}) = \bar{n}_{s\ell} | N(C_{s,\ell-1}) = \bar{n}_{s,\ell-1}, X, Z_n = 0)$ as $P_{0N_\ell}(|\mathcal{C}_{s\ell}|)$.

Note $P(\boldsymbol{n}_s | D, Z_n = 0)$ in (4.6) can then be used to compute the conditional expectation for the counts in each sub-interval of $\mathcal{A}_s$ since

$$\eta_{sk}^{(0)} = E(n_{sk} | D, Z_n = 0) = \sum_{n_{sk}=0}^{n_s} n_{sk} P(n_{sk} | D, Z_n = 0) \; , \tag{4.10}$$

where

$$P(n_{sk} | D, Z_n = 0) = \sum_{\boldsymbol{n}_s \in \mathcal{N}_s} I(N_{sk} = n_{sk}) P(\boldsymbol{n}_s | D, Z_n = 0)$$

for any $k \in \mathcal{K}^s$.

The conditional expectation of the time at risk in each $\mathcal{C}_{rk}$ when $Z_n = 0$ can be obtained by following similar idea. Since $t_n \in \mathcal{A}_s$, it is easy to see that when $r < s$,

$$\omega_{rk}^{(0)} = E(w_{rk}(t_n)|D, Z_n = 0) = |\mathcal{C}_{rk}|$$

for all $k \in \mathcal{K}^r$, and when $r > s$

$$\omega_{rk}^{(0)} = E(w_{rk}(t_n)|D, Z_n = 0) = 0$$

for all $k \in \mathcal{K}^r$ by the definition of $\mathcal{A}_s$. When $r = s$, $\mathcal{K}^s = \{k_\ell^s, \ell = 1, \ldots, q_s\}$, for a given count vector $\boldsymbol{n}_s = (n_{sk}, k \in \mathcal{K}^s)$, we can further find a ℓ such that $t_n \in \mathcal{C}_{s\ell}$. Once again, for $k \in \mathcal{K}^s$, when $k < \ell$, we have

$$E(w_{sk}(t_n)|D, Z_n = 0) = |\mathcal{C}_{sk}| \ ,$$

and when $k > \ell$,

$$E(w_{sk}(t_n)|D, Z_n = 0) = 0$$

by the definition of $\mathcal{C}_{s\ell}$. When $r = s$ and $k = \ell$,

$$w_{s\ell}(t_n) = \int_0^{t_n} I(u \in \mathcal{C}_{s\ell})du = \int_{\mathcal{C}_{s\ell}} I(t_n > u)du = \int_{\mathcal{C}_{s\ell}} I(N(u) < n)du \ ,$$

and for $u \in \mathcal{C}_{s\ell}$, the later expression can be helpful since

$$
\begin{aligned}
& P(N(u) < n | D, Z_n = 0, \boldsymbol{n}_s) \\
= & \sum_{n - n_{s\ell} \leq j < n} P(N(u) = j | D, Z_n = 0, \boldsymbol{n}_s) \\
= & \sum_{n - n_{s\ell} \leq j < n} \frac{P(N(u) = j | N(C_{s,\ell-1}) = n - n_{s\ell}, X, Z_n = 0) P(N(C_{s\ell}) = n | N(u) = j, X, Z_n = 0)}{P(N(C_{s\ell}) = n | N(C_{s,\ell-1}) = n - n_{s\ell}, X, Z_n = 0)} \\
= & \sum_{j=0}^{n_{s\ell}-1} \frac{P(N(u) = j | N(C_{s,\ell-1}) = 0, X, Z_n = 0) P(N(C_{s\ell}) = n_{s\ell} | N(u) = j, X, Z_n = 0)}{P(N(C_{s\ell}) = n_{s\ell} | N(C_{s,\ell-1}) = 0, X, Z_n = 0)} \qquad (4.11)
\end{aligned}
$$

by the Markov property. Note that over $\mathcal{C}_{s\ell}$, we again have a continuous time Markov process with a time homogenous transition intensity $\exp(\alpha_\ell + X'\beta)$ and transitions only from ℓ to $\ell + 1$ for $\ell = 0, \ldots, N_\ell - 1$ where $N_\ell = n_{s\ell}$, we can therefore use (4.9) to obtain the values of the individual items in (4.11) for given $u \in \mathcal{C}_{s\ell}$, and obtain

$$
E(w_{s\ell}(t_n) | D, Z_n = 0, \boldsymbol{n}_s) = \int_{\mathcal{C}_{s\ell}} P(N(u) < n | D, Z_n = 0, \boldsymbol{n}_s) du
$$

via numerical integration. Finally, for any $k \in \mathcal{K}^s$, we could calculate

$$
\omega_{sk}^{(0)} = E(w_{sk}(t_n) | D, Z_n = 0) = \sum_{\boldsymbol{n}_s \in \mathcal{N}_s} E(w_{sk}(t_n) | D, Z_n = 0, \boldsymbol{n}_s) P(\boldsymbol{n}_s | D, Z_n = 0) \ ,
$$

where $P(\boldsymbol{n}_s | D, Z_n = 0)$ is given in (4.6).

Finally we consider $\zeta = E(Z_n | D)$. Note that $P(Z_n = 1 | D)$ can be written as

$$
\frac{P(D | Z_n = 1) P(Z_n = 1 | H(t_n^-), \bar{Z}_{n-1} = 1_{n-1}, X)}{P(D | Z_n = 1) P(Z_n = 1 | H(t_n^-), \bar{Z}_{n-1} = 1_{n-1}, X) + P(D | Z_n = 0) P(Z_n = 0 | H(t_n^-), \bar{Z}_{n-1} = 1_{n-1}, X)} \ ,
$$

where

$$P(D|Z_n = 1) = P(n_1, \ldots, n_R | a_1, \ldots, a_R, X, Z_n = 1) = \prod_{r=1}^{R} \frac{\mu_r^{n_r} e^{-\mu_r}}{n_r!} \ ,$$

with

$$\mu_r = \int_{\mathcal{A}_r} \rho(u|X) du = \sum_{k \in \mathcal{K}^r} \mu_{rk}$$

denoting the cumulative intensity over $\mathcal{A}_r$, while

$$\begin{aligned} P(D|Z_n = 0) &= P(n_1, \ldots, n_R | a_1, \ldots, a_R, X, Z_n = 0) \\ &= \prod_{r=1}^{s-1} \frac{\mu_r^{n_r} e^{-\mu_r}}{n_r!} \cdot P(N(a_s) = n(a_s) | N(a_{s-1}) = n(a_{s-1}), X, Z_n = 0) \ , \end{aligned}$$

where $P(N(a_s) = n(a_s)|N(a_{s-1}) = n(a_{s-1}), X, Z_n = 0)$ is given in (4.8). Therefore, ζ is an fraction with numerator

$$\frac{\mu_s^{n_s} e^{-\mu_s}}{n_s!} \prod_{r=s+1}^{R} e^{-\mu_{sr}} \cdot P(Z_n = 1 | H(t_n^-), \bar{Z}_{n-1} = 1_{n-1}, X) \tag{4.12}$$

and denominator given by the sum of (4.12) and

$$P(N(a_s) = n | N(a_{s-1}) = n(a_{s-1}), X, Z_n = 0) \cdot P(Z_n = 0 | H(t_n^-), \bar{Z}_{n-1} = 1_{n-1}, X) \ ,$$

where $P(Z_n = 1|H(t_n^-), \bar{Z}_{n-1} = 1_{n-1}, X)$ follows some particular model assumption such as

$$P(Z_n = 1 | H(t_n^-), \bar{Z}_{n-1} = 1_{n-1}, X) = \text{expit}(\gamma_0 + \gamma_1 n + \gamma_2 X) \ .$$

4.2.3 The EM Algorithm

For the EM algorithm, we have

$$Q(\theta;\widetilde{\theta}) = Q_1(\theta_1;\widetilde{\theta}) + Q_2(\theta_2;\widetilde{\theta}) \ ,$$

where $Q(\theta;\widetilde{\theta}) = E(\ell_C(\theta)|D,\widetilde{\theta})$ and $Q_j(\theta_j;\widetilde{\theta}) = E(\ell_C j(\theta_j)|D,\widetilde{\theta})$, $j = 1,2$, $\theta_1 = (\alpha',\beta')'$, $\theta_2 = \gamma$ and $\theta = (\theta_1',\theta_2')'$.

