UNIVERSITY OF CALIFORNIA,
IRVINE

A Statistical Approach to Detecting Patterns in Behavioral Event Sequences

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Statistics

by

Kevin Andrew Heins

Dissertation Committee:
Professor Hal Stern, Chair
Professor Daniel Gillen
Professor Padhraic Smyth

2014

DEDICATION

To Jessica

TABLE OF CONTENTS

Page

LIST OF FIGURES **v**

LIST OF TABLES **viii**

ACKNOWLEDGMENTS **xi**

CURRICULUM VITAE **xii**

ABSTRACT OF THE DISSERTATION **xiii**

1 Introduction **1**
- 1.1 Pattern Detection Models . . . 4
  - 1.1.1 Pattern Detection in Discrete Time . . . 4
  - 1.1.2 Pattern Detection in Continuous Time . . . 6
- 1.2 Maternal Data . . . 7
- 1.3 Outline . . . 10

2 A Renewal Process Model for Pattern Detection **11**
- 2.1 Event History Analysis . . . 12
- 2.2 Model Description . . . 13
  - 2.2.1 Background Processes . . . 14
  - 2.2.2 Sequence Processes . . . 18
  - 2.2.3 Combining Background and Sequence Processes . . . 22
- 2.3 Inference . . . 24
  - 2.3.1 Prior Distributions . . . 25
  - 2.3.2 MCMC . . . 27
  - 2.3.3 Model Selection . . . 28
  - 2.3.4 Convergence . . . 30
- 2.4 Simulations . . . 32
  - 2.4.1 Proof Of Concept . . . 32
  - 2.4.2 Realistic Scenarios . . . 35
  - 2.4.3 Effect of Sample Size . . . 39
  - 2.4.4 Detecting Longer Sequences . . . 40
  - 2.4.5 Effect of Prior Distribution on Time Parameter . . . 42
- 2.5 Maternal Behavior Analysis . . . 45

2.6 Discussion 51

3 Population Level Pattern Analysis **53**
3.1 A Hierarchical Renewal Process Model 54
3.2 Population Distributions for Sequence Processes 57
3.3 Inference 59
3.4 Experiments 61
3.4.1 Efficacy of the Hierarchical Renewal Process Model 61
3.4.2 Detecting Sequences with Varying Levels of Prevalence 65
3.5 Maternal Behavior Analysis Revisited 66
3.5.1 Rodent Maternal Behavior Analysis 67
3.5.2 Human Maternal Behavior Analysis 70
3.6 Discussion 72

4 Renewal Processes with General Probability Distributions **74**
4.1 Parametric Renewal Processes 76
4.2 Non Parametric Renewal Processes 79
4.3 Experiments 85
4.3.1 Effectiveness of the Nonparametric Model for Point Process Data . . 85
4.3.2 Fitting Different Distributions 86
4.3.3 Robustness Against Misspecified Sequence Processes 91
4.4 Maternal Behavior Analysis Revisited 93
4.4.1 Weibull Model Results 93
4.4.2 Nonparametric Model Results 98
4.5 Discussion 102

5 Conclusion **104**

Bibliography **106**

LIST OF FIGURES

Page

1.1 Displays of behavior data. There are 15 events types, here denoted by integers, which are indicated on the vertical axis. The horizontal axis gives time in seconds. Each dot represents a recorded event. Top and bottom panels are the same, with patterns identified in the lower plot. 3

2.1 An example containing three background processes. The figure on the right shows the three competing events at time T_{i-1}. The figure on the left shows the three competing events at time T_i, including the newly drawn interarrival time for event A. 16

2.2 An example of the data generating process for a model that includes the sequence $A \to B$. On the left, the event type A just occurred at time T_i, which triggers the sequence $A \to B$. On the right, a new time increment is drawn for both event type A as well as event type B, which belongs to the sequence. 19

2.3 An example of simulated data for the proof of concept simulation. Three different events are present, with patterns identified by lines. 33

2.4 An example of simulated data consistent with a mother who is fragmented and unpredictable. There are 20 possible events, with three short patterns. 35

2.5 An example of simulated data consistent with a mother who is consistent and predictable. There are 20 possible events, with six patterns. 36

2.6 Density plots of the prior and posterior distributions corresponding to four choices of prior distributions for the maximum time parameter τ. Note that the prior distribution for both plots on the bottom have been scaled for visibility. 43

2.7 An example of maternal behavior data. The three patterns seen in this example include New Toy $\to$ Manipulating Toy, Instructive Speech $\to$ Instructive Speech, Positive Speech $\to$ Positive Speech, and New Toy $\to$ Look at Toy $\to$ Look at Baby. 46

