Contributions to the Analysis of Multistate and Degradation Data

by

Yang Yang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2011

Doctoral Committee:

Assistant Professor Yves A. Atchadé, Co-Chair
Professor Vijayan N. Nair, Co-Chair
Professor John D. Kalbfleisch
Associate Professor Kerby A. Shedden

ACKNOWLEDGEMENTS

This thesis would not have been possible without the guidance of my committee members and the support from my family and friends.

I owe my deepest gratitude to my two advisors, Dr. Vijayan Nair and Dr. Yves Atchadé for their excellent guidance, encouragement and patience. I would like to thank Dr. Vijayan Nair, who introduced me to multistate and degradation data modeling and carefully guided me to formulate and develop the research problems present in this dissertation. I am grateful to Dr. Yves Atchadé, who gave me extensive suggestions on my thesis work. My interests in statistical computing started from our numerous discussions about the problem. I would also like to thank my other two committee members for their support and advice. It is an honor to have Dr. Kalbleisch on my committee and have his expertise and insights on my reseach work. Special thanks to Dr. Shedden for his constant support doing my graduate studies and research.

I would like to give special thanks to my parents. Without their support and love, I would not have been able to go this far. Also, I would like to thank my fiance, Chenling Huang, who has always cared for me and accompanied me through good times and bad. I would also like to thank my close friends, Jing Wang, Eric Laber and Joel Vaughan. And I offer my thanks and blessings to everyone who supported me during the completion of this dissertation.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

ABSTRACT

Contributions to the Analysis of Multistate and Degradation Data

by

Yang Yang

Co-Chairs: Yves A. Atchadé and Vijayan N. Nair

Traditional methods in survival, reliability, actuarial science, risk, and other event-history applications are based on the analysis of time-to-occurrence of some event of interest, generically called "failure". In the presence of high-degrees of censoring, however, it is difficult to make inference about the underlying failure distribution using failure time data. Moreover, such data are not very useful in predicting failures of specific systems, a problem of interest when dealing with expensive or critical systems. As an alternative, there is an increasing trend towards collecting and analyzing richer types of data related to the states and performance of systems or subjects under study. These include data on multistate and degradation processes. This dissertation makes several contributions to the analysis of multistate and degradation data.

The first part of the dissertation deals with parametric inference for multistate processes with panel data. These include interval, right, and left censoring, which arise naturally as the processes are not observed continuously. Most of the literature in this area deal with Markov models, for which inference with censored data

can be handled without too much difficulty. The dissertation considers progressive semi-Markov models and develops methods and algorithms for general parametric inference. A combination of Markov Chain Monte Carlo techniques and stochastic approximation methods are used. A second topic deals with the comparison of the traditional method and the process method for inference about the time-to-failure distribution in the presence of multistate data. Here, time-to-failure is the time when the process enters an absorbing state. There is limited literature in this area. The gains in both estimation and prediction efficiency are quantified for various parametric models of interest.

The second part of the dissertation deals with the analysis of data on continuous measures of performance and degradation with missing data. In this case, time-to-failure is the time at which the degradation measure exceeds a certain threshold or performance level goes below some threshold. Inference problems about the mean and variance of the degradation and the imputation of the missing are studied under different settings.

CHAPTER I

Introduction

This dissertation consists of two parts. Part I deals with two different aspects of multistate modeling with censored data: a) parametric inference for multistate semi-Markov processes with panel data; and b) comparison of the efficiencies of two methods for inference about the time-to-failure – one based on just the time-to-failure data and the second based on modeling the entire multi-state semi-Markov process and using the results to make inference about "failure" or time-to-absorption of the underlying process. Both estimation and prediction efficiencies are studied.

Part II deals with analysis of degradation data in the presence of complex missing patterns. There is an extensive literature on missing data in longitudinal studies, but the problems studied here are different. Several topics are addressed in this part: a) estimation of the underlying mean and variance functions which are specified in a non-parametric form but the time-dependence structure of the degradation data takes on different possible models, including comparison of efficiencies and robustness; b) development of inference for functional regression and ANOVA techniques in the presence of missing data; c) imputing the missing data and developing uncertainty bounds; and d) predicting time-to-failure and developing associated prediction intervals. This research problem originated as part of an applied project on the

degradation of road pavements in Michigan.

While the two parts are distinct, they are related in that both deal with "degradation data". This connection is discussed below as part of the motivation for the research presented in this dissertation.

Chapters II, III, and IV have been written up as individual papers which are being prepared for submission or have been submitted (Yang, Atchade and Nair 2011, Yang and Nair 2011a, and Yang and Nair 2011b). Chapter III has been tentatively accepted for publication. Since the papers are self contained, there is some repetition of the motivation, notation, etc. in the different chapters.

1.1 Motivation

Traditional survival and reliability analysis are based on the time to occurrence of some event of interest. We denoted this event generically as "failure" in this dissertation. The event can be: death of a patient, onset of a disease, failure of a device, completion of a task such as repair of failed equipment or servicing a customer, default of a bank loan, marriage, divorce, having a first child, graduation, etc. There are only two states in this situation – either the event has happened or not, and information about intermediate states or condition of a unit is not available or taken into account.

The modeling and analysis of time-to-failure data is a mature area with a huge literature. See, for example, the excellent books by Andersen *et al.* (1992), Klein and Moeschberger (2003), Kalbfleisch and Prentice (2002), Lawless (2003) and Meeker and Escobar (1998). The literature covers homogeneous populations as well as heterogeneous situations, regression analysis with proportional hazard models, acceler-

ated failure time models, and so on. There is extensive use of these techniques to applications in health and medicine, reliability engineering, actuarial science, risk analysis, and social sciences.

Data from reliability and health studies, including reliability test programs, clinical trials, warranty data, and other types of field data, are subject to various forms of censoring, leading to possibly extensive incompleteness in the available information. For example, in engineering and manufacturing applications, there has been considerable emphasis in recent years on increasing the quality and reliability of products to be globally competitive. As reliability increases, one observes very few failures, and most of the units are censored. During the product design and development stage, the amount of time available for reliability estimation and assessment can be in the order of a few months. If the products are designed with high reliability, few units will fail in a 3-6 month window, leading to very high degree of right censoring. It is not uncommon to have no failures during this period. Similar issues arise in clinical trials. Hence, it will be very difficult to make reasonable inference and assess product reliability in such situations.

Several approaches have been developed to get around these problems. In engineering applications, accelerated life testing (ALT) is commonly used to induce early failures. One analyzes the time-to-failure at accelerated conditions and uses acceleration transform models to extrapolate to nominal conditions. See, for example, Nelson (2004) and Meeker and Escobar (1998) for an overview of these methods. The main difficulty with the use of ALT techniques is that most of the common models are empirical (or heuristic) in nature and reduce to a linear model after a logarithmic transformation (see Section 2.11 in Nelson 2004). The adequacy of these models for extrapolation to design conditions is often questionable. Other approaches that have

developed include the collection of extensive information on covariates and surrogate variables, and using this information to improve estimation and prediction efficiency. See, for example, Davies (1998) and Cox (1999).

A different direction that is becoming more common is the collection and analysis of richer data on the condition or performance of the systems/subjects under study, beyond time-to-failure. An early instance of this was the three-state illness-death model where a subject can move from the state of being "well" to an intermediate state of being "ill", and not just "death". The use of multistate models has become increasingly popular, both in biostatistics and engineering. In a multistate model, the system moves among different "states" with each state representing the "health" of the system. Multistate models have been used to study many different applications. In the case of survival and reliability analysis, the states typically represent stages of degradation, such as worsening health, and are often progressive (i.e., the system/subject moves only in one direction). Further, one of the states will be an absorbing state denoting the occurrence of the event of interest ("failure"). In such cases, failure can be viewed as the end point of the underlying multi-state process (see Aalen, Borgan and Gjessing 2008).

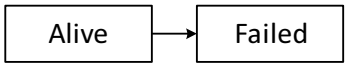
The primary advantage of multistate data is that information about intermediate states is available even when time-to-failure is censored at the end of the study. In addition to increasing the efficiency of estimating the failure distribution, the data can also be used to predict the failure of particular systems/subjects that were censored at the end of the study. The analysis of multistate data is the subject of Chapter II and Chapter III.

A different type of degradation data deals with continuous evolution of the state or performances of the system in terms of the condition of the system or degradation

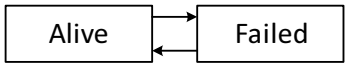
in performance. Within the engineering context, advances in sensing technologies are making it feasible to collect extensive amounts of data on performance-related measures and "degradation" associated with components, systems, and equipment. Davies (1998) discusses a variety of engineering applications, types of degradation data, and recent developments in the area of condition monitoring and system maintenance. These cover data from vibration and acoustical monitoring, thermography, lubricant and wear debris analysis, etc. Meeker and Escobar (1998, Chapter 14 and 21 and references therein) describe applications in fatigue crack growth, luminosity of light bulbs, corrosion of batteries, semiconductor (MOS) devices, etc. Modeling and analysis of degradation data is also receiving increasing attention. Chapter IV deals with flexible methods for inference in some degradation models in the presence of complicated missing patterns. My interest in this research problem was motivated by an application to road pavements that I was involved in, and the project was funded by the Michigan Department of Transportation (MDOT). This is described in more detail in Chapter IV.

1.2 Review and Summary of Contributions

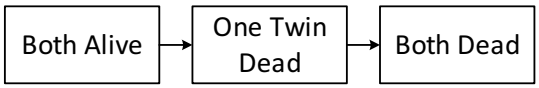
This section provides general background and references to the problems being addressed and summarizes the contributions in the different chapters. The first subsection focuses on multistate modeling, applications to survival and reliability analysis, and the contributions in Chapters II and III. The second subsection deals with the analysis of degradation data and Chapter IV.



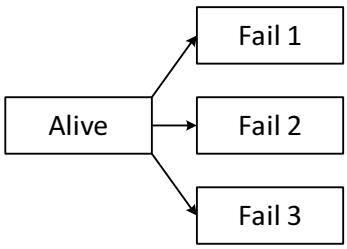
(a) Traditional 2-state model



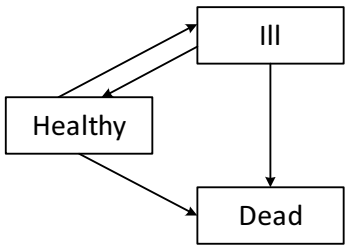
(b) Recurrent Model



(c) Twin model



(d) Competing Risk



(e) Illness-death model

Figure 1.1: Popular Multistate Models Used in Applications

1.2.1 Multistate Modeling

Various types of multistate models have been proposed to analyze survival, reliability, and other types of event history data. Hougaard (1999) provides a review of some models; Commenges (1999) emphasizes applications in epidemiology. See also Chapter 8 of Kalbfleisch and Prentice (2002) and Chapter 11 of Lawless (2003). The discussion below is based on these sources.

A multistate process can be described by a stochastic process $Z(t)$ which takes values in a countably infinite state space. In this dissertation, the state space is assumed to be finite. The process $Z(t)$ spends a random amount of time, characterized by some distribution, in a given state (called sojourn or occupancy times) and then moves to another state according to some transition probabilities. Figure 1.1 gives some examples. The traditional case can be viewed as a two-state model with an "alive" state and a "failed" state. The "failed" state is the absorbing state since further transitions are not possible once this state is reached. The failure time is the duration (sojourn) time in the "alive" state before the subject moves to the absorbing state. Several other traditional problems can also be viewed as special cases of multistate processes. The situation with recurrent failures can be treated as a multistate model as shown in Figure 1.1(b). The recurrent model is used to represent repairable systems where the system fails, is repaired, and then fails again. In this case, the "failed" state is not absorbing. The competing risks model can also be viewed as a multistate model with multiple absorbing states as shown in Figure 1.1(d). The more interesting, non-trivial extension of the two-state survival model is the illness-death model with two transient states – "healthy" and "ill" – and one absorbing state "dead". As shown in Figure 1.1(e), a patient can move from "healthy" to "ill" and then to "dead". It is also possible for an ill patient to get better and

move back to "healthy".

This dissertation deals with progressive multistate models (see, for example, the one-step "twin" model in Figure 1.1(c)). We assume that the states are ordered in an appropriate (progressive) manner and the system moves from left to right only, i.e., from state i to state j with $j > i$. Further, it is assumed that the right-most state is an absorbing ("failure") state and is the end-point of the process. Thus, the traditional time-to-failure can be viewed as the time-to-reach the absorbing state. Such progressive models are useful in capturing situations where there is monotone degradation and the system cannot get better with time. Cases where a patient has a particular disease or event (such as a stroke) and then gets back to being "normal" can be handled by adding an additional state that distinguishes someone who has always been healthy from another person who had such an event and then has returned to being normal. While this would increase the number of states, the analysis of progressive models is considerably easier due to the finite number of possible transitions. A special case of progressive models is the one-step progressive case (Figure 1.1(c)), where one can move from one state only to the one that is immediately to its right.

Despite the increasing attention on multistate models, most of the literature on statistical inference has focused on cases with simple structure. The Markov model is clearly the most popular one, and it assumes that the sojourn times are all exponentially distributed. See Kalbfleisch and Lawless (1985), the books and review papers cited earlier for a discussion of inference under Markov models. See also Kay (1986) and Andersen, Esbjerg and Sørensen (2000) who discuss the use of Markov models in medical applications.

Part of the popularity of Markov models is due to the fact that inference under

these models are relatively simple even in the presence of censoring. Extensions to non-Markovian cases have dealt either with a small number of states or under restrictions on the model structure or types of censoring. Sternberg and Satten (1999) introduced a nonparametric estimator for semi-Markov multistate models with interval-censored observations, but the underlying process is one-step progressive. Foucher *et al.* (2007) proposed a likelihood estimation algorithm for interval-censored data, but for semi-Markov multistate models with Weibull durations only. The possibility of missing transitions was not considered there either, largely simplifying the problem. The goal of Chapter II is to develop general parametric inference methods for progressive semi-Markov models.

The presence of interval censoring arises naturally with multistate data as the underlying process can rarely be monitored continuously. For example, the health status of patients can be recorded only when they visit the doctor for periodic check-ups; the time of onset of a disease is often unknown. In other words, the state information is observed only at some discrete time points, leading to interval censoring. (This discretely-observed state information is also called panel data.) Further, the data can be subject to the usual left and right censoring patterns. Right censoring is present when the process has not reached the absorbing state (event of interest) by the end of the study. This also arises with traditional time-to-failure data. Despite the fact that the process is being observed at discrete time, one could still observe the exact time to absorption in some cases – for example in many health situations, the time of death of a patient is available.

The presence of all these types of censoring complicates the analysis of multistate data. To illustrate this, consider the following example. Suppose the process is progressive, and the states are ordered, i.e., if the subject is in state i, it can only

move forward to state j with $j > i$. Further assume that it started in state 1. We observe the process at time points t_1, t_2, t_3, and t_4 and the states at these times are: $Z(t_1) = 1, Z(t_2) = 2, Z(t_3) = 4, Z(t_4) = 7$. Denote i_k the state the process visits at the $k-$th transition, and τ_k the sojourn time spent in state I_{k-1}, $k = 1, 2, 3, \ldots$. Then, at time t_1, we know that no transition has occurred and that τ_1, the sojourn time in state 1 is at least t_1. Since $Z(t_2) = 2$, we know that the system transitioned from state 1 to state 2 at some time between t_1 and t_2 but we do not know exactly when. Further, we only know that τ_1 is censored in the interval $(t_1, t_2]$. The next observation is $Z(t_3) = 4$. From this, we cannot tell if the system moved from state 2 to 3 and then 4 ($i_1 = 2, i_2 = 3, i_3 = 4$) or if it jumped straight from state 2 to 4 ($i_1 = 2, i_2 = 4$). Both possibilities must be considered. As far as τ_2 (sojourn time in state 2) is concerned, we do not know when the system moved to state 2 (left censored). This is not an issue with Markov models since the exponential distribution is memoryless, but it cannot be ignored for other distributions. So all we know is that either $t_2 < \tau_1 + \tau_2 + \tau_3 < t_3$ or $t_2 < \tau_1 + \tau_2 < t_3$ depending on which set of transitions occurred. Another complication is that now τ_1 and τ_2 are related because of the censoring in data. One can see that the problem can get rather complex with even moderate number of states. It would become intractable if we did not restrict attention to progressive structure and allowed cycles (one can go from any state to any other state) as the number of possible transitions is then infinite.

It is possible to develop inference methods for specific parametric distributions with special multistate structure. For example, the problem is not too difficult for one-step progressive models where the sojourn times are closed under convolution, such as gamma with the same scale parameter and are independent across states (as assumed in semi-Markov models). Then, the cumulative amount of time spent

before reaching any state is a sum of gamma random variables with the same scale parameter and hence is also gamma. The goal in Chapter II is to develop general methods for parametric models with panel data.

An obvious approach is to treat this as a missing data problem (with the unobserved/censored sojourn times and transition patterns as complete data). It turns out, however, that the E-step is rather complex and even Monte-Carlo based methods do not work well. Chapter II develops a computationally intensive approach, based on a combination of MCMC (Gibbs and Reversible Jump MCMC) sampling and stochastic approximation methods for likelihood-based inference.

Chapter III deals with a related problem. Suppose we have multistate process with an absorbing state and the time-to-absorption can be considered as the time-to-failure. Further, suppose we are interested only in inference about the time-to-failure (estimation of the failure distribution as well as prediction of the time-to-failure for systems that have not failed). There are two ways in which we can deal with this: a) analyze the entire multistate data and use the results for inference about the failure time distribution; or b) use just the time-to-absorption (failure) data with all of the possible censoring issues.

As an example, consider the case of consumers who have bank loans and pay them off periodically. This can be mortgage payments, credit card payments, etc. A customer starts in the "current" state (C) – meaning s/he has been paying at least the minimum amount required by the bank. Over time, s/he may stay in that state or move to a different state (1-month delinquent (D_{30}), 2-month delinquent (D_{60}), etc.). Here delinquency means the minimum payment has not been made. Banking institutions typically consider 6-month delinquency to be default (D) or failure state, and the time-to-default distribution is of one of the quantities of interest.

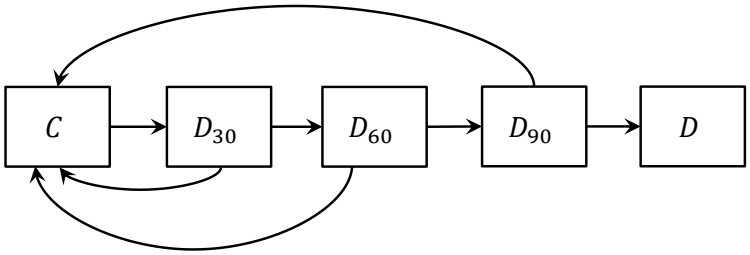


Figure 1.2: Delinquency Model in Credit Risk Modeling

There are clearly many other advantages to modeling the entire multistate process. In the consumer loan example, one may want to study separately the behavior of customers who are delinquent and then go back to being current versus those who go straight through to default, what are the different demographic and socio-economic characteristics of the different groups and so on. Such information can be very useful to the banks in developing future loan programs. But suppose the primary quantity of interest is just inference about the failure time distribution. A natural question is whether it is worthwhile to analyze the multistate data or focus just on the time-to-default data.

There are several different considerations in answering this question. From a statistical efficiency point of view, it seems clear that modeling the entire multistate data has to be at lease as good as the time-to-failure data since the former includes the latter. But it is worth quantifying the gains in efficiency. One reason is that, in non-Markovian models, the analysis of multistate panel data is complex. The methods developed in Chapter II are computationally intensive and do not scale up to situations with a large number of states. Further, the developments for semi-Markov processes are based on model and independence assumptions that have to hold for all the states. On the other hand, analysis of time-to-failure data, even with very complex censoring patterns, is a mature area, and there is considerable literature and

software available. Further, the assumptions about the time-to-failure model have to (approximately) hold (only for the time-to-absorption and not the intermediate states. Therefore, it is worth investigating and quantifying the efficiency gains from using the entire multistate data.

Chapter III studies this problem for specific parametric cases. The general problem is rather involved since the cumulative sojourn times (sums of individual sojourn times in each state) do not have closed form expressions even in the simple one-step case where they are just convolutions. In the more general multistep case, they are mixture distributions. Therefore, the study focuses on specific distributions and examines the performance through asymptotic analysis and simulation studies. The gains in estimation efficiency are substantial, often more than 2-3 times, even with small to moderate number of states. Further insights into when the efficiency gains are high and low are also obtained. Not surprisingly, the bigger gain is with prediction efficiency. This deals with the ability to predict the failure of a particular system given the data at time of censoring. With time-to-failure data, the only information available is that the system was working at the time the study ended and that it is right censored. So only the population-level residual time distribution can be used to predict failure. With multistate data, however, we know the state of the system at the end of the study. This individual-level information is clearly a lot more informative in prediction. This gain in efficiency is much higher if the system is close to the absorbing state at the time of censoring vs being in an early state. Chapter III studies these and related issues, and quantifies the gains under several different scenarios.

The idea that failure can be viewed as the end point of a process has gained attention recently. The concept of using Markov models to study the absorbing time of

the system was discussed in Neuts (1981), in which the phase-type distributions was introduced. Aalen (1995) considered the use of phase-type distributions in survival analysis. Aalen and Gjessing (2001) further elaborated on this and examined the shapes of hazard rates that can be obtained from times-to-absorption of different processes. Asmussen, Nerman and Olsson (1996) and Olsson (1996) applied E-M algorithm to estimate phase type distributions when only the censored absorbing times are observed. In the reliability and operations-research literature, finite-state Markov processes and semi-Markov processes have also been studied, although most of the work relates to optimization and not inference. For example, there is an extensive literature on partially observed Markov decision processes. Limnios and Oprisan (2001) considered application of semi-Markov processes in the field of reliability. More recently, reliability techniques are being used to analyze financial data. D'Amico, Janssen and Manca (2005) applied homogeneous Markov model to investigate the company-rating status.

1.2.2 Analysis of Degradation Data with Complex Missing Patterns

As noted earlier, advances in sensing and measurement technologies have made it possible to collect and analyze detailed information about continuous measures of degradation and performance of systems. In the engineering context, these measures include data from vibration and acoustical monitoring, thermography, lubricant and wear debris analysis, etc. Meeker and Escobar (1998) describes applications in fatigue crack growth, luminosity of light bulbs, corrosion of batteries, semiconductor (MOS) devices, etc. In health and medical applications, data on patient's health conditions can be viewed as degradation data. Modeling and analysis of degradation

data has received considerable attention. In this case, the time-to-failure is usually viewed as the time at which the degradation measure exceeds a certain threshold or performance level goes below some threshold.

My interest in this research problem was motivated by a project about analyzing the degradation of highway road pavement, sponsored by the Michigan Department of Transportation (MDOT). The specific degradation measure was distress index (DI). MDOT collects visual images of road conditions by videotaping highway pavement surfaces using a van equipped with cameras and driven at regular speeds. These videotapes are then sent to a central location where they are viewed and scored by the type, extent, severity, and other types of pavement defects. Points are assigned for each distress type depending on the severity and quantity of each distress based on pre-established algorithm, leading to a distress index for each 0.1 mile segment of pavement. These indices are often aggregated to assign more crude measures to larger segments of the road. The DI should be zero for a pavement that has no distress; if the DI is 50 or more, that segment of road is a candidate for rehabilitation.

The DI scoring is done subjectively, so there is usually a lot of measurement error and random effects. There were many instances where the DI scores were drastically lower as the pavement aged. (Some of this could be due to the fact that parts of the road segments could have been repaired and that information was not available.) In addition, there was a lot of missing data. For some pavements, no records were available before certain time period (missing to the left in the time-to-failure context), data were missing for certain consecutive years (missing in an interval) and data were not collected after certain years (missing to the right). In fact, few of the pavements had complete data for the period of study.

MDOT had several goals, and the primary one was to assess the effect of different

pavement designs and materials on the life of the pavement. It turned out that we could not fit a reasonable parametric model to the mean degradation curve over time. So we decided to use functional ANOVA of the form $Y(t) = X\beta(t)$ where $Y(t)$ is the degradation data over time, X is the design matrix corresponding to the types of pavement design and materials and $\beta(t)$ was a time-varying measure of the effect of the design factors. See Ramsay and Silverman (2002), Ramsay and Silverman (2005) for a discussion of functional regression and applications. But the extensive nature and types of missing data made the inference difficult. In particular, we needed to develop inference procedures for the functional regression coefficients $\widehat{\beta}(t)$ in the presence of missing data. In addition, it was also of interest to see how the profile of the pavements for particular areas would look like if the data had not been missing (especially in the tails or right missing). It was also of interest to predict the time-to-failure of pavements so that one can plan resources for road repairs.

Our report to MDOT was based on some heuristic analysis but it led to a more systematic investigation as described in Chapter IV. Several inference problems are studied in this chapter:

1. Estimating the mean and variance of the degradation data $Y(t)$: $\mu(t)$ and $\sigma^2(t)$ based on a random sample of units in the presence of different types of missing data and for several common models for the error structure.

2. Extend the results to functional regression or ANOVA and develop appropriate test procedures for testing various hypothesis of interest.

3. Impute the degradation at the missing values and obtain uncertainty bounds.

4. Assuming a parametric form for $\mu(t)$ and $\sigma^2(t)$, predict the time-to-failure and obtain prediction intervals.

5. The above were all done under normal error structure, so another goal was to examine non-normal error structures and also robustness to normality and other assumptions.

These are done under a combination of missing data patterns.

Degradation data analysis has been discussed in literature. Lu and Meeker (1993) proposes a parametric model for the degradation data, and use the estimated sample path to predict the corresponding time-to-failure whenever the measurement passes certain threshold. Lawless and Crowder (2004) suggested a gamma process for the modeling of the underlying degradation process. Nonparametric estimation for degradation data has also been studied in Wang (2005), Wang (2009) and Wang (2010). See Nelson (2004, Chapter 11) for the review of the literature of degradation. Even though missing data issue has not been systematically addressed under the degradation setting, the literature about missing in the longitudinal study is extensive. See Laird (1988), Lann and Robins (2003).

CHAPTER II

Parametric Inference for Multistate Semi-Markov Models with Panel Data

2.1 Introduction

The use of multi-state models to analyze survival, risk, reliability and other event history data is becoming increasingly common. An early application was a three-state illness-death model in which the patient can be well, become ill, get well again or die. Our interest in this problem arose from applications in credit risk analysis, where lending institutions use intermediate delinquency states to develop insights into the behavior of customers who are likely to default. There are also applications in reliability engineering where the degradation data are modeled through multistate models. See Andersen (1988) and Hougaard (1999) for a general discussion of multistate models and comparisons with the traditional survival analysis based on just lifetime data.

Interval-censored data arise naturally in this context since the underlying multi-state process can rarely be monitored continuously. For instance, the health status of patients can be recorded only when they visit a health-care center. Such data, often called panel data, exhibit a combination of interval, left and right-censoring. Most of the papers in the literature focus on Markov models where statistical inference

with panel data can be handled without too much difficulty (see Kay 1986, Andersen, Esbjerg and Sørensen 2000). Non-Markovian situations have received only limited attention. Sternberg and Satten (1999) introduced a nonparametric estimator with interval-censored observations under some restricted conditions. Foucher *et al.* (2007) developed an algorithm for likelihood estimation with Weibull sojourn distributions but assumed that all the transitions are observed, drastically simplifying the problem. More recently, Titman and Sharples (2010) developed results for panel data with phase-type sojourn distributions. Lagakos, Sommer and Zelen (1978) developed likelihood-based estimation methods for situations where the multistate process is observed continuously with possible right censoring.

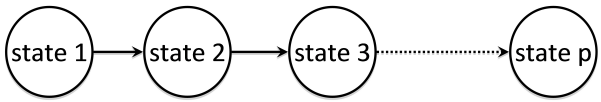
This chapter develops a general approach and algorithms to make parametric inference for semi-Markov processes with panel data. We restrict attention to models with progressive (or acyclic) structure although this assumption can be relaxed. The chapter is organized as follows: Section 2.2 provides the problem formulation. The challenges of doing parametric estimation for semi-Markov models with panel data are discussed in Section 2.3. The next two sections develop the details of the estimation algorithm involved in parametric inference for semi-Markov models with panel data. In Section 2.4, we focus on likelihood-based inference. Section 2.4.2 describes the use of stochastic approximation to compute the maximum likelihood estimate (MLE) and the observed information matrix for likelihood-based inference. Section 2.4.3 discusses the algorithms for imputing (sampling) the complete data from the incomplete panel data. In the simple one-step progressive case, this can be done through data augmentation. For the more general multi-step progressive case, two algorithms are studied: conditional sampling and reversible jump MCMC sampling (RJMCMC). The latter is shown to be computationally more efficient. Extensions

to Bayesian inference and (static) covariates are indicated in Section 2.5. Section 2.6 describes simulation studies to evaluate the performance of the proposed algorithms. The results are illustrated on a heart transplant data set in Section 2.7. This chapter concludes with some remarks.

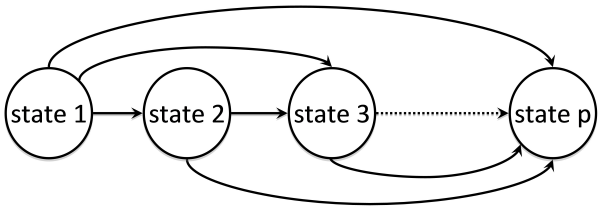
2.2 Formulation

Let $\{Z(t),\ t \geq 0\}$ be a stochastic process where $Z(t)$ denotes the state of the subject at time t. The process $\{Z(t),\ t \geq 0\}$ takes values in the finite state space $E = \{1, 2, \ldots, p\}$. Throughout this chapter, we take p to be an absorbing state, denoted as "failure". Also, we assume that the process starts in state 1, i.e., $Z(0) = 1$. This can be relaxed without too much difficulty. Further, we assume that the states are ordered in some natural way, and that the subject moves from left to right only. We refer to this as a progressive (acyclic) model. Note that only a finite number of transitions can occur in this case. Progressive models are quite natural in health settings where there is monotone degradation of health status over time. Cases where a patient is ill and then recovers can be handled by adding a new state rather than allowing the patient to return to the initial healthy state. Even though this increases the number of states, it is useful to distinguish a patient who had recovered from a health problem from someone who has always been healthy.

Figure 2.1 shows two examples of progressive multistate models. The top panel deals with the one-step progressive model, where the system can move only to the immediate right sternberg:99seem, for example, . The bottom panel allows for multiple steps and is the more general case we are interested in. This general progressive structure is called multi-step progressive.