Given the expressions in Section 4.2.2, adopting notations as

$$\begin{aligned}
\widetilde{\eta}_{irk}^{(1)} &= E(n_{irk}|D_i, Z_{in_i} = 1;\widetilde{\theta}) \\
\widetilde{\eta}_{irk}^{(0)} &= E(n_{irk}|D_i, Z_{in_i} = 0;\widetilde{\theta}) \\
\widetilde{\omega}_{irk}^{(1)} &= E(w_{irk}(C_i)|D_i, Z_{in_i} = 1;\widetilde{\theta}) \\
\widetilde{\omega}_{irk}^{(0)} &= E(w_{irk}(t_{in_i})|D_i, Z_{in_i} = 0;\widetilde{\theta}) \\
\widetilde{\zeta}_i &= E(Z_{in_i}|D_i;\widetilde{\theta})
\end{aligned}$$

we'd have $Q_1(\theta_1;\widetilde{\theta})$ to be maximized with respect to the event process as

$$\begin{aligned}
Q_1(\theta_1;\widetilde{\theta}) &= \sum_{i=1}^{m}\left\{\widetilde{\zeta}_i \sum_{k=1}^{K}\sum_{r=1}^{R_i}\left[\widetilde{\eta}_{irk}^{(1)}(\alpha_k + X_i'\beta) - \exp(\alpha_k + X_i'\beta + \log\widetilde{\omega}_{irk}^{(1)})\right] + \right. \\
&\left. (1-\widetilde{\zeta}_i)\sum_{k=1}^{K}\sum_{r=1}^{R_i}\left[\widetilde{\eta}_{irk}^{(0)}(\alpha_k + X_i'\beta) - \exp(\alpha_k + X_i'\beta + \log\widetilde{\omega}_{irk}^{(0)})\right]\right\} \ .
\end{aligned}$$

One can construct a pseudo-data frame for each individual as in Table 4.1, which in fact can be written more concisely as in Table 4.2 and use

$$\mathrm{glm(n \sim X + factor(k) + offset(\log w), weight = wt, family = poisson, link = log)}$$

to obtain the updated estimates of θ_1 after some manipulation on the coefficients of the model fitted.

Table 4.1: Pseudo-data frame for recurrent event process for one subject using the proposed method, where r is the index of the inspection interval, k is the index of the piece of the baseline rate, X is the covariate, w is the expected time at risk, n is the expected number of events, and wt is the weight

r	k	X	w	n	wt
1	k_1^1	x	$\widetilde{\omega}^{(1)}_{1k_1^1}$	$\widetilde{\eta}^{(1)}_{1k_1^1}$	$\widetilde{\zeta}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	$k_{q_1}^1$	x	$\widetilde{\omega}^{(1)}_{1k_{q_1}^1}$	$\widetilde{\eta}^{(1)}_{1k_{q_1}^1}$	$\widetilde{\zeta}$
1	k_1^1	x	$\widetilde{\omega}^{(0)}_{1k_1^1}$	$\widetilde{\eta}^{(0)}_{1k_1^1}$	$1-\widetilde{\zeta}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	$k_{q_1}^1$	x	$\widetilde{\omega}^{(0)}_{1k_{q_1}^1}$	$\widetilde{\eta}^{(0)}_{1k_{q_1}^1}$	$1-\widetilde{\zeta}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
R	k_1^R	x	$\widetilde{\omega}^{(1)}_{Rk_1^R}$	$\widetilde{\eta}^{(1)}_{Rk_1^R}$	$\widetilde{\zeta}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
R	$k_{q_R}^R$	x	$\widetilde{\omega}^{(1)}_{Rk_{q_R}^R}$	$\widetilde{\eta}^{(1)}_{Rk_{q_R}^R}$	$\widetilde{\zeta}$
R	k_1^R	x	$\widetilde{\omega}^{(0)}_{Rk_1^R}$	$\widetilde{\eta}^{(0)}_{Rk_1^R}$	$1-\widetilde{\zeta}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
R	$k_{q_R}^R$	x	$\widetilde{\omega}^{(0)}_{Rk_{q_R}^R}$	$\widetilde{\eta}^{(0)}_{Rk_{q_R}^R}$	$1-\widetilde{\zeta}$

Table 4.2: Pseudo-data frame for recurrent event process for one subject using the proposed method (simplified version), where r is the index of the inspection interval, k is the index of the piece of the baseline rate, X is the covariate, w is the expected time at risk, n is the expected number of events, and wt is the weight

r	k	X	w	n	wt
1	k_1^1	x	$\widetilde{\omega}_{1k_1^1}^{(1)}$	$\widetilde{\eta}_{1k_1^1}^{(1)}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	$k_{q_1}^1$	x	$\widetilde{\omega}_{1k_{q_1}^1}^{(1)}$	$\widetilde{\eta}_{1k_{q_1}^1}^{(1)}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$s-1$	k_1^{s-1}	x	$\widetilde{\omega}_{s-1,k_1^{s-1}}^{(1)}$	$\widetilde{\eta}_{s-1,k_1^{s-1}}^{(1)}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$s-1$	$k_{q_{s-1}}^{s-1}$	x	$\widetilde{\omega}_{s-1,k_{q_{s-1}}^{s-1}}^{(1)}$	$\widetilde{\eta}_{s-1,k_{q_{s-1}}^{s-1}}^{(1)}$	1
s	k_1^s	x	$\widetilde{\omega}_{sk_1^s}^{(1)}$	$\widetilde{\eta}_{sk_1^s}^{(1)}$	$\widetilde{\zeta}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
s	$k_{q_s}^s$	x	$\widetilde{\omega}_{sk_{q_s}^s}^{(1)}$	$\widetilde{\eta}_{sk_{q_s}^s}^{(1)}$	$\widetilde{\zeta}$
s	k_1^s	x	$\widetilde{\omega}_{sk_1^s}^{(0)}$	$\widetilde{\eta}_{sk_1^s}^{(0)}$	$1-\widetilde{\zeta}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
s	$k_{q_s}^s$	x	$\widetilde{\omega}_{sk_{q_s}^s}^{(0)}$	$\widetilde{\eta}_{sk_{q_s}^s}^{(0)}$	$1-\widetilde{\zeta}$
$s+1$	k_1^{s+1}	x	$\widetilde{\omega}_{s+1,k_1^{s+1}}^{(1)}$	$\widetilde{\eta}_{s+1,k_1^{s+1}}^{(1)}$	$\widetilde{\zeta}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$s+1$	$k_{q_{s+1}}^{s+1}$	x	$\widetilde{\omega}_{s+1,k_{q_{s+1}}^{s+1}}^{(1)}$	$\widetilde{\eta}_{s+1,k_{q_{s+1}}^{s+1}}^{(1)}$	$\widetilde{\zeta}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
R	k_1^R	x	$\widetilde{\omega}_{Rk_1^R}^{(1)}$	$\widetilde{\eta}_{Rk_1^R}^{(1)}$	$\widetilde{\zeta}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
R	$k_{q_R}^R$	x	$\widetilde{\omega}_{Rk_{q_R}^R}^{(1)}$	$\widetilde{\eta}_{Rk_{q_R}^R}^{(1)}$	$\widetilde{\zeta}$