3.1 The population level prior distribution that we consider for the sequence process rates. It is a mixture model with two components, one near zero indicating no pattern is present, and one centered around a nonzero rate. 59

3.2 Estimates (posterior mean) for the background rate parameter of the type 1 event. The figure includes results from the individual level analysis (left), results from the hierarchical model (middle), and the true parameters (right). Most parameters display some level of shrinkage under the hierarchical model. The number of total events, as well as the number of type 1 events, are indicated on the far left. 64
3.3 Estimates (posterior mean) for the sequence rate parameter for the $3 \rightarrow 4$ sequences. The figure includes results from the individual level analysis (left), results from the hierarchical model (middle), and the true parameters (right). Most parameters display some level of shrinkage under the hierarchical model. The number of total events, as well as the number of patterns, are indicated on the far left. 65
3.4 Estimates (posterior mean) of the background rate parameter for Instructive Speech. The figure includes results from the individual level analysis (left) and results from the hierarchical model (right). Most parameters display some level of shrinkage under the hierarchical model. 71
3.5 Estimates (posterior mean) of the sequence rate parameter for the pattern New Toy $\rightarrow$ Play with Toy. The figure includes results from the individual level analysis (left) and results from the hierarchical model (right). The rates are generally smaller for the hierarchical model than the individual model. The mothers who did not appear to display the sequence from the individual model results have estimated rates that indicate the pattern might actually occur, just not as often. 72

4.1 Simulated data from three distributions. 75
4.2 The different shapes of the Weibull hazard function: Constant, Decreasing, and Increasing. 77
4.3 The different shapes of the gamma hazard function: Constant, Decreasing, and Increasing. 78
4.4 The different shapes of the lognormal hazard function: Constant, Decreasing, and Increasing. 79
4.5 Simulated step functions plotted for various values of the smoothness parameter c. 83
4.6 The results of fitting our nonparametric model for varying hazard functions. The red line indicates the true hazard. The black lines are sampled draws from the posterior distribution of the hazard function. 87
4.7 Simulated data that incorporates a variety of background processes with differing hazard functions, along with four different sequence processes. 88
4.8 The distribution of the number of patterns per mother for each of the three models we consider in this section: the exponential, Weibull, and nonparametric RPM. 94
4.9 The distribution of the maximum pattern length per mother for each of the three models we consider in this section: the exponential, Weibull, and nonparametric RPM. 95

4.10 The data for an example mother, including lines indicating patterns found with the nonparametric model. The patterns include : New Toy (New) → Manipulating Toy (Manip), Smiling (Sm) → Laughing (L), Set Down Baby (Set) → New Toy (New), and Affectionate Touch (Aff) → Instructive Speech (Ins). 100
4.11 Samples from the posterior distributions of the hazard function for two background processes. Instructive speech shows an increasing hazard, while positive speech shows a decreasing hazard. 101
4.12 Samples from the posterior distributions of the hazard function for the background process corresponding to the event Laughing. The event Laughing generally comes in bursts, but always with a small gap between the individual events, causing a spike in the hazard function. 102

LIST OF TABLES

Page

2.1 The parameter values for each of the five parameters in our first simulation experiment. The estimated mean is the average over the posterior means across the simulations. Standard error (S. E.) and coverage probability (Cov.) are also included for each parameter. 33
2.2 Performance summaries for $M = 100$ simulated data sets 34
2.3 Results for background process rate parameters in the second simulation scenario (three simple patterns). The estimated mean is the average over the posterior means across the simulations. Standard error (S. E.) and coverage probability (Cov.) are also included for each parameter. 37
2.4 Results for sequence process rate and time parameters in the second simulation scenario (three simple patterns). The estimated mean is the average over the posterior means across the simulations. Standard error (S. E.) and coverage probability (Cov.) are also included for each parameter. 37
2.5 Results for background process rate parameters in the third simulation scenario (six patterns, including longer ones). The estimated mean is the average over the posterior means across the simulations. Standard error (S. E.) and coverage probability (Cov.) are also included for each parameter. 37
2.6 Results for sequence process rate and time parameters in the third simulation scenario (six patterns, including longer ones). The estimated mean is the average over the posterior means across the simulations. Standard error (S. E.) and coverage probability (Cov.) are also included for each parameter. . . 38
2.7 The results of our sample size experiment. The true parameter values of the rates for the 10 sequence processes are given. For each sample size, the percentage of times that a pattern is identified, out of 100 simulations each, is also given. At the bottom, the total number of detected patterns and the number of false positives for each sample size are given. 40
2.8 Percentage of simulated data sets where the model correctly identified the pattern. While longer patterns can be difficult to detect, there is some dependence on the size of the data set. 41
2.9 On the left: The parameters for each of the six prior distributions that we consider for the τ parameter. On the right: the posterior mode and 95% credible interval for each of the posterior distributions. 42
2.10 The possible event types in the maternal behavior data set. There are a total of 24 different event types. 46