(a) One-step Progressive Case



(b) Multi-step Progressive Case

Figure 2.1: Progressive multistate Models

The focus in this chapter is on time-homogenous semi-Markov processes (SMP) which are generalizations of time-homogeneous Markov processes. It is convenient to define an SMP in terms of its equivalent Markov renewal process as in Janssen and Manca (2006). Consider a time-homogeneous Markov renewal process (MRP) $\{(I_s, \tau_s); s = 0, 1, \ldots\}$. That is $(I_s, \tau_s) \in E \times (0, \infty)$ and for any $n \geq 1$, $(j, t) \in E \times (0, \infty)$,

$$\begin{aligned}\mathbb{P}\left(I_s = j, \tau_s \leq t | (I_u, \tau_u),\ 1 \leq u \leq s-1\right) &= \mathbb{P}\left(I_s = j, \tau_s \leq t | I_{s-1}\right) \\ &= \mathbb{P}\left(\tau_s \leq t | I_s = j, I_{s-1}\right) \mathbb{P}\left(I_s = j | I_{s-1}\right) = \mathbb{P}\left(\tau_1 \leq t | I_1 = j, I_0\right) \mathbb{P}\left(I_1 = j | I_0\right).\end{aligned}$$

The last equation arises from the time-homeogeneous assumption. Here I_s is the state the process visits at the $s-$th transition, and τ_s is the sojourn time spent in state I_{s-1} before the $s-$th transition occurs. A MRP can be characterized by (i) the transition probability

$$p_{ij} := \mathbb{P}(I_s = j | I_{s-1} = i), \quad i, j \in E, \tag{2.1}$$

and (ii) the conditional distribution of the sojourn times, which is given by

$$F_{ij}(t) \ := \ \mathbb{P}(\tau_s \leq t | I_{s-1} = i, I_s = j), \quad i, j \in E, \, t > 0. \tag{2.2}$$

Equivalently, a MRP can be characterized by its semi-Markov kernel

$$Q_{ij}(t) \ := \ \mathbb{P}(\tau_s \leq t, I_s = j | I_{s-1} = i) = p_{ij} F_{ij}(t), \quad i, j \in E, \, t > 0. \tag{2.3}$$

Note that $Q_{ij}(\infty) = p_{ij}$. This implies that the marginal process $\{I_s, \ s = 0, 1, \ldots\}$ is a Markov chain with transition matrix $P = (p_{ij})$, also known as the embedded Markov chain of the MRP. We assume that $F_{ij}(\cdot)$ and $Q_{ij}(\cdot)$ are both differentiable and denote their derivatives as $f_{ij}(\cdot)$ and $q_{ij}(\cdot)$ respectively. We will use $P = (P_{ij})$ and $\boldsymbol{f}(\cdot) = (f_{ij}(\cdot))$ interchangeably with $\boldsymbol{q}(\cdot) = (q_{ij}(\cdot))$ to characterize the MRP. An important property of MRP that we will use is that the sojourn times $\{\tau_s, \ s = 0, 1, \ldots\}$ are conditionally independent given the embedded Markov chain. For the MRP as defined above, the number of transitions to absorption (failure) is the random variable $\sigma := \inf\{s \geq 1 : \ I_s = p\}$. Notice that in the present case of a progressive MRP, $\sigma \in \{1, \ldots, p-1\}$. We assume $I_0 = 1$ and set $\tau_0 = 0$. Denote $\mathbf{Z} = \{(I_s, \tau_s), \ s = 1, \ldots, \sigma\}$ the complete history of the MRP.

Let $S_0 = \tau_0$, and $S_n = \tau_1 + \cdots + \tau_n$, the cumulative time to the $n-$th transition. Then, we define the associated SMP $\{Z(t), \ t \geq 0\}$ as $Z(t) = I_n$ for $t \in [S_n, S_{n+1})$. $Z(t)$ degenerates to a traditional time-to-failure analysis if $p = 2$. The Markov model is the special case when $F_{ij}(\cdot) = F_i(\cdot)$, and are exponential distributions.

For a SMP (or MRP) with an absorbing state, an important quantity of interest is the time-to-absorption (failure)

$$T \ := \ \sum_{j=1}^{\sigma} \tau_j.$$

Since the sojourn times are conditionally independent given the embedded Markov chain, it can be shown that in the multi-step progressive case, the cumulative distribution function of T is given by

$$F_T \;=\; \sum_j p_{i_0,i_1,\ldots,i_k} \left(F_{i_0,i_1} \star F_{i_1,i_2} \star \cdots \star F_{i_{k-1},i_k} \right),$$

where $p_{i_0,i_1,\ldots,i_k} = p_{i_0,i_1} \times p_{i_1,i_2} \times \cdots \times p_{i_{k-1},i_k}$, $\star$ denotes the convolution operator, and the summation is over all possible paths such that $\{i_0 = 1 < i_1 < \cdots < i_k = p\}$.

In this chapter we are interested in modeling and analyzing the individual distributions F_{ij}, understanding their hazard behavior, and so on. We consider a parametric framework, assuming $F_{ij}(t) = F_{ij}(t; \boldsymbol{\phi})$ where $\boldsymbol{\phi}$ is a vector of unknown parameters. Then, the parameter of interest for inference is $\boldsymbol{\theta} = (\boldsymbol{\phi}, P)$ where P is the transition matrix defined in (2.1). Note that P is upper diagonal for the progressive SMP. Further, for the one-step SMP, $p_{i,i+1} = 1$, so P is completely known. In this case, $\sigma = p - 1$ and the time-to-failure can be expressed simply as

$$T \;=\; \sum_{j=1}^{p-1} \tau_j,$$

and its distribution is a convolution of the individual sojourn time distributions.

2.3 Challenges from Censoring

If the SMP $\{Z(t),\ t \geq 0\}$ is observed continuously, then we have complete information on all the states visited, the times of transition, and hence the sojourn times in all the states. In this case, inference about the underlying parameters $\boldsymbol{\theta}$ is straightforward, especially because of the conditional independence of the sojourn times in the MRP. Lagakos, Sommer and Zelen (1978) developed the likelihood-based

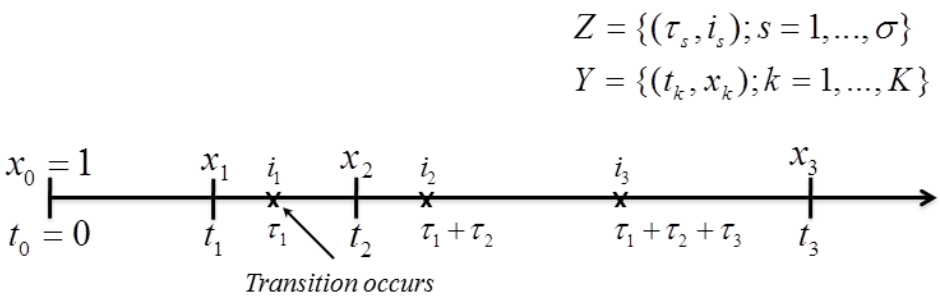


Figure 2.2: **Z** Unobserved Complete History, **Y** Observed Panel Data

estimation methods for such situation with the presence of right censoring. In practice, however, the process is rarely observed continuously. Rather, data are collected periodically, leading to interval censoring. This type of data is sometimes referred to as panel data. The inference problem is much more challenging in this situation.

In Figure 2.2, one realization of $\mathbf{Z} = \{(\tau_s, i_s);\ s = 1, \ldots, \sigma\}$ is shown along with a typical example of a sequence of observed data. The observed data **Y** consist of recording the system at time t_k and observing the states x_k at that time, for $k = 0, \ldots, K$. Here we assume $t_0 = 0$ and $X_0 = I_0$, i.e., the origin of the process is always observed (so later k starts from 1). Suppose the underlying model is one-step progressive, and we observe $(x_1 = 1, x_2 = 1, x_3 = 2, x_4 = 4)$. Then, it is clear that $i_1 = 1, i_2 = 2, i_3 = 3$, i.e., a transition must have occurred between times t_2 and t_3 and two transitions must have occurred between times t_3 and t_4. But we do not know the exact sojourn times in the states. All we know is that $t_1 < \tau_1 \leq t_3$ and $t_3 < \tau_1 + \tau_2 < \tau_1 + \tau_2 + \tau_3 \leq t_4$. The estimation of the underlying distributions is still not too difficult if the distributions are exponential, as we shall see. But in the non-memoryless case, it becomes relatively difficult even in this case. The situation becomes more difficult in the multi-step case. For example, suppose again we observe $(x_1 = 1, x_2 = 1, x_3 = 2, x_4 = 4)$. It is clear that no transition could have taken place at time t_2 and that one transition took place between t_2 and t_3. However, without

knowing the complete data, we do not know if the system jumped from state 2 to state 4 directly (one transition between times t_3 and t_4) or it moved from 2 to 3 and then 4 as is the case indicated in Figure 2.2. Both situations are possible and we have to integrate the complete data likelihood over both scenarios. As we can see from this case, the dimension of the complete data $\mathbf{Z}$ varies, depending on the number of transitions.

Suppose, instead of $x_4 = 4$, we had observed $x_4 = 7$. Then, the number of possible paths is considerably larger. The problem gets rather involved as the number of states, number of observation times, and number of systems being observed increase. The situation would be even a lot more complex if we allowed non-progressive models as, in principle, there could be an infinite number of possible transitions between the observation points.

For a single system that has been observed at times $t_0, \ldots, t_K$, the likelihood function of the panel data can be written as

$$L(\boldsymbol{\theta}|\mathbf{y}) \;=\; \mathbb{P}(Z(t_0) = x_0, Z(t_1) = x_1, \ldots, Z(t_K) = x_K).$$

When the process $\{Z(t),\ t \geq 0\}$ is Markov, we can rewrite the likelihood as

$$\begin{aligned} L(\boldsymbol{\theta}|\mathbf{y}) &= \mathbb{P}(Z(t_0) = x_0) \prod_{k=1}^{K} \mathbb{P}(Z(t_k) = x_k \,|\, Z(t_{k-1}) = x_{k-1}) \\ &= \mathbb{P}(Z(t_0) = x_0) \prod_{k=1}^{K} \mathbb{P}(Z(t_k - t_{k-1}) = x_k \,|\, Z(0) = x_{k-1}). \end{aligned} \tag{2.4}$$

The last equation arises if the Markov process is time-homogeneous. The conditional probabilities in (2.4) can be calculated from the transition probability

$$\mathbb{P}(Z(t)|Z(0)) \;=\; \exp(Qt) \;:=\; \sum_{k=0}^{\infty} \frac{(Qt)^k}{k!},$$

where Q is the transition rate matrix of the time-homogeneous Markov process. The matrix exponential $\exp(Qt)$ defined above can be approximated numerically.

Therefore, the likelihood function can be maximized by Newton type methods (see Kalbfleisch and Lawless 1985).

When the process is not Markov, the computation of the probabilities (and hence the likelihood) becomes very difficult. It involves integrating the (often high-dimensional) complete data likelihood over all possible sample paths and sample spaces that could have given rise to the observed data. The situation becomes even more involved if we have a combination of left and right censored data and exact failure (time-to-absorption) for some of the systems.

2.4 Parametric Inference Procedures

In this section, we discuss likelihood-based inference. This problem falls under the framework of the classical missing data problem, and one could use the E-M approach to compute the MLEs. The major challenge is the E-step, which involves obtaining the complete data likelihood given the observed data as discussed. Monte Carlo EM has been proposed to address such situations. In this case, however, the Monte Carlo EM does not work well, so we resort to other methods. A similar parametric inference procedure from Bayesian perspective is given as an extension in next section.

2.4.1 Observed Likelihood

The presence of interval censoring arises naturally with panel data as the underlying process are monitored discretely. In addition to interval censoring, right censoring is present when the process has not reached the absorbing state by the end of the study. Besides, despite the fact that the process is being observed at discrete

time, one could still observe the exact time to absorption in some cases – for example in many health situations, the time of death of a patient is available. We call this exact failure case.

Here we will derive the observed likelihood based on a random sample of N units (systems/subjects). Denote the observed panel data as $\mathbf{y}_{1:N} = (\mathbf{y}_1, \ldots, \mathbf{y}_N)$, where $\mathbf{y}_n = \{(t_{kn}, x_{kn})\}$ and the $n-$th sample is observed at times t_{kn}, $k = 1, \ldots, K_n$ and $n = 1, \ldots, N$. Let $\{Z_n(t),\ t \geq 0\}$ be the SMP associated to the n-th sample. We can write the likelihood function of the observed sample as

$$L(\boldsymbol{\theta}|\mathbf{y}_{1:N}) \ = \ \prod_{n=1}^{N} \mathbb{P}\left(Z_n(t_{1n}) = x_{1n}, Z_n(t_{2n}) = x_{2n}, \ldots, Z_n(t_{K_n,n}) = x_{K_n,n}\right).$$

This likelihood function can be re-expressed as a continuous mixture by introducing the distribution of the complete history $\mathbf{Z} = \{(I_s, \tau_s),\ s = 1, \ldots, \sigma\}$. Let $f(\cdot; \boldsymbol{\theta})$ be the density of $\mathbf{Z}$ and for $\mathbf{y} = \{(t_k, x_k),\ k = 1, \ldots, K\}$, define

$$g(\mathbf{y}|\mathbf{z}) \ = \ \mathbb{P}\left(Z(t_1) = x_1, \ldots, Z(t_K) = x_K | \mathbf{Z} = \mathbf{z}\right),$$

the conditional distribution of the observed random vector $(Z(t_1), \ldots, Z(t_K))$ given the complete history $\mathbf{Z}$. Then by marginalization, the likelihood can be written as

$$L(\boldsymbol{\theta}|\mathbf{y}_{1:N}) \ = \ \int_{\mathcal{Z}} g(\mathbf{y}_{1:N}|\mathbf{z}_{1:N}) f(\mathbf{z}_{1:N}; \boldsymbol{\theta}) d\mathbf{z}_{1:N}, \tag{2.5}$$

where $\mathcal{Z}$ is the sample space of the complete history $\mathbf{Z}$ and

$$g(\mathbf{y}_{1:N}|\mathbf{z}_{1:N}) \ = \ \prod_{n=1}^{N} g(\mathbf{y}_n|\mathbf{z}_n), \quad \text{and} \quad f(\mathbf{z}_{1:N}; \boldsymbol{\theta}) \ = \ \prod_{n=1}^{N} f(\mathbf{z}_n; \boldsymbol{\theta}).$$

To complete the description of the likelihood, we need the expression of $f(\mathbf{z}; \boldsymbol{\theta})$ and $g(\mathbf{y}|\mathbf{z})$. We first introduce some notations. Let $\mathcal{I}$ be the set of all sequences $\boldsymbol{i} = (i_0, i_1, \ldots, i_\sigma)$, such that $i_0 = 1 < i_1 < i_2 < \cdots < i_\sigma = p$. We use $\sigma(\boldsymbol{i})$ to represent the number of jumps along $\boldsymbol{i}$. The set $\mathcal{I}$ is the sample space of the embedded

Markov chain of the SMP, and $\sigma(\boldsymbol{i})$ corresponds to the number of transitions along path $\boldsymbol{i}$.

For a given sample path $\boldsymbol{i}$ of the embedded Markov chain, let $\mathcal{T}_{\boldsymbol{i}} = (0,\infty)^{\sigma(\boldsymbol{i})}$ denote the sample space of the sojourn times along path $\boldsymbol{i}$. The sample space of the complete history $\mathbf{Z}$ is $\mathcal{Z} = \cup_{\boldsymbol{i}\in\mathcal{I}}\{\boldsymbol{i}\} \times \mathcal{T}_{\boldsymbol{i}}$. A generic element of $\mathcal{Z}$ is $\mathbf{z} = (\boldsymbol{i}, \boldsymbol{\tau_i})$ where $\boldsymbol{i}$ represents a possible path of the embedded Markov chain and $\boldsymbol{\tau_i} = (\tau_1, \ldots, \tau_{\sigma(\boldsymbol{i})})$ represents a possible set of sojourn times along that path. Notice that there are possibly different number of sojourn times along different paths. As we will see, this particularity of the model complicates the inference.

The density of the complete history $\mathbf{Z}$ is obtained from the definition of the MRP and is given by

$$f(\mathbf{z};\boldsymbol{\theta}) \;=\; f(\boldsymbol{i},\boldsymbol{\tau_i};\boldsymbol{\theta}) \;=\; \prod_{s=1}^{\sigma(\boldsymbol{i})} q_{i_{s-1},i_s}(\tau_s;\boldsymbol{\theta}), \tag{2.6}$$

where $q_{ij}(\tau;\boldsymbol{\theta})$ is the derivative of the semi-Markov kernel in (2.3).

Define $N(t) := \max\{s \geq 0 : \sum_{j=0}^{s} \tau_j \leq t\}$. Then, it is easy to see that

$$\begin{aligned} g(\mathbf{y}|\boldsymbol{i},\boldsymbol{\tau_i}) \quad &= \quad \mathbb{P}\left(Z(t_1) = x_1, \ldots, Z(t_k) = x_k | \mathbf{Z} = (\boldsymbol{i},\boldsymbol{\tau_i})\right) \\ &= \quad \prod_{k=1}^{K} \mathbf{1}\{i_{N(t_k)} = x_k\}. \end{aligned} \tag{2.7}$$

For a single observation $\mathbf{y}$, the likelihood is then obtained by integrating out $\mathbf{z}$ from the joint density of $(\mathbf{Y},\mathbf{Z})$:

$$L(\boldsymbol{\theta}|\mathbf{y}) \;=\; \int_{\mathcal{Z}} g(\mathbf{y}|\mathbf{z}) f(\mathbf{z};\boldsymbol{\theta}) d\mathbf{z} \;=\; \sum_{\boldsymbol{i}\in\mathcal{I}} \int_{\mathcal{T}_{\boldsymbol{i}}} g(\mathbf{y}|\boldsymbol{i},\boldsymbol{\tau_i}) \prod_{s=1}^{\sigma(\boldsymbol{i})} q_{i_{s-1},i_s}(\tau_s;\boldsymbol{\theta})\; d\boldsymbol{\tau_i}.$$

For general SMPs, this observed likelihood involves high dimension integrals and cannot be simplified further.

The formulation of the problem presented above and the expression of the likelihood derived in (2.5) is very general and automatically accommodates right censoring, that is observation $\mathbf{y} = \{(t_k, x_k),\ k = 1, \ldots, K\}$ for which $x_K < p$. In many

other cases, the exact failure time is additionally observed. This is the case of many studies (as in the example presented in Section 2.7) where the time of death of patient is available. In this case the observed data are $\mathbf{y} = \{t, (t_k, x_k), k = 1, \ldots, K\}$, where t is the observed absorbing time of the system. This case is easily handled by modifying $g(\mathbf{y}|\mathbf{z})$, the conditional distribution of the observed random vector given the complete history from (2.7) to

$$g(\mathbf{y}|\boldsymbol{i}, \boldsymbol{\tau_i}) \;=\; \mathbf{1}\{\sum_{j=1}^{\sigma(\mathbf{i})} \tau_j = t\} \;\prod_{k=1}^{K} \mathbf{1}\{i_{N(t_k)} = x_k\}.$$

2.4.2 Likelihood-Based Inference Using Stochastic Approximation and MCMC Sampling

For general SMPs, the likelihood function $L(\boldsymbol{\theta}|\mathbf{y}_{1:N})$ given in (2.5) is intractable. Therefore, computing the MLE is not straightforward. One possible approach is the E-M (or Monte Carlo EM) algorithm. For the E-M, the Q function is given by

$$Q(\boldsymbol{\theta}'|\boldsymbol{\theta}) \;=\; \int \log\left\{g(\mathbf{y}_{1:N}|\mathbf{z}_{1:N}) f(\mathbf{z}_{1:N}; \boldsymbol{\theta}')\right\} p(\mathbf{z}_{1:N}|\boldsymbol{\theta}, \mathbf{y}_{1:N}) d\mathbf{z}_{1:N},$$

where $p(\mathbf{z}_{1:N}|\boldsymbol{\theta}, \mathbf{y}_{1:N}) \;=\; \prod_{n=1}^{N} p(\mathbf{z}_n|\boldsymbol{\theta}, \mathbf{y}_n)$ and $p(\mathbf{z}_n|\boldsymbol{\theta}, \mathbf{y}_n)$ denotes the conditional distribution of $\mathbf{z}_n$ given $\mathbf{y}_n$. That is,

$$p(\mathbf{z}_n|\boldsymbol{\theta}, \mathbf{y}_n) \;\propto\; f(\mathbf{z}_n; \boldsymbol{\theta}) g(\mathbf{y}_n|\mathbf{z}_n).$$

In the above equation, $f(\mathbf{z}_n; \boldsymbol{\theta})$ is the density of the complete history given in (2.6) and $g(\mathbf{y}_n|\mathbf{z}_n)$ is the conditional density of $\mathbf{y}$ given $\mathbf{z}$. In the E-step of the E-M algorithm, one evaluates the function $Q(\boldsymbol{\theta}'|\boldsymbol{\theta})$ and in the M-step this function is maximized to yield the next estimate of the MLE. These two steps are repeated until convergence. But, in the present case, the $Q(\cdot|\boldsymbol{\theta})$ function is intractable and the

E-M algorithm cannot be implemented in an analytical form. A common approach in such cases is the use of the Monte Carlo EM algorithm in which the exact calculation of $Q(\cdot|\boldsymbol{\theta})$ in the E-step is replaced by an approximation using Monte Carlo simulation from $p(\mathbf{z}_{1:N}|\boldsymbol{\theta}, \mathbf{y}_{1:N})$, see for example Wei and Tanner (1990), Delyon (1999). Unfortunately, as we will see in Section 2.4.3, in general the conditional distribution $p(\mathbf{z}_{1:N}|\boldsymbol{\theta}, \mathbf{y}_{1:N})$ is very difficult to sample from and we have to use MCMC techniques. Thus, each iteration of the Monte Carlo EM algorithm requires a full-fledged MCMC simulation from $p(\mathbf{z}_{1:N}|\boldsymbol{\theta}, \mathbf{y}_{1:N})$. Hence, it is too expensive to use Monte Carlo EM.

An alternative is the use of stochastic approximation (SA) together with MCMC sampling. In fact, and as explained, for example, in Atchadé (2010), these two algorithms are closely related. SA becomes computationally more attractive in cases (such as the one dealt with in this chapter) where MCMC is needed to evaluate the Q function in the E-M algorithm. Thus in what follows, we shall use SA to compute the MLE. The idea of using stochastic approximation algorithms to deal with intractable likelihood functions goes back at least to Younes (1988). A more general treatment is given by Gu and Kong (1998). See also Cappé, Moulines and Ryden (2005) for applications to state space models.

Define the score function $h(\boldsymbol{\theta}|\mathbf{y}_{1:N}) := \nabla_\theta \log L(\boldsymbol{\theta}|\mathbf{y}_{1:N})$. It is easy to see that $h(\boldsymbol{\theta}|\mathbf{y}_{1:N})$ can be written as

$$h(\boldsymbol{\theta} \,|\, \mathbf{y}_{1:N}) \;=\; \int H(\boldsymbol{\theta}, \mathbf{z}_{1:N}) p(\mathbf{z}_{1:N}|\boldsymbol{\theta}, \mathbf{y}_{1:N}) d\mathbf{z}_{1:N}, \tag{2.8}$$

where $H(\boldsymbol{\theta}, \mathbf{z}_{1:N}) := \sum_{n=1}^{N} \nabla_\theta \log f(\mathbf{z}_n; \boldsymbol{\theta})$. The expression of $H(\boldsymbol{\theta}, \mathbf{z}_{1:N})$ is simple to evaluate even for general SMPs. Let $\mathcal{K}_{\boldsymbol{\theta}}$ be a transition kernel on $\mathcal{Z}^N$ with invariant distribution $p(\mathbf{z}_{1:N}|\boldsymbol{\theta}, \mathbf{y}_{1:N})$, where we omit the dependence of $\mathcal{K}_{\boldsymbol{\theta}}$ on $\mathbf{y}_{1:N}$. Designing such Markov kernel $\mathcal{K}_{\boldsymbol{\theta}}$ with invariant distribution $p(\mathbf{z}_{1:N}|\boldsymbol{\theta}, \mathbf{y}_{1:N})$ holds some chal-

lenges. This is because the complete history $\mathbf{z}_n$ lives in $\cup_{\boldsymbol{i}\in\mathcal{I}}\{\boldsymbol{i}\}\times\mathcal{T}_{\boldsymbol{i}}$, a union of spaces of varying dimensionality. We give more detail on building $\mathcal{K}_{\boldsymbol{\theta}}$ in Section 2.4.3.

Let $m \geq 1$ be a given integer, $\{\gamma_\ell,\ \ell \geq 1\}$ a sequence of positive numbers, and Γ a $r \times r$ positive definite matrix, where r is the dimension of $\boldsymbol{\theta}$. The SA algorithm to compute the MLE generates a random process $\{(\boldsymbol{\theta}_\ell, \boldsymbol{\zeta}_\ell),\ \ell \geq 0\}$ in $\Theta \times \mathcal{Z}^N$ as described in Algorithm 2.4.1.

Algorithm 2.4.1 MLE approximation by SA

Given $(\boldsymbol{\theta}_{\ell-1}, \boldsymbol{\zeta}_{\ell-1}) \in \Theta \times \mathcal{Z}^N$:

1. Set $\boldsymbol{\zeta}^{(0)} = \boldsymbol{\zeta}_{\ell-1}$. For $s = 1, \cdots, m$,

$$\text{generate} \quad \boldsymbol{\zeta}^{(s)} | \boldsymbol{\zeta}^{(s-1)} \sim \mathcal{K}_{\boldsymbol{\theta}_{\ell-1}}\left(\boldsymbol{\zeta}^{(s-1)}, \cdot\right), \quad \text{and set} \quad \boldsymbol{\zeta}_\ell = \boldsymbol{\zeta}^{(m)}.$$

2. Compute the new estimate:

$$\boldsymbol{\theta}_\ell \ = \ \boldsymbol{\theta}_{\ell-1} + \gamma_\ell \Gamma^{-1} \left(\frac{1}{m} \sum_{s=1}^{m} H\left(\boldsymbol{\theta}_{\ell-1}, \boldsymbol{\zeta}^{(s)}\right) \right). \tag{2.9}$$

These two steps are then iterated until convergence.

Equation (2.8) suggests that

$$\bar{H}_m(\boldsymbol{\theta}) \ = \ \frac{1}{m} \sum_{s=1}^{m} H(\boldsymbol{\theta}, \boldsymbol{\zeta}^{(s)})$$

is an empirical estimate of $h(\boldsymbol{\theta}|\mathbf{y}_{1:N})$. Thus, (2.9) is a sort of stochastic quasi-Newton method to solve the normal equation $h(\boldsymbol{\theta}|\mathbf{y}_{1:N}) = 0$. If, for a particular example, exact sampling from $p(\mathbf{z}_{1:N}|\boldsymbol{\theta}, \mathbf{y}_{1:N})$ is feasible, the MCMC simulation step of the algorithm (Step 1) can be replaced by exact simulation from $p(\mathbf{z}_{1:N}|\boldsymbol{\theta}, \mathbf{y}_{1:N})$; however this is rarely the case in our context.

In Algorithm 2.4.1, m is a fixed integer introduced to improve numerical stability. In particular, m need not be large and can be set to $m = 1$ if the mixing of the kernels $\mathcal{K}_{\boldsymbol{\theta}_{\ell-1}}\left(\boldsymbol{\zeta}^{(s-1)}, \cdot\right)$ is reasonably good. In our simulations, we use a conservative value

of $m = 100$. The step size γ_ℓ is a decreasing positive sequence such that

$$\sum_\ell \gamma_\ell = \infty, \quad \sum_\ell \gamma_\ell^2 < \infty.$$

In our simulations, we use $\gamma_\ell \propto \ell^{-\lambda}$ for $\lambda = 2/3$. The matrix Γ is introduced in order to properly scale the algorithm, particularly in the vicinity of the MLE. Ideally, we would like to use $\Gamma = I(\hat{\boldsymbol{\theta}}|\mathbf{y}_{1:N})$, the observed information matrix evaluated at the MLE $\hat{\boldsymbol{\theta}}$. Denote the observed information matrix at $\boldsymbol{\theta}$ as $I(\boldsymbol{\theta}|\mathbf{y}_{1:N}) = -\nabla^2_{\boldsymbol{\theta}} \log L(\boldsymbol{\theta}|\mathbf{y}_{1:N})$. By the missing information principle (as in Louis 1982), $I(\boldsymbol{\theta}|\mathbf{y}_{1:N})$ has the representation

$$\begin{aligned} I(\boldsymbol{\theta}|\mathbf{y}_{1:N}) &= h(\boldsymbol{\theta}|\mathbf{y}_{1:N}) h(\boldsymbol{\theta}|\mathbf{y}_{1:N})' \\ &\quad - \int \left\{ \nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}, \mathbf{z}_{1:N}) + H(\boldsymbol{\theta}, \mathbf{z}_{1:N}) H(\boldsymbol{\theta}, \mathbf{z}_{1:N})' \right\} p\left(\mathbf{z}_{1:N}|\boldsymbol{\theta}, \mathbf{y}_{1:N}\right) d\mathbf{z}_{1:N}. \end{aligned} \quad (2.10)$$

For a proof of (2.10), see Proposition 10.1.6 in Cappé, Moulines and Ryden (2005) or Lemma 1 in Gu and Kong (1998). In view of this expression, and following Gu and Kong (1998), we introduce the function

$$S(\boldsymbol{\theta}, \mathbf{z}_{1:N}) = -\nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}; \mathbf{z}_{1:N}) - H(\boldsymbol{\theta}, \mathbf{z}_{1:N}) H(\boldsymbol{\theta}, \mathbf{z}_{1:N})'.$$

Hence, we can estimate $I(\hat{\boldsymbol{\theta}} \,|\, \mathbf{y}_{1:N})$ recursively (along the iterations of Algorithm 2.4.1) using

$$\Gamma_\ell = \Gamma_{\ell-1} + \gamma_\ell \left\{ \bar{S}_m(\boldsymbol{\theta}_{\ell-1}) + \bar{H}_m(\boldsymbol{\theta}_{\ell-1}) \bar{H}_m(\boldsymbol{\theta}_{\ell-1})' - \Gamma_{\ell-1} \right\}, \quad (2.11)$$

where

$$\begin{aligned} \bar{H}_m(\boldsymbol{\theta}_{\ell-1}) &= \frac{1}{m} \sum_{s=1}^{m} H\left(\boldsymbol{\theta}_{\ell-1}, \boldsymbol{\zeta}^{(s)}\right), \\ \bar{S}_m(\boldsymbol{\theta}_{\ell-1}) &= \frac{1}{m} \sum_{s=1}^{m} S\left(\boldsymbol{\theta}_{\ell-1}, \boldsymbol{\zeta}^{(s)}\right). \end{aligned}$$

Under mild conditions, the random process $\boldsymbol{\theta}_\ell$ generated from Algorithm 2.4.1 (with $\Gamma = \Gamma_\ell$ estimated recursively using equation 2.11) converges to the MLE $\hat{\boldsymbol{\theta}}$ and Γ_ℓ converges to the observed information matrix $I(\hat{\boldsymbol{\theta}}|\mathbf{y}_{1:N})$, evaluated at the MLE $\hat{\boldsymbol{\theta}}$. The convergence of stochastic approximation algorithms have been extensively studied in the literature. We mention Gu and Kong (1998) and Cappé, Moulines and Ryden (2005), where further references can also be found.

$\boldsymbol{\theta}_0$ and Γ_0 are needed to initialize the algorithm. From our simulation study, the choice of Γ_0 is not critical and the identity matrix works fine. Note that $S_m(\boldsymbol{\theta}_{\ell-1}) + \bar{H}_m(\boldsymbol{\theta}_{\ell-1})\bar{H}_m(\boldsymbol{\theta}_{\ell-1})'$ might not be positive definite when ℓ is small. In that case, one can use $m^{-1}\sum_{k=1}^{m} \nabla_{\boldsymbol{\theta}} H(\boldsymbol{\theta}_{\ell-1}; \boldsymbol{\zeta}^{(k)})$ as an approximation of the information matrix. In addition, $\boldsymbol{\theta}_0$ needs to be chosen reasonable. If the interval censoring is not too severe, we can ser $\sum_{n=1}^{s} \tau_n$ as the mid-point of the interval in which the $s-$th transition occurred. Then, rough estimation can be done for each transition (from state i to state j) separately by regular survival analysis. If censoring is severe, we suggest running a short chain by data augmentation approach proposed in Section 2.5.1 and use the approximate marginal modes of $\boldsymbol{\theta}$ as initials. This is feasible since the data augmentation method is not so picky about $\boldsymbol{\theta}_0$.