In addition, $Q_2(\theta;\widetilde{\theta})$ to be maximized with respect to the mover-stayer model is

$$\begin{aligned} Q_2(\theta;\widetilde{\theta}) &= \sum_{i=1}^{m}\left[\sum_{j=0}^{n_i-1}\log P(Z_{ij}=1|H(t_{in_i}^-),\bar{Z}_{i,j-1}=1_{j-1},X_i)\right. \\ &\quad +\widetilde{\zeta}_i\log P(Z_{in_i}=1|H(t_{in_i}^-),\bar{Z}_{i,n_i-1}=1_{n_i-1},X_i) \\ &\quad \left.+(1-\widetilde{\zeta}_i)\log P(Z_{in_i}=0|H(t_{in_i}^-),\bar{Z}_{i,n_i-1}=1_{n_i-1},X_i)\right] . \end{aligned}$$

Again we could construct a pseudo-data frame as in Table 4.3 and use

$$\mathrm{glm}(\mathrm{Z}\sim\mathrm{j}+\mathrm{X},\mathrm{weight}=\mathrm{wt},\mathrm{family}=\mathrm{quasi}-\mathrm{binomial},\mathrm{link}=\mathrm{logit})$$

to obtain updated estimates of θ_2.

Table 4.3: Pseudo-data frame for mover-stayer model for one subject using the proposed method, where Z is the mover-stayer indicator, j is the number of events, X is the covariate, wt is the weight

Z	j	X	wt
1	0	x	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	$n-1$	x	1
1	n	x	$\widetilde{\zeta}$
0	n	x	$1-\widetilde{\zeta}$

Quite often the long periods of time without recurrence at the end of follow-up is unnoticed and thus the latent mover-stayer process is ignored in data analysis. Naively treating all subjects as movers will lead to biased estimates of treatment effect and underestimated estimates of event rate; and the association between disease resolution and number of events and treatment is taken as zero. Generally speaking, we could use these naive estimates as the initial value in our proposed EM algorithm. We repeat E-step and M-step above until some pre-specified convergence criterion is met and the final estimate of θ is obtained. Repeat the proposed EM algorithm over multiple datasets and report empirical bias (EBIAS) and empirical standard error (ESE).

Note that for the naive method where we assume $Z_{in_i} = 1$ for all subjects, according to Lawless and Zhan (1998), the log-likelihood that we need to maximize after one E-step is

$$\sum_{i=1}^{m}\sum_{k=1}^{K}\sum_{r=1}^{R_i}\left[\widetilde{\eta}_{irk}^{(1)}(\alpha_k + X_i'\beta) - \exp(\alpha_k + X_i'\beta + \log\widetilde{\omega}_{irk}^{(1)})\right] ,$$

That is, for subject i, we could create a pseudo-data frame as in Table 4.4 and use

$$\mathrm{glm(n \sim X + factor(k) + offset(\log w), family = poisson, link = log)}$$

to obtain the updated estimates of θ_1 after some manipulation. Repeat until the preset convergence criterion is met.

4.3 Simulation Studies

In this section, a simulation study is conducted to evaluate the performance of the method we proposed to deal with interval-censored recurrent event data with disease resolution.

Here for subject i, we let the treatment X_i be binary and follow a Bernoulli distribution

Table 4.4: Pseudo-data frame for recurrent event process for one subject using the naive method, where r is the index of the inspection interval, k is the index of the cut-interval, X is the covariate, w is the expected time at risk, n is the expected number of events

r	k	X	w	n
1	k_1^1	x	$\widetilde{\omega}^{(1)}_{1k_1^1}$	$\widetilde{\eta}^{(1)}_{1k_1^1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	$k_{q_1}^1$	x	$\widetilde{\omega}^{(1)}_{1k_{q_1}^1}$	$\widetilde{\eta}^{(1)}_{1k_{q_1}^1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
R	k_1^R	x	$\widetilde{\omega}^{(1)}_{Rk_1^R}$	$\widetilde{\eta}^{(1)}_{Rk_1^R}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
R	$k_{q_R}^R$	x	$\widetilde{\omega}^{(1)}_{Rk_{q_R}^R}$	$\widetilde{\eta}^{(1)}_{Rk_{q_R}^R}$

with equal probability of being on either one of the two treatments. We then generate the initial mover-stayer indicators Z_{ij}, $j = 0, \cdots, n_i$, following the model

$$\text{logit} P(Z_{ij} = 1 | \mathcal{H}(t_{ij}^-), dN(t_{ij}) = 1, X_i) = \gamma_0 + \gamma_1 j + \gamma_2 X_i \ ,$$

where n_i is the total number of events observed over the entire study period $(0, \tau]$ if there is administrative censoring only. For simplicity we set $\tau = 1$ for all subjects. γ_1 and γ_2 are set as $\log 0.95$ and $\log 0.75$ respectively, so that for given treatment X_i, the odds of being a mover decreases by 5% with the occurrence of each additional event, and the odds of being a mover are 25% lower in the $X_i = 1$ group compared to the $X_i = 0$ group if the number of events that occurred are the same. For the purpose of illustration, we assume a homogenous Poisson process, where gap times follow an Exponential distribution with rate

$\lambda \exp(X_i\beta)$, where λ is the baseline intensity and β is the treatment effect on the event rate. We let $\beta = \log 0.75$ so that there is a 25% reduction in event rate if the movers are on treatment $X_i = 1$. Then we can solve for λ if the expected number of events among movers up to time τ is specified, as 6 or 12 for example, we could then solve for γ_0 given the averaged expected number of events among the mixed sample of movers and stayers, say 1.5, 3 or 6.

We could let the number of assessments follow a Poisson distribution with rate specified to allow assessment times to vary as is often the case in observation studies. For simplicity, we assume each subject has same number of assessments with $R = 4$ or $R = 8$ as evenly pre-scheduled and precisely followed clinic visits. For the interval-censored recurrent event data, we only obtain the number of events between clinic visits and the treatment that the subject is on.

As for the convergence criterion, we could let $\vartheta = (\alpha_1, \alpha_2 - \alpha_1, \cdots, \alpha_K - \alpha_1, \beta, \gamma_0, \gamma_1, \gamma_2)$ and the EM procedure is stopped when $\max(|\vartheta^{new} - \vartheta^{old}|) < \epsilon$, where ϵ is some pre-specified value, say $\epsilon = 10^{-6}$.

We conduct 2000 simulations with respect to each set of parameter values and let the sample size be $m = 500$ or 2000. We then apply our proposed EM algorithm while adopting a piecewise constant baseline hazard model. The entire observation period $(0, \tau]$ is evenly divided into $K = 3$ intervals on which the baseline intensity is constant. Both empirical biases (EBIAS) and empirical standard errors (ESE) are reported to summarize the performance of the estimates.

As we can see from the simulation results in Table 4.5 and Table 4.6, the empirical biases

Table 4.5: Simulation results where the gap times follow Exponential distribution by adopting a piecewise constant model when the data are subject to administrative censoring only, $m = 500$ or $m = 2000$, $nsim = 2000$, $E(N(1)|Z_n = 1) = 6$

		R=4				R=8			
		m=500		m=2000		m=500		m=2000	
Prmt	Value	EBIAS	ESE	EBIAS	ESE	EBIAS	ESE	EBIAS	ESE
					$E(N(1)) = 1.5$				
γ_0	0.7091	0.0041	0.1009	0.0001	0.0499	0.0039	0.1005	0.0002	0.0497
γ_1	-0.0513	-0.0014	0.0510	0.0005	0.0256	-0.0022	0.0481	0.0000	0.0237
γ_2	-0.2877	-0.0038	0.1237	0.0004	0.0628	-0.0040	0.1224	0.0001	0.0624
α_1	1.9253	0.0009	0.0654	0.0004	0.0329	0.0003	0.0621	0.0000	0.0306
α_2	1.9253	-0.0113	0.1035	-0.0023	0.0504	-0.0036	0.0897	-0.0006	0.0441
α_3	1.9253	-0.0095	0.2668	-0.0069	0.1310	-0.0086	0.2052	-0.0043	0.1016
β	-0.2877	0.0005	0.0896	-0.0019	0.0431	0.0005	0.0871	-0.0017	0.0410
					$E(N(1)) = 3$				
γ_0	1.7331	0.0100	0.1174	0.0038	0.0584	0.0105	0.1165	0.0037	0.0580
γ_1	-0.0513	0.0008	0.0424	-0.0007	0.0206	0.0002	0.0398	-0.0006	0.0192
γ_2	-0.2877	-0.0068	0.1395	-0.0030	0.0684	-0.0076	0.1376	-0.0029	0.0677
α_1	1.9253	-0.0017	0.0500	-0.0014	0.0252	-0.0013	0.0474	-0.0012	0.0240
α_2	1.9253	-0.0004	0.0661	0.0009	0.0326	0.0001	0.0589	0.0010	0.0291
α_3	1.9253	-0.0021	0.1163	0.0003	0.0594	-0.0014	0.1003	-0.0006	0.0499
β	-0.2877	-0.0004	0.0641	0.0002	0.0313	-0.0001	0.0631	0.0002	0.0307

are generally small with acceptable empirical standard errors. When the parameter settings are all the same, both the empirical bias and the empirical standard error decrease as the sample size increases. When the sample sizes are the same, the estimators overall result in smaller empirical biases and smaller empirical standard errors with more frequent clinic assessments given all the other parameters are the same, which agrees with our intuition.

Table 4.6: Simulation results where the gap times follow Exponential distribution by adopting a piecewise constant model when the data are subject to administrative censoring only, $m = 500$ or $m = 2000$, $nsim = 2000$, $E(N(1)|Z_n = 1) = 12$

		R=4				R=8			
		m=500		m=2000		m=500		m=2000	
Prmt	Value	EBIAS	ESE	EBIAS	ESE	EBIAS	ESE	EBIAS	ESE
				$E(N(1)) = 3$					
γ_0	1.4123	0.0032	0.0944	-0.0002	0.0484	0.0036	0.0938	-0.0002	0.0481
γ_1	-0.0513	-0.0018	0.0221	-0.0003	0.0113	-0.0022	0.0211	-0.0003	0.0108
γ_2	-0.2877	0.0010	0.1089	0.0004	0.0542	0.0005	0.1085	0.0003	0.0541
α_1	2.6184	0.0032	0.0447	0.0001	0.0221	0.0024	0.0415	0.0000	0.0205
α_2	2.6184	-0.0026	0.0779	-0.0000	0.0389	0.0007	0.0654	0.0004	0.0329
α_3	2.6184	-0.0039	0.2066	0.0004	0.0999	-0.0004	0.1426	0.0004	0.0701
β	-0.2877	-0.0042	0.0592	-0.0012	0.0289	-0.0040	0.0560	-0.0011	0.0274
				$E(N(1)) = 6$					
γ_0	2.4275	0.0037	0.1145	0.0003	0.0575	0.0046	0.1125	0.0005	0.0568
γ_1	-0.0513	-0.0005	0.0187	-0.0001	0.0091	-0.0009	0.0175	-0.0002	0.0086
γ_2	-0.2877	0.0041	0.1194	0.0007	0.0599	0.0039	0.1186	0.0007	0.0591
α_1	2.6184	0.0020	0.0342	0.0006	0.0168	0.0013	0.0321	0.0005	0.0157
α_2	2.6184	-0.0008	0.0482	0.0003	0.0234	0.0005	0.0409	0.0004	0.0204
α_3	2.6184	-0.0011	0.0783	-0.0011	0.0386	0.0004	0.0623	-0.0006	0.0312
β	-0.2877	-0.0026	0.0410	-0.0008	0.0197	-0.0024	0.0402	-0.0008	0.0192

When the expected numbers of events among the movers are the same, the covariate effects on the recurrent event rate has smaller empirical biases and smaller empirical standard errors when the marginal expected number of events among the mixture of movers and stayers is bigger, given the same sample size and the same number of clinical examinations. In such cases, the empirical standard errors in event rate estimation are also smaller with insignificant biases. If the marginal expected numbers of events are the same, both the association between explanatory variables and the mover-stayer indicator and the covariate effect on event rate are better assessed with smaller standard errors and trivial biases when the expected number of events among movers is larger.

4.4 Modeling Joint Damage in a Psoriatic Arthritis Cohort Study

HLA-B27 (human leukocyte antigen B27) is a protein found on the surface of white blood cells. It can be tested from a blood sample. Its prevalence in general population varies significantly. It is found to be strongly associated with inflammatory disease such as ankylosing spondylitis and psoriatic arthritis. We are interested in examining the effect of the genetic marker HLA-B27 on the development of damage as measured by radiographic examination among patients with psoriatic arthritis. The damage is assessed in each of 42 joints, including 30 hand joints (wrists, metacarpophalangeals, proximal interphalangeals and distal interphalangeals) and 12 foot joints (metatarsophalangeals and interphalangeal fist toes). Each joint is scored as 0 (normal), 1 (soft tissue swelling), 2 (surface erosions), 3 (joint space narrowing), 4 (disorganization, including subluxation, pencil-in-cup deformity and ankylosis) or 5 (requiring surgery). A joint scoring 2 or higher is counted damaged.

We consider a sub-cohort of patients with psoriatic arthritics from University of Toronto Psoriatic Arthritis Clinic. These 207 selected patients have disease onset time and HLA-B27 information available. They entered the clinic and were followed-up between 1978 and 2013. The observation period is limited to be within 30 years after disease onset. The reported age of disease onset is taken as the time origin and dates of radiological assessments and numbers of new damaged joints were recorded at the following assessment visits. The average time since disease onset to first radiological assessment is 5.54 years (S.D. 6.02, range 0.03 to 27.23). The average number of radiological assessments within 30 years of disease onset is 3.63 (S.D. 2.83, range 1 to 13). A total of 32 (15.5%) patients are HLA-B27 positive.