2.4.3 Sampling the Complete History

We now discuss how to sample the complete history from $p(\mathbf{z}_{1:N}|\boldsymbol{\theta}, \mathbf{y}_{1:N})$ by MCMC sampling. Since $\mathbf{z}_n$ only depends on $\mathbf{y}_n$, we have

$$p(\mathbf{z}_{1:N}|\boldsymbol{\theta}, \mathbf{y}_{1:N}) \;=\; \prod_{n=1}^{N} p(\mathbf{z}_n|\boldsymbol{\theta}, \mathbf{y}_n).$$

Sampling from $p(\mathbf{z}|\boldsymbol{\theta}, \mathbf{y})$ is essential to the proposed stochastic approximation algorithm.

For notation convenience, let δ be the numbering of $\boldsymbol{i}$. For instance, if $\boldsymbol{i}$ can take two paths $(1,2,3)$ and $(1,3)$ in $\mathcal{I}$. We name $(1,2,3)$ path 1 with $\delta = 1$, and $(1,3)$ path 2 with $\delta = 2$. There is a 1-to-1 relationship between $\boldsymbol{i}$ and δ, and we will use $\mathbf{z} = (\delta, \boldsymbol{\tau}_\delta)$ and $\mathbf{z} = (\boldsymbol{i}, \boldsymbol{\tau}_{\boldsymbol{i}})$ interchangeably later. Further, denote d the number of consistent paths of $\mathbf{y}$, i.e., the number of elements in $\mathcal{I}$ s.t. $g(\mathbf{y}|\boldsymbol{i}, \boldsymbol{\tau}_{\boldsymbol{i}}) = 1$.

From Bayes' Theorem, given $\mathbf{y}$, the density of the corresponding complete data $\mathbf{z}$ is proportional to the joint density of $(\mathbf{y}, \mathbf{z})$:

$$p(\mathbf{z}|\boldsymbol{\theta}, \mathbf{y}) \;\propto\; g(\mathbf{y}|\mathbf{z}) f(\mathbf{z}; \boldsymbol{\theta}) \;=\; \sum_{j=1}^{d} g(\mathbf{y}|\delta = j, \boldsymbol{\tau}_\delta) f(\delta = j, \boldsymbol{\tau}_\delta; \boldsymbol{\theta}) \tag{2.12}$$

Note that sampling from $p(\mathbf{z}|\boldsymbol{\theta}, \mathbf{y})$ is not trivial, since the normalizing constant of this density is difficult to calculate.

When $d = 1$, there is only one consistent path of $\mathbf{y}$. Thus, $\boldsymbol{i}$ is known and write $\delta = 1$. Note that, when one-step progressive model is assumed, this is always the case. Under this scenario, $\mathbf{z} = (\delta = 1, \boldsymbol{\tau}_\delta)$ has density

$$p(\mathbf{z}|\boldsymbol{\theta}, \mathbf{y}) \;\propto\; g(\mathbf{y}|\delta = 1, \boldsymbol{\tau}_\delta) f(\delta = 1, \boldsymbol{\tau}_\delta; \boldsymbol{\theta}). \tag{2.13}$$

Note that, when $\sigma(\boldsymbol{i}) > 1$, for each $k \in \sigma(\boldsymbol{i})$, τ_k and $\boldsymbol{\tau}_{-k}$ are dependent because of $g(\mathbf{y}|\delta = 1, \boldsymbol{\tau}_\delta)$. For instance, assume the SMP is a 4-state one-step progressive model. If $\mathbf{y} = (t_1, 1, t_2, 1, t_3, 2, t_4, 4)$, $\delta = 1$ with $\boldsymbol{i} = (1,2,3,4)$ and

$$g(\mathbf{y}|\delta = 1, \boldsymbol{\tau}_\delta) \;=\; \mathbf{1}\{t_2 < \tau_1 < t_3, t_3 < \tau_1 + \tau_2 < t_4, t_3 < \tau_1 + \tau_2 + \tau_3 < t_4\}.$$

Therefore, τ_1, τ_2 and τ_3 are dependent since

$$\tau_1 \in (\max(t_2, t_3 - \tau_2),\, \min(t_3, t_4 - \tau_2 - \tau_3)],$$

$$\tau_2 \in (t_3 - \tau_1, t_4 - \tau_1 - \tau_3], \text{ and } \tau_3 \in (0, t_4 - \tau_1 - \tau_2].$$

The conditional distribution of τ_k given $\boldsymbol{\tau}_{-k}$ is readily present through (2.13), so Gibbs Sampling can be used to draw $\boldsymbol{\tau}_\delta$ if $\sigma(\boldsymbol{i}) > 1$.

When $d > 1$, δ is random. Hence, $\mathbf{z} = (\delta, \boldsymbol{\tau}_\delta)$ has a discrete component δ and a continuous component $\boldsymbol{\tau}_\delta$ with varying dimensions. Since the distribution of sojourn times $\boldsymbol{\tau}_\delta$ is determined when δ is known, we can rewrite the joint density of $\mathbf{z} = (\delta, \boldsymbol{\tau}_\delta)$ as a mixture

$$p(\delta, \boldsymbol{\tau}_\delta|\boldsymbol{\theta}, \mathbf{y}) \;=\; \sum_{j=1}^{d} a_j \, p(\boldsymbol{\tau}_\delta|\boldsymbol{\theta}, \mathbf{y}, \delta = j), \tag{2.14}$$

where $a_j := \mathbb{P}(\delta = j|\boldsymbol{\theta}, \mathbf{y})$ and

$$p(\boldsymbol{\tau}_\delta|\boldsymbol{\theta}, \mathbf{y}, \delta = j) \;\propto\; g(\mathbf{y}|\delta = j, \boldsymbol{\tau}_\delta) f(\delta = j, \boldsymbol{\tau}_\delta; \boldsymbol{\theta}).$$

Note that δ follows a multinomial distribution with probability vector $\boldsymbol{a} = \{a_j, j = 1, \ldots, d\}$, denoted as Multinomial($\boldsymbol{a}$). Therefore, a natural approach is to sample $\mathbf{Z}$ by the conditional sampling shown in Algorithm 2.4.2.

Algorithm 2.4.2 Conditional Sampling

Given $\boldsymbol{\theta}$ and $\mathbf{y}$, sample $\mathbf{z} = (\delta, \boldsymbol{\tau}_\delta)$ via

1. Calculate the probability vector $\boldsymbol{a}$ as follows:

$$a_j = \frac{\int g(\mathbf{y}|\delta = j, \boldsymbol{\tau}_\delta) f(\delta = j, \boldsymbol{\tau}_\delta; \boldsymbol{\theta}) d\boldsymbol{\tau}_\delta}{\sum_{j'=1}^{d} \int g(\mathbf{y}|\delta = j', \boldsymbol{\tau}_\delta) f(\delta = j', \boldsymbol{\tau}_\delta; \boldsymbol{\theta}) d\boldsymbol{\tau}_\delta}, \quad j = 1, \ldots, d,$$

2. Generate $\delta \sim$ Multinomial($\boldsymbol{a}$).
3. Generate $\boldsymbol{\tau} \sim p(\boldsymbol{\tau}_\delta|\boldsymbol{\theta}, \mathbf{y}, \delta)$.

In the algorithm, Step 3 is similar to (2.13) and Gibbs sampling can be used when $\boldsymbol{\tau}_\delta$ includes more than one sojourn time. In Step 1, to get the marginal probability vector $\boldsymbol{a}$, we need to integrate over $\boldsymbol{\tau}_\delta$ (usually of high dimension). For general SMPs, these integrals are numerically approximated and the computation of a_j is

very time-consuming. We will show later in the simulation study that this is the major drawback of using conditional sampling.

Now we introduce a better sampling procedure, i.e., the reversible jump MCMC sampling (RJMCMC). The idea is that, if sampling $(\delta, \boldsymbol{\tau}_\delta)$ together from the joint (2.12) is possible, the computational difficulty caused by sampling δ marginally in (2.14) can be spared. Regular Metropolis-Hasting sampling or Gibbs sampling is not suitable, since the dimension of $\boldsymbol{\tau}_\delta$ changes when δ varies. Green (1995) proposed the RJMCMC procedure which handles the sampling of distribution with variable dimensions. Refer to Green (1995) for convergence proof and details about the sampling procedure. To relate our notation to Green's, we refer to $(\delta, \boldsymbol{\tau}_\delta)$ as the model $\mathcal{M}_\delta$ in his paper. Then, the joint distribution is the mixture of $\{\mathcal{M}_\delta; \delta = 1, \ldots, d\}$. The RJMCMC procedure is given in Algorithm 2.4.3.

Algorithm 2.4.3 RJMCMC Sampling

From the current draw $(\delta, \boldsymbol{\tau}_\delta)$, the new proposal can be stated as follows:

1. Select model $\mathcal{M}_{\delta'}$ with probability $\pi_{\delta \to \delta'}$.
2. Generate u from some distribution $\psi_{\delta \to \delta'}(u)$.
3. Obtain new draw $(\delta', \boldsymbol{\tau}_{\delta'})$ through a dimension matching transformation T s.t.

$$(\boldsymbol{\tau}_{\delta'}, \nu) \;=\; T_{\delta \to \delta'}(\boldsymbol{\tau}^\delta, u).$$

   The dimension of $(\boldsymbol{\tau}^{\delta'}, \nu)$ and the dimension of $(\boldsymbol{\tau}^\delta, u)$ are equal.

4. Accept $(\delta', \boldsymbol{\tau}_{\delta'})$ with probability

$$\alpha_{\delta \to \delta'} \;=\; \min(1, \frac{f(\delta', \boldsymbol{\tau}_{\delta'}; \boldsymbol{\theta}) g(\mathbf{y}|\delta', \boldsymbol{\tau}_{\delta'}) \pi_{\delta' \to \delta} \psi_{\delta' \to \delta}(\nu)}{f(\delta, \boldsymbol{\tau}_\delta; \boldsymbol{\theta}) g(\mathbf{y}|\delta, \boldsymbol{\tau}_\delta) \pi_{\delta \to \delta'} \psi_{\delta \to \delta'}(u)} |J|), \tag{2.15}$$

   where $|J| = \left| \frac{\partial T_{\delta \to \delta'}(\boldsymbol{\tau}^\delta, u)}{\partial (\boldsymbol{\tau}^\delta, u)} \right|$ is the Jacobian factor.

Note that the acceptance probability defined in (2.15) is easy to evaluate and

there is no need to calculate normalizing constants. Depending on the dimension of $\boldsymbol{\tau}_\delta$ and $\boldsymbol{\tau}_{\delta'}$, u and ν might be randomly generated from the proposal $\psi_{\delta\to\delta'}$ or they can be deterministic as long as the dimension matches. If $\delta' = \delta$, the move is the regular within-model sampling. The procedure coincides with the Metropolis-Hastings algorithm.

Now we illustrate how to use RJMCMC to sample the complete history. The detailed sampling procedure depends on the underlying multistate structure, and the form of the panel data, i.e., whether they are only interval-censored, or including right censoring or exact failure. Here, we focus on the interval censoring case. The illustration of sampling along with right censoring or exact failure is given in Appendix.

Consider a 3-state multi-step progressive model with 2 possible paths: $(1,2,3)$ and $(1,3)$. The challenge is then to sample $\mathbf{Z}$ with observed $\mathbf{y} = (t_1, 1, t_2, 3)$. There are 2 consistent paths of $\mathbf{y}$ and let $\delta = 1$ when $\boldsymbol{i} = (1,2,3)$, $\delta = 2$ when $\boldsymbol{i} = (1,3)$. Denote $\tau_s^{(ij)}$ as the $s-$th transition made from state i to state j. The distribution of $\mathbf{z} = (\delta, \boldsymbol{\tau}_\delta)$ given $\mathbf{y}$ follows (2.12) with

$$p(\mathbf{z}|\boldsymbol{\theta}, \mathbf{y}) \;\propto\; \sum_{j=1}^{2} g(\mathbf{y}|\delta = j, \boldsymbol{\tau}_\delta) f(\delta = j, \boldsymbol{\tau}_\delta; \boldsymbol{\theta})$$

where

$$\begin{aligned} g(\mathbf{y}|\delta = 1, \boldsymbol{\tau}_\delta) f(\delta = 1, \boldsymbol{\tau}_\delta; \boldsymbol{\theta}) &= \mathbf{1}\{\tau_1, \tau_1 + \tau_2 \in (t_1, t_2)\} \times p_{12} f_{12}(\tau_1) f_{23}(\tau_2), \\ g(\mathbf{y}|\delta = 2, \boldsymbol{\tau}_\delta) f(\delta = 2, \boldsymbol{\tau}_\delta; \boldsymbol{\theta}) &= \mathbf{1}\{\tau_1 \in (t_1, t_2]\} \times p_{13} f_{13}(\tau_1). \end{aligned}$$

Assume the current iterate takes $\delta = 2$, i.e., $\boldsymbol{\tau}_\delta = \tau_1^{(13)}$. To calculate the acceptance probability $\alpha_{2\to1}$, we use the following dimension matching transformation $(\tau_1^{(12)}, \tau_2^{(23)}) = T_{2\to1}(\tau_1^{(13)})$:

$$\tau_1^{(12)} = \tau_1^{(13)}, \quad \tau_2^{(23)} = u(t_2 - \tau_1^{(13)}), \tag{2.16}$$

where $u \sim \mathcal{U}(0,1)$. The Jacobian factor of this transformation is $|J| = t_2 - \tau_1^{(13)}$. Then, the probability of accepting the new draw $(\tau_1^{(12)}, \tau_2^{(23)})$ is

$$\alpha_{2\to 1} = \min\left(1, \frac{p_{12} f_{12}(\tau_1^{(12)}) p_{23} f_{23}(\tau_2^{(23)}) \times \pi_{2\to 1}}{p_{13} f_{13}(\tau_1^{(13)}) \times \pi_{1\to 2}} |J|\right).$$

If the current iterate takes $\delta = 1$, we use the dimension matching transform $\tau_1^{(13)} = T_{1\to 2}(\tau_1^{(12)}, \tau_2^{(23)})$: $\tau_1^{(13)} = \tau_1^{(12)}$, and the acceptance probability $\alpha_{1\to 2}$ can be derived similarly. Note that other dimension matching transformations and distributions of u can be used as long as the interval censoring constraint is met.

Now consider the update within each path, i.e., the new iterate takes the same path as the old one. To update for path (1,2,3), denote the current iterate as $\mathbf{z}_{\ell-1} = (\tau_{\ell-1,1}^{(12)}, \tau_{\ell-1,2}^{(23)})$. $\tau_{\ell,1}^{(12)}$ and $\tau_{\ell,2}^{(23)}$ can be sampled from $p_{12}(\cdot)$ and $p_{23}(\cdot)$ respectively

$$\begin{aligned} p_{12}(\tau_1^{(12)}) &\propto \mathbf{1}\{\tau_1^{(12)} \in (t_1, t_2] \cap (t_1 - \tau_{\ell-1,2}^{(23)}, t_2 - \tau_{\ell-1,2}^{(23)})\} \times f_{12}(\tau_1^{(12)}), \\ p_{23}(\tau_2^{(23)}) &\propto \mathbf{1}\{\tau_2^{(23)} \in (t_1 - \tau_{\ell,1}^{(12)}, t_2 - \tau_{\ell,1}^{(12)}]\} \times f_{23}(\tau_2^{(23)}). \end{aligned}$$

To update for path (1,3) is trivial.

The sampling procedure of $\mathbf{Z}$ depends on how the state space is visited. When multi-step progressive structure is assumed (possibly $d > 1$ for some observed $\mathbf{y}$), the sampling procedure is more involved than that of the one-step progressive model (always $d = 1$ for any observed $\mathbf{y}$). We have proposed two algorithms to sample for any progressive multistate model. For the general multistate model with cyclic structure, the joint (2.12) will now have infinity element in $\mathcal{I}$ with $d = \infty$ because of the recurrent property. For instance, even if the state information has not changed at 2 consecutive time points, there is positive probability that the process has transit out and back into this state when recurrent transition is allowed. Since RJMCMC is feasible even when the number of the mixture components is unknown, the extension

can be done in principle. However, the setup of the procedure is more complex and may not be an efficient algorithm to pursue. Thus, we will not go further in this chapter.

2.5 Extensions

2.5.1 Bayesian Inference Using Data Augmentation

Even though we focus on likelihood-based inference in Section 2.4, Bayesian inference can be made using a related data augmentation approach. Here we briefly discuss how to make Bayesian inference under the same setting. See Tanner and Wong (1987) and Hobert (2009) for more details about data augmentation.

Let $p(\boldsymbol{\theta})$ be the prior distribution of the parameter $\boldsymbol{\theta}$. The posterior distribution given the observed data $\mathbf{y}_{1:N}$ is

$$p(\boldsymbol{\theta}|\mathbf{y}_{1:N}) \ \propto \ p(\boldsymbol{\theta})p(\mathbf{y}_{1:N}|\boldsymbol{\theta}).$$

Note that $p(\mathbf{y}_{1:N}|\boldsymbol{\theta})$ is the likelihood function given in Section 2.4.1, so $p(\boldsymbol{\theta}|\mathbf{y}_{1:N})$ is still intractable. For general SMPs, the exact sampling of $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|\mathbf{y}_{1:N})$ is difficult, so we augment the complete history $\mathbf{z}_{1:N}$ by data augmentation. Here we assume that the complete history $\mathbf{z}_{1:N}$ can be sampled from $p(\mathbf{z}_{1:N}|\boldsymbol{\theta}, \mathbf{y}_{1:N})$ by MCMC sampling from a transition kernel $\mathcal{K}_{\boldsymbol{\theta}}$ with invariant distribution $p(\mathbf{z}_{1:N}|\boldsymbol{\theta}, \mathbf{y}_{1:N})$. The details of the sampling procedures have been discussed in Section 2.4.3. It is obvious that when augmenting $\mathbf{z}_{1:N}$,

$$p(\boldsymbol{\theta}|\mathbf{y}_{1:N}) \ = \ \int p(\boldsymbol{\theta}|\mathbf{y}_{1:N}, \mathbf{z}_{1:N})p(\mathbf{z}_{1:N}|\mathbf{y}_{1:N})d\mathbf{z}_{1:N} \ = \ \int p(\boldsymbol{\theta}, \mathbf{z}_{1:N}|\mathbf{y}_{1:N})d\mathbf{z}_{1:N}.$$

One iteration of the data augmentation algorithm can be describes as follows:

Algorithm 2.5.1 Data Augmentation

Given $(\boldsymbol{\theta}_{\ell-1}, \boldsymbol{\zeta}_{\ell-1}) \in \Theta \times \mathcal{Z}^N$:

1. Generate $\boldsymbol{\zeta}_\ell | \boldsymbol{\zeta}_{\ell-1} \sim \mathcal{K}_{\boldsymbol{\theta}_{\ell-1}}(\boldsymbol{\zeta}_{\ell-1}, \cdot)$, where $\mathcal{K}_{\boldsymbol{\theta}}$ denotes a Markov kernel with invariant distribution $p(\mathbf{z}_{1:N}|\boldsymbol{\theta}, \mathbf{y}_{1:N})$.
2. Generate $\boldsymbol{\theta}_\ell | \mathbf{y}_{1:N}, \boldsymbol{\zeta}_\ell \sim \mathcal{K}_{\boldsymbol{\zeta}_\ell}(\boldsymbol{\theta}_{\ell-1}, \cdot)$, where $\mathcal{K}_{\boldsymbol{\zeta}}$ a Markov kernel with invariant distribution $p(\boldsymbol{\theta}|\mathbf{z}_{1:N}, \mathbf{y}_{1:N})$.

Iterate through the two steps and keep $\boldsymbol{\theta}_\ell$ after a burn-in period. With proper slicing, the draws $\boldsymbol{\theta}_\ell$ should be samples from the posterior $p(\boldsymbol{\theta}|\mathbf{y}_{1:N})$. Credible intervals can be formed based on the samples to make inference on the underlying SMP. Notice that no multiple $\boldsymbol{\zeta}$ is needed, i.e., $m = 1$ as in Tanner and Wong (1987). Contrary to sampling from $p(\boldsymbol{\theta}|\mathbf{y}_{1:N})$, the sampling from $p(\boldsymbol{\theta}|\mathbf{z}_{1:N}, \mathbf{y}_{1:N})$ is straightforward, and for some distributions, exact sampling is feasible.

2.5.2 Covariate Analysis

The discussion so far has been focused on parametric estimation based on a random sample. It is straightforward to extend the proposed algorithms to include static covariate effects. For Markov models, the proportional hazards model is usually imposed on the transition rate functions, i.e.,

$$q_{ij} \;=\; q_{ij}^{(0)} \exp(\boldsymbol{\omega}' \boldsymbol{\beta}_{ij}),$$

where $q_{ij}^{(0)}$ is the baseline and $\boldsymbol{\omega}$ the covariate vector and $\boldsymbol{\beta}_{ij}$ the associated covariate effect. Hence, for sample with covariate $\boldsymbol{\omega}$,

$$\lambda_i^\omega \;:=\; -q_{ii} \;=\; \sum_{j \neq i} q_{ij}^{(0)} \exp(\boldsymbol{\omega}' \boldsymbol{\beta}_{ij})$$

is the new rate for the exponential sojourn time in state i.

For general SMPs, the proportional hazards model is imposed on the conditional distribution of the sojourn time. For instance, assume the sojourn time in state i before moving to state j follows a Weibull distribution with shape parameter γ_{ij} and scale parameter $\eta_{ij} = 1/\lambda_{ij}$. Then, for the sample with covariate $\boldsymbol{\omega}$, its hazard function for transition from state i to state j is

$$\lambda_{ij}^{\boldsymbol{\omega}}(t) \;=\; \lambda_{ij}^{\gamma_{ij}}\gamma_{ij}t^{\gamma_{ij}-1}\exp(\boldsymbol{\omega}'\boldsymbol{\beta}_{ij}) \;:=\; {\lambda_{ij}^{\boldsymbol{\omega}}}^{\gamma_{ij}}\gamma_{ij}t^{\gamma_{ij}-1}, \tag{2.17}$$

where $\lambda_{ij}^{\boldsymbol{\omega}} = \lambda_{ij}\exp(\boldsymbol{\omega}'\boldsymbol{\beta}_{ij}/\gamma_{ij})$. This sample follows a Weibull distribution with still γ_{ij} parameter but a new scale parameter $\lambda_{ij}^{\boldsymbol{\omega}}$. Therefore, the density of the complete history $\mathbf{Z}$ in (2.6) needs to be adjusted accordingly when covariates are included.

2.6 Simulation Study

This section presents simulation studies of parametric estimation based on panel data. Here we focus on likelihood-based inference. Bayesian inference can be similarly done, and we will not duplicate the work here. First for Markov models, the estimates by conditional sampling and the estimates by RJMCMC sampling are compared to the MLE obtained by direct likelihood estimation. The properties of the two sampling procedures are investigated. Then, the estimation for semi-Markov models with panel data is performed using the proposed algorithms. The advantage of RJMCMC sampling over conditional sampling is overwhelming and is recommended for use in practice. All the computation is done on a desktop with 4-core i5 CPU @ 2.67 GHz.

2.6.1 Markov Models

We first evaluate the convergence performance of the proposed SA algorithm, i.e., whether $\boldsymbol{\theta}_\ell$ and Γ_ℓ converges to the MLE and the observed information matrix. For illustration purpose, we assume the panel data $\mathbf{y}_{1:N}$ are random samples with interval censoring only. We will introduce right censoring, exact failure and covariate effect in Section 2.7.

For Markov models, the likelihood-based inference can be done through numerical method as in Kalbfleisch and Lawless (1985). The estimated MLE and the observed information are used as reference. To construct panel data $\mathbf{y}_{1:N}$, we generate the complete history $\mathbf{z}_{1:N}$ first. Then, a number of observation time points $\{t_i : t_i = i \times \Delta t\}$ are picked in advance and only state information at those time points are recorded. The value of Δt controls the severity of the interval censoring. Fixed observation points t_i are mainly used here for the convenience of simulation.

Consider a 3-state multi-step progressive Markov model with the parameters of interest $\boldsymbol{\theta} = (\lambda_{12}, \lambda_{13}, \lambda_{23})$. The corresponding conditional distribution $\boldsymbol{F}(t) = (F_{12}(t), F_{13}(t), F_{23}(t))$ and the transition probability are

$$F_{12} = F_{13} = Exp(\lambda_{12} + \lambda_{13}),\ F_{23} = Exp(\lambda_{23}),\ p_{12} = \lambda_{12}/(\lambda_{12} + \lambda_{13}). \tag{2.18}$$

The interval-censored samples $\mathbf{y}_{1:100}$ are generated from (2.18) with $\boldsymbol{\theta} = (1, 1, 1)$ and $\Delta t = 0.4$.

To fulfill stochastic approximation updates, we set $m = 100$ and $\gamma_\ell = 1/(10+\ell^{2/3})$. Figure 2.3 shows two traceplots of 2000 iterates from stochastic approximation, one by conditional sampling (dot) and one by RJMCMC sampling (solid). They have the same initials. The MLEs obtained by direct likelihood estimation (grey) are also shown in the figure as reference. The convergence of the SA updates is visible. In

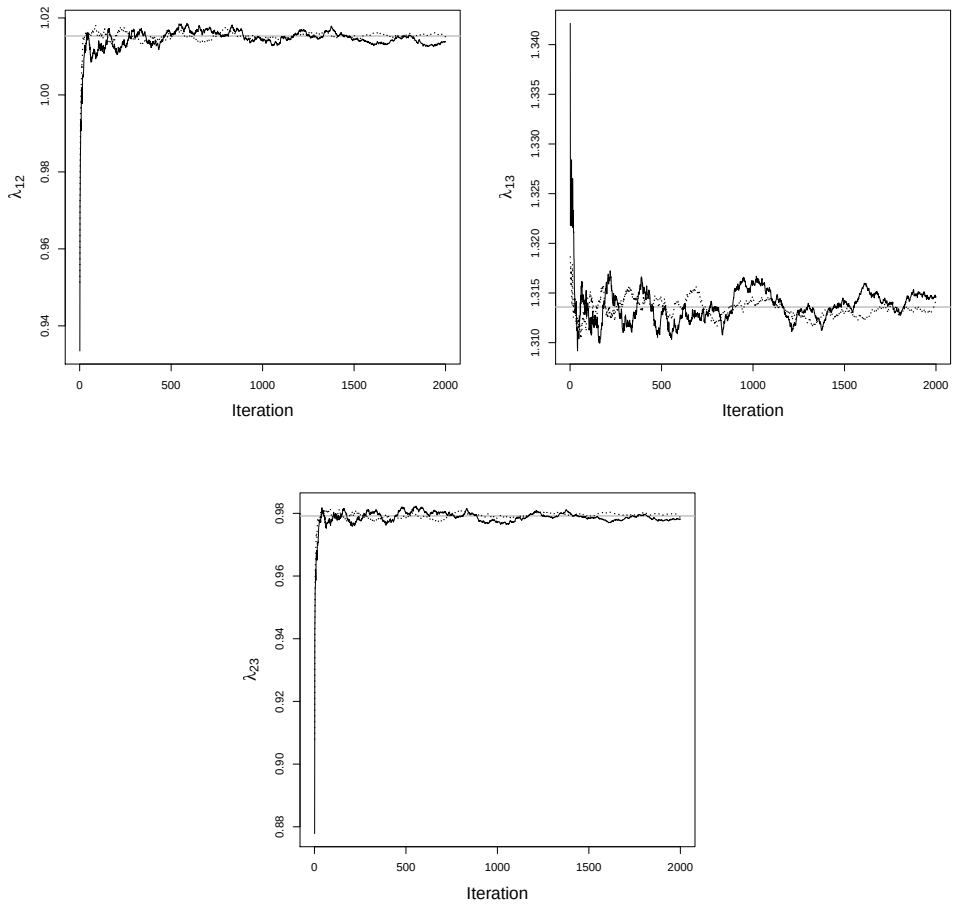


Figure 2.3: Markov Model: Iterates from Stochastic Approximation by Conditional Sampling (dots), by Reversible Jump MCMC Sampling (solid). The Maximum Likelihood Estimates (grey line) are given as reference.

the RJMCMC sampling proposal, $\pi_{12} = \pi_{21} = 0.5$ and $T_{2\to1}$ defined by (2.16) in Section 2.4.3 is used. To investigate the acceptance rate of the RJMCMC sampling, we perform data augmentation with uniform prior for $\boldsymbol{\theta}$. The average acceptance rate of the proposal is 61%, which indicates good mixing of MCMC sampling.

Table 2.1 shows the last iterate $\boldsymbol{\theta}_{2000}$ and its variance estimates from Γ_{2000} by conditional sampling and RJMCMC sampling respectively, and compared to the MLE $\hat{\boldsymbol{\theta}}$ from direct likelihood estimation. Clearly, the estimates from the proposed procedure converge to the right limit. Note that, for Markov models, conditional sampling is

exact sampling, since the weights for the mixture component $\boldsymbol{a} = (a_1, \ldots, a_d)$ shown in (2.12) have explicit forms and no numerical approximation is needed. That is why the estimates are slightly closer to the MLE than those by RJMCMC.

	MLE	Conditional	RJMCMC
λ_{12}	1.015 (0.663, 1.367)	1.015 (0.664, 1.367)	1.014 (0.666, 1.364)
λ_{13}	1.314 (0.931, 1.697)	1.314 (0.931, 1.698)	1.315 (0.933, 1.696)
λ_{23}	0.980 (0.653, 1.306)	0.979 (0.653, 1.305)	0.978 (0.655, 1.303)

Table 2.1: Parameter Estimates Along With Approximate 95% Confidence Interval

We also compare the performance of the two sampling procedures. Simulated samples are generated under three different settings as shown in Table 2.2. With the stopping criterion that the last 10 iterates satisfy $\|\boldsymbol{\theta}_\ell - \hat{\boldsymbol{\theta}}\|_1 < 0.0005$, the average iterations needed are recorded for the two sampling procedures and are shown in Table 2.2. Averaging over 500 samples, the number of iterations by RJMCMC sampling is about twice many as the ones by conditional sampling under all settings. This is consistent with the acceptance rate obtained in the data augmentation procedure. However, we will see later that the number of iterations is not the major issue that matters in the estimation. The more important problem is the computation time needed to iterate.