The data suggested that some patients experience remission during the follow-up. We fit the proposed algorithm on the interval-censored recurrent event data to study the occurrence of joint damage. Piecewise constant baseline rate functions are adopted to model the recurrent event process not subject to resolution with one fixed covariate HLA-B27 ($X = 1$ if HLA-B27 positive, $X = 0$ if HLA-B27 negative); The canonical baseline rate are assumed to be constant for every 10 years. A dynamic mover-stayer model is fitted with the number of damaged joints (j) and HLA-B27 (X) being the explanatory variables in the logistic regression. A recurrent event model treating all patients as susceptible for joint damage is fitted for comparison, in which a piecewise constant baseline rate model is assumed as well and the effect of HLA-B27 on event rate is also of interest.

The estimation results, presented in Table 4.7, demonstrate that the event rate is noticeably underestimated when all the patients are assumed to experience joint damage over the entire observation period. The effect of HLA-B27 on event rate is also underestimated though it is not significant is either model. The increased number of damaged joints is associated with higher odds of continuing to have new damaged joints.

Table 4.7: Results of fitting piecewise constant baseline rate model and dynamic mover-stayer model to study the occurrence of joint damage among patients with psoriatic arthritis whose follow-ups are within 30 years of disease onset; $m = 207$, standard errors based on 100 bootstraps

	Recurrent Event Model		Dynamic Mover-Stayer Model	
	EST	SE	EST	SE
Mover-Stayer Model				
γ_0	—	—	1.6360	0.2566
γ_1	—	—	0.1525	0.1066
γ_2	—	—	-0.2011	0.3173
Recurrent Event Model				
α_1	-0.4672	0.0971	-0.0115	0.1094
α_2	-0.8794	0.1487	-0.5087	0.1882
α_3	-0.9524	0.3647	-0.1330	0.5366
β	-0.1629	0.2693	-0.0693	0.3569

4.5 Remarks

We developed an EM algorithm to analyze interval-censored recurrent event data. A dynamic mover-stayer model was fitted to handle the feature that disease process may resolve at some point and the latent event intensity is assumed to be piecewise constant for baseline function. Some calculation is done under the framework of a progressive time nonhomogeneous finite-numbered multistate Markov process with an absorbing state. Our method worked well in producing small bias despite the difficulty that the exact event times are unobserved, whether the disease process has resolved and the precise resolution time are undetected as well.

The computational challenges associated with models featuring latent processes is made relatively easy through the specification of an EM algorithm which can exploit existing software at the maximization step. While there is a wide class of intensity functions that can be adopted for the right-censored setting, when event times are interval-censored the simplicity of the Poisson assumption for the conditional event process makes it much more attractive than other models.

Chapter 5

Further Research

These three topics of statistical research focus on developing appropriate methods for the analysis of different incomplete lifetime data that are easy to implement with the help of existing software packages. They are all likelihood based approaches and use the EM algorithm to handle the unobserved information in the dataset. Parametric, weakly parametric, non-parametric and semiparametric models are utilized in different settings. The proposed methods are shown to work well empirically and application is done on corresponding motivating studies. It would be interesting to explore these topics further.

5.1 Incomplete Covariates with Left Truncation

In Chapter 2, i.e. Shen and Cook (2013b), we have focused on the setting with two binary covariates for which specification of the population covariate distribution is easy. More complex settings could involve incomplete categorical or continuous covariates and similarly more complex observed covariates. Specification of a model for the joint distribution of the covariates in these settings would be considerably more challenging and indeed one may be willing to give up the potential efficiency gains from the proposed method in order to ensure robustness of the findings. We have also focused on the simplest kind of missing data mechanism, where missingness is driven by a covariate that is always observed. More elaborate missing data mechanisms may require modeling of the missing data process. Standard software can also be used to obtain point estimates of regression coefficients from Cox regression models with incomplete covariates via inverse probability weighted estimating equations. This approach has been considered by several authors (Robins et al., 1994; Lipsitz et al., 1999) and it is of interest to explore this approach in the context of left-truncated data.

In addition to the two settings described so far, truncated data arise naturally in studies of multistate Markov processes. Consider a progressive multistate process comprised of three states with transitions possible from state 1 to state 2 and from state 2 to state 3. The transition time from state 2 to state 3 is typically treated as left-truncated because of the delayed entry time to state 2. When incomplete covariate data arise from such processes likelihoods may have a different form from those considered here depending on the selection process. For example, individuals may be observed from the start of the process, or may be selected for follow-up based on being in state 2; the latter would be more similar to the problem considered in this paper.

Covariates are often imprecisely observed due to misclassification for discrete covariates or measurement error for continuous covariates and there is a large literature on methods for fitting regression models with covariate measurement error (Carroll et al., 2006). When a structural modeling approach is taken models for the latent covariate are adopted, and such models would again require one to specify these models in such a way that the covariate distribution addressed the selection effects arising due to left truncation; this would be necessary for an analysis based on either the observed data likelihood or an EM algorithm.

5.2 Dynamic Mover-Stayer Model for Recurrent Events

The formulation in Section 3.1 is quite flexible given the general form of the latent intensity. We have emphasized simple latent Markov models in our derivations and simulations. Natural extensions include the use of baseline rates which stratify on the cumulative number of events, latent semi-Markov models, or models with hybrid time scales. The expectation-maximization algorithm was described for parametric and semiparametric baseline rates within the latent Markov family of models, but adaptations to these other intensities are relatively straightforward. The introduction of random effects to offer a further avenue for explaining heterogeneity, while possible, may require large sample sizes to ensure convergence. Price and Manatunga (2001) illustrate the interplay between cure rate models and frailty models and Yu (2008) describes a mixture cure model with the latent mover-stayer and frailty variables realized at the time origin. Aalen (1992) discusses the use of a compound Poisson random effect distribution as a means of accommodating a fraction of nonsusceptible individuals as well as heterogeneity in risk among susceptible individuals. More general dynamic mover-stayer models can be specified by building upon these static latent variable models. Issues of estimability arise and become more challenging the more flexible the model components become and examination of profile likelihood contours can be instructive when investigating reasons for convergence problems.

Model assessment is challenging in settings with latent variables and this is particularly true of mixture models of this type. A particular issue of concern is the fact that there may be multiple configurations of the baseline intensity and the mover-stayer model which render similar mean functions. Clear ideas regarding which component of the model covariates are to be placed can help circumvent this challenging problem. Model expansion could be investigated using a likelihood ratio test. Cross-validation is important when the main goal is prediction.

In many settings with recurrent events, the events are not observed but only known to occur between to assessment times. In cohort studies of patients with osteoporosis for example, asymptomatic fractures may be detected upon periodic radiographic examination. Establishment of suitable medications or other changes in lifestyle and diet may minimize risk of further fractures, but it can be difficult to determine if these changes have taken place. The dynamic mover-stayer model offers a way of describing this phenomenon but adaptations to enable model fitting with interval-censored data are required. Cook et al. (2002) offer one such approach in the content of parametric Markov models.

5.3 Interval-Censored Recurrent Processes Subject to Resolution

Some degree of robustness to misspecification is achieved through use of a piecewise constant baseline rate function, but extensions to deal with semiparametric models would be worthy of development. Possible avenues include adapting the pseudo-likelihood estimator proposed by Sun and Kalbfleisch (1995) for the mean function, or the semiparametric maximum likelihood approach of Wellner and Zhang (2000). Here, however, one might expect more challenges in maximization of the observed data likelihood whether by direct maximization or an extension of the algorithm we present here.

Cook et al. (2002) describe a generalized mover-stayer model for multistate data under interval censoring, which is somewhat similar in spirit to what we have described. In this model, conditional on the mover-stayer indicators, subjects move according to time-homogeneous Markov transition intensities. Here however, the first time an individual enters a state, a latent mover-stayer indicator is realized which can render it an absorbing state. Thus individuals can make transitions between a number of states before finally entering their absorbing state.

Often recurrent events arise in settings where the event process is terminated by some event. For example in transplant studies recurrent graft rejection episodes arise when recipients are experiencing graft versus host disease (Cole et al., 1994). This condition resolves at a latent time when the graft is fully accepted, but the process can also end in severe cases by total graft rejection or death of the patient. Adapting these methods to handle this situation is feasible but may again be more naturally addressed by casting the process into a multistate framework as in Conlon et al. (2013). Extensions of these methods would be useful for this setting as well.

References

Aalen, O. O. (1992). Modelling heterogeneity in survival analysis by the compound Poisson distribution. *The Annals of Applied Probability*, 2(4):951–972.

Afifi, A. A. and Elashoff, R. M. (1966). Missing observations in multivariate statistics I. review of the literature. *Journal of the American Statistical Association*, 61(315): 595–604.

Andersen, P. and Gill, R. (1982). Cox's regression model for counting processes: a large sample study. *The Annals of Statistics*, 10(4):1100–1120.

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag, New York.

Anderson, T. W. (1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing. *Journal of the American Statistical Association*, 52(278):200–203.

Balakrishnan, N. and Zhao, X. (2009). New multi-sample nonparametric tests for panel count data. *The Annals of Statistics*, 37(3):1112–1149.

Barber, C. E., Geldenhuys, L. and Hanly, J. (2006). Sustained remission of lupus nephritis. *Lupus*, 15(2):94–101.

Begg, C. B. and Gray, R. J. (1987). Methodology for case-control studies with prevalent cases. *Biometrika*, 74(1):191–195.

Bergeron, P. J., Asgharian, M. and Wolfson, D. B. (2008). Covariate bias induced by length-biased sampling of failure times. *Journal of the American Statistical Association*, 103(482):737–742.

Berkson, J. and Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, 47(259):501–515.

Blumen, I. M., Kogan, M. and McCarthy, P. J. (1955). *The Industrial Mobility of Labor as a Probability Process*, volume 6. Cornell University, Ithaca.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11 (1):15–53.

Bourassa, M. G., Gurné, O., Bangdiwala, S. I., Ghali, J. K., Young, J. B., Rousseau, M., Johnstone, D. E. and Yusuf, S. (1993). Natural history and patterns of current practice in heart failure. *Journal of the American College of Cardiology*, 22(4):A14–A19.

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

Byar, D., Kaihara, S., Sylvester, R., Freedman, L., Hannigan, J., Koiso, K., Oohashi, Y. and Tsugawa, R. (1986). Statistical analysis techniques and sample size determination for clinical trials of treatments for bladder cancer. *Progress in Clinical and Biological Research*, 221:49–64.

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall, Boca Raton, second edition.

Chen, B. E., Cook, R. J., Lawless, J. F. and Zhan, M. (2005). Statistical methods for multivariate interval-censored recurrent events. *Statistics in Medicine*, 24(5):671–691.

Chen, H. Y. and Little, R. J. A. (1999). Proportional hazards regression with missing covariates. *Journal of the American Statistical Association*, 94(447):896–908.

Chen, H. Y. and Little, R. J. A. (2001). A profile conditional likelihood approach for the semiparametric transformation regression model with missing covariates. *Lifetime Data Analysis*, 7:207–224.

Cheng, S. C. and Wei, L. J. (2000). Inferences for a semiparametric model with panel data. *Biometrika*, 87(1):89–97.

Cole, E. H., Cattran, D. C., Farewell, V. T., Aprile, M., Bear, R. A., Pei, Y. P., Fenton, S. S., Tober, J. A. L. and Cardella, C. J. (1994). A comparison of rabbit antithymocyte serum and OKT3 as prophylaxis against renal allograft rejection. *Transplantation*, 57 (1):60–67.

Cole, E. H., Farewell, V. T., Aprile, M., Cattran, D. C., Pei, Y. P., Fenton, S. S., Zaltzman, J. and Cardella, C. J. (1995). Renal transplantation in older patients: the University of Toronto experience. *Geriatric Nephrology and Urology*, 5(2):85–92.

Conlon, A. S. C., Taylor, J. M. G. and Sargent, D. J. (2013). Multi-state models for colon cancer recurrence and death with a cured fraction. *Statistics in Medicine*. doi: 10.1002/sim.6.

Cook, R. J. and Bergeron, P.-J. (2011). Information in the sample covariate distribution in prevalent cohorts. *Statistics in Medicine*, 30(12):1397–1409.

Cook, R. J. and Lawless, J. F. (1997). Marginal analysis of recurrent events and a terminating event. *Statistics in Medicine*, 16(8):911–924.

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

Cook, R. J. and Lawless, J. F. (2013). Concepts and tests for trend in recurrent event processes. *Journal of Iranian Statistical Society*, 12(1):35–69.

Cook, R. J., Kalbfleisch, J. D. and Yi, G. Y. (2002). A generalized mover-stayer model for panel data. *Biostatistics*, 3(3):407–420.

Cox, D. R. (1961). Tests of separate family of hypotheses. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 105–123.

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

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

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

Fang, H. B., Li, G. and Sun, J. (2005). Maximum likelihood estimation in a semiparametric logistic/proportional-hazards mixture model. *Scandinavian Journal of Statistics*, 32(1): 59–75.

Farewell, V. T. (1982). The use of mixture models for the analysis of survival data with long-term survivors. *Biometrics*, 38(4):1041–1046.

Farewell, V. T. (1986). Mixture models in survival analysis: Are they worth the risk? *Canadian Journal of Statistics*, 14(3):257–262.

Fielding, S., Fayers, P. M., McDonald, A., McPherson, G. and Campbell, M. K. (2008). Simple imputation methods were inadequate for missing not at random (MNAR) quality of life data. *Health and Quality of Life Outcomes*, 6(57):1–57.

Frydman, H. (1984). Maximum likelihood estimation in the mover-stayer model. *Journal of the American Statistical Association*, 79(387):632–638.

Fuchs, C. and Greenhouse, J. B. (1988). The EM algorithm for maximum likelihood estimation in the mover-stayer model. *Biometrics*, 44(2):605–613.

Gennari, A., Sormani, M. P., Pronzato, P., Puntoni, M., Colozza, M., Pfeffer, U. and Bruzzi, P. (2008). HER2 status and efficacy of adjuvant anthracyclines in early breast cancer: a pooled analysis of randomized trials. *Journal of the National Cancer Institute*, 100(1):14–20.

Ghosh, D. and Lin, D. Y. (2000). Nonparametric analysis of recurrent events and death. *Biometrics*, 56(2):554–562.