Setting	Conditional	RJMCMC
$N = 100$, $\Delta t = 0.4$	496	790
$N = 500$, $\Delta t = 0.4$	117	213
$N = 500$, $\Delta t = 0.2$	52	94

Table 2.2: Iterations Needed For Convergence (Averaging Over 500 Datasets)

2.6.2 Semi-Markov Models

When $Z(t)$ is assumed to be a SMP, direct likelihood estimation is not feasible. Here, we further compare the two sampling procedures. The computation time needed for update by conditional sampling deteriorates when the underlying process is assumed to be semi-Markov. $\mathbf{y}_{1:100}$ are now generated under the following 3 settings with $\Delta t = 0.3$:

(1) $F_{12} = LN(-1, 1),\ F_{13} = LN(-1, 1),\ F_{23} = LN(-1.5, 1),\ p_{12} = 0.6$

(2) $F_{12} = W(1.25, 1.2),\ F_{13} = W(1.25, 1.2),\ F_{23} = W(1, 2),\ p_{12} = 0.4$

(3) $F_{12} = W(1.25, 1.2),\ F_{13} = LN(-1, 1),\ F_{23} = \text{Exp}(2),\ p_{12} = 0.6$

Here, $LN(\mu, \sigma^2)$ is a lognormal distribution characterized by its mean μ and variance σ^2; $W(\eta, \gamma)$ denotes a Weibull distribution with shape parameter γ and scale parameter η. For conditional sampling, the weights for the mixture components $\boldsymbol{a} = (a_1, \ldots, a_d)$ needs to be approximated numerically. The integrals involved in calculating $\boldsymbol{a}$ are approximated by function *adapt* with $eps = 0.01$ in R programming system. To save computation time, only 3000 iterations are run.

Figure 2.4 shows two traceplots of estimates $\boldsymbol{\theta}$ by stochastic approximation based on sample generated by Model (1). The dot lines are updated by conditional sampling, and the solid ones are updated by RJMCMC sampling. The convergence behavior is visible in the figure. Besides, the variation of the iterates from two sampling approaches is comparable. In Table 2.3, the computation times of 3000 iterations by conditional sampling and RJMCMC sampling are given for the three models. The advantage of using RJMCMC sampling is obvious. Stochastic approximation by conditional sampling takes much longer time to iterate than that by RJMCMC sampling (at least 15 times longer).

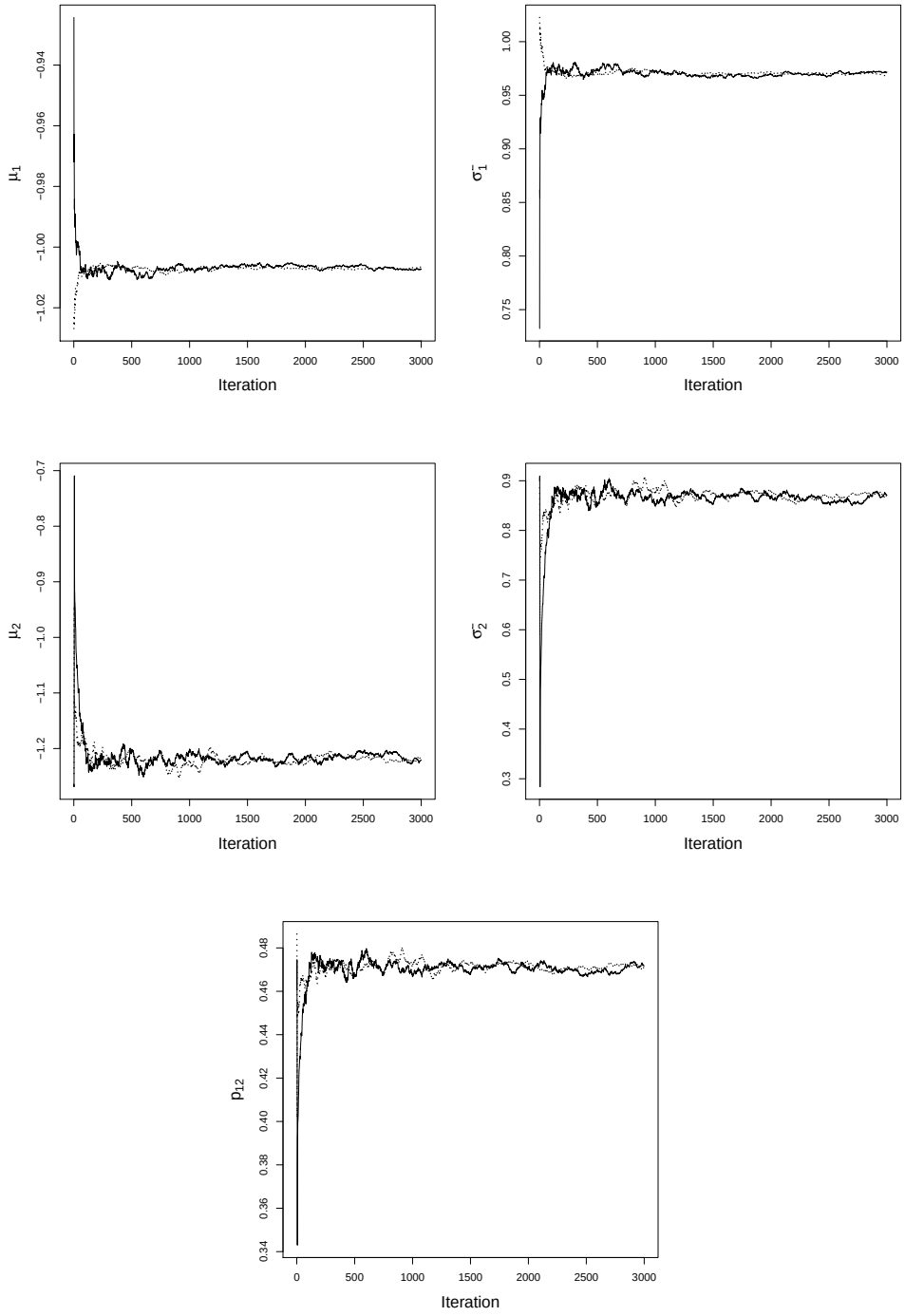


Figure 2.4: Semi-Markov Model with LogNormal Sojourn Times: Iterates from Stochastic Approximation by Conditional Sampling (dot), by Reversible Jump MCMC Sampling (solid)

	Conditional	RJMCMC
Model (1)	152	10
Model (2)	302	11
Model (3)	203	11

Table 2.3: Computation Time of Running 3000 Iterations (Rounded by Minute)

Besides, the computation time needed by RJMCMC sampling is about the same no matter what distribution $F_{ij}(\cdot)$ is. When conditional sampling is used, $F_{ij}(\cdot)$ does matter (determines how hard it is to approximate a_j, $j = 1, \ldots, d$). For instance, Model (2) takes almost twice amount of time to iterate as Model (1).

2.7 Coronary Allograft Vasculopathy Example

In this section, we use our proposed alogithm to analyze a real data set. This example data set is provided in R package *msm*, called *cav*. Under Markov assumption, Sharples (2003) used this data set to study the progression of coronary allograft vasculopathy (CAV), a post-transplant deterioration of the arterial walls.

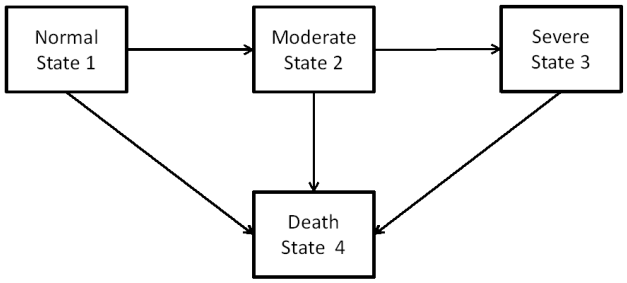


Figure 2.5: multistate Model Constructed to Analyze the Progression of Coronary Allograft Vasculopathy Disease (CAV)

Between August 1979 and January 2002, 622 patients underwent angiography annually or biennially. At each checkup, the patients were classified into 4 states according to their angiography results: (1) normal coronary angiography, (2) $0-70\%$

senosis, (3) $>$ 70% stenosis, and (4) death (absorbing state). All patients were assumed to be CAV free at transplantation (origin of the study), then followed up until death or until their most recent coronary angiography if alive at the time of data analysis. The death of the patient was recorded with accuracy, and right censoring is present for some patients. Refer to Appendix for details regarding sampling for panel data with right censoring and exact failures. The structure of the underlying model is given in Figure 2.5 with the following 2 assumptions:

1. CAV is not spontaneously reversible. For instance, any patient with a pattern of measurements of the form 1-2-1 must have been misreported. In the data set, 58 of 622 patients had reversed senosis status. For illustrative purpose, we remove them in our analysis.

2. If any patient has an observed path 1-3, he/she must have gone through state 2 somewhere in between. (This must be true based on the definition of the states.)

Path	1-2-3-4	1-3-4	1-2-4	1-4	1	1-2	1-2-3	1-3
Count	32	17	34	95	239	55	8	13

Table 2.4: The Counts of Observed Sample Paths in CAV data

Among the rest 564 patients, We pick 491 patients, whose primary diagnostic information are available (for covariate analysis later). The 491 patients have one of the 8 paths shown in Table 2.4. Based on our proposed stochastic approximation algorithm, a semi-Markov model can be estimated based on this panel data set. Here we assume Weibull sojourn times, i.e. $F_{ij}(t) = W(\gamma_{ij}, \eta_{ij})$. Denote $\hat{\boldsymbol{\theta}}_{SM}$ and $\hat{\Gamma}_{SM}$ the estimated MLE and the observed information evaluated at the MLE obtained by SA. For comparison purpose, a Markov model is also fitted and is actually a nested

model of SMP with $\gamma_{ij} = 1$ and $\eta_{ij} = \eta_i$. Under Markov assumption, denote its MLE as $\hat{\boldsymbol{\theta}}_M$ and the estimated information matrix as $\hat{\Gamma}_M$. We approximate numerically the log-likelihood at $\hat{\boldsymbol{\theta}}_M$, i.e. l_M, the log-likelihood at $\hat{\boldsymbol{\theta}}_{SM}$, i.e. l_{SM} and the corresponding deviance, i.e. $2(l_{SM} - l_M)$.

	Markov	semi-Markov	Deviance	df	p-value
No Covariate	-1248	-1221	54	7	0
With Covariate	-1239	-1207	64	7	0

Table 2.5: The Log-likelihood Evaluated at the Maximum Likelihood Estimates Under Markov and Semi-Markov Assumptions, Fitted with Covariate or No Covariate, Along with the Deviance

In Table 2.5, the results in "No Covariate" row are obtained by assuming $y_{1:491}$ are a random sample; while the results in "With Covariate" row are obtained by introducing covariate effect. The included covariate is IHD primary diagnosis. IHD is the most common cause of death in most Western countries, and a major cause of hospital admissions. This covariate is binary, i.e., IHD = ischaemic heart disease (coded as 1), IDC = idiopathic dilated cardiomyopathy (coded as 0). Here we assume proportional hazard model as (2.17). Under these two settings, the p-value for the corresponding likelihood ratio tests are almost 0. This indicates semi-Markov model fits the CAV data set significantly better than Markov model. In addition to test Markov assumption, we can also perform likelihood ratio test to study whether including the covariate improves the fit or not. Under Markov and semi-Markov assumption, the p-value for testing covariate effect (not shown in table) are correspondingly 0.003 and 0.0003 ($df = 5$). Judging by the deviances, however, the improvement in the fit is larger by assuming a SMP than including the binary covariate into Markov model.

Now that it has been shown that semi-Markov model fits this data set better,

Table 2.6: Approximates of Maximum Likelihood Estimates and their 95% Confidence Interval from Stochastic Approximation by Reversible Jump MCMC Sampling Under Semi-Markov Assumption

(a) Without Covariate

	shape γ_{ij}	scale η_{ij}	probability p_{ij}
$1 \to 2$	1.28 (1.10, 1.47)	7.43 (5.95, 8.90)	0.67 (0.40, 0.932)
$1 \to 4$	1.90 (1.52, 2.28)	9.67 (7.35, 11.9)	
$2 \to 3$	1.21 (0.92, 1.50)	2.10 (1.57, 2.62)	0.76 (0.11, 1.00)
$2 \to 4$	2.05 (0.19, 3.91)	4.47 (2.42, 6.53)	
$3 \to 4$	1.03 (0.69, 1.37)	3.72 (2.48, 4.96)	

(b) With Covariate

	shape γ_{ij}	scale η_{ij}	covariate β_{ij}	probability p_{ij}
$1 \to 2$	1.32 (1.14, 1.50)	9.80 (7.56, 12.0)	0.69 (0.34, 1.04)	0.68 (0.43, 0.94)
$1 \to 4$	1.97 (1.55, 2.39)	9.40 (6.44, 12.3)	-0.11 (-0.80, 0.57)	
$2 \to 3$	1.30 (0.98, 1.61)	1.81 (1.09, 2.52)	-0.40 (-1.00, 0.19)	0.78 (0.36, 1.00)
$2 \to 4$	5.44 (1.83, 9.05)	5.84 (4.85, 6.83)	7.58 (2.13, 13.0)	
$3 \to 4$	1.06 (0.83, 1.28)	2.94 (1.96, 3.92)	-0.41 (-0.86, 0.03)	

the MLEs obtained under SMP setting are studied. In Table 2.6(a), the MLEs and their corresponding 95% confidence intervals (by normal approximation) are shown. The shape column gives the estimates of the shape parameters associated with each transition. Clearly, transition $1 \to 2$ and transition $1 \to 4$ both have their shape estimates significantly larger than 1, which indicates that Markov assumption is not adequate. The scale column show the estimates of the scale parameters. The larger the scale estimate is, the bigger the variance of the transition is. Among all, transition $1 \to 4$ has the largest scale estimate and the widest CI. This is due to the fact that many right-censored (at state 1) observations are included in the data. The estimates of the transition probabilities are given as well. The variance associated with those estimates are large though. For comparison purpose, the estimates of the semi-Markov models including one binary covariate are also given in Table 2.6(b).

The estimated shape and scale parameters are consistent to those in Table 2.6(a); however, the estimates associated with transitions starting from state 2 changes a bit, especially $\hat{\gamma}_{24}$ now significantly bigger than 1. Besides, the scale and covariate effect estimates of transition $2 \to 4$ also have large values and large variance. This may be due to the fact that only 34 samples present with path 2-4, and among these 34 samples, it is also possible that state 3 may be visited. Other than the transition $2 \to 4$, we find the covariate effect associated with transition $1 \to 2$ is significantly larger than zero. This indicates that the patient with IHD diagnosed before the transplant has a high risk of advancing to the moderate state (state 2) than those who have not.

With the estimated parameters in Table 2.6(a), we can estimate a_1 and a_2 in (2.14) for the 34 observations with observed path 1-2-4. Let $\delta = 1$ indicate $\boldsymbol{i} = (1, 2, 3, 4)$, and $\delta = 2$ indicate $\boldsymbol{i} = (1, 2, 4)$. Denote t'_1 and t'_3 the observation time points when the patient was last seen in state 1 and state 2. Similarly, let t'_2 and t'_4 be the observation time points when the patient was first seen in state 2 and state 4. Here t'_1, t'_2, t'_3, t'_4 are not nessarily consecutive observation points. Three observations are shown in Table 2.7. Their estimated probabilities of going through state 3 are given as well, i.e. $\hat{a}_1$. If $t'_4 - t'_3$ is large, the probability of going through state 3 increases! The probability of the third observation is the highest in the data set.

Obs	t'_1	t'_2	t'_3	t'_4	$\hat{a}_1$
1	5.99	7.08	10.42	10.67	0.08
2	0	2.00	2.00	2.54	0.50
3	0	2.02	2.02	14.14	0.99

Table 2.7: The Conditional Probability of Going Through State 3 of the Censored Observation with Observed Path 1-2-4

2.8 Conclusions

Semi-Markov models provide a more general structure for modeling and analyzing multistate data than Markov models. However, in the presence of censoring, the inference problem becomes challenging. This chapter has developed a general methodology for parametric estimation of progressive semi-Markov multistate processes with interval censored observations, possibly coupled with exact failure and right censoring. The methods can be used to make likelihood-based inference and Bayesian inference. From simulation stuyies, we find sampling complete history $\mathbf{Z}$ efficiently is essential for the estimation, and recommend using RJMCMC sampling for reliable estimation and shorter computation time. By the proposed algorithm, we fit a semi-Markov model on CAV data, and it fits much better than a Markov model. The methods proposed here are computationally intensive and work for moderate number of states.

CHAPTER III

Inference for Time-to-Failure in Multistate Semi-Markov Models: A Comparison of Marginal and Process Approaches

The traditional approach to reliability and survival analysis is based on collecting and analyzing time-to-failure (TTF) data. In many situations, however, failure is the end point of an underlying multistate process: a system (subject, equipment, etc.) moves among different "states" before reaching an absorbing state. There are many examples of such multistate processes in medical, engineering, finance, and social science applications (see, for example, Kay 1986, Andersen 1988, Andersen *et al.* 1992, Andersen, Esbjerg and Sørensen 2000, Aalen 1995, Commenges 1999, Limnios and Oprisan 2001, D'Amico, Janssen and Manca 2005, Foucher *et al.* 2007, Kang and Lagakos 2007). Development of statistical inference methods for these models has also received considerable attention (in addition to the above references, see Lagakos, Sommer and Zelen 1978, Kalbfleisch and Lawless 1985, Andersen *et al.* 1992, Hougaard 1999, Sternberg and Satten 1999, Aalen and Gjessing 2001, Janssen and Manca 2006, Liquet and Commenges 2010, Titman and Sharples 2010).

This chapter deals with a specific inference problem associated with multistate models. Let $\{Z(t),\ t \geq 0\}$ denote the multistate process, taking values in a finite

state space $E = \{1, 2, \ldots, p\}$. Throughout, state p is assumed to be an absorbing or failure state (the end point of the process). We will use the terms "absorption" and "failure" interchangeably in this chapter. The data are obtained by recording $Z(t)$, the state of the system. If the recording is continuous until time-to-absorption, we will know the states of the system at all times and the transition times, including when it moved to the absorbing state, i.e., the exact TTF. This is rarely done in practice. It is more common to observe the system periodically and record the state it is in. This results in interval censoring – we know only the states of the system at the observation times but not the when the transitions occurred. In addition, we may not know the number of transitions that occurred during the interval. Further, if we stop observing the system before the TTF, the data are right censored. The resulting data are often called panel data (see Kalbfleisch and Lawless 1985). It is quite difficult to make inferences based on panel data, and we will discuss the challenges shortly.

The goal here is to make inference about the system's TTF: estimation of the TTF distribution and prediction intervals for the TTF given that the system has not failed by time t. We compare two different approaches to this inference problem. One is based on just the TTF data of N systems, subject to different forms of censoring. We will refer to it as the *marginal* method as it is based on just the TTF data and ignores additional information about the various states that the system has gone through. The other approach, which we call the *process* approach, involves modeling the entire multistate process of the system, estimating the underlying parameters and using this to make inference. For prediction, the process approach will use additional information on which state the system is in at the time of right censoring and how long it has been in that state. We restrict attention to progressive semi-Markov

processes (defined formally in Section 3.1). The problem studied in this chapter has also been discussed, albeit not as formally, in the literature (see, for example, Andersen 1988, Liquet and Commenges 2010).

It may seem intuitive that the process-based method should be more efficient, especially for the prediction problem. The goal of this chapter is to examine and quantify these advantages in selected parametric models. This is an important problem for the following reasons. Parametric inference for multistate processes is challenging in the presence of censored data. Several authors have developed methods in special cases (see, for example, Lagakos, Sommer and Zelen 1978, Sternberg and Satten 1999, Foucher *et al.* 2007). More recently, Titman and Sharples (2010) developed results for general panel data with phase-type sojourn distributions. Chapter II developed computationally-intensive methods for general parametric sojourn distributions. The general techniques in Chapter II do not scale up to large number of states. On the other hand, there are well-developed methods in the literature for doing marginal inference based on TTF data. So, if the efficiency loss in doing the marginal inference is small, one can bypass the process approach if one is interested in just inference related to the TTF.

However, the results in this chapter show that inference based on the process method is considerably more efficient, with estimation efficiencies being 2-3 times more than the marginal method even with relatively small number of states. The gain in prediction efficiency is also considerable. The latter is especially important if we are interested in predicting failure at the individual system level, which would be the case with expensive or highly-critical systems.

The rest of the chapter is organized as follows. Section 3.1 provides the background and problem formulation. Section 3.2 deals with estimation efficiency for

parameters such as the mean, variation and quantiles of the TTF distribution. The analysis for estimation efficiency will be done under the following data collection schemes:

- Complete: sample paths/TTFs continuously observed until absorption;
- Right-censored: sample paths/TTFs continuously observed until some censoring time point;
- Interval-censored: sample paths/TTFs observed periodically at discrete time points.

Section 3.3 compares the prediction efficiency of the two methods in terms of the width and coverage probabilities of the prediction interval for the TTF of a system given that it has survived up to some time t (therefore right-censored at t).

For reasons that will become apparent, we restrict attention for the most part to the gamma sojourn times for the parametric model, although the inverse Gaussian and normal are briefly discussed as well.

3.1 Formulation and Background

We consider the time-homogeneous semi-Markov multistate models (SMP) $Z(t)$ as defined in Section 2.3. The process $Z(t)$ is fully determined by the transition matrix P as (2.1) and the conditional distributions $F_{ij}(t)$, $i, j \in E$ defined in (2.2) or the semi-Markov kernels $Q_{ij}(t)$ as (2.3). Assume $F_{ij}(x) = F_0(x; \boldsymbol{\phi})$ where F_0 is some known baseline distribution and $\boldsymbol{\phi}$ denotes the unknown parameters. Thus, the unknown parameters consist of $\boldsymbol{\theta} = (\boldsymbol{\phi}, P)$. For notation convenience later, we denote the random transition time occurred from state i to state j as X_{ij}. Therefore,

X_{ij} has the cumulative distribution function $F_{ij}(\cdot)$.

We restrict attention to progressive models in this chapter. Specifically, we assume that the states are arranged in some naturally increasing order and that the system can move only from left to right, i.e., no cycling back to previous states. Such models have been considered by many authors for different applications in the literature. This is a reasonable assumption for characterizing situations with monotone degradation. In this case, only the upper-diagonal elements of transition matrix can be nonzero. A special case that we will consider in some detail is the one-step progressive model where one can move from state i to state $(i+1)$ only, i.e., $p_{i,i+1} = 1$, and hence the transition matrix P is known.

3.1.1 Time-to-Failure Distribution

The distribution of T, the TTF of the multistate process, is introduced here. We can write T as a function of the transitions and X_{ij}'s as follows.

For the one-step progressive case,

$$T \;=\; X_{12} + X_{23} + \cdots + X_{p-1,p}. \tag{3.1}$$

The marginal distribution of T is given by the convolution of the distributions of the individual X_{ij}'s. Therefore, the mean and the variance of T are given by the sums of the means and variances of the individual components. Common distributions, such as Weibull and lognormal, for X_{ij}'s are not closed under convolution, so the distribution of T in these cases have complex forms. In order to do some analytical calculations, we will focus for the most part on gamma distributions with the same scale parameter. The special case of the inverse Gaussian with this convolution property will also be briefly considered.

For the multi-step progressive case,

$$T \;=\; \sum_j I_{i_0,i_1,\ldots,i_k}\left(X^{(j)}_{i_0,i_1} + X^{(j)}_{i_1,i_2} + \cdots + X^{(j)}_{i_{k-1},i_k}\right), \tag{3.2}$$

where $i_0 = 1$, $i_k = p$ and the sum j is over all possible paths $\{i_0 = 1 < i_1 < \ldots < i_k = p\}$ from the initial state 1 to the absorbing state p. Note that there are $J = 2^{p-2}$ possible such paths. Further, the choice of the path depends on the transition probabilities of $Z(t)$.

As an example, consider the case with 4 states. There are four possible paths: $\{1,2,3,4\}$, $\{1,2,4\}$, $\{1,3,4\}$, and $\{1,4\}$, so

$$T \;=\; I_{\{1,2,3,4\}}(X^{(1)}_{12}+X^{(1)}_{23}+X^{(1)}_{34})+I_{\{1,2,4\}}(X^{(2)}_{12}+X^{(2)}_{24})+I_{\{1,3,4\}}(X^{(3)}_{13}+X^{(3)}_{34})+I_{\{14\}}X^{(4)}_{14}. \tag{3.3}$$

Note that $X^{(1)}_{12}$ and $X^{(2)}_{12}$ in the first and second components of the sum in equation (3.3) are independent copies of the random variable X_{12}. The reason is that a given system can take only one of the 4 possible paths, and different systems are independent of each other.

In this multi-step case, the distribution of the TTF is a finite mixture distribution with $J = 2^{p-2}$ elements. We can write the density explicitly as

$$f_T(t) \;=\; \sum_j p_{i_0,i_1,\ldots,i_k} f_{i_0,i_1} * f_{i_1,i_2} * \cdots * f_{i_{k-1},i_k}(t), \tag{3.4}$$

where, as before, the sum j is over all possible paths $\{i_0 = 1 < i_2 < \ldots < i_k = p\}$. Here, $p_{i_0,i_1,\ldots,i_k} := \mathbb{E}(I_{i_0,i_1,\ldots,i_k}) = p_{i_0,i_1}p_{i_1,i_2}\cdots p_{i_{k-1},i_k}$, $f_{i_s,i_{s+1}}(t)$ is the density of $X_{i_s,i_{s+1}}$ and $*$ denotes the convolution operator.

From standard properties of mixture distributions (see Mclachlan and Peel (2000)), we get the mean of the TTF distribution as

$$\mu_T \;:=\; \mathbb{E}(T) \;=\; \sum_j p_{i_0,i_1,\ldots,i_k}\left(\mu_{i_0,i_1} + \mu_{i_1,i_2} + \cdots + \mu_{i_{k-1},i_k}\right), \tag{3.5}$$

where $\mu_{i_s,i_{s+1}}$ is the mean of $X_{i_s,i_{s+1}}$. From the independence of the sojourn distributions under the semi-Markov property, we get the variance of T as

$$\begin{aligned}\sigma_T^2 \;:=\; \mathrm{Var}(T) \;&=\; \sum_j p_{i_0,i_1,\ldots,i_k}\left(\sigma^2_{i_0,i_1}+\sigma^2_{i_1,i_2}+\cdots+\sigma^2_{i_{k-1},i_k}\right)\\ &+\; \sum_j p_{i_0,i_1,\ldots,i_k}\left(\mu^2_{i_0,i_1}+\mu^2_{i_1,i_2}+\cdots+\mu^2_{i_{k-1},i_k}\right)-\mu^2,\end{aligned}$$

where $\sigma^2_{i_s,i_{s+1}}$ is the variance of $X_{i_s,i_{s+1}}$.

3.1.2 Challenges with Process-based Inference with Censored Data

We briefly describe the challenges in inference for multistate processes based on interval-censored or panel data. Recall that this is one of the reasons for considering the marginal approach.

Consider a multistate process with $p = 5$ states. Suppose the process is observed at times $t_i,\ i = 1, 2, 3$, and the states of the system at these times are $y_i,\ i = 1, 2, 3$ with $y_1 = 2,\ y_2 = 4,\ y_3 = 5$. (As before, we assume the system starts at time 0 in state 1.) So, we know that the system transitioned from state 1 state 2 at some time during the interval $(0, t_1]$ but we do not know when. Further, we know that the system moved from state 2 to state 4 during the interval $(t_1, t_2]$ but we do not know if it moved first to state 3 and then state 4 or straight from state 2 to state 4. In addition, the exact times of the transitions are unknown.

For exponential distributions, this type of incompleteness does not pose a major problem. The likelihood can be written down and maximized using numerical methods (see Kalbfleisch and Lawless 1985). For semi-Markov models with non-exponential distributions, however, the inference problem becomes quite challenging. Chapter II develops a computationally intensive approach for doing parametric inference in such cases. However, the computational burden involved is quite high for

situations with a large number of states. See also Titman and Sharples (2010) for inference with phase-type sojourn distributions.

3.2 Comparison in Estimation Efficiency

In this section, we will compare the estimation efficiency of the process estimator and the marginal estimator. As we have mentioned before, most TTF distributions are not closed under convolution. In order to do some analytical calculations, we discuss the illustrative normal case, the gamma case and the inverse Gaussian case.

3.2.1 Illustrative Case: Normal Distributions

We start with the normal distributions as an illustrative case; they are not commonly used to model TTF data which are non-negative random variables. Consider just the one-step progressive situation with N complete observations. Therefore, the process data X_{ij}^n and the absorbing time T_n are observed exactly for $n = 1, \ldots, N$.

Let $X_{j,j+1} \sim N(\mu_j, \sigma_j^2)$, $j = 1, \ldots, p-1$. Assume that $\mu_j > 0$ and μ_j/σ_j is sufficiently large so that the probability of a negative value is very small. In such a case, the normal distribution can be used to model (non-negative) TTF data. Note that $T = \sum_{j=1}^{p-1} X_{j,j+1} \sim N(\mu_T, \sigma_T^2)$, where $\mu_T = \sum_{j=1}^{p-1} \mu_j$ and $\sigma_T^2 = \sum_{j=1}^{p-1} \sigma_j^2$.

The loglikelihood function based on N event history data is:

$$l_N(P) = -\frac{N(p-1)}{2}\log(2\pi) - \sum_{j=1}^{p-1}\frac{N}{2}\log\sigma_j^2 - \sum_{j=1}^{p-1}\frac{1}{2\sigma_j^2}\sum_{n=1}^{N}(X_{j,j+1}^n - \mu_j)^2 \qquad (3.6)$$

Maximum likelihood estimate satisfies $\forall j = 1, \ldots, p-1$,

$$\frac{\partial l_N(P)}{\partial \mu_j} = -\frac{1}{\sigma_j^2}\sum_{n=1}^{N}(\mu_j - X_{j,j+1}^n) = 0, \tag{3.7}$$

$$\frac{\partial l_N(P)}{\partial \sigma_j^2} = -\frac{n}{2\sigma_j^2} + \frac{1}{2\sigma_j^2}\sum_{i=1}^{n}(X_{j,j+1}^n - \mu_j)^2 = 0. \tag{3.8}$$

The inverse of Fisher information matrix for the parameters is

$$I_N^{-1}(P) = \frac{1}{N}\,\mathrm{diag}\{\sigma_1^2, \cdots, \sigma_{p-1}^2, 2\sigma_1^4, \cdots, 2\sigma_{p-1}^4\}. \tag{3.9}$$

If we instead model $T = \sum_{j=1}^{p-1} X_{j,j+1}$ directly, the maximum likelihood estimator then satisfies

$$\frac{\partial l_N(M)}{\partial \mu_T} = -\frac{1}{\sigma_T^2}\sum_{n=1}^{N}(\mu_T - T_n) = 0 \tag{3.10}$$

$$\frac{\partial l_N(M)}{\partial \sigma_T^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma_T^2}\sum_{n=1}^{N}(T_n - \mu_T)^2 = 0. \tag{3.11}$$

The inverse of Fisher information for (μ, σ^2) is

$$I_N^{-1}(M) = \frac{1}{N}\,\mathrm{diag}\{\sigma_T^2, 2\sigma_T^4\}. \tag{3.12}$$

For μ_T, the maximum likelihood estimator (MLE) based on process data is $\hat{\mu}_T(P) = \sum_{j=1}^{p-1}\bar{X}_{j,j+1}$ where $\bar{X}_{j,j+1} = \sum_{n=1}^{N} X_{j,j+1}^n/N$. This is easily seen to be identical to the MLE based on marginal TTF data $\hat{\mu}_T(M) = \sum_{n=1}^{N} T_n/N$. So there is no efficiency gain in using the process data.