Ghosh, D. and Lin, D. Y. (2002). Marginal regression models for recurrent and terminal events. *Statistica Sinica*, 12(3):663–688.

Gladman, D. D., Farewell, V. T. and Nadeau, C. (1995). Clinical indicators of progression in psoriatic arthritis: multivariate relative risk model. *The Journal of Rheumatology*, 22 (4):675–679.

Gladman, D. D., Hing, E. N., Schentag, C. T. and Cook, R. J. (2001). Remission in psoriatic arthritis. *The Journal of Rheumatology*, 28(5):1045–1048.

Goodman, L. A. (1961). Statistical methods for the mover-stayer model. *Journal of the American Statistical Association*, 56(296):841–868.

Grossman, R., Mukherjee, J., Vaughan, D., Cook, R., LaForge, J., Lampron, N. and Eastwood, C. (1998). A 1-year community-based health economic study of ciprofloxacin vs usual antibiotic treatment in acute exacerbations of chronic bronchitis: the Canadian Ciprofloxacin Health Economic Study Group. *CHEST Journal*, 113(1):131–141.

Healey, L. A. (1984). Long-term follow-up of polymyalgia rheumatica: evidence for synovitis. In *Seminars in Arthritis and Rheumatism*, volume 13, pages 322–328.

Heckman, J. J. and Walker, J. R. (1987). Using goodness of fit and other criteria to choose among competing duration models: A case study of Hutterite data. *Sociological Methodology*, 17:247–307.

Heddle, N. M., Cook, R. J., Webert, K. E., Sigouin, C. and Rebulla, P. (2003). Methodologic issues in the use of bleeding as an outcome in transfusion medicine studies. *Transfusion*, 43(6):742–752.

Herring, A. H., Ibrahim, J. G. and Lipsitz, S. R. (2004). Non-ignorable missing covariate data in survival analysis: a case-study of an International Breast Cancer Study Group trial. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 53(2):293–310.

Hocking, R. R. and Smith, W. B. (1968). Estimation of parameters in the multivariate normal distribution with missing observations. *Journal of the American Statistical Association*, 63(321):159–173.

Hortobagyi, G. N., Theriault, R. L., Porter, L., Blayney, D., Lipton, A., Sinoff, C., Wheeler, H., Simeone, J. F., Seaman, J., Knight, R. D., Heffernan, M., Reitsma, D. J., Kennedy, I., Allan, S. G. and Mellars, K. (1996). Efficacy of pamidronate in reducing skeletal complications in patients with breast cancer and lytic bone metastases. *New England Journal of Medicine*, 335(24):1785–1792.

Hortobagyi, G. N., Theriault, R. L., Lipton, A., Porter, L., Blayney, D., Sinoff, C., Wheeler, H., Simeone, J. F., Seaman, J. J., Knight, R. D., Heffernan, H., Mellars, K. and Reitsma, D. J. (1998). Long-term prevention of skeletal complications of metastatic breast cancer with pamidronate. Protocol 19 Aredia Breast Cancer Study Group. *Journal of Clinical Oncology*, 16(6):2038–2044.

Horton, N. J. and Laird, N. M. (1999). Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research*, 8(1):37–50.

Huang, C. Y., Wang, M. C. and Zhang, Y. (2006). Analysing panel count data with informative observation times. *Biometrika*, 93(4):763–775.

Ibrahim, J., Chen, M.-H. and Kim, S. (2008). Bayesian variable selection for the Cox regression model with missing covariates. *Lifetime Data Analysis*, 14:496–520.

Ibrahim, J. G. (1990). Incomplete data in generalized linear models. *Journal of the American Statistical Association*, 85(411):765–769.

S-PLUS 8 Guide to Statistics (2007). Insightful Corporation, Seattle, WA.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, New York, second edition.

Kessing, L., Mortensen, P. B. and Bolwig, T. G. (1998). Clinical consequences of sensitisation in affective disorder: A case register study. *Journal of Affective Disorders*, 47(1-3): 41–47.

Kessing, L. V., Hansen, M. G. and Andersen, P. K. (2004*a*). Course of illness in depressive and bipolar disorders naturalistic study, 1994-1999. *The British Journal of Psychiatry*, 185(5):372–377.

Kessing, L. V., Hansen, M. G., Andersen, P. K. and Angst, J. (2004*b*). The predictive effect of episodes on the risk of recurrence in depressive and bipolar disorders - a life-long perspective. *Acta Psychiatrica Scandinavica*, 109(5):339–344.

King, G., Honaker, J., Joseph, A. and Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. In *American Political Science Association*, volume 95, pages 49–69. Cambridge University Press.

Klein, J. P. and Moeschberger, M. L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, New York, first edition.

Kuk, A. Y. C. and Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, 79(3):531–541.

Kvist, K., Gerster, M., Andersen, P. K. and Kessing, L. V. (2007). Non-parametric estimation and model checking procedures for marginal gap time distributions for recurrent events. *Statistics in Medicine*, 26(30):5394–5410.

Lawless, J. F. (1987*a*). Negative binomial and mixed Poisson regression. *Canadian Journal of Statistics*, 15(3):209–225.

Lawless, J. F. (1987*b*). Regression methods for Poisson process data. *Journal of the American Statistical Association*, 82(399):808–815.

Lawless, J. F. (1995). The analysis of recurrent events for multiple subjects. *Applied Statistics*, 44(4):487–498.

Lawless, J. F. (2002). *Statistical Models and Methods for Lifetime Data.* John Wiley & Sons, Hoboken, second edition.

Lawless, J. F. and Nadeau, C. (1995). Some simple robust methods for the analysis of recurrent events. *Technometrics*, 37(2):158–168.

Lawless, J. F. and Zhan, M. (1998). Analysis of interval-grouped recurrent-event data using piecewise constant rate functions. *Canadian Journal of Statistics*, 26(4):549–565.

Lin, D. Y., Wei, L. J., Yang, I. and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):711–730.

Lipsitz, S. R. and Ibrahim, J. G. (1996). Using the EM-algorithm for survival data with incomplete categorical covariates. *Lifetime Data Analysis*, 2:5–14.

Lipsitz, S. R. and Ibrahim, J. G. (1998). Estimating equations with incomplete categorical covariates in the Cox model. *Biometrics*, 54(3):1002–1013.

Lipsitz, S. R., Ibrahim, J. G. and Zhao, L. P. (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *Journal of the American Statistical Association*, 94(448):1147–1160.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis With Missing Data.* John Wiley & Sons, Hoboken, second edition.

Liu, L., Wolfe, R. A. and Huang, X. (2004). Shared frailty models for recurrent events and a terminal event. *Biometrics*, 60(3):747–756.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 44(2):226–233.

Moher, D., Schulz, K. F. and Altman, D. G. (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet*, 357(9263):1191–1194.

Musil, C. M., Warner, C. B., Yobas, P. K. and Jones, S. L. (2002). A comparison of imputation techniques for handling missing data. *Western Journal of Nursing Research*, 24(7):815–829.

Park, D.-H., Sun, J. and Zhao, X. (2007). A class of two-sample nonparametric tests for panel count data. *Communications in Statistics - Theory and Methods*, 36(8):1611–1625.

Pascual, J., Falk, R., Piessens, F., Prusinski, A., Docekal, P., Robert, M., Ferrer, P., Luria, X., Segarra, R. and Zayas, J. (2000). Consistent efficacy and tolerability of almotriptan in the acute treatment of multiple migraine attacks: results of a large, randomized, double-blind, placebo-controlled study. *Cephalalgia*, 20(6):588–596.

Peng, Y. and Dear, K. (2000). A nonparametric mixture model for cure rate estimation. *Biometrics*, 56(1):237–243.

Peng, Y., Dear, K. B. G. and Denham, J. W. (1998). A generalized F mixture model for cure rate estimation. *Statistics in Medicine*, 17(8):813–830.

Peters, R. J., Mehta, S. R., Fox, K. A., Zhao, F., Lewis, B. S., Kopecky, S. L., Diaz, R., Commerford, P. J., Valentin, V., Yusuf, S. et al. (2003). Effects of aspirin dose when used alone or in combination with clopidogrel in patients with acute coronary syndromes observations from the clopidogrel in unstable angina to prevent recurrent events (cure) study. *Circulation*, 108(14):1682–1687.

Pledger, G. W., Sackellares, J. C., Treiman, D. M., Pellock, J. M., Wright, F. S., Mikati, M., Sahlroot, J. T., Tsay, J. Y., Drake, M. E., Olson, L. et al. (1994). Flunarizine for treatment of partial seizures results of a concentration-controlled trial. *Neurology*, 44 (10):1830–1836.

Prentice, R. L., Williams, B. J. and Peterson, A. V. (1981). On the regression analysis of multivariate failure time data. *Biometrika*, 68(2):373–379.

Price, D. L. and Manatunga, A. K. (2001). Modelling survival data with a cured fraction using frailty models. *Statistics in Medicine*, 20(9-10):1515–1527.

R: A Language and Environment for Statistical Computing (2013). R Foundation for Statistical Computing. URL `http://www.R-project.org`.

Riggs, B. L., Wahner, H. W., Dunn, W. L., Mazess, R. B., Offord, K. P. and Melton 3rd, L. J. (1981). Differential changes in bone mineral density of the appendicular and axial skeleton with aging: relationship to spinal osteoporosis. *Journal of Clinical Investigation*, 67(2):328–335.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89(427):846–866.

Rotnitzky, A. and Wypij, D. (1994). A note on the bias of estimators with missing data. *Biometrics*, 50(4):1163–1170.

Salvarani, C., Cantini, F., Boiardi, L. and Hunder, G. G. (2002). Polymyalgia rheumatica and giant-cell arteritis. *New England Journal of Medicine*, 347(4):261–271.

Base SAS 9.2 Procedures Guide (2009). SAS Institute Inc., Cary, NC: SAS Institute Inc.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8(1):3–15.

Shen, H. and Cook, R. J. Analysis of interval-censored recurrent processes subject to resolution. *submitted*.

Shen, H. and Cook, R. J. (2013*a*). A dynamic mover-stayer model for recurrent event process subject to resolution. *Lifetime Data Analysis*. doi: 10.1007/s10985-013-9271-7.

Shen, H. and Cook, R. J. (2013*b*). Regression with incomplete covariates and left-truncated time-to-event data. *Statistics in Medicine*, 32(6):1004–1015.

Spilerman, S. (1972). Extensions of the mover-stayer model. *American Journal of Sociology*, 78(3):599–626.

Sun, J. and Fang, H.-B. (2003). A nonparametric test for panel count data. *Biometrika*, 90(1):199–208.

Sun, J. and Kalbfleisch, J. (1995). Estimation of the mean function of point processes based on panel count data. *Statistica Sinica*, 5:279–290.

Sun, J. and Wei, L. (2000). Regression analysis of panel count data with covariate-dependent observation and censoring times. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(2):293–302.

Sun, J. and Zhao, X. (2013). *Statistical Analysis of Panel Count Data*. Springer, New York.

Sun, J., Tong, X. and He, X. (2007). Regression analysis of panel count data with dependent observation times. *Biometrics*, 63(4):1053–1059.

Sy, J. P. and Taylor, J. M. G. (2000). Estimation in a Cox proportional hazards cure model. *Biometrics*, 56(1):227–236.

Taylor, J. M. G. (1995). Semi-parametric estimation in failure time mixture models. *Biometrics*, 51(3):899–907.

Thall, P. F. and Lachin, J. M. (1988). Analysis of recurrent events: Nonparametric methods for random-interval count data. *Journal of the American Statistical Association*, 83(402): 339–347.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3):290–295.

Wang, C. Y. and Chen, H. Y. (2001). Augmented inverse probability weighted estimator for Cox missing covariate regression. *Biometrics*, 57(2):414–419.

Webert, K., Cook, R. J., Sigouin, C. S., Rebulla, P. and Heddle, N. M. (2006). The risk of bleeding in thrombocytopenic patients with acute myeloid leukemia. *Haematologica*, 91 (11):1530–1537.

Wellner, J. A. and Zhang, Y. (2000). Two estimators of the mean of a counting process with panel count data. *Annals of Statistics*, 28(3):779–814.

Wellner, J. A. and Zhang, Y. (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *The Annals of Statistics*, 35(5):2106–2142.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, 50(1):1–25.

Winokur, G. (1975). The Iowa 500: heterogeneity and course in manic-depressive illness (bipolar). *Comprehensive Psychiatry*, 16(2):125–131.

Wolfson, C., Wolfson, D., Asgharian, M., M'Lan, C., Østbye, T., Rockwood, K. and Hogan, D. (2001). A reevaluation of the duration of survival after the onset of dementia. *New England Journal of Medicine*, 344(15):1111–1116.

Yamaguchi, K. (1992). Accelerated failure-time regression models with a regression model of surviving fraction: an application to the analysis of "permanent employment" in Japan. *Journal of the American Statistical Association*, 87(418):284–292.

Yamaguchi, K. (1994). Some accelerated failure-time regression models derived from diffusion process models: An application to a network diffusion analysis. *Sociological Methodology*, 24:267–300.

Yamaguchi, K. (1998). Mover-stayer models for analyzing event nonoccurrence and event timing with time-dependent covariates: An application to an analysis of remarriage. *Sociological Methodology*, 28(1):327–361.

Yamaguchi, K. (2003). Accelerated failure-time mover-stayer regression models for the analysis of last-episode data. *Sociological Methodology*, 33(1):81–110.

Ye, Y., Kalbfleisch, J. D. and Schaubel, D. E. (2007). Semiparametric analysis of correlated recurrent and terminal events. *Biometrics*, 63(1):78–87.

Yu, B. (2008). A frailty mixture cure model with application to hospital readmission cata. *Biometrical Journal*, 50(3):386–394.

Yusuf, S., Wittes, J., Probstfield, J. and Tyroler, H. A. (1991). Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. *The Journal of the American Medical Association*, 266(1):93–98.

Zhang, Y. (2002). A semiparametric pseudolikelihood estimation method for panel count data. *Biometrika*, 89(1):39–48.

Zhang, Y. (2006). Nonparametric k-sample tests with panel count data. *Biometrika*, 93 (4):777–790.

Zhao, X. and Tong, X. (2011). Semiparametric regression analysis of panel count data with informative observation times. *Computational Statistics & Data Analysis*, 55(1): 291–300.

Zochling, J. and Braun, J. (2006). Remission in ankylosing spondylitis. *Clinical and Experimental Rheumatology*, 24(6):88–92.