Consider next the estimation of the variance parameter σ_T^2. The usual unbiased estimator of $\hat{\sigma}_j^2(P) = \sum_{n=1}^{N}\left(X_{j,j+1}^n - \bar{X}_{j,j+1}\right)^2/(N-1) \sim \sigma_j^2\chi^2(N-1)/(N-1)$, where $\chi^2(N-1)$ denotes the chi-square distribution with $N-1$ degree of freedom. From these, we get the process estimator $\hat{\sigma}_T^2(P)$ as the sum of the individual estimators for each state. On the other hand, the marginal estimator is $\hat{\sigma}_T^2(M) = \sum_{n=1}^{N}\left(T_n - \bar{T}\right)^2/(N-1) \sim \sigma_T^2\chi^2(N-1)/(N-1)$.

The relative efficiency (RE) of the marginal estimator to the process estimator, given by the ratio of the variance of the process estimator over the marginal estimator, is $\sum_{j=1}^{p-1} \sigma_j^4/\sigma_T^4$. This is always less than one and reduces to $1/(p-1)$ when the σ_j^2's are all the same. The equal σ_j^2's is the worst-case scenario. For example, if $(\sigma_1^2, \ldots, \sigma_4^2) = (1, 2, 3, 4)$, the RE is 0.3 compared to 0.25 in the equal σ_j^2's case. If one of the σ_j^2's is much larger than all the others, the RE will be close to one. In general, the marginal estimator can be considerably less efficient than the process estimator, especially when there is a relatively large number of states involved.

The quantiles of the distribution are of more interest in practice. The $u-$th normal quantile of the TTF distribution can be expressed as $t_u = \mu_T + \sigma_T z_u$ where z_u is the $u-$th quantile of standard normal. We can get the asymptotic variances (aVar) of $\hat{\sigma}_T$ using Taylor series as: $\text{aVar}[\hat{\sigma}_T(M)] = \dfrac{\sigma_T^2}{2N}$ and $\text{aVar}[\hat{\sigma}_T(P)] = \dfrac{\sum_{j=1}^{p-1} \sigma_j^4/\sigma_T^2}{2N}$. From this, we see that the asymptotic relative efficiency (ARE) of the marginal estimator to the process estimator of t_u is

$$\text{ARE}(t_u) \;=\; \frac{1 + z_u^2/2 \sum_{j=1}^{K} \sigma_j^4/(\sigma_T^2)^2}{1 + z_u^2/2}. \tag{3.13}$$

As $|z_u|$ ranges between 0 (the median/mean) and ∞ (extreme quantiles), this ARE varies from that of the median/mean estimator to that of the standard deviation estimator (which is the same as the variance estimator). So the ARE for the quantiles varies between 1 and $\sum_{j=1}^{p-1} \sigma_j^4/(\sigma_T^2)^2$. For extreme upper and lower quantiles, the marginal estimator can suffer from the same type of severe inefficiency as the variance estimator.

3.2.2 Gamma Distributions

Now suppose that the duration times X_{ij} have gamma distributions, which is a natural family used for TTF data. Specifically, let $F_{ij} = \mathcal{G}(\kappa_{ij}, \theta)$, $\forall i, j$, i.e., gamma distribution with shape parameter κ_{ij} and common scale parameter θ. Under this setting, the gamma family is closed under convolution.

One-Step Progressive Case

The TTF under a one-step progressive model is also from the gamma family. Here, assume $X_{j,j+1} \sim \mathcal{G}(\kappa_j, \theta)$. Then, $T \sim \mathcal{G}(\sum_{j=1}^{p-1} \kappa_j, \theta)$. Let $\kappa_T = \sum_{j=1}^{p-1} \kappa_j$ and $\psi(x)$ be the logarithm of gamma function.

Consider estimation with complete observations (no censoring case) first. For $n = 1, \ldots, N$, we observe $X_{j,j+1}^n, j = 1, \ldots, p-1$ and the TTF's T_n. The log-likelihood based on N random samples is

$$l_N(P) = \sum_{j=1}^{p-1}(\kappa_j - 1)\sum_{n=1}^{N}\log(X_{j,j+1}^n) - \frac{1}{\theta}\sum_{n=1}^{N} T_n - N\kappa_T \log\theta - N\sum_{j=1}^{p-1}\psi(\kappa_j),$$

where $\kappa_T = \sum_{j=1}^{p-1} \kappa_j$ and $\psi(x)$ is the logorithm of gamma function. Solving score functions and MLE satisfies

$$\frac{\partial l_N(P)}{\partial \kappa_j} = \sum_{n=1}^{N}\log X_{j,j+1}^n - N\log\theta - N\psi'(\kappa_j) = 0, \quad j = 1, \ldots, p-1 \tag{3.14}$$

$$\frac{\partial l_N(P)}{\partial \theta} = \frac{1}{\theta^2}\sum_{n=1}^{N} T_n - \frac{N\kappa_T}{\theta} = 0, \tag{3.15}$$

The corresponding Fisher information matrix is

$$I_N(P) = N\begin{pmatrix} \psi''(\kappa_1) & 0 & \cdots & 0 & 1/\theta \\ 0 & \psi''(\kappa_2) & \cdots & 0 & 1/\theta \\ & \cdots & \cdots & \cdots & \\ 0 & 0 & \cdots & \psi''(\kappa_{p-1}) & 1/\theta \\ 1/\theta & 1/\theta & \cdots & 1/\theta & \kappa_T/\theta^2 \end{pmatrix}. \tag{3.16}$$

This can be inverted to get the asymptotic covariance matrix of MLE.

From this, we get the asymptotic variances of κ_T and θ as

$$\begin{aligned} \text{aVar}_P(\kappa_T) &= \frac{\kappa_T\Delta_P - \Delta_P^2 + \Delta}{N(\kappa_T - \Delta_P)} = \frac{\kappa_T\Delta_P}{N(\kappa_T - \Delta_P)} - \frac{\Delta_P^2 - \Delta}{N(\kappa_T - \Delta_P)}, \quad (3.17) \\ \text{aVar}_P(\theta) &= \frac{\theta^2}{N(\kappa_T - \Delta_P)}, \quad (3.18) \end{aligned}$$

where $\Delta_P = \sum_{j=1}^{p-1} 1/\psi''(\kappa_j)$ and $\Delta = \sum_{j=1}^{p-1} 1/\psi''(\kappa_j)^2$.

For marginal data, the log-likelihood function for κ_T and θ is

$$l_N(M) = (\kappa_T - 1)\sum_{n=1}^{N}\log(T_n) - \frac{N}{\theta}\sum_{n=1}^{N} T_n - N\kappa_T \log\theta - N\psi(\kappa_T). \quad (3.19)$$

The Fisher information matrix is

$$I_N(M) = N\begin{pmatrix} \psi''(\kappa_T) & 1/\theta \\ 1/\theta & \kappa_T/\theta^2 \end{pmatrix}. \quad (3.20)$$

From this, we get the asymptotic covariance matrix of the marginal MLE as

$$\frac{1}{N(\kappa_T - \Delta_M)}\begin{pmatrix} \kappa_T\Delta_M & -\theta\Delta_M \\ -\theta\Delta_M & \theta^2 \end{pmatrix}, \quad (3.21)$$

where $\Delta_M = 1/\psi''(\kappa_T)$.

Therefore, the ARE's, given by the ratio of the asymptotic variance of the process estimator to that of the marginal estimator, are:

$$\begin{aligned} \text{ARE}(\kappa_T) &= \frac{\kappa_T - \Delta_M}{\kappa_T - \Delta_P}\cdot\left(\frac{\Delta_P}{\Delta_M} - \frac{\Delta_P^2 - \Delta}{\kappa_T\Delta_M}\right), \quad (3.22) \\ \text{ARE}(\theta) &= \frac{\kappa_T - \Delta_M}{\kappa_T - \Delta_P}. \quad (3.23) \end{aligned}$$

Note that these ARE's do not depend on the scale parameter θ.

The expression in (3.22) is less than $\dfrac{\kappa_T - \Delta_M}{\kappa_T - \Delta_P}\cdot\dfrac{\Delta_P}{\Delta_M}$ since $\Delta_P^2 > \Delta$ for any $\kappa_j > 0, j = 1, \ldots, p-1$. In addition, it can be shown that $\Delta_P < \Delta_M$. Thus, both ARE's are smaller than 1, so the marginal estimators of κ_T and θ are always less

efficient than the corresponding process estimators. This loss is largest when the κ_i's are small (close to zero). When all the κ_i's are large, Δ_P/Δ_M is close to one.

Since parameters of the models are themselves not of direct interest in actual applications, we consider estimation efficiency of the mean, standard deviation and quantiles of the TTF distribution. We first consider the asymptotic relative efficiencies and then examine finite-sample relative efficiencies.

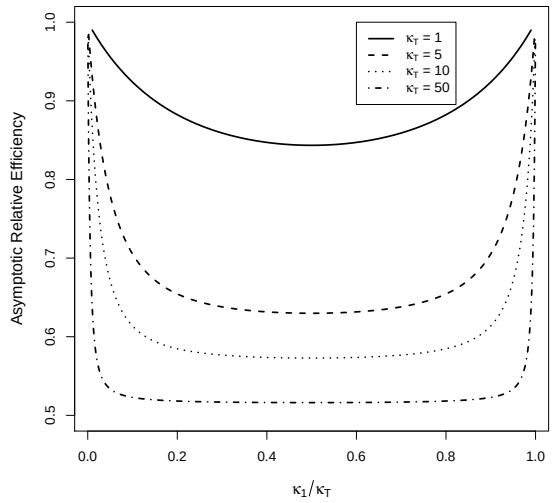
Recall that the mean of T is $\mu_T = \theta\kappa_T$ in this gamma case. It can be shown that $\hat{\mu}_T(P) = \hat{\mu}_T(M)$, i.e., the MLEs of the mean using the process data is the same as that from the marginal data. Thus, for the mean, the two approaches are equally efficient, even though both $\hat{\theta}(P)$ and $\hat{\kappa}_T(P)$ are more efficient than $\hat{\theta}(M)$. Therefore, there is no gain in using process data, just as in the normal case.

The standard deviation of T is $\sigma_T = \theta\sqrt{\kappa_T}$. With some algebra, the ARE(σ_T) can be computed as

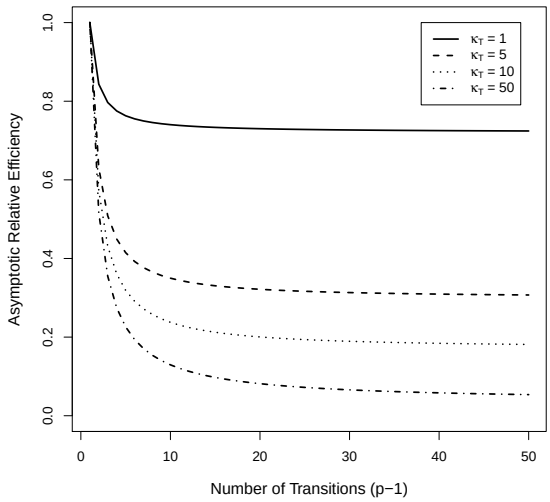
$$\text{ARE}(\sigma_T) = \frac{\kappa_T - \Delta_M}{\kappa_T - \Delta_P} \cdot \frac{\Delta_P + 4(\kappa_T - \Delta_P)}{\Delta_M + 4(\kappa_T - \Delta_M)}. \tag{3.24}$$

Figure 3.1 shows the ARE(σ_T) under different settings. The left panel shows the comparisons for a 3-state model. The horizontal-axis corresponds to κ_1/κ_T with $\kappa_T = \kappa_1 + \kappa_2$. We see that as κ_1/κ_T goes from zero to 1/2, ARE(σ_T) decreases from 1 to some lower bound that depends on κ_T. It decreases with the magnitude of κ_T. The right panel studies the effect of increasing the number of states but keeping the κ_j's equal. As expected, the ARE(σ_T) decreases as the number of states increases, but it stabilizes after about 20 states. Again, the ARE(σ_T) decreases with the magnitude of κ_T.

Denote the $u-$th quantile of $\mathcal{G}(\kappa_T, \theta)$ by $Q_u(\kappa_T, \theta)$. The asymptotic variances of the quantile estimators have no closed form expressions, and we calculated them



(a) 3-state model with $\kappa_T = \kappa_1 + \kappa_2$



(b) p-state model with $\kappa_j = \kappa_T/(p-1)$

Figure 3.1: Asymptotic Relative Efficiency of the Marginal Estimator to the Process Estimator for σ_T Under Two Settings

numerically. The AREs of the quantiles are also given in Table 3.1 and will be discussed later.

Now let us consider the more realistic case with censoring. Suppose that the observations now are right or interval-censored. For the TTF data, the computations

of the asymptotic variances of the estimators using the marginal approach have been studied extensively and can be handled using numerical methods. The corresponding computations for the process method is much more involved. We describes the issues briefly but omit the details. Now, suppose N systems with a mixture of exact failures and right-censored data are available. The log-likelihood function of the process data is

$$\ell_N(P) \;=\; \sum_{j=1}^{p-1} \sum_{n=1}^{N_{j,j+1}} \delta_{j,n} \log f_{j,j+1}(\tilde{x}^n_{j,j+1}) + (1-\delta_{i,n}) \log S_j(\tilde{x}^n_{j,j+1}), \tag{3.25}$$

where $\delta_{j,n} = 0$ if $\tilde{x}^n_{j,j+1}$ is right-censored, and $= 1$ otherwise. Further $N_{j,j+1}$ is the number of transitions from state j to $j+1$, and $f_{j,j+1}(x)$ is the density of $\mathcal{G}(\kappa_j, \theta)$. We can use numerical methods to calculate the Hessian matrix of $\ell_N(P)$ and use this to obtain the asymptotic variances of the quantities of interest.

If interval-censored process data are present, the log-likelihood function is

$$\ell_N(P) \;=\; \sum_{n=1}^{N} \log \left(\int_{R_n} \prod_{j=1}^{p-1} f_{j,j+1}(x^n_{j,j+1}) \prod_{j=1}^{p-1} dx^n_{j,j+1} \right), \tag{3.26}$$

where $R_n = \{(a^n_j, b^n_j]; j = 1, \ldots, K \text{ s.t. } a^n_j < x^n_{j,j+1} \leq b^n_j\}$ denotes the interval censoring constraints of the $n-$th system. Direct evaluation of this log-likelihood is difficult. Instead, we use the methods in Chapter II to approximate the information matrix and then get the asymptotic variances.

Table 3.1 shows the AREs for a 4-state model with $(\kappa_1, \kappa_2, \kappa_3, \theta) = (1, 2, 3, 2)$. The values for right censoring were generated randomly according to exponential distributions. The parameters of the distributions were chosen so that the required right-censoring percentages were achieved on average. We considered two situations: 12% right censoring and 67% right censoring. The end points for interval censoring were also chosen randomly from exponential distributions with mean $\bar{C}$. The interval censoring is more severe when $\bar{C}$ is large.

Parameter	μ_T	σ_T	$Q_{.10}$	$Q_{.25}$	$Q_{.50}$	$Q_{.75}$	$Q_{.90}$
True Values	12.00	4.90	6.30	8.44	11.34	14.85	18.55
No censoring	1.00	0.47	0.55	0.75	0.98	0.95	0.79
12% Right-censored	0.98	0.46	0.56	0.76	0.97	0.92	0.77
67% Right-censored	0.76	0.33	0.61	0.81	0.82	0.65	0.52
Interval-censored with $\bar{C} = 0.33$	1.00	0.49	0.56	0.76	0.98	0.94	0.80
Interval-censored with $\bar{C} = 10$	0.96	0.58	0.63	0.74	0.92	0.93	0.81

Table 3.1: One-Step Progressive Case: Asymptotic Relative Efficiencies (of the Marginal Estimators to the Process Estimators) for the Mean, Standard Derivation and Quantiles

The AREs of μ_T are close to one except when there is severe right censoring. For the standard deviation, the AREs range from 0.58 to 0.33, indicating (as was the case with complete data) a substantial loss with the use of the marginal approach. Turning to the quantiles Q_u, the AREs for the medians and even for $Q_{.75}$ are not too bad, except perhaps for the case with high right censoring. The AREs for $Q_{.25}$ are generally lower than those for $Q_{.75}$. This is due both to the asymmetric nature of the gamma distribution and less right censoring at $Q_{.25}$. This pattern is also present for $Q_{.90}$ versus $Q_{.10}$. In general, the marginal method performs quite poorly in estimating the lower quantiles compared to the process-based method. Note also that, the loss in efficiency for severe interval censoring is not as bad as that for severe right censoring. This may be due to the fact that there is corresponding efficiency loss for process data also with interval censoring.

Table 3.2 shows the finite-sample relative efficiencies for sample size $N = 100$. The values were obtained through simulation. We chose a moderate value since there is very high censoring in some cases and the calculations requite a reasonable sample size. While the numbers vary from the asymptotic case, the conclusions are qualitatively the same. In particular, the REs for the mean are reasonably good

Parameter	μ_T	σ_T	$Q_{.10}$	$Q_{.25}$	$Q_{.50}$	$Q_{.75}$	$Q_{.90}$
True Values	12.00	4.90	6.30	8.44	11.34	14.85	18.55
No censoring	1.00	0.60	0.50	0.68	0.93	1.02	0.93
12% Right-censored	0.96	0.51	0.53	0.73	0.93	0.92	0.80
67% Right-censored	0.64	0.33	0.53	0.67	0.67	0.58	0.49
Interval-censored with $\bar{C} = 0.33$	0.98	0.52	0.55	0.74	0.97	0.96	0.82
Interval-censored with $\bar{C} = 10$	1.00	0.47	0.69	0.87	0.99	0.88	0.69

Table 3.2: One-Step Progressive Case: Finite Sample Relative Efficiencies (of the Marginal Estimators to the Process Estimators) for the Mean, Standard Derivation and Quantiles

except when there is severe right censoring. The marginal method performs poorly in estimating the standard deviation. Its performance is again poor for the lower quantiles compared to the upper ones.

Multi-Step Progressive Case

Now consider the general multi-step progressive case. As noted in (3.2), the TTF follows a finite mixture distribution. For simplicity, we rewrite its density in (3.4) as

$$f(t) \;=\; \sum_{j=1}^{J} p_j f_j(t), \tag{3.27}$$

where $p_j = p_{i_0,i_1,\ldots,i_k}$, $f_j = f_{i_0,i_1} * \cdots * f_{i_{k-1},i_k}$ and $J = 2^{p-2}$ is the number of all possible paths. Here f_{ij} is the density of $\mathcal{G}(\kappa_{ij}, \theta)$.

When there is no censoring, likelihood-based inference based on the entire process data is straightforward since we know all the transitions and the sojourn times in different states. The log-likelihood for the marginal approach based on TTF data is obtained from the corresponding mixture model. Inference for parametric finite mixture models has been studied extensively (see, for example, Mclachlan and Peel 2000) using the EM algorithm. In this section, we restrict the performance comparisons to

estimators of the mean, standard deviation and quantiles.

The computations become a lot more involved when there is censoring. We will omit the details here and refer readers to Chapter II for details with right-censored and interval-censored process data.

Consider a 4-state model with $(\kappa_{12}, \kappa_{13}, \kappa_{14}, \kappa_{23}, \kappa_{24}, \kappa_{34}) = (8, 2, 2, 10, 8, 6)$, $\theta = 1$ and $p_{12} = 0.4$, $p_{13} = 0.3$, $p_{23} = 0.5$. These result in a different set of values for the mean, standard deviation, and quantiles, compared to the one-step case. But the censoring set up in Tables 3.3 and 3.4 is mostly the same as the one-step progressive case. The only difference is that $\bar{C} = 2$ instead of 10 for the second interval censoring case.

Parameter	μ_T	σ_T	$Q_{.10}$	$Q_{.25}$	$Q_{.50}$	$Q_{.75}$	$Q_{.90}$
True Values	11.00	8.76	1.19	3.12	8.83	17.52	24.08
No censoring	0.98	0.82	0.93	0.80	0.78	0.74	0.68
10% Right-censored	0.98	0.79	0.97	0.88	0.85	0.85	0.75
55% Right-censored	0.71	0.45	1.04	0.86	0.67	0.63	0.46
Interval-censored with $\bar{C} = 0.33$	1.00	0.85	0.95	0.88	0.88	0.89	0.80
Interval-censored with $\bar{C} = 2$	0.96	0.90	0.58	0.60	0.42	0.77	0.36

Table 3.3: Multi-Step Progressive Case: Asymptotic Relative Efficiencies (of the Marginal Estimators to the Process Estimators) for the Mean, Standard Deviation and Quantiles

Tables 3.3 and 3.4 show, respectively, the AREs and finite-sample ($N = 100$) REs. The behavior for the mean is very similar to the one-step case. Interestingly, the efficiency values for the standard deviation are higher. Of course, the computations here have been done for only one value of σ_T, so it is difficult to generalize. Since the relative efficiencies of the standard deviation are now higher, so are those of the lower quantiles. Another interesting observation is the lower figures for the upper quantiles compared to those for the one-step case. Again, we cannot generalize these

conclusions due to the limited nature of the investigations. Our main point here is to just provide an indication of the efficiency losses in specific cases.

Parameter	μ_T	σ_T	$Q_{.10}$	$Q_{.25}$	$Q_{.50}$	$Q_{.75}$	$Q_{.90}$
True Values	11.00	8.76	1.19	3.12	8.83	17.52	24.08
No censoring	1.00	0.84	0.97	0.85	0.82	0.89	0.80
10% Right-censored	0.98	0.75	0.96	0.84	0.85	0.85	0.72
55% Right-censored	0.46	0.19	1.00	0.81	0.60	0.45	0.20
Interval-censored with $\bar{C} = 0.33$	1.03	0.88	0.96	0.90	0.86	0.92	0.88
Interval-censored with $\bar{C} = 2$	1.01	0.92	0.83	0.89	0.88	0.96	0.83

Table 3.4: Multi-Step Progressive Case: Finite Sample Relative Efficiencies (of the Marginal Estimators to the Process Estimators) for the Mean, Standard Deviation and Quantiles with Finite Sample

3.2.3 Inverse Gaussian Distributions

We consider a special case of the inverse Gaussian distribution with the closure property under convolution. Recall that, if $X_{j,j+1} \sim \mathcal{IG}(\mu w_j, \lambda w_j^2)$, $j = 1, \ldots, p-1$ (inverse Gaussian with location μw_j and scale λw_j^2), the sum is also inverse Gaussian, i.e.,

$$\sum_j X_{j,j+1} \ \sim \ \mathcal{IG}(\mu \sum_j w_j, \lambda (\sum_j w_j)^2). \tag{3.28}$$

To ensure identifiability, we take $w_1 = 1$.

Consider first the one-step progressive case. Inference for the TTF under the marginal case has been studied extensively (see, for example, Seshadri 1999). For the process data, the log-likelihood function of N systems with no censoring is

$$l_N(P) = \frac{p-1}{2} \log \frac{\lambda}{2\pi} + N \sum_{j=2}^{p-1} \log w_j - \frac{3}{2} \sum_{j=1}^{p-1} \sum_{n=1}^{N} \log X_{j,j+1}^n - \sum_{j=1}^{p-1} \sum_{n=1}^{N} \frac{\lambda (X_{j,j+1}^n - \mu w_i)^2}{2\mu X_{j,j+1}^n}. \tag{3.29}$$

One can use this to compute the information matrix and the asymptotic variances. We omit the details.

The mean of the TTF distribution is given by $\mu_T = \mu \sum_j w_j$. It can (again) be shown that MLE of μ_T is the same under the process and marginal cases. So there is no gain in efficiency, following the same pattern for the gamma and normal cases.

The asymptotic variance of the scale parameter $\lambda_T = \lambda(\sum_j w_j)^2$ under both approaches can also be computed directly. For example, for $p = 4$, the asymptotic variance for the process-based MLE is

$$2\lambda^2(1+w_2)^4 \times \frac{2\mu + 2\mu w_2^2 + \lambda w_2 + \lambda w_2^2}{2\mu + 2\mu w_2^2 + 2\lambda w_2 + 2\lambda w_2^2 + 2\mu w_2}. \tag{3.30}$$

The corresponding value for the marginal MLE is $2\lambda^2(1+w_2)^4$. Computing the ratio of these two values show that scale estimator from the process data is always more efficient.

Table 3.5: Inverse Gaussian Durations: Asymptotic Relative Efficiencies (of the Marginal Estimators to the Process Estimators) for the Mean, Standard Derivation and Quantiles

(a) One-step case: $(w_2, w_3) = (2, 3)$

Parameter	μ_T	σ_T	$Q_{.10}$	$Q_{.25}$	$Q_{.50}$	$Q_{.75}$	$Q_{.90}$
True Value	12.00	4.00	7.50	9.12	11.37	14.20	17.31
No censoring	1.00	0.57	0.57	0.75	0.97	0.97	0.85
44% Right-censored	0.87	0.45	0.57	0.75	0.89	0.80	0.67
88% Right-censored	0.48	0.20	0.61	0.72	0.58	0.40	0.31

(b) Multi-step case: $(w_{13}, w_{14}, w_{23}, w_{24}, w_{34}) = (2, 3, 3, 4, 1)$, $p_{12} = 0.6$, $p_{13} = 0.3$, $p_{23} = 0.5$

Parameter	μ_T	σ_T	$Q_{.10}$	$Q_{.25}$	$Q_{.50}$	$Q_{.75}$	$Q_{.90}$
True Value	8.40	3.88	3.93	5.55	7.83	10.57	13.54
No censoring	1.00	0.61	0.70	0.67	0.76	0.88	0.82
33% Right-censored	0.95	0.49	0.67	0.66	0.74	0.85	0.70
76% Right-censored	0.68	0.24	0.59	0.62	0.68	0.63	0.38

We used numerical calculations to compute the AREs of the standard deviation and quantiles in complete and censored data situations. The details are tedious but

straightforward. The results for a 4-state case for the one-step progressive model are given in Table 3.5(a). The corresponding results for multi-step progressive case given in Table 3.5(b). The overall conclusions are qualitatively the same as those for the gamma case.

3.3 Comparison in Prediction Efficiency

We compare the prediction efficiencies of the marginal and the process approaches under various scenarios. Specifically, the goal is to predict the failure time of a system given that it has not failed by a specified time t. For the marginal method, we do not have any information about the system except that it is alive at the right censoring time t. For the process case, however, we know that the system was in a particular state (say $s < p$) at the censoring time t. This can be quite informative in predicting the failure time of the system.

We will use two quality metrics for the prediction intervals to assess the two approaches. They are: a) the lengths of the prediction intervals, and b) the (conditional) coverage probability given that a system was in some specific state at the last observation time. The prediction comparison is intrinsically biased towards the process-based approach since it uses more information than the marginal approach. But this is of course the basic goal here – to quantify the gain in using the process-based approach.

Since the underlying parameters of the multistate model are unknown, one would first have to estimate them based on data from N systems. These estimated values will be used to develop prediction intervals for a new system or perhaps even one of the N systems that has not failed. So, in reality, one would have to incorpo-

rate this estimation uncertainty in the prediction intervals in making the efficiency comparisons. However, this makes the problem extremely challenging in most cases except the normal. Since our goal is to develop a qualitative understanding and we have already covered the estimation efficiency in the previous section, we will assume that the parameters are known in this section. This simplifies the calculations considerably and allows us to focus directly on the prediction problem. Since the process-based approach is more efficient in terms of efficiency, we can treat the relative efficiencies here as lower bounds when one also wants to incorporate estimation uncertainty.

We will consider only one-step progressive models with gamma durations in this section. Assume, as before, that $X_{j,j+1} \sim \mathcal{G}(\kappa_j, \theta)$, $j = 1, \ldots, p-1$, so $T = \sum_{j=1}^{p-1} X_{j,j+1} \sim \mathcal{G}(\kappa_T, \theta)$, $\kappa_T = \sum_{j=1}^{p-1} \kappa_j$. We use the notation $\mathcal{G}_t(\kappa_T, \theta)$ for the conditional distribution $[T|T > t]$, i.e., a gamma distribution truncated from below by t.

We restrict attention to symmetric prediction intervals even though they are not the shortest intervals. The $100(1-\alpha)\%$ symmetric prediction interval based on the marginal data, $I_M = (l_M, u_M)$, satisfies

$$\frac{\int_{u_M}^{\infty} f_T(x)dx}{\int_t^{\infty} f_T(x)dx} = 1 - \alpha/2 \quad \text{and} \quad \frac{\int_t^{l_M} f_T(x)dx}{\int_t^{\infty} f_T(x)dx} = \alpha/2, \tag{3.31}$$

where f_T is the density of $\mathcal{G}(\kappa_T, \theta)$.

For the process-based approach, we have complete information of the system until time t. Then, we know that, the system is at state s $(< p)$ at time t, and that it entered the state s at time t_0. Further, define $S_{i:j}(\cdot)$ as the survival function of $X_{i,i+1} + \cdots + X_{j-1,j}$. Denote the $100(1-\alpha)\%$ symmetric prediction interval for the process method as $I_P = (l_P, u_P)$. We discuss the computation of the upper prediction

interval u_P for various cases of interest. The lower prediction interval is obtained similarly.

(1) $s = p-1$: Then, $[X_{p-1,p}|X_{p-1,p} > t - t_0] \sim \mathcal{G}_{t-t_0}(\kappa_{p-1}, \theta)$, so

$$\frac{\int_{u_P - t_0}^{\infty} f_{p-1,p}(x)dx}{\int_{t-t_0}^{\infty} f_{p-1,p}(x)dx} = 1 - \alpha/2; \tag{3.32}$$

(2) $s = 1$: Here t_0 is 0, so we must have $\mathbb{P}(X_{12} + \cdots + X_{p-1,p} > u_P | X_{12} > t) = 1 - \alpha/2$. This gives

$$\frac{\int_{t}^{\infty} S_{2:p}(u_P - x) f_{1,2}(x)dx}{\int_{t}^{\infty} f_{1,2}(x)dx} = 1 - \alpha/2; \tag{3.33}$$

(3) $1 < s < p-1$: In this case,

$$\begin{aligned} &\mathbb{P}(X_{12} + \cdots + X_{p-1,p} > u_P | X_{12} + \cdots + X_{s-1,s} = t_0, X_{12} + \cdots + X_{s,s+1} > t) \\ &= \mathbb{P}(X_{s,s+1} + \cdots + X_{p-1,p} > u_P - t_0 | X_{s,s+1} > t - t_0) = 1 - \alpha/2. \end{aligned} \tag{3.34}$$

Hence,

$$\frac{\int_{t-t_0}^{\infty} S_{s+1:p}(u_P - t_0 - x) f_{s,s+1}(x)dx}{\int_{t-t_0}^{\infty} f_{s,s+1}(x)dx} = 1 - \alpha/2. \tag{3.35}$$

To study the prediction efficiency, we compare the width of I_M with that of I_P and examine the coverage probability of I_M which need not equal $1 - \alpha$ when we know that the system was in state s at the time of right censoring. Recall that the coverage of I_P is $1 - \alpha$ by construction. The ratio of the widths $r = I_P/I_M$ depends on time t, t_0 and the censored state s. In order to examine their influence on the outcome, we selected t, t_0 in the following way. Suppose the system is in state s at time t:

- pick t_0 as the $p_1\%$ quantile of $\mathcal{G}(\sum_{i=1}^{s-1} \kappa_i, \theta)$, i.e., the distribution of $X_{12} + \cdots X_{s-1,s}$;

- pick z as the $p_2\%$ quantile of $\mathcal{G}(\kappa_s, \theta)$, i.e. the distribution of $X_{s,s+1}$, and let $t = t_0 + z$.

We discuss the results for a 6-state model with scale parameter $\theta = 2$ and shape parameters $\kappa_1 = 5$, $\kappa_2 = 4$, $\kappa_3 = 3$, $\kappa_4 = 2$, $\kappa_5 = 1$. The 95% prediction intervals I_M and I_P were obtained for different values of t_0, t, and s. The ratios $r = I_P/I_M$ and the coverage probabilities I_M (in parentheses) are given in Table 3.6. There are three blocks in the table, corresponding to three states s at the time of right censoring. Within each block, t_0 was chosen to represent the 0.1, 0.5 and 0.9 quantiles of $\mathcal{G}(\sum_{i=1}^{s-1} \kappa_i, \theta)$ (top to bottom respectively). Further, z was taken to be the 0.1, 0.5 and 0.9 quantiles of $\mathcal{G}(\kappa_s, \theta)$ (left to right).

$s = 5$	$z = 0.21$	$z = 1.39$	$z = 4.61$
$t_0 = 18.9$	0.27 (0.62)	0.28 (0.68)	0.30 (0.80)
$t_0 = 27.3$	0.34 (0.87)	0.35 (0.88)	0.38 (0.91)
$t_0 = 37.9$	0.44 (0.93)	0.45 (0.93)	0.48 (0.94)
$s = 3$	$z = 2.20$	$z = 5.35$	$z = 10.6$
$t_0 = 10.9$	0.61 (0.94)	0.60 (0.99)	0.65 (1.00)
$t_0 = 17.3$	0.68 (1.00)	0.71 (1.00)	0.79 (0.99)
$t_0 = 26.0$	0.88 (0.97)	0.90 (0.97)	0.97 (0.96)
$s = 1$	$z = 4.87$	$z = 9.34$	$z = 15.99$
	0.98 (0.96)	0.93 (0.95)	0.93 (0.87)

Table 3.6: Prediction Efficiencies and Coveragese of 95% Prediction Intervals Constructed by the Marginal Method with $\kappa_1 = 5$, $\kappa_2 = 4$, $\kappa_3 = 3$, $\kappa_4 = 2$, $\kappa_5 = 1$

When the system is censored at state 5 (close to the absorbing state), the marginal prediction intervals are considerably wider than the process intervals (as to be expected) with values as low as 0.27. This is so despite the fact that their (conditional) coverage probabilities are lower than 0.95, and much lower in some cases. When the system is censored at state 3 (middle state), the marginal intervals are wider than

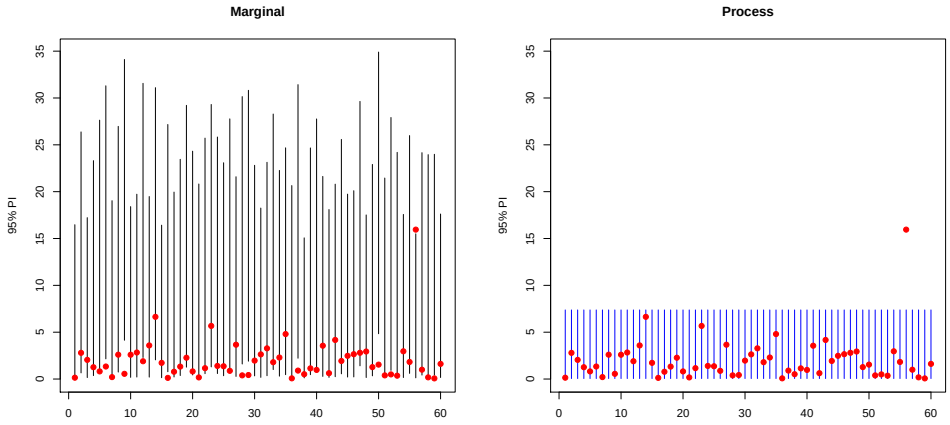
the process intervals but much less so than before (again as to be expected). The conditional coverage probabilities are now higher than 0.95 in most cases. When the system is at state 1 at the time of right censoring, I_M and I_P are comparable in width. Note, though, that the conditional coverage probability of I_M decreases with z. In summary, the loss in prediction efficiency for the marginal method can be very substantial. As to be expected, this loss is especially big when the system is close to the absorbing state at the time of censoring.

Table 3.7 shows similar results for a different set of parameters: $\kappa_1 = 1$, $\kappa_2 = 2$, $\kappa_3 = 3$, $\kappa_4 = 4$, $\kappa_5 = 5$. The efficiency loss of the marginal method is less severe in this case. This can be explained by the fact that the sojourn times in the states closer to the absorbing state have bigger means($\kappa_5 = 5$ now compared to $\kappa_5 = 1$ before), so being closer to the absorbing state is not as predictive as before.

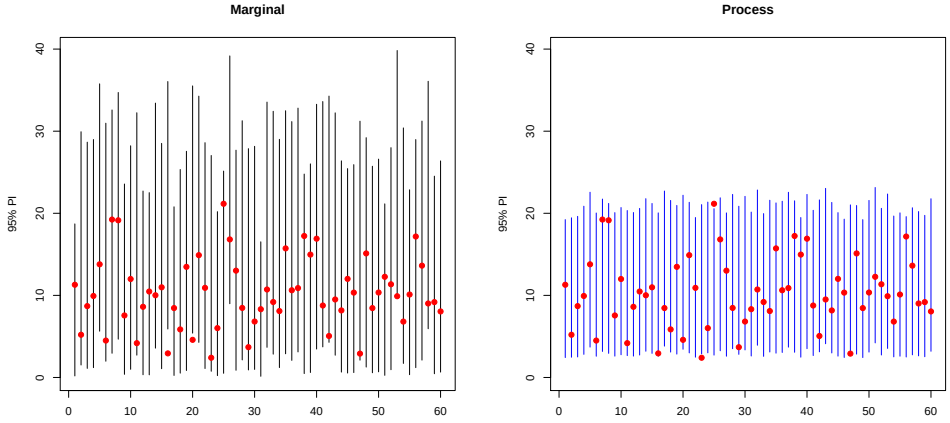
$s = 5$	$z = 4.87$	$z = 9.34$	$z = 16.0$
$t_0 = 12.4$	0.55 (0.87)	0.52 (0.89)	0.53 (0.93)
$t_0 = 19.3$	0.66 (0.97)	0.63 (0.95)	0.62 (0.95)
$t_0 = 28.4$	0.85 (0.98)	0.78 (0.97)	0.74 (0.96)
$s = 3$	$z = 2.20$	$z = 5.35$	$z = 10.6$
$t_0 = 2.20$	0.88 (0.94)	0.86 (0.98)	0.85 (0.97)
$t_0 = 5.35$	0.88 (0.98)	0.86 (0.97)	0.88 (0.93)
$t_0 = 10.6$	0.88 (0.95)	0.89 (0.92)	0.99 (0.81)
$s = 1$	$z = 0.21$	$z = 1.39$	$z = 4.61$
	1.00 (0.95)	1.00 (0.95)	1.00 (0.93)

Table 3.7: Prediction Efficiencies and Coveragese of 95% Prediction Intervals Constructed by the Marginal Method with $\kappa_1 = 1$, $\kappa_2 = 2$, $\kappa_3 = 3$, $\kappa_4 = 4$, $\kappa_5 = 5$

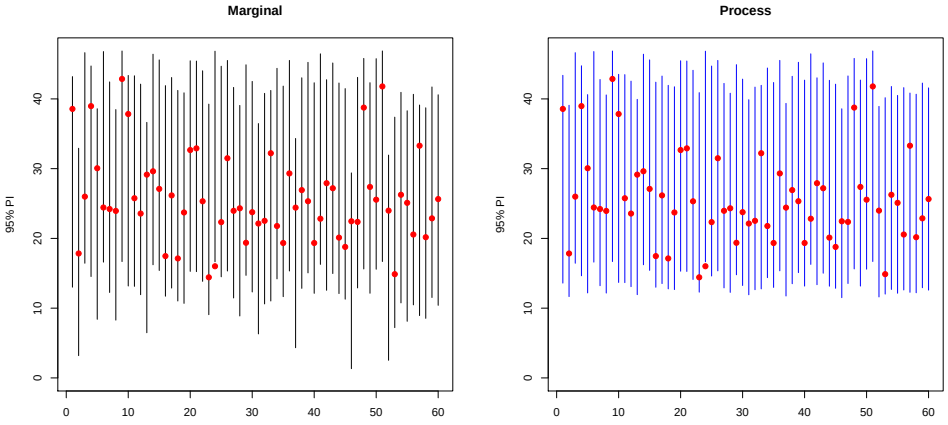
Figure 2 provides a different view of the comparisons. We simulated TTFs from a 6-state model with $\theta = 2$ and $\kappa_1 = 5$, $\kappa_2 = 4$, $\kappa_3 = 3$, $\kappa_4 = 2$, $\kappa_5 = 1$. These TTFs were randomly right censored with censoring times t ($\sim$ Exp(0.1)). If the units had



(a) Censored at state 5



(b) Censored at state 3



(c) Censored at state 1

Figure 3.2: Prediction Intervals Constructed by the Marginal and the Process Methods and the Hold-out TTFs

not occurred yet at time t, we constructed 95% prediction intervals based on the marginal and process methods. Sixty prediction intervals of the remaining life $T - t$ are shown in each panel. Additionally, we kept the exact TTFs to check how well the prediction intervals did.

We see that, when the data are censored at state 5, the I_M's are much wider than I_P. In fact, most of the "true" remaining life times are located at the lower part of the interval. This explains why I_M has lower coverage than nominal. If the middle point of the interval (median residual life) is used as a point predictor, the prediction from the marginal method will substantially overestimate the residual TTF. When the observations are censored at state 3, most of the intervals I_M cover the points, but the width of I_M is still wider than I_P.

On the other hand, when the observations are censored at state 1, I_M's are only slightly wider than I_P. Notice that the lower bound for I_M is usually smaller than the lower bound for I_P. This suggests that the point prediction from the marginal method will under-predict the residual TTF. These results are consistent with the findings in Table 3.6.

3.4 Concluding Remarks

Even if the time-to-event ("failure") is an end-point of an underlying process, it is tempting to ignore the process data and analyze just the TTF data. There are several reasons for this: a) the analysis of TTF data has been extensively studied and there are many existing methods and software packages for analyzing such data; and b) the analysis of the process data, on the other hand, is more involved, especially when in the presence of censoring. So practitioners may not go to the extra length

to model and analyze the process data. The results in this chapter show that the efficiency losses that one can incur in ignoring the additional information in the process data is very high. This is especially so when one is interested in predicting failures at individual system level.

It is also worth noting that inference for multistate processes in an active area of research. See Titman and Sharples (2010) for inference for general panel data under phase-type sojourn distributions that appears to scale up to large number of states. We are also currently exploring approximate algorithms that will allow the general methods in Chapter II to scale up. Thus, it is quite likely that the challenges with inference for multistate models with panel data will be addressed to a large extent.

This chapter has focused exclusively on comparisons in terms of estimation and prediction efficiencies in this chapter. Of course, there are many other reasons for considering an analysis of multistate data, including the development of much better insights into the behavior of the underlying process. For example, in the context of a particular disease, Andersen (1988) noted that more biological insight was gained by analyzing the steps in the disease process.

Finally, we note that the analysis of multistate data rely on more assumptions than that based just on TTF data. The former involves specification of models for the sojourn times in each state, for their dependence structure (independence for semi-Markov process), for time dependence structure (assumed to be time-homogeneous in this chapter), etc. It would be of interest to also take into account robustness considerations in comparing the performance of the two methods.

CHAPTER IV

Modeling and Analysis of Degradation Data with Missing Patterns

4.1 Background

4.1.1 Degradation Data

Recent advances in sensing and measurement technologies are making it feasible to collect extensive amounts of data on degradation and performance-related measures associated with components, systems, and manufacturing equipment. Davies (1998) discusses a variety of engineering applications, types of degradation data, and recent developments in the area of condition monitoring and system maintenance. These cover data from vibration and acoustical monitoring, thermography, lubricant and wear debris analysis, etc. Meeker and Escobar (1998, Chapter 14 and 21 and references therein) describes applications in fatigue crack growth, luminosity of light bulbs, corrosion of batteries, semiconductor devices, etc.

Our interest in this research problem was motivated by an application on the degradation of road pavement data collected by the Michigan Department of Transportation (MDOT). This is discussed in more detail in the next section.

With degradation or performance data, time-to-failure is viewed as the first time point when the degradation level exceeds a certain threshold (or performance level

goes below a threshold). This threshold is usually defined by subject-matter considerations. For example, with MDOT's road pavement data, the engineers classify the pavement as having "failed" and in need of repair if the distress index (DI) exceeded 50. Such a definition of failure works only for certain types of failure modes, and degradation or performance data are not useful with catastrophic failures.

4.1.2 Road Pavement Data and Distress Index

This research was stimulated by a project, funded by the Michigan Department of Transportation (MDOT), to analyze degradation of road pavement data on highways to determine if certain types of designs and materials led to longer life of pavements than others. An overview of the design and management of road pavements can be found in Peterson (1987), Haas, Hudson and Zaniewski (1994). In particular, there are different choices for design and materials used in pavement construction.

The specific degradation variable that was available was a measure called distress index (DI). MDOT collects visual images of road conditions by videotaping highway pavement surfaces using a van equipped with cameras and driven at regular speeds. These videotapes are then sent to a central location where they are viewed and scored by the type, extent, severity, and other types of pavement defects. Points are assigned for each distress type depending on the severity and quantity of each distress based on pre-established algorithm, leading to a distress index for each 0.1 mile segment of pavement. These indices are often aggregated to assign more crude measures to larger segments of the road. The DI should be zero for a pavement that has no distress; if the DI is 50 or more, that segment of road is a candidate for rehabilitation.

The DI scoring is done subjectively, so there is usually a lot of variability due to different individuals who do the scoring. The individuals were not the same from year to year, and it was common to find DI scores that were not monotone. In addition, there was a lot of incomplete data. For some pavements, no records were available before certain time period (missing to the left), data were missing for certain consecutive years (missing in an interval) and data were not collected after certain years (missing to the right). In fact, few of the records had complete data for the period of study.

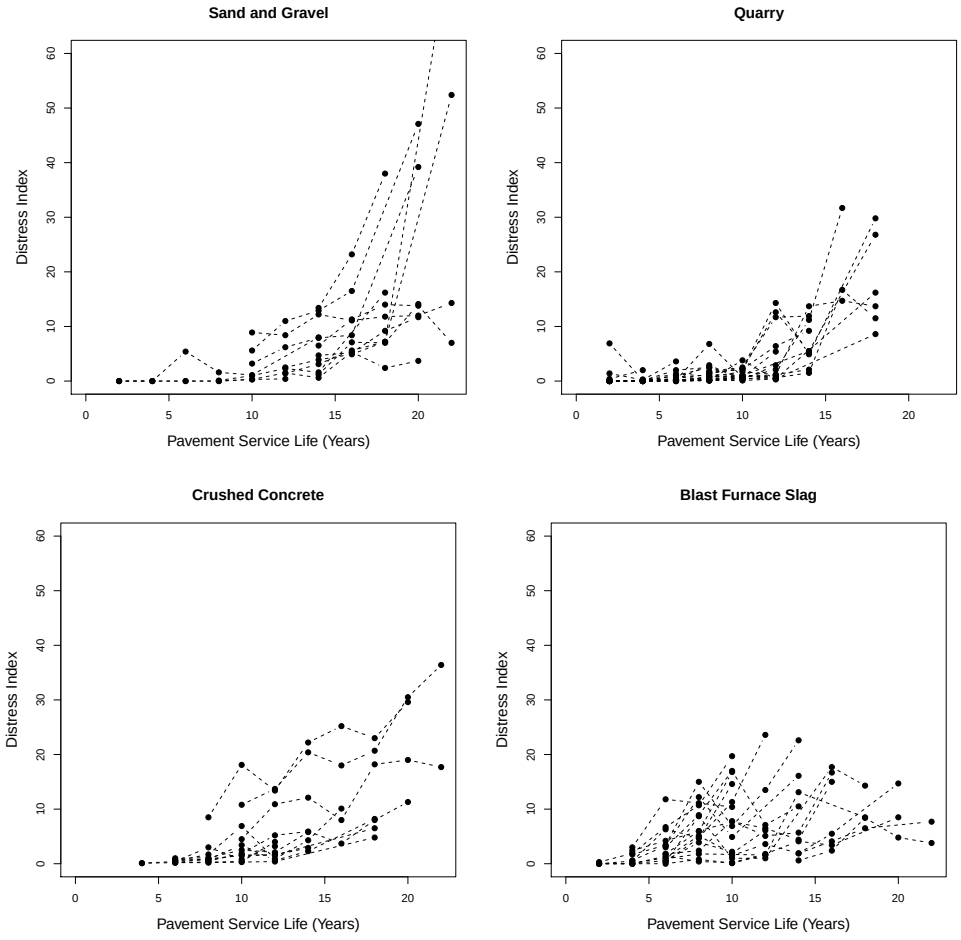


Figure 4.1: The Distress Indices of Highway Pavements (dots) with Different Coarse Aggregate Types (Joined by Pavement)

Figure 4.1 shows the DI values for different pavements recorded over several years for different types of pavement designs. Each connected curve corresponds to one pavement. The x-axis is age (service life) of the pavement and y-axis is the DI. We see that different pavements were measured at different ages, and some measurements are missing. Besides, the degradation paths of the pavements constructed with different design display different patterns and there is a lot of variability even within groups. There is also a lot of missing data.

Several questions were of interest in this application. Are there differences in the design factors and pavement materials in terms of performance? Can we predict when a particular pavement needs repair? Distress indices have been studied in the literature. See, for example, Abu-Lebdeh *et al.* (2003) for a comprehensive summary of recent developments in pavement distress index models. Most of the models in the literature assume that the data follow a parametric form (typically a sigmoidal function) over time. In our case, however, there is no clear parametric form for the degradation data. This is in part due to the poor quality of data. So we resorted to flexible specifications of the mean and variance functions in our analysis.

4.1.3 Missing Patterns

The DI values shown in Figure 4.1 are given by years. Notice that the records for most pavements have missing data. For instance, if the pavement was constructed before the inspection starts, the observations at the beginning of the degradation paths would be missing. All the pavements with Crushed Concrete design fall into this category. Secondly, the inspection period for some pavement ended earlier than others, resulting in the missing at the end. Moreover, measurements for some years

are not collected or not available, leading to missing data in the middle.

Table 4.1 shows the recorded DI values for 4 highway pavements. The observation window for a complete degradation path ranges from age 2 to age 18, and the measurements were usually taken every other year. The $n-$th row stands for the $n-$th pavement, and the $t-$th column represents all the DIs recorded at pavement age $2t$. NA denoted missing DI values. There are quite a few missing data in these 4 pavements.

Pavement	2	4	6	8	10	12	14	16	18
1	0	0	0	0.1	0.9	1.1	13.7	14.7	13.7
2	NA	NA	0.2	0.3	1.1	0.6	1.5	11.5	16.7
3	NA	NA	0.1	0.1	0.1	0.4	NA	NA	16.2
4	0.1	0.2	1.3	0.9	0.7	NA	NA	NA	NA

Table 4.1: The Distress Indices of 4 Pavements ('NA' Means Missing)

Notice also that the DI values for pavements decrease occasionally. This is usually due to measurement error and variations in the scoring. In some cases, the non-monotonicity could be due to partial road repairs which were not taken into account in the DI data.

The missing data issues here have been considered extensively in the literature. However, the problem formulation and research questions in this chapter are somewhat different from those in the existing literature, as we will see later.

4.1.4 Missing Data: Literature Review

Degradation data analysis has been considered by many authors in the literature. Lu and Meeker (1993) used degradation measures to estimate the corresponding time-to-failure distribution parametrically. Lawless and Crowder (2004) suggested a

gamma process as the model of the underlying degradation process. Nonparametric estimation for degradation data has also been studied in Nair and Wang (2011), Wang (2005), Wang (2009) and Wang (2010). See Meeker and Escobar (1998) for additional references and discussion. However, inference with missing degradation data has not been considered much.

There is an extensive literature on missing data relating to longitudinal studies. Methods have been developed under various concepts of "missingness" (see, for example, Laird and Ware 1982, Laird 1988, Lann and Robins 2003, Little 1995, Little and Rubin 2002, Schafer and Graham 2002, Schafer 1997). This refers to the underlying mechanism that causes the missing values (Rubin 1976). If the probability of missing is completely independent of the measurements (observed or unobserved), it is called missing completely at random (MCAR). If the missing mechanism depends on the observed measurements but not the unobserved ones, it is missing at random (MAR). When MCAR and MAR are assumed, the distribution of missing can be ignored for likelihood-based (including Bayesian) inference on the measurements; they are sometimes referred to as "ignorable" missing. When the data are not missing at random, the missing data cannot be ignored and one has to develop a model for including the data in the analysis.

In engineering applications, such as the one discussed here, the data are often missing because they were not collected due to lack of resources or because they were not archived. So it is reasonable to assume that they are missing at random. A possible exception may be data missing in the right tail (sometimes referred to as "dropout"). This may be pavements that had been repaired, so the missingness may be informative. We will not consider this problem in the current chapter.

Most of the inference with missing data in the literature have been developed

under a multivariate normal model (Little 1976). Some of the following structures for the variance-covariance matrix have been considered: identity, exchangeable, certain types of patterned matrices, and autoregressive (Liang and Zeger 1986, Jennrich and Schluchter 1985, and references therein). Most of the popular models used to analyze longitudinal data, such as the regression model, marginal model and random effects model, assume multivariate normality (see Diggle, Liang and Zeger 1994). Both maximum-likelihood estimators (MLEs) and and restricted MLEs have been considered. The robustness of normal models was studied in Little (1988). Liang and Zeger (1986) and others since then have developed generalized estimation equations (GEE) to analyze longitudinal data. GEE is based on quasi-likelihood ideas and does not require a full specification of the likelihood function, and only mean, variance and autocorrelation structure are needed. In addition to the above references, see the excellent books by Diggle, Liang and Zeger (1994), Fitzmaurice, Laird and Ware (2004), Daniels and Hogan (2008) and the references therein.

With incomplete data, the EM algorithm (Dempester, Laird and Rubin 1977) and its Monte Carlo variants have been commonly used for parameter estimation in longitudinal data analysis. For normal models, Jennrich and Schluchter (1985) studied likelihood estimation with unbalanced data (induced by the missing patterns) by using Newton-Raphson, Fisher scoring, and generalized EM algorithms. Laird, Lange and Stram (1987) studied a random-effects model using EM algorithms with arbitrary covariance structure and missing data, again under normal assumption. Calvin (1993) used EM algorithm to do restricted maximum likelihood estimation (REML) in unbalanced multivariate variance-components models. GEE provides consistent estimators when the data are MCAR and not with MAR (Laird 1988). As an extension, the weighed GEE was introduced to correct the bias and give

consistent estimates when the data are MAR (Robins, Rotnitzky and Zhao 1994, 1995). Nonignorable missingness has also been extensively studied, especially in the context of dropouts in surveys. See, for example Little (1995), Diggle and Kenward (1994), Joseph and Molenberghs (2009) and references therein.

There are also software packages for analysis with ignorable missing: SAS procedures PROC MIXED, PROC GLIMMIX and PROC NLMIXED, and the Splus/R functions `lme`, `nlme`, etc.

4.1.5 Formulation

In this section, we present a formulation of the models for the problems we are interested in. Denote the degradation data as $Y^c(t)$, $t = 1, \ldots, T$, where T is the end point of the data collection period. Consider first the case where the N units can be viewed as iid. For $n = 1, \ldots, N$,

$$Y_n^c(t) \;=\; \mu(t) + \epsilon_n(t), \qquad t = 1, \ldots, T,$$

where n denotes the records corresponding to the $n-$th unit in the sample, $\mu(t)$ is the mean of $Y_n(t)$, and $\epsilon_n(t)$ is the random error with zero mean and variance $\sigma^2(t)$. The dependence structure of $Y_n^c(t)$ over time will be specified later.

Let T_n be a subset of $1, \ldots, T$ indicating the time points for which we have available data for the $n-$th unit. Missing to the left would imply that $T_n = \{\ell, \ell + 1, \ldots, T\}$ for some $\ell > 1$. Missing to the right means $T_n = \{1, 2, \ldots, u-1, u\}$ for some $u < T$. Missing in an interval means that $T_n = \{1, 2, \ell, u, \ldots, T\}$ for some $\ell < u+1$, i.e., no observations were recorded during the period $\ell+1, \ldots, u-1$. Any particular record can have a combination of such missing patterns. Throughout this chapter, we assume that the data are missing completely at random.

We will denote the complete data as $Y_n^c(t)$, $t = 1, \ldots, T$ and the observed data (with missing patterns) as $Y_n(t)$, $t \in T_n$, i.e., the subset of $Y_n^c(t)$ that corresponds to T_n. Later, we will use $\boldsymbol{Y}_n^c$ and $\boldsymbol{Y}_n$ to indicate $Y_n^c(t), t = 1, \ldots, T$ and $Y_n(t), t \in T_n$.

We have several inference goals of interest:

1. Develop inference methods for $\mu(t)$ and $\sigma^2(t)$ based on N units and compare their efficiency under different dependence structures.

2. Impute $Y_n^c(t)$ at the missing values and obtain uncertainty bounds.

3. Predict the failure time (time-to-exceed a threshold) for a particular $Y_n(t)$, under some parametric form for $\mu(t)$ and $\sigma^2(t)$.

4. Consider heterogeneous data such as that from a regression study or designed experiments. Using a functional regression or ANOVA model, develop inference procedures for the regression coefficients $\beta(t)$ and develop appropriate test procedures for testing various hypothesis of interest.

In this chapter, topics (1), (2) and (4) are studied. Topic (3) will be pursued as part of future work.

4.1.6 Organization of the Chapter

Section 4.2 deals with inference for $\mu(t)$ and $\sigma^2(t)$. Results are obtained for various dependence structures under a multivariate normal model. Two special cases are considered. Direct likelihood estimation and EM algorithms are examined. Inference under non-normal (gamma) setting is presented in Section 4.3. The normal theory methodology is generalized to one-way ANOVA and regression problems in Section 4.4. Imputation of missing data and development of associated uncertainty bounds

are considered in Section 4.5.

4.2 Inference for $\mu(t)$ and $\sigma^2(t)$ under Normality

We describe some general results here and, for the sake of completeness, include some known results in the literature. The complete data $\boldsymbol{Y}_n^c$ follows a T-dimensional multivariate normal distribution, i.e.,

$$\mathbf{Y}_n^c \ \sim \ \mathcal{MVN}_T(\boldsymbol{\mu}, \Sigma_Y),$$

where $\boldsymbol{\mu} = (\mu(1), \ldots, \mu(t))$ and Σ_Y captures both the variances and the dependence structure over time.

4.2.1 Maximizing the Observed Data Likelihood Function

The observed data (with missing patterns) $\mathbf{Y}_n$ is also multivariate normal but with dimension T_n. Let the number of elelments in T_n be p_n, i.e., there are p_n observed data points. Denote by C the $T \times T$ identity matrix, and C_n the $p_n \times T$ submatrix of C, i.e., the $t-$th row of C is in C_n only when $Y_n^c(t)$ is observed. For example, if $T = 5$ and $\mathbf{Y}_n = (Y_3, Y_4)$, $p_n = 2$, and $C_n = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix}$. We then have

$$\mathbf{Y}_n \ \sim \ \mathcal{MVN}_{p_n}(\boldsymbol{\mu}_n, \Sigma_n),$$

where $\boldsymbol{\mu}_n = C_n\boldsymbol{\mu}$, and $\Sigma_n = C_n\Sigma_Y C_n'$.

As discussed before, when the data are missing completely at random, one can get unbiased estimators by ignoring the missing data. It can be shown that the resulting estimators are optimal for the model where the data are not correlated over time.

The observed likelihood function based on N observed data $\mathbf{y} = \{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ is

$$\ell(\boldsymbol{\mu}, \Sigma_Y; \mathbf{y}) \;=\; -\sum_{n=1}^{N} \left\{ \frac{p_n}{2} \log(2\pi) + \frac{1}{2} \log |\Sigma_{Y,n}| \right\} - \sum_{n=1}^{N} \left\{ \frac{1}{2} (\mathbf{y}_n - \boldsymbol{\mu}_n)' \Sigma_{Y,n}^{-1} (\mathbf{y}_n - \boldsymbol{\mu}_n) \right\}.$$

The corresponding score function for $\boldsymbol{\mu}$ is

$$\frac{\partial \ell}{\partial \boldsymbol{\mu}} \;=\; \sum_{n=1}^{N} C_n' \Sigma_{Y,n}^{-1} (\mathbf{y}_n - C_n \boldsymbol{\mu}).$$

This is of the form $\boldsymbol{w} - M\boldsymbol{\mu}$ for some vector $\boldsymbol{w}$ which is a function of the data and and the matrix M which depends on Σ_Y. So, if Σ_Y is known or can be estimated, we can use this score function to solve for $\boldsymbol{\mu}$.

The score function for Σ_Y is difficult to solve in general. Jennrich (1986) proposed several algorithms using Newton-Raphson method and a generalized EM approach.

In the next few subsections, we develop explicit expressions for the EM algorithm for the special variance structures that we are interested for our application. These provide faster convergence than the Newton-Raphson or generalized EM algorithm.

4.2.2 Models for the Increment $Z^c(t) = Y^c(t) - Y^c(t-1)$

It is more natural to consider $Z^c(t) = Y^c(t) - Y^c(t-1)$, the incremental degradation. Define $Y_n^c(0) = 0$ and let $Z_n^c(t) = Y_n^c(t) - Y_n^c(t-1)$, $t = 1, 2, \ldots, T$ be the increments.

If the original data are multivariate normal, the $\mathbf{Z}_n^c$ are also multivariate normal:

$$\mathbf{Z}_n^c \;\sim\; \mathcal{MVN}_T(\boldsymbol{\theta}, \Sigma_Z),$$

with the mean vector $\boldsymbol{\theta} = (\theta(1), \ldots, \theta(T))$ and the covariance matrix Σ_Z. Correspondingly, we can write the mean of $Y^c(t)$ as $\mu(t) = \sum_{k=1}^{t} \theta(k)$ and $\Sigma_Y = B \Sigma_Z B'$,

where

$$B = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 0 \\ & & & \cdots & \\ 1 & 1 & 1 & \cdots & 1 \end{pmatrix}.$$

One advantage of studying increments, which is well known in the literature, is that (additive) random effects that are attributable to individual units can be removed by this differencing. Another reason is that the mean degradation increases with time. Even though we are using a normal model here, we will assume the signal to noise ratio $\theta(t)/\tau(t)$, where $\tau^2(t)$ is the variance of $Z^c(t)$, will be large so that the probability of the Z^c's being negative will be small. This will ensure that the Y's are generally increasing.

Later, we will consider two models for the independent increments, i.e., Σ_Z is diagonal with elements $\text{diag}\{\tau^2(1), \ldots, \tau^2(T)\}$. Write $\boldsymbol{\tau}^2 = (\tau^2(1), \ldots, \tau^2(T))$. Denote $t_n = T_n \cup \{0\}$. The number of elements in t_n is $p_n + 1$. For ease of notation, write $t_n = \{s_1, s_2, \ldots s_{p_n+1}\}$. We partition the observed $\mathbf{Y}_n$ based on $s_1, s_2, \ldots, s_{p_n+1}$, i.e.,

$$G_n(t) \ := \ (G_n(s_1 : s_2), G_n(s_2 : s_3), \ldots, G_n(s_{p_n} : s_{p_n+1})),$$

with $G_n(i : j) := Y_n^c(j) - Y_n^c(i)$. $\mathbf{G}_n := G_n(t_n)$ denotes the observed differences. When independent increments are assumed, any two components of $\mathbf{G}_n$ are independent and $\mathbf{G} = \{\mathbf{G}_n, n = 1, \ldots, N\}$ are sufficient statistics for the observed likelihood. The MLE's can be obtained by optimizing of the likelihood based on $\mathbf{G}$ directly.

4.2.3 Model 1: No relationship between mean and variance

If $Z_n^c(t) \sim \mathcal{N}(\theta(t), \tau^2(t))$, we have

$$G_n(i : j) \ \sim \ \mathcal{N}(\theta_{i:j}, \tau_{i:j}^2),$$

where $\theta_{i:j} := \sum_{t=i+1}^{j} \theta(t)$ and $\tau^2_{i:j} := \sum_{t=i+1}^{j} \tau^2(t)$. The observed likelihood function is:

$$\ell(\boldsymbol{\theta}, \boldsymbol{\tau}^2; \mathbf{G}) \;=\; -\frac{1}{2} \sum_{1 \le i < j \le T} \sum_{n=1}^{N_{ij}} \left\{ \log(2\pi) + \log \tau^2_{i:j} + \frac{(G_n(i:j) - \theta_{i:j})^2}{\tau^2_{i:j}} \right\},$$

where i and j run over all the possible values such that $G_n(i:j)$ are fully observed, and N_{ij} is the number of observations such that $Y_n(i)$ and $Y_n(j)$ are observed (so $G_n(i:j)$ as well). Then, the score functions are

$$\frac{\partial \ell}{\partial \theta(t)} \;=\; \sum_{i,j} \frac{\partial \ell}{\partial \theta_{i:j}} \cdot \frac{\partial \theta_{i:j}}{\partial \theta(t)} \;=\; \sum_{i,j: i < t \le j} \sum_{n=1}^{N_{ij}} \frac{G_n(i:j) - \theta_{i:j}}{\tau^2_{i:j}}, \tag{4.1}$$

$$\frac{\partial \ell}{\partial \tau^2(t)} \;=\; \sum_{i,j} \frac{\partial \ell}{\partial \tau^2_{i:j}} \cdot \frac{\partial \tau^2_{i:j}}{\partial \tau^2(t)} \;=\; -\frac{1}{2} \sum_{i,j: i < t \le j} \left[\frac{N_{ij}}{\tau^2_{i:j}} - \sum_{n=1}^{N_{ij}} \frac{(G_n(i:j) - \theta_{i:j})^2}{(\tau^2_{i:j})^2} \right], \tag{4.2}$$

for $t = 1, \ldots, T$. Notice that the gradient $\nabla_{\boldsymbol{\theta}} \ell$ is linear in $\boldsymbol{\theta}$. After some algebra, it can be shown

$$\nabla_{\boldsymbol{\theta}} \ell \;=\; \Sigma_Z^{-1}(\boldsymbol{b} - A\boldsymbol{\theta}).$$

Therefore, the MLE $\hat{\boldsymbol{\theta}}$ satisfies $\hat{\boldsymbol{\theta}} = A^{-1}\boldsymbol{b}$. Here $A = (a_{tk})$ is a $T \times T$ matrix and $\boldsymbol{b} = (b_1, \ldots, b_T)^T$ is a T-dim column vector with

$$a_{tk} \;=\; \sum_{i=0}^{\min\{k-1, t-1\}} \sum_{j=\max\{k,t\}}^{T} N_{ij}^{[t]} \cdot \frac{\tau^2(t)}{\tau^2_{i:j}}, \tag{4.3}$$

$$b_t \;=\; \sum_{i=0}^{t-1} \sum_{j=t}^{T} \left(\sum_{n=1}^{N_{ij}^{[t]}} G_n(i:j) \cdot \frac{\tau^2(t)}{\tau^2_{i:j}} \right), \tag{4.4}$$

where $N_{ij}^{[t]}$ is N_{ij} if $i < t \le j$, 0 otherwise.

With constant variance, i.e., $\tau^2(1) = \cdots = \tau^2(T) = \tau^2$, A and $\boldsymbol{b}$ do not depend on τ^2 anymore. The MLE $\hat{\boldsymbol{\theta}}$ can be obtained immediately. In addition, $\tau^2_{i:j} = (j-i)\tau^2$. Plugging $\hat{\boldsymbol{\theta}}$ in (4.2), the MLE $\hat{\tau}^2$ can be obtained directly as well. The observed information at $\hat{\boldsymbol{\theta}}$ is $\hat{\Sigma}_Z^{-1} A$, where $\hat{\Sigma}_Z$ is evaluated at $\hat{\tau}^2$.

In general, A and $\boldsymbol{b}$ are functions of $\boldsymbol{\tau}^2$. $\nabla_{\boldsymbol{\theta}}\ell$ depends on $\boldsymbol{\tau}^2$ and $\nabla_{\boldsymbol{\tau}^2}\ell = 0$ cannot be solved explicitly. But they can be obtained by solving the score functions numerically. Another approach is to optimize over the profile likelihood of $\boldsymbol{\tau}^2$. This is valid, since $\boldsymbol{\theta}$ is fully determined by $\boldsymbol{\tau}^2$. In our experience, the likelihood and profile likelihood estimations are slower than the proposed EM algorithm, especially when T is large and when we are under the latter ANOVA setting.

Now we consider the EM algorithm. To iterate through the E-step and the M-step of EM algorithm, the following function is defined. Given the current estimate $\boldsymbol{\theta}_0$ and $\Sigma_{Z,0}$,

$$Q(\boldsymbol{\theta}, \Sigma_Z | \boldsymbol{\theta}_0, \Sigma_{Z,0}) \ := \ \mathbb{E}_{\mathbf{Z}^c|\boldsymbol{\theta}_0,\Sigma_{Z,0},\mathbf{y}} \left[\log L(\boldsymbol{\theta}, \Sigma_Z; \mathbf{Z}^c)\right],$$

where $L(\boldsymbol{\theta}, \Sigma_Z; \mathbf{Z}^c)$ is the likelihood of $\mathbf{Z}^c = \{\mathbf{Z}_n^c; n = 1, \ldots, N\}$. The expectation is taken over $\mathbf{Z}^c$ conditioning on $\mathbf{y}$ and the current estimate $\boldsymbol{\theta}_0$ and $\Sigma_{Z,0}$. The E-step derives the form of $Q(\boldsymbol{\theta}, \Sigma_Z | \boldsymbol{\theta}_0, \Sigma_{Z,0})$, while the M-step maximizes the Q function over $\boldsymbol{\theta}$ and Σ_Z to get the new update. Fact IV.1 is used for the derivation later.

Fact IV.1. *Let* $\mathbf{Z}$ *have a multivariate normal distribution. Partition* $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2)$ *with mean vector* $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$*, covariance matrix*

$$\Sigma \ = \ \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_{22} \end{pmatrix}$$

Suppose Σ_{22} *is nonsingular. Then the conditional distribution of* $\mathbf{Z}_1$ *given* $\mathbf{Z}_2 = \mathbf{z}_2$ *is* $N(\boldsymbol{\theta}_1 + A(\mathbf{z}_2 - \boldsymbol{\theta}_2), \Sigma^*)$ *for* $A = \Sigma_{12}\Sigma_{22}^{-1}$, $\Sigma^* = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}'$.

Under the independence increment assumption, Σ_Z is determined by $\boldsymbol{\tau}^2$. Given $\boldsymbol{\theta}_0$ and $\boldsymbol{\tau}_0^2$, we have

$$Q(\boldsymbol{\theta}, \boldsymbol{\tau}^2 | \boldsymbol{\theta}_0, \boldsymbol{\tau}_0^2) \ = \ -\sum_{t=1}^{T}\sum_{n=1}^{N}\left\{\frac{1}{2}\log(2\pi\tau^2(t)) + \frac{1}{2\tau^2(t)}\left[S_{n,2}(t) - 2\theta(t)S_{n,1}(t) + \theta^2(t)\right]\right\},$$

where $S_{n,1}(t) := \mathbb{E}(Z_n^c(t)|\mathbf{y}_n, \boldsymbol{\theta}_0, \boldsymbol{\tau}_0^2)$, $S_{n,2}(t) := \mathbb{E}((Z_n^c(t))^2|\mathbf{y}_n, \boldsymbol{\theta}_0, \boldsymbol{\tau}_0^2)$. $S_{n,1}(t)$ and $S_{n,2}(t)$ need to be calculated in the E-step if $Z_n^c(t)$ is not observed. Note that the expectation $S_{n,2}(t)$ does not necessarily equal to $S_{n,1}^2(t)$. From Fact IV.1, we have

(1) Missing to the left: $y_n(u)$ for $t = 1, \ldots, u-1$ are missing with only $y_n(u)$ observed. This can be taken as a special case of interval missing to be discussed in (2).

(2) Missing in an interval: $y_n(\ell)$, $y_n(u)$ are observed, but $y_n(t)$, $t = \ell+1, \ldots, u-1$ are missing. For $t = \ell+1, \ldots, u$, we have

$$\begin{aligned} S_{n,1}(t) &= \theta_0(t) + \frac{\tau_0^2(t)}{\tau_0^2(\ell+1) + \cdots + \tau_0^2(u)} \left[y_n(u) - y_n(\ell) - \sum_{k=\ell+1}^{u} \theta_0(k) \right], \\ S_{n,2}(t) &= S_{n,1}^2(t) + \left(1 - \frac{\tau_0^2(t)}{\tau_0^2(\ell+1) + \cdots + \tau_0^2(u)} \right) \tau_0^2(t). \end{aligned}$$

(3) Missing to the right: $y_n(t)$ for $t = \ell+1, \ldots, T$ are missing with only $y_n(\ell)$ observed. Then, for $t = \ell+1, \ldots, T$,

$$S_{n,1}(t) = \theta_0(t), \quad S_{n,2}(t) = \theta_0^2(t) + \tau_0^2(t).$$

For ease of notation, if $Z_n^c(t)$ is observed, write $S_{n,1}(t) = Z_n^c(t)$ and $S_{n,2}(t) = (Z_n^c(t))^2$. The expectations for data with missing to the left and missing in an interval are dependent on the increment $y_n(u) - y_n(\ell)$. Take a simple example. If $\ell = 0$ and $u = 2$, the conditional expectation of $Z_n^c(1)$ is adjusted by $\tau^2(1)/(\tau^2(1) + \tau^2(2)) \times [y_n(2) - \mathbb{E}Y_n^c(2)]$ comparing to its unconditional mean. Therefore, if $y_n(2)$ is larger than $\mathbb{E}Y_n^c(2)$, the conditional expected value of $Z_n^c(1)$ should also be larger for this subject. In addition to the change of mean, the conditional variance of $Z_n^c(1)$ is shrunk by $\tau^2(2)/(\tau^2(1) + \tau^2(2))$ comparing to the unconditional one. On the other hand, for missing to the right data, the expected values are not dependent on its last observed $y_n(\ell)$.

The maximization of $Q(\boldsymbol{\theta}, \boldsymbol{\tau}^2 | \boldsymbol{\theta}_0, \boldsymbol{\tau}_0^2)$ over $\boldsymbol{\theta}$ and $\boldsymbol{\tau}^2$ in the M-step is straightforward. For $t = 1, \ldots, T$, the update $\boldsymbol{\theta}_1$ and $\boldsymbol{\tau}_1^2$ satisfies

$$\theta_1(t) \;=\; \bar{S}_{N,1} \quad \text{and} \quad \tau_1^2(t) \;=\; \bar{S}_{N,2}(t) - \theta_1^2(t),$$

where $\bar{S}_{N,1}(t) = \sum_{n=1}^{N} S_{n,1}(t)/N$ and $\bar{S}_{N,2}(t) = \sum_{n=1}^{N} S_{n,2}(t)/N$. Iterate through the E-step and the M-step until convergence and the MLEs are obtained.

In general, with missing data, the expression of $\hat{\boldsymbol{\theta}}$ is quite involved. Here, we show its form under a simple but nontrivial setting. Assume constant variance over time and $T = 3$. We observe 2 units, with $\mathbf{y}_2$ fully observed, and $\mathbf{y}_1$ having one component missing,

1. if $y_1(1)$ missing: $\hat{\theta}(1) = y_2(1) + \delta$, $\hat{\theta}(2) = y_2(2) - y_2(1) + \delta$, $\hat{\theta}(3) = \bar{y}(3) - \bar{y}(2)$, where $\delta = \frac{1}{4}(y_1(2) - y_2(2))$ and $\bar{y}(t)$ denotes the average at time t.

2. if $y_1(2)$ missing: $\hat{\theta}(1) = \bar{y}(1)$, $\hat{\theta}(2) = y_2(2) - y_2(1) + \delta$, $\hat{\theta}(3) = y_2(3) - y_2(2) + \delta$, where $\delta = \frac{1}{4}[(y_1(3) - y_1(1)) - (y_2(3) - y_2(1))]$.

3. if $y_1(3)$ missing: $\hat{\theta}(1) = \bar{y}(1)$, $\hat{\theta}(2) = \bar{y}(2) - \bar{y}(1)$, $\hat{\theta}(3) = y_2(3) - y_2(2)$.

The three cases respectively indicate missing to the left, missing in an interval and missing to the right. Notice that, when there is missing to the left or missing in an interval, an adjustment δ is added to the usual unbiased estimate. For instance, consider the missing to the left case, if $y_1(2) > y_2(2)$, we have $\hat{\theta}(1) > y_2(1)$. This indicates, the mean estimate at time 1 is adjusted upward if the unit $\mathbf{y}_1$ (with missing) degrades faster than $\mathbf{y}_2$ at time 2. For missing to the right, however, this adjustment is not displayed.

4.2.4 Model 2: Mean-Variance Structure

We consider now the case where the mean and variance are related by

$$\tau^2(t) \;=\; \tau^2\theta^\alpha(t).$$

In this case, Σ_Z is diagonal with elements $\text{diag}\{\tau^2\theta^\alpha(1),\ldots,\tau^2\theta^\alpha(T)\}$. As the value of α varies, this covers a broad range of mean-variance relationshipss. In particular, if $\alpha = 0$, we get the simple increments model with constant variance $\tau^2(t) = \tau^2$. Under this assumption,

$$G_n(i:j) \;\sim\; \mathcal{N}(\theta_{i:j}(\alpha), \tau^2\theta_{i:j}(\alpha)),$$

where $\theta_{i:j}(\alpha) := \sum_{t=i+1}^{j}\theta^\alpha(t)$.

The observed likelihood function can be expressed in terms of $\mathbf{G}$:

$$\ell(\boldsymbol{\theta},\tau^2,\alpha;\mathbf{G}) \;=\; -\frac{1}{2}\sum_{1\le i<j\le T}\sum_{n=1}^{N_{ij}}\left\{\log(2\pi)+\log\tau^2+\log\theta_{i:j}(\alpha)+\frac{(G_n(i:j)-\theta_{i:j})^2}{\tau^2\theta_{i:j}(\alpha)}\right\},$$

Then, the score functions are

$$\begin{aligned}
\frac{\partial\ell}{\partial\theta(t)} &= \sum_{i,j}\frac{\partial\ell}{\partial\theta_{i:j}}\cdot\frac{\partial\theta_{i:j}}{\partial\theta(t)}+\frac{\partial\ell}{\partial\theta_{i:j}(\alpha)}\cdot\frac{\partial\theta_{i:j}(\alpha)}{\partial\theta(t)}\\
&= -\sum_{i,j:i<t\le j}\left\{\sum_{n=1}^{N_{ij}}\frac{\theta_{i:j}-G_n(i:j)}{\tau^2\theta_{i:j}(\alpha)}+\left[\frac{N_{ij}}{\theta_{i:j}(\alpha)}-\sum_{n=1}^{N_{ij}}\frac{(G_n(i:j)-\theta_{i:j})^2}{\tau^2\theta_{i:j}^2(\alpha)}\right]\cdot\frac{\alpha\theta^{\alpha-1}(t)}{2}\right\}\\
\frac{\partial\ell}{\partial\tau^2} &= -\frac{1}{2}\sum_{i,j}\left[\frac{N_{ij}}{\tau^2}-\frac{1}{\tau^4}\sum_{n=1}^{N_{ij}}\frac{(G_n(i:j)-\theta_{i:j})^2}{\theta_{i:j}(\alpha)}\right].\\
\frac{\partial\ell}{\partial\alpha} &= \sum_{i,j}\frac{\partial\ell}{\partial\theta_{i:j}(\alpha)}\cdot\frac{\partial\theta_{i:j}(\alpha)}{\partial\alpha}\\
&= -\frac{1}{2}\sum_{i,j:i<t\le j}\left\{\frac{N_{ij}}{\theta_{ij}(\alpha)}-\sum_{n=1}^{N_{ij}}\frac{(G_n(i:j)-\theta_{i:j})^2}{\tau^2\theta_{i:j}^2(\alpha)}\right\}\sum_{t=i+1}^{j}\log\theta(t)\theta^\alpha(t)
\end{aligned}$$

The MLEs do not have closed form expressions, but they can be obtained by solving these score functions numerically. The Hessian of ℓ can be evaluated to get variance estimates.

Now consider the corresponding EM algorithm. Given the current estimate $\boldsymbol{\theta}_0$, α_0 and τ_0^2, we have

$$Q(\boldsymbol{\theta}, \alpha, \tau^2 \,|\, \boldsymbol{\theta}_0, \alpha_0, \tau_0^2) \;=\; -\sum_{t=1}^{T}\sum_{n=1}^{N}\left\{\frac{1}{2}\log(2\pi\tau^2\theta^\alpha(t)) + \frac{1}{2\tau^2\theta^\alpha(t)}\left[S_{n,2}(t) - 2\theta(t)S_{n,1}(t) + \theta^2(t)\right]\right\},$$

where $S_{n,1}(t) := \mathbb{E}(Z_n^c(t)|\mathbf{y}_n, \boldsymbol{\theta}_0, \alpha_0, \tau_0^2)$ and $S_{n,2}(t) := \mathbb{E}((Z_n^c(t))^2|\mathbf{y}_n, \boldsymbol{\theta}_0, \alpha_0, \tau_0^2)$. We will see later that $S_{n,1}(t)$ only depends on $\boldsymbol{\theta}_0$ and α_0. The EM algorithm is similar to that of the previous independent increment case with only mild modification. In the E-step, given an observation with missing in an interval $\ell+1, \ldots, u$,

$$\begin{aligned} S_{n,1}(t) &= \theta_0(t) + \frac{\theta_0^\alpha(t)}{\theta_0^\alpha(\ell+1) + \cdots + \theta_0^\alpha(u)}\left(y_n(u) - y_n(\ell) - \sum_{k=\ell+1}^{u}\theta_0(k)\right) \\ S_{n,2}(t) &= S_{n,1}^2(t) + \left(1 - \frac{\theta_0^\alpha(t)}{\theta_0^\alpha(\ell+1) + \cdots + \theta_0^\alpha(u)}\right)\tau_0^2\theta_0^{\alpha_0}(t). \end{aligned}$$

Missing to the left is a special case of with $\ell = 0$. When there is missing to the right, it is the trivial case with $S_{n,1}(t) = \theta_0(t)$ and $S_{n,2}(t) = \theta_0^2(t) + \tau_0^2\theta_0^{\alpha_0}(t)$. It is obvious that the expectations are not linear in $\boldsymbol{\theta}$ anymore.

On the other hand, in the M-step, the update $\boldsymbol{\theta}_1$, α_1 and τ_1^2 satisfies

$$\alpha_1\tau_1^2 = \frac{\alpha_1}{\theta_1^{\alpha_1}(t)}\bar{S}_{N,2}(t) - \frac{2(\alpha_1-1)}{\theta_1^{\alpha_1-1}(t)}\bar{S}_{N,1}(t) + \frac{\alpha_1-2}{\theta_1^{\alpha_1-2}(t)}, \quad t = 1, \ldots, T, \tag{4.5}$$

$$\tau^2 = \frac{1}{T}\sum_{t=1}^{T}\frac{\bar{S}_{N,2}(t) - 2\bar{S}_{N,1}(t)\theta_1(t) + \theta_1^2(t)}{\theta_1^{\alpha_1}(t)}, \tag{4.6}$$

$$\sum_{t=1}^{T}\log\theta_1(t) = \frac{1}{\tau_1^2}\sum_{t=1}^{T}\log\theta_1(t)\cdot\frac{\bar{S}_{N,2}(t) - 2\bar{S}_{N,1}(t)\theta_1(t) + \theta_1^2(t)}{\theta_1^{\alpha_1}(t)}, \tag{4.7}$$

Combining the T equations in (4.5) and plugging in (4.6), we have

$$\sum_{t=1}^{T}\frac{\bar{S}_{N,1}(t)}{\theta_1^{\alpha_1-1}(t)} = \sum_{t=1}\frac{1}{\theta_1^{\alpha_1-2}(t)},$$

Especially, when $\alpha = 2$, we have

$$\begin{aligned}\sum_{t=1}^{T} \frac{\bar{S}_{N,1}(t)}{\theta_1(t)} &= T, \\ \frac{\bar{S}_{N,2}(i)}{\theta_1^2(i)} - \frac{\bar{S}_{N,1}(i)}{\theta_1(i)} &= \frac{\bar{S}_{N,2}(j)}{\theta_1^2(j)} - \frac{\bar{S}_{N,1}(j)}{\theta_1(j)}, \quad \forall i, j \in \{1, \ldots, T\},\end{aligned}$$

Iterate through the E-step and the M-step until convergence and the MLEs are obtained.

4.3 Inference under Gamma Model

The procedures proposed in the last section can be generalized to the exponential family. Here we will briefly discuss the gamma case, i.e., $Z_n^c(t)$'s are gamma distributed with shape parameter $\kappa(t)$ and common scale parameter θ. When the increments $Z_n^c(t)$'s are assumed to be gamma distributed, we cannot model directly the missing $Y_n(t)$'s, and we focus on the missing $Z_n(t)$'s. We consider the independent increment model here, i.e., $Z_n^c(t) \sim \mathcal{G}(\kappa(t), \theta)$. Write $\boldsymbol{\kappa} = (\kappa(1), \ldots, \kappa(T))$. Then, we have $G_n(i:j) \sim \mathcal{G}(\kappa_{ij}, \theta)$, where $\kappa_{i:j} = \sum_{t=i+1}^{j} \kappa(t)$.

The observed likelihood function can be expressed in terms of $\mathbf{G}$:

$$\ell(\boldsymbol{\kappa}, \theta; \mathbf{G}) = \sum_{1 \le i < j \le T} \sum_{n=1}^{N_{ij}} \left\{ (\kappa_{i:j} - 1) \log(G_n(i:j)) - \frac{1}{\theta} G_n(i:j) - \kappa_{i:j} \log \theta - \psi(\kappa_{i:j}) \right\},$$

where $\psi(\cdot)$ is the logarithem of gamma function. The score functions are

$$\begin{aligned}\frac{\partial \ell}{\partial \kappa(t)} &= \sum_{i,j} \frac{\partial \ell}{\partial \kappa_{i:j}} \cdot \frac{\partial \kappa_{i:j}}{\partial \kappa(t)} = \sum_{i,j;\, i < t \le j} \sum_{n=1}^{N_{ij}} \log G_n(i:j) - N_{ij} \log \theta - N_{ij} \psi'(\kappa_{i:j}), \\ \frac{\partial \ell}{\partial \theta} &= \sum_{i,j} \left\{ \frac{1}{\theta^2} \sum_{n=1}^{N_{ij}} G_n(i:j) - \frac{N_{ij} \kappa_{i:j}}{\theta} \right\}.\end{aligned}$$

Even though the MLEs do not have closed form expressions, they can be obtained by solving these score functions numerically. The Hessian of ℓ can be evaluated to get variance estimates. However, the convergence could be slow when T is large.

Consider the EM algorithm now. Given the current estimate $\boldsymbol{\kappa}_0$ and θ_0,

$$Q(\boldsymbol{\kappa}, \theta | \boldsymbol{\kappa}_0, \theta_0) \;=\; \sum_{t=1}^{T} \sum_{n=1}^{N} \left\{ (\kappa(t) - 1) S_{n,3}(t) - \frac{1}{\theta} S_{n,1}(t) - \kappa(t) \log \theta - \psi(\kappa(t)) \right\},$$

where $S_{n,3}(t) := \mathbb{E}(\log Z_n^c(t) | G_n(i:j), \boldsymbol{\kappa}_0, \theta_0)$ and $S_{n,1}(t) = \mathbb{E}(Z_n^c(t) | G_n(i:j), \boldsymbol{\kappa}_0, \theta_0)$. Since independent increments are assumed, the expectation of the right missing observations $Z_n^c(t)$ follows its unconditional distribution, i.e., $\mathcal{G}(\kappa(t), \theta)$. Thus, $S_{n,1}(t) = \kappa(t)\theta$. The nontrivial case is to find the conditional distribution of interval missing observations (left missing included). Given $G_n(i:j)$, we have

$$\frac{Z_n^c(t)}{G_n(i:j)} \;\sim\; \mathcal{B}(\kappa(t), \kappa_{i:j} - \kappa(t)), \qquad \text{for } t = i+1, \ldots, j,$$

which follows a Beta distribution. Therefore, $S_{n,1}(t) = G_n(i:j) \times \kappa_t / \kappa_{i:j}$. $S_{n,3}(t)$ does not have a close form expression based on this distribution, but can be approximated by Monte Carlo methods.

As for the M-step, the update cannot be obtained explicitly and is found by using numerical method such as Newton-Raphson.

Iterating through the E-step and the M-step, the MLEs are obtained at the convergence. Under this setting, the variance estimate of MLEs cannot be obtained as the normal case, so are approximated by numerically evaluate the Hessian of ℓ at the MLEs.

4.4 Functional Regression and Analysis of Variance

In Section 4.2, the inference problem based on iid samples was discussed. Now we study the extension to regression problem for normal independent increment models, i.e.,

$$Z_n^c(t) \;=\; \boldsymbol{x}_n' \boldsymbol{\beta}(t) + \epsilon_n(t), \qquad \text{where } \epsilon_n(t) \sim \mathcal{N}(0, \tau^2(t)),$$

where $\boldsymbol{x}_n$ is the d-dim covariates associated with the n-th observation, and $\boldsymbol{\beta}(t)$ the corresponding covariate effects at time t. The inference here includes the estimation of the covariate effect $\boldsymbol{\beta}(t)$, $t = 1, \ldots, T$. Without confusion, later denote $\boldsymbol{\beta}$ as $\{\boldsymbol{\beta}(t), t = 1, \ldots, T\}$, a $T \times d$ matrix. Therefore,

$$\mathbf{Z}_n^c \ \sim \ \mathcal{MVN}_T(\boldsymbol{x}_n'\boldsymbol{\beta}, \Sigma_Z). \tag{4.8}$$

The direct likelihood estimation can be done as before. For $\mathbf{y}_n$, its covariance matrix still Σ_Y, but the mean is now $\boldsymbol{\mu}_n$ with $\mu_n(t) = \sum_{k=1}^{t} \boldsymbol{x}_n'\boldsymbol{\beta}(k)$. As for the EM algorithm, the E-step is mostly unchanged; however, the expectations $S_{n,1}(t)$ and $S_{n,2}(t)$ are calculated under (4.8). The M-step update is obtained by performing a least square fit on $S_{n,1}(t)$ over a $N \times d$ design matrix X, with $\boldsymbol{x}_n$ in its $n-$th row. Then, the update satisify

$$\begin{aligned} \boldsymbol{\beta}(t) &= (X^{'}X)^{-1}X^{'}S_1(t), \\ \tau^2(t) &= \frac{1}{N}\sum_{n=1}^{N_t}\sum_{i=1}^{K}\left\{S_{n,2}(t) - [\boldsymbol{x}_n'\boldsymbol{\beta}(t)]^2\right\}, \end{aligned}$$

where $S_1(t) = \{S_{1,1}(t), \ldots, S_{N,1}(t)\}$.

The ANOVA problem is a special case here. Assume we have K samples with

$$Z_{n,i}^c(t) \ = \ \theta(t) + \eta_i(t) + \epsilon_n(t), \quad i = 1, \ldots, K, \quad n = 1, \ldots, N_i$$

where i indicates the ith subgroup, N_i the number of observations present in the group, $\epsilon_n(t) \sim N(0, \tau^2(t))$ and we impose the constraint that $\sum_{i=1}^{K} \eta_i(t) = 0$, $\forall t$. Hence, $\eta_K(t)$ is determined by $\eta_1(t), \ldots, \eta_{K-1}(t)$ at each $t = 1, \ldots, T$.

Now denote $\boldsymbol{\eta}_i = (\eta_i(1), \ldots, \eta_i(T))$ and $\boldsymbol{\beta} = (\boldsymbol{\theta}, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_{K-1})$. Then, the EM algorithm proposed for the regression problem can be readily applied. In the M-step, the design matrix X used in the least square fit is changed: at time t, pool the observations from the K samples together, abuse the notation a bit and write

them as $S_{n,1}(t)$. If $S_{n,1}(t)$ is from the $i-$th sample, with $i < K$, the $n-$th row of X is $(1, 0, \ldots, 1, \ldots, 0)$. If $S_{n,1}(t)$ is from the $K-$th sample, the $n-$th row of X is $(1, -1, -1, \ldots, -1)$.

For the ANOVA problem, the MLE of $\hat{\boldsymbol{\beta}}$ satisfies

$$\begin{aligned} \hat{\boldsymbol{\theta}} &= (\sum_{i=1}^{K} A_i)^{-1} \sum_{i=1}^{N} \boldsymbol{b}_i, \\ \hat{\boldsymbol{\eta}}_i &= A_i^{-1}\boldsymbol{b}_i - \hat{\boldsymbol{\theta}}, \quad i = 1, \ldots, K-1, \end{aligned}$$

where A_i and $\boldsymbol{b}_i$ are the coefficient matrix and vector defined in Section 4.2, for the $i-$th sub-sample (iid's within). Besides, the observed information $(KT \times KT)$ associate with $\hat{\boldsymbol{\beta}}$ is

$$\begin{pmatrix} D^{-1}\sum_{i=1}^{K} D^{-1}A_i & D^{-1}A_K & \cdots & D^{-1}A_K \\ D^{-1}A_K & D^{-1}(A_1 + A_K) & \cdots & D^{-1}A_K \\ \cdots & & & \\ D^{-1}A_K & D^{-1}A_K & \cdots & D^{-1}(A_{K-1} + A_K) \end{pmatrix}$$

With the estimation procedure proposed, we can test the hypothesis

$$H_0 : \eta_1(t) = \eta_2(t) = \cdots = \eta_K(t), \quad t = 1, \ldots, T$$

by using the variance estimate of $\hat{\boldsymbol{\eta}}_i$, i.e., if the constructed confidence band of $\hat{\boldsymbol{\eta}}_i$ includes 0 at all time, H_0 cannot be rejected.

The one-way ANOVA is performed on the pavement distress indices data. The estimate $\hat{\boldsymbol{\eta}}_i$, $i = 1, \ldots, 4$ and their confidence bands are plotted in Figure 4.2. Note that they are on the increment scale instead of the original distress index scale. The 4 subgroups represent 4 different materials used to construct the pavement, i.e. Sand and Gravel, Quarry, Crushed Concrete and Blast Furnace Slag. We see the pavement group constructed by Quarry degrades significantly faster than the grand mean around year 10, while pavements constructed with Crushed Concrete behave significantly better around then. Besides, pavements constructed with Sand and

Gravel degrades significantly faster than the grand mean $\boldsymbol{\theta}$ after year 20. For the other 3 groups, the variation is too large to judge whether the trends are real at the end.

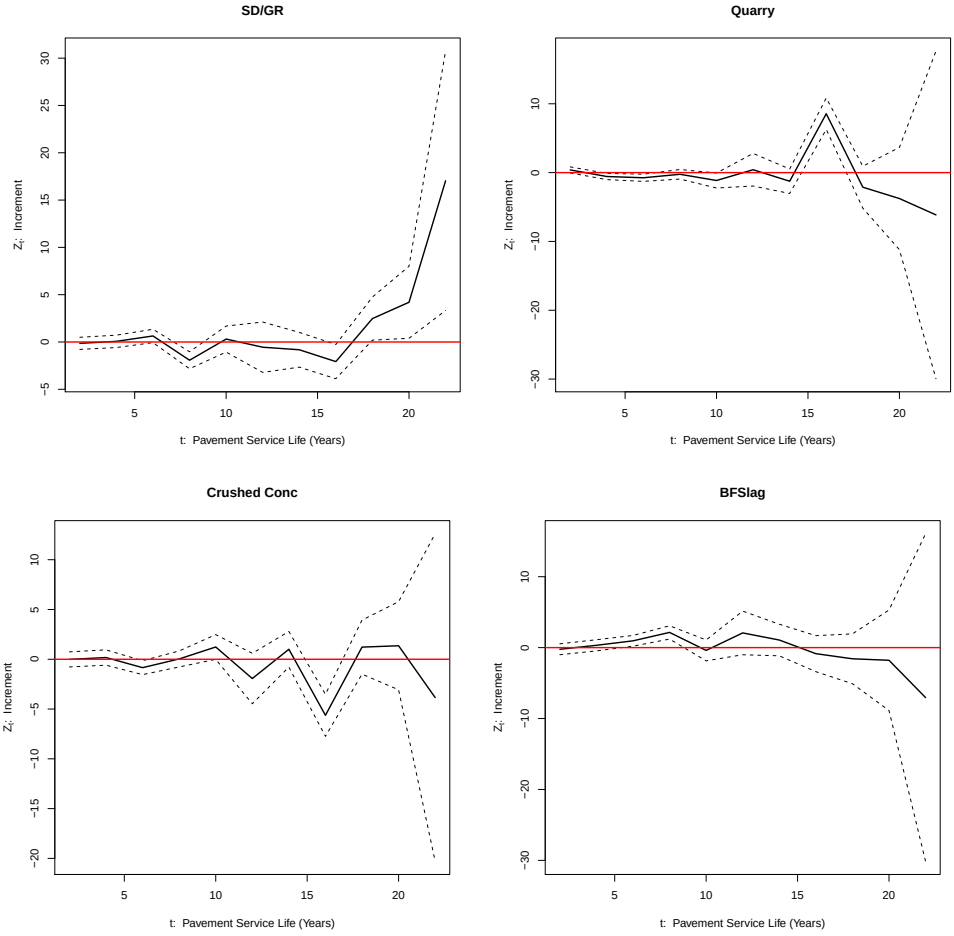


Figure 4.2: Estimate of the Design Effect on Pavement and 90% Confidence Band

4.5 Imputation of the Missing Degradation Data

We have discussed inference for $\mu(t)$ and $\sigma^2(t)$ based on incomplete degradation data under various settings. Now, we are interested in imputing the missing $\mathbf{Y}_n^c$ based on the estimates from the proposed procedures. This can be done by using

the calculation of $S_{n,1}(t)$ in the E-step of the proposed EM algorithm to fill in the missing.

Consider the simple case, i.e., independent increments with constant variance. Suppose we obtain the MLE of the mean $\hat{\boldsymbol{\theta}}_n$ and the variance $\hat{\tau}^2$ of $\mathbf{Z}_n^c$ (no matter from an iid sample or regression) and their variance estimates $\hat{V}(\hat{\boldsymbol{\theta}}_n) = \hat{\Sigma}_Z^{-1}\hat{A}$ and $\hat{V}(\hat{\tau}^2)$. If $\mathbf{y}_n$ is missing in between ℓ and u, we have, for $t = \ell+1, \ldots, u$,

$$\hat{S}_{n,1}(t) \;=\; \hat{\theta}(t) + \frac{1}{u-\ell}\left[y_n(u) - y_n(\ell) - \sum_{k=\ell+1}^{u}\hat{\theta}(k)\right].$$

If $\mathbf{y}_n$ is missing to the right with the right-most observation at ℓ, it is easy to see $\hat{S}_{n,1}(t) = \hat{\theta}(t),\ t = \ell+1, \ldots, T$. Once we fill in the missing of $\mathbf{Z}_n^c$ with $\hat{S}_{n,1}(t)$, the imputed values of $\mathbf{Y}_n^c$ is easy to get. The missing to the left case is a special case of missing in the interval by letting $y_n(\ell) = 0$. Besides, the confidence band can be obtained exactly here, since the expectation of $\hat{S}_{n,1}(t)$ is linear in $\hat{\boldsymbol{\theta}}$, i.e., $\hat{S}_{n,1}(t) = \boldsymbol{c}_n'\hat{\boldsymbol{\theta}}$. Thus, $\hat{V}(\hat{S}_{n,1}(t)) = \boldsymbol{c}_n'\hat{V}(\hat{\boldsymbol{\theta}})\boldsymbol{c}_n = \boldsymbol{c}_n'\hat{\Sigma}_Z^{-1}\hat{A}\boldsymbol{c}_n$, and consequently the confidence band obtained by imputing $Y_n^c(t)$ for $t \notin T_n$. Notice that here $\hat{V}(\tau^2)$ is not needed for the computation.

Under other settings, the imputation of the missing $\mathbf{Y}_n^c$ can be similarly done but with different expectation forms for the missing $\mathbf{Z}_n^c$. For instance, 1) if $Z_n^c(t) \sim \mathcal{N}(\theta(t), \tau^2\theta^\alpha(t))$. Given $(\hat{\boldsymbol{\theta}}, \hat{\tau}^2, \hat{\alpha})$. the interval missing observations with its left-most observation $y_n(\ell)$ and right-most observation $y_n(u)$ has

$$\hat{S}_{n,1}(t) \;=\; \hat{\theta}(t) + \frac{\hat{\theta}^{\hat{\alpha}}(t)}{\hat{\theta}^{\hat{\alpha}}(\ell+1) + \cdots + \hat{\theta}^{\hat{\alpha}}(u)}\left(y_n(u) - y_n(\ell) - \sum_{k=\ell+1}^{u}\hat{\theta}(k)\right),$$

for $t = \ell+1, \ldots, u$. The left missing is a special case of the interval missing with $\ell = 0$. When there is missing to the right, we have $\hat{S}_{n,1}(t) = \hat{\theta}(t)$. 2) For the gamma model, i.e., $Z_n^c(t) \sim \mathcal{G}(\kappa(t), \theta)$. Given $(\hat{\boldsymbol{\kappa}}, \hat{\theta})$ and $G_n(i:j)$, for interval/left missing

$$\frac{Z_n^c(t)}{G_n(i:j)} \;\sim\; \mathcal{B}(\hat{\kappa}(t), \hat{\kappa}_{i:j} - \hat{\kappa}(t)), \qquad \text{for } t = i+1, \ldots, j.$$

Therefore, $\hat{S}_{n,1}(t) = G_n(i:j) \times \hat{\kappa}(t)/\hat{\kappa}_{i:j}$. If there is right missing, $\hat{S}_{n,1}(t) = \hat{\kappa}(t)\hat{\theta}$. Besides, the confidence band can be approximated by delta method.

Figure 4.3 gives two imputed degradation paths from the pavement data. We estimate the parameters by assuming the increments are independent over time. The black points are the censored $\mathbf{y}_n$, joined if consecutive data are available. The missing is quite severe for these pavement data. The red dash line gives the imputation, along with its 95% confidence band in black dash lines. The observation in the left panel suffers from left, interval and also right missing. The one in the right panel has right missing early on. Comparatively, the missing is more severe in the left panel, the variation accumulates and results in the wide confidence band associated with the imputation after year 20.

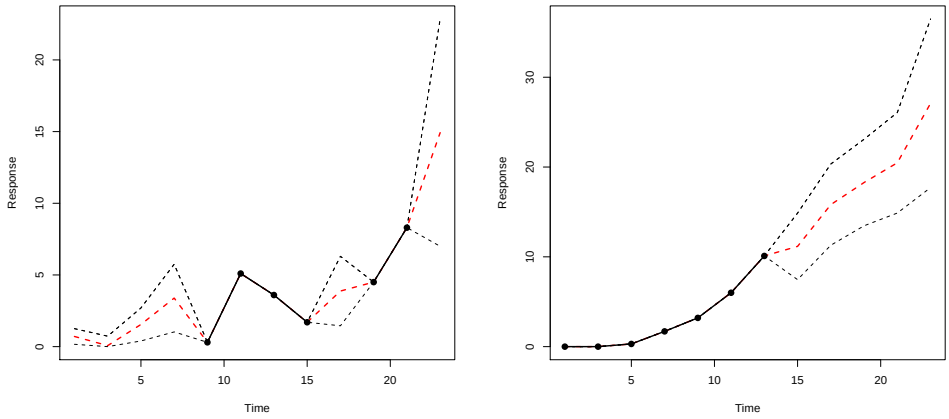


Figure 4.3: Imputed Degradation Path Along with 95% Confidence Band

The imputation of the right missing observations (as in the right panel) can be viewed as the prediction of the degradation in the future. This helps to predict how the unit will degrade in the future, and whether the failure may occur in a certain time period; however, note that the prediction can only be done for time smaller than T, since we estimate $\mu(t)$ nonparametrically. For these two projects in

particular, either of them reaches 50 at the end of inspection. In order to do a long term prediction, parametric form should be imposed on $\mu(t)$.

CHAPTER V

Future Work

This chapter describes some areas and ideas of future work in each of the three topics discussed in the previous three chapters.

I. Inference for Multi-State Models with Censored Data

1. A major disadvantage of the algorithms developed in Chapter II is that they are computationally intensive and work only for moderate number of states. They do not scale up to situations with a large number of states. So, further research is needed in several directions: a) speeding up the algorithms, by writing more efficient code, parallelizing the code, or possibly using approximations; b) develop other estimating equations that are easier to handle computationally and led to reasonably efficient etimators.

2. Initial examination of non-parametric maximum likelihood estimation methods indicated that there were severe non-uniqueness problems. Additional work is needed to understand and develop appropriate constraints to resolve the non-uniqueness problems or characterize all possible solutions.

3. Once a non-parametric approach is available, it will be of interest to develop

graphical and goodness-of-fit methods for selecting appropriate parametric families and doing parametric inference.

4. The consumer credit loan application that motivated our research does not fit into the formulation in this research. The time-to-move to another state is not random but fixed by the definition of the state: one-month delinquent, two-month delinquent, etc. Further, cycling back from being $n-$month delinquent to being current is allowed, so it is not strictly progressive. One way to address the non-progressive nature of the problem is to add additional states, such as current but previously $n$month delinquent once, or more than once, etc. While this will increase the number of states, the problem is still manageable because customers are allowed only at most 6 months of delinquency before defaulting. There is an extensive amount of data available, including socio-demographic characteristics, to model these multi-state data and address specific questions of practical interest.

5. Another area includes inference with time-varying covariates.

II. Comparing Two Methods for Analyzing Time-to-Failure

The main goal of this research was to quantify and document the gains to be made from using the multi-state process data to analyze time-to-failure. A much more extensive documentation of the results for other distributions is probably not of much value, as this research already makes a compelling case. One interesting area for future research is comparison of the estimators in non-parametric and semi-parametric situations.

III. Analysis of Degradation Data with Missing Patterns

This work is in its early stages and there are many avenues for future research.

1. There is extensive work in modeling and analysis of growth curves and longitudinal data, including a discussion of the use of marginal and conditional approaches as well as random effect models. There is also extensive use of generalized estimating equations (GEE). We propose to examine the connections to these areas, especially on the use of estimating equations, development of robust methods and comparisons to the normal-theory based methods.

2. Even within the context of normal-theory models and a general variance-covariance matrix Σ, the dimension of the matrix grows with the number of observations T. It would be interesting to examine the use of sparse methods of inference for Σ, with some structure since the data are ordered over time.

APPENDIX

0.1 Panel Data: Extension to Right Censoring and Exact Failure

Since the sampling procedure of the complete history $\mathbf{Z}$ depends on the censoring pattern of the observed $\mathbf{y}$, we illustrate how to sample for right-censored/exact failured $\mathbf{y}$ along with interval censoring. First suppose that exact failure is also observed in $\mathbf{y}$. Consider a 3-state progressive model with 2 possible paths: $(1,2,3)$ and $(1,3)$. Now the exact failure time is observed along with panel data and sample $\mathbf{Z}$ based on $\mathbf{y} = (t_1, 1, t_2, 3) \cup \{T = t_2\}$. We need to augment δ here. Let $\delta = 1$ when $\boldsymbol{i} = (1,2,3)$, $\delta = 2$ when $\boldsymbol{i} = (1,3)$, and $\tau_s^{(ij)}$ as the $s-$th transition made from state i to state j. The distribution of $\mathbf{z} = (\delta, \boldsymbol{\tau}_\delta)$ given $\boldsymbol{\theta}$ and $\mathbf{y}$ follows (2.12) with

$$g(\mathbf{y}|\delta = 1, \boldsymbol{\tau}_\delta) f(\delta = 1, \boldsymbol{\tau}_\delta|\boldsymbol{\theta}) \; = \\ \mathbf{1}\{\tau_1 \in (t_1, t_2) \cup \tau_2 = t_2 - \tau_1\} \times p_{12} f_{12}(\tau_1) p_{23} f_{23}(t_2 - \tau_1), \quad (0.1)$$

$$g(\mathbf{y}|\delta = 2, \boldsymbol{\tau}_\delta) f(\delta = 2, \boldsymbol{\tau}_\delta|\boldsymbol{\theta}) = \mathbf{1}\{\tau_1 = t_2\} \times p_{13}. \quad (0.2)$$

To propose new jump from path 2 to path 1, the following dimension transformation transform is used, i.e., $(\tau_1^{(12)}, \tau_2^{(23)}) = T_{2\to1}^{Exact}(\tau_1^{(13)})$:

$$\tau_1^{(12)} = t_1 + u(\tau_1^{(13)} - t_1), \quad \tau_2^{(23)} = \tau_1^{(13)} - \tau_1^{(12)},$$

where $u \sim \mathcal{U}(0,1)$. If the current iterate takes path 1, the following transformation is used, i.e., $\tau_1^{(13)} = T_{1\to2}^{Exact}(\tau_1^{(12)}, \tau_2^{(23)})$: $\tau_1^{(13)} = \tau_1^{(12)} + \tau_2^{(23)}$. As for the update within each path, since we observe the exact absorbing time of the system, there is no need to sample for path 2. To sample for path 1, Metropolis-Hastings sampling is used. With the 3-state model, $\tau_2^{(23)}$ is determined by $\tau_1^{(12)}$, so no Gibbs sampling is needed. Based on (0.1), we sample $\tau_1^{(12)}$ from density $p_{12}^E(\cdot)$:

$$p_{12}^E(\tau_1^{(12)}) \; \propto \; \mathbf{1}\{\tau_1^{(12)} \in (t_1, t_2)\} \times f_{12}(\tau_1^{(12)}) f_{23}(t_4 - \tau_1^{(12)}).$$

For general distribution families, we cannot sample exactly from $p_{12}^E(\cdot)$ (even without truncation). Random walk Metropolis-Hasting is used to do the update. Here, because of the truncation, the new draw is rejected if outside of (t_1, t_2). The transition kernel used for this case, is a one-dimension normal distribution $\mathcal{N}(\tau_{n-1,1}^{(12)}, (\frac{c}{2}(t_2 - t_1))^2)$, where c is a scaling parameter. Note that, if the path involves more than 2 transitions, Metropolis-within-Gibbs is needed.

When right censoring is present, the complete history $\mathbf{z} = (\delta, \boldsymbol{\tau}_\delta)$ denotes the SMP observed till the last observation point and $i_{l(\boldsymbol{i})} < p$. Now consider a 4-state model, with possible paths $(1,2,3,4)$, $(1,2,4)$, $(1,3,4)$ and $(1,4)$. Given $\mathbf{y} = (t_1, 1, t_2, 3)$, the distribution of $\mathbf{z}$ given $\boldsymbol{\theta}$ and $\mathbf{y}$ follows (2.12) with

$$g(\mathbf{y}|\delta = 1, \boldsymbol{\tau}_\delta) f(\delta = 1, \boldsymbol{\tau}_\delta|\boldsymbol{\theta}) \ \propto \\ \mathbf{1}\{\tau_1, \tau_1 + \tau_2 \in (t_1, t_2)\} \times p_{12} f_{12}(\tau_1) p_{23} f_{23}(\tau_2) S_3(t_2 - \tau_1 - \tau_2),$$

$$g(\mathbf{y}|\delta = 2, \boldsymbol{\tau}_\delta) f(\delta = 2, \boldsymbol{\tau}_\delta|\boldsymbol{\theta}) \ \propto \ \mathbf{1}\{\tau_1 \in (t_1, t_2)\} \times p_{13} f_{13}(\tau_1) S_3(t_2 - \tau_2),$$

where $\delta = 1$ if $\boldsymbol{i} = (1,2,3)$, $\delta = 2$ if $\boldsymbol{i} = (1,3)$. Here we denote $S_i(t) = \sum_j Q_{ij}(t)$ the survival function for the sojourn time in state i. In this case, we have $S_3(t) = 1 - F_{34}(t)$ since the system can only go to state 4 afterwards. The dimension matching transformation $T_{2\to1}$ defined in (2.16) can be used with the new $p(\tau_\delta|\boldsymbol{\theta}, \mathbf{y}, \delta)$ plugged into $\alpha_{2\to1}$. As for the update within each path, Metropolis-Hastings algorithm is needed as the exact failuare case.

BIBLIOGRAPHY

Aalen, O. (1995). Phase type distribution in survival analysis. *Scandinavian Jounral of Statistics*, **22**, 447–463.

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

Aalen, O. O., Borgan, O. and Gjessing, H. K. (2008). Survival and event history analysis: a process point of view. *Springer Series in Statistics for Biology and Health*.

Abu-Lebdeh, G., Lyles, R., Baladi, G. and Ahmed, K. (2003) Development of alternative pavement distress index models. Technical Report

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

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1992). Statistical models based on counting processes. *Springer*, New York.

Andersen, P., Esbjerg, S., Sørensen, T. (2000). Multi-state models for bleeding episodes and mortality in liver cirrhosis. *Statistics in Medicine*, **19**, 587–599.

Asmussen, S., Nerman, O. and Olsson, M. (1996). Fitting phase-type distributions via the EM algorithm. *Scand. J. Statist.*, **23**, 419–441.

Atchadé, Y. (2010). A computational framework for empirical Bayes inference. *Statistics and Computing*, 1–1110.1007/s11222-010-9182-3.

Barndoff-Neilsen, O., Blaesild, P. and Halgreen. C. (1978). First hitting times models for the generalized inverse Gaussian Distribution. *Stochastic Processes and Applications*, 49–54.

Benveniste, A., Métivier, M. and Priouret, P. (1999). Adaptive algorithms and stochastic approximations. *Springer*, **22**.

Bickel, P and Doksum, K. (2001) Mathematical statistics: basic ideas and selected topics, Vol I. *Prentice Hall*.

Blossfield, H., Hamerle, A. and Mayer K. U. (1989). Event history analysis: statistical theory and application in the social sciences. *Lawence Elbaum Associates*, Hillsdale, NJ.

Bogdanoff, J. L. and Kozin, F. (1985). Probabilistic models of cumulative damage. *Wiley*, New York.

Breslow, N. E. (1975). Analysis of survival data under the proportional hazards model. *International Statistical Review*, **43**(1), 45–57.

Calvin, J. A. (1993). REML estimation in unbalanced multivariate variance components models using an EM algorithm. *Biometrics*, **49**(3), 691–701.

Cappé, O., Moulines, E. and Ryden, T. (2005). Inference in hidden Markov models. *Springer*.

Chhikara, R. S. and Folks, J. L. (1989). The inverse Gaussian distribution: theory, methodology, and applications. *Marcell Dekker*, New York.

Commenges, D. (1999). Multi-state models in epidemiology. *Lifetime Data Analysis*, **5**, 315–327.

Cox, D. R. (1999). Some remarks on failure-times, surrogate markers, degradation, wear, and the quality of Life. *Lifetime Data Analysis*, 307–314.

Cox, D. R. and Oakes, D. (1984) Analysis of survival data. *Chapman & Hall*, London.

Daniels, M. J. and Hogan, J. W. (2008). Missing data in longitudinal studies. *Chapman & Hall*, London.

Davies, A. (1998). Handbook of conditional monitoring: techniques and methodology. *Chapman and Hall*, London.

Delyon, B., Lavielle, M. and Moulines, E. (1999) Convergence of a stochastic approximation verison of the EM algorithm. *The Annals of Statistics*, **27**, 94–128.

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

Diggle, P. J. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Journal of the Royal Statistical Society*, Series C, **43**(1), 49–93.

Diggle, P. J., Liang, K. Y. and Zeger, S. L. (1994). Analysis of longitudinal data. *Clarendom Press*, Oxford.

Doksum, K. and Hoyland, A. (1992). Models for variable stress accelerated life testing experiments based on Wiener process and the inverse Gaussian distribution. *Technometrics*, **34**, 74–82.

D'Amico, G., Janssen, J. and Manca, R. (2005). Homogeneous semi-Markov reliability models for credit risk management. *Decisions in Economics and Finance*, **28**, 79–93.

Esary, J. D., Marshall, A. W., and Proschan, F. (1973). Shock models and wear processes. *Annals of Probability*, **1**, 627–649.

Fitzmaurice, G., Laird, N. M. and Ware, J. H. (2004). Applied longitudinal analysis. *Wiley*, New York.

Foucher, Y., Giral, M., Soulillou, J. and Daures J. (2007). A semi-Markov model for multistate and interval-censored data with multiple terminal events. Applications in renal transplantation. *Statistics in Medicine*, **26**, 5381–5393.

Frydman, H. (1995). Nonparametric Estimation of a Markov 'illness-death' process from interval-censored observations with application to diabetes survival data. *Biometrika*, 773–89.

Gill, R. D. (1980). Nonparametric Estimation Based on Censored Observations of a Markov Renewal process *Z. Wahrscheinlichkeitstheire*, 97–116.

Green, P. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Gu, M. and Li, S. (1998). A stochastic approximation algorithm for maximum-likelihood estimation with incomplete data. *The Canadian Journal of Statistics*, **26**, 567–582.

Gu, M. and Kong, F. (1998). A stochastic approximation algorithm with Markov chain Monte-Carlo method for incomplete data estimation problems. *Proceedings of the National Academy of Sciences* USA, **95**, 7270–7274.

Haas, R., Hudson, W. R., and Zaniewski, J. (1994). Modern pavement management. *Krieger Publishing Co.*, Malablar FL.

Hardin, J. and Hilbe, J. (2003). Generalized estimating equations. *Chapman and Hall CRC*, London.

Hobert, J. (2009). The data augmentation algorithm: theory and methodology. *Handbook of Markov Chain Monte Carlo*, S. Brooks, A. Gelman, G. Jones and X.-L. Meng, eds. Chapman & Hall CRC Press.

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

Janssen, J. and Manca R. (2006). Applied semi-Markov processes. *Springer*.

Jennich, R. I. and Schluchter, M. D. (1985). Unbalanced repeated measures models with structured covariance matrices. *Biometrics*, **42**(4), 805–820.

Jewell, N. P. and Kalbfleisch, J. D. (1996). Marker processes in survival analysis. *Lifetime Data Analysis*, 15–29.

Joseph, G. I. and Molenberghs, G. (2009). Missing data methods in longitudinal studies: A review. *Test*, **18**, 1–43.

Kalbfleisch, J. and Lawless, J. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, **80**, No. 392 (Dec., 1985), 863–871.

Kalbfleisch, J. D. And Prentice, K. L. (2002). The statistical analysis of failure time data. (Second Edition). *Wiley Series in Probability and Statistics*.

Kang, M. and Lagakos, S. (2007). Statistical methods for panel data from a semi-Markov process, with application to HPV. *Biostatistics*, **8**, 2, 252–264.

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

Kenward, M. G. and Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science*, **13**(3), 236–247.

Klein, J.P. and Moeschberger, M. L., (2003). Survival analysis: Techniques for censored and truncated data. *Springer Series in Statistics for Biology and Health*.

Lagakos, S., Sommer, C. and Zelen, M. (1978). Semi-Markov models for partially censored Data. *Biometrika*, **65**, 311–317.

Laird, N. M. (1988) Missing data in longitudinal studies. *Statistics in Medicine*, **7**, 305–315.

Laird, N. M., Lange, N. and Stram, D. (1987). Maximum likelihood computation with repeated measures: Applications of the EM algorithm. *Journal of the American Statistical Association*, **82**(397), 97–105.

Laird, N. M. and Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Lann, M. J., Robins, J. M. (2003). Unified methods for censored longitudinal data and causality series. *Springer Series in Statistics*.

Lawless, J. F. (2003). Statistical models and methods for lifetime data. (Second Edition). *Wiley Series in Probability and Statistics*.

Lawless, J. and Crowder, M. (2004). Covariates and random effects in a gamma process model with application to degradation and failure. *Lifetime Data Analysis*, **10**, 213–227.

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

Limnios, N. and Oprisan, G. (2001). Semi-Markov processes and reliability. *Statistics for industry and technology.*

Liquet, B. and Commenges, D. (2010). Choice of estimators based on different observations: modified AIC and LCV criteria. *Scandinavian Journal of Statistics*, in press.

Little. R. J. A. (1976). Inference about means from incomplete multivariate data. *Biometrika*, **63**, 593–604.

Little, R. J. A. (1988). Robust estimation of the mean and covariance matrix from data with missing values. *Journal of the Royal Statistical Society*, Series C, **37**(1), 23–38.

Little, R.J. (1995). Modeling the dropout mechanism in repeated-measures studies. Journal of the American Statistical Association, **90**, 1112–1121.

Little, R. J. A. and Rubin, D. B. (2002). Statistical analysis with missing Data. (Second edition). *Wiley*, New York.

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

Lu, C. J. and Meeker, W. Q. (1993). Using degradation measures to estimate a time-to-failure Distribution. *Technometrics*, **35(2)**, 161–174.

Lu, J., Park, J. and Yang, Q. (1997). Statistical inference of a time-to-failure distribution derived from linear degradation data. *Technometrics*, **39(4)**, 391–400.

Mclachlan, G and Peel, D (2000) Finite mixture models. *Wiley.*

Meeker. W. Q. and Escobar L. (1998). Statistical methods for reliability data. *Wiley*, New York.

Nair, V. and Wang, X. (2011) Inference for a Class of Nonhomegeneous Gaussian Processes for Degradation Data. *preprint.*

Nelson, W. (2004). Accelerated testing: statistical models, test plans, and data analysis. *Wiley Series in Probability and Statistics.*

Aalen, O. (1995). Matrix-geometric solutions in stochastic models: an algorithm approach. *The Johns Hopkins University Press*, Baltimore.

Olsson, M. (1996). Estimation of phase-type distributions from censored data. *Scand. J. Statist.*, **23**, 443–460.

Peterson, D. E. (1987). Pavement management practices. *NCHRP Synthesis 135.* Washington, D.C.: Transportation Research Board, NRC.

Phelan, M. J. (1990). Estimating the transition probabilities from censored Markov renewal processes. *Statistics and Probability Letters*, 43–47.

Puterman, M. L. (1994). Markov Decision Processes. *Wiley*, New York.

Pyke, R. (1961). Markov renewal processes with finitely many states. *The Annals of Mathematical Statistics*, **32**, 1243–1259.

Ramsay, J. and Silverman B. (2002). Applied functional data analysis:methods and case studies. *Springer Series in Statistics*.

Ramsay, J. and Silverman B. (2005). Functional data analysis. *Springer Series in Statistics*.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, **22**, 400–407.

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**, 846–866.

Robins,J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, **90**, 106–121.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, **63**(3), 581–592.

Schafer, J. L. (1997). Analysis of incomplete multivariate data. *Chapman & HallCRC Press*, London.

Schafer, J. L. and Graham, J. W. (2002). Missing data: our view of the state of the art. *Psychological Methods*, **7**(2), 147–177.

Seshadri, V. (1999). The inverse Gaussian distribution. *Springer*, New York.

Sharples, L., Jackson, C., Parameshwar, J. and Wallwork, J. and Large, S. (2003). Diagnostic accuracy of coronary angiopathy and risk factors for post-heart-transplant cardiac allograft vasculopathy. *Transplantation*, **76(4)**, 679–682.

Shiryaev, A. N. (1977), Optimal Stopping Rules. *Springer-Verlag*, New York.

Simgpurwalla, N.D. (1995). Survival in dynamic environments. *Statistical Science*, **1**, 86–103.

Sternberg, M. and Satten, G. (1999). Discrete-time nonparametric estimation for semi-Markov models of chain-of-events data subject to interval censoring and truncation. *Biometrics*, **55**, 514–522.

Tanner, M. and Wong, W. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528–550.

Themeau, T. M. and Grambsch, P. M. (2000). Modeling survival data: extending the Cox Model. *Springer*.

Titman, A. C. and Sharples L. D. (2010). Semi-Markov models with phase-type sojourn distributions. *Biometrics*, **66**, 742–752.

Verbeke, G. and Molenberghs, G. (2000). Linear mixed models for longitudinal data. *Springer*, New York.

Wang, X. (2005) Nonparametric inference with applications to astronomy and degradation modeling. *PhD Thesis*.

Wang, X. (2009). Nonparametric estimation of the shape function in a gamma process for degradation data. *Canadian Journal of Statistics*, **37**, 101–118.

Wang, X. (2010). Wiener processes with random effects for degradation data. *Journal of Multivariate Analysis*, **101**, 340–351.

Wei, G. C. G. and Tanner, M. A. (1990). A monte carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, **85**, 699–704.

Whitmore, G. A. (1995). Estimating degradation by a Wiener diffusion process subject to measurement error. *Lifetime Data Analysis*, 307–319.

Whitmore, G. A. and Schenkelberg, F. (1997). Modelling accelerated degradation data using Wiener diffusion with a time scale transformation. *Lifetime Data Analysis*, 27–43.

Whitmore, G. A., Crowder, M.J. and Lawless, J.F. (1998) Failure Inference from a Marker Process Based on a Bivariate Wiener Model. *Lifetime Data Analysis*, 229–251.

Yang, Y., Atchadé, A., and Nair, V. (2011). Paramteric inference for multistate semi-Markov processes with censored data. (preprint).

Yang, Y. and Nair, V. (2011). Inference for time-to-failure with multistate data: A comparison of traditional and process-based approaches. *Canadian Journal of Statistics* (tentatively accepted).

Yang, Y. and Nair, V. (2011). Analysis of degradation data with missing patterns (in preparation).

Younes, L. (1988). Estimation and annealing for gibbsian fields. *Annales de l'Institut Henri Poincaré. Probabilité et Statistiques*, **24**, 269–294.