2.11 The number of mothers that display various patterns as identified with the descriptive statistics method. Results are shown for three values of τ. 48
2.12 Distribution of mothers who exhibited a given number of patterns. 48
2.13 Distribution of mothers who had a given maximum pattern length. 49
2.14 Distribution of mothers who had a given percentage of patterns that were longer than two events. The highest percentage was 60%. 49
2.15 The most common patterns discovered in the maternal behavior data using the Renewal Process Model. The number of mothers who exhibited a given pattern is also provided (out of a total of 121 mothers). 49

3.1 Population parameters for the four patterns. All four patterns are assumed to occur for each simulated mother, hence no prevalence parameter. 62
3.2 Population parameters for the four patterns. All four patterns are no longer expected to occur for each mother, thus we estimate the prevalence of each pattern. 63
3.3 For each of nine different simulation scenarios, each with a different prevalence level, 100 simulations were run. The percentage of times a pattern was discovered out of total of 100 simulations is displayed. with varying levels of prevalence. 66
3.4 The population-level background rate parameters for all events in both control and fragmented rats. Both the posterior mean and 95% credible interval are included. There is overlap in the credible intervals for all events but nest building, indicating that the rates are relatively similar. 68
3.5 The population-level sequence process rate parameters for all patterns in both control and fragmented rats. Both the posterior mean and 95% credible interval are included. The rate parameters for the control rats are uniformly larger, indicating that the patterns occur more often. 69
3.6 The population-level sequence process time parameters for all patterns in both control and fragmented rats. Time indicates the maximum amount of time between events in a sequence in seconds. Both the posterior mean and 95% credible interval are included. 69
3.7 Each of the six patterns discovered by the hierarchical model. The prevalence column is the posterior mean of the prevalence parameter. The individual model percentage refers to the percentage of mothers, out of 121, who displayed that pattern according to the individual level RPM from Chapter 2. . 70

4.1 For each of the exponential, Weibull, and nonparametric models, the percentage of 100 simulations where a given pattern was successfully detected. 89
4.2 The estimated shape and scale parameter determined using a Weibull RPM for each of three background processes. Posterior mean refers to the average of the posterior means over the 100 simulations. Standard error was calculated directly from the posterior means, and coverage was calculated using the 95% credible intervals. 90

4.3 The proportion of simulations where each model successfully identified each pattern. 92
4.4 Distribution of mothers who had a given percentage of patterns that were longer than two events under the Weibull model. The highest percentage was 60%. 95
4.5 The most common patterns discovered in the mothers using the Weibull model. On the right is the number of mothers who exhibited a given pattern, out of 121, which are noticeably smaller than in the exponential model. 96
4.6 Distribution of mothers who had a given percentage of patterns that were longer than two events under the nonparametric model. The highest percentage was 60%. 99
4.7 The most common patterns discovered from the mothers using the nonparametric model. On the right is the number of mothers who exhibited each pattern, out of a total of 121 mothers. 99

ACKNOWLEDGMENTS

I would like to thank my adviser Hal Stern for all the help and guidance that he has provided me over the last several years. I greatly appreciate all the time and effort he has given me. This dissertation would not be possible without him.

I would also like to thank Tallie Z. Baram, Curt Sandman, Elysia Davis, and everyone else at the UCI Conte Center on Brain Programming in Adolescent Vulnerabilities for our interesting collaborations and providing us with wonderful research problems and data. Furthermore, this research would not be possible without the support of NIH grant MH096889.

I would like to thank all my professors at UCI for their instruction and guidance, especially during my first years at UCI, and both Rosemary Busta and Lisa Stieler for their help and administrative support over the years. I would also like to thank all of my fellow grad students, who have provided me with ceaseless help, distractions, and friendship.

I would like to thank both Susan Paddock and John Adams for their guidance and help as my mentors during my summer at RAND.

I would like to thank my mom and dad, as well as the rest of my family for their unconditional love and support.

Finally, I would like to thank my wife Jessica for always being by my side throughout grad school, and giving me the motivation to finish this dissertation.

CURRICULUM VITAE

Kevin Andrew Heins

EDUCATION

Doctor of Philosophy in Statistics **2014**
University of California, Irvine *Irvine, California*

Master of Science in Statistics **2014**
University of California, Irvine *Irvine, California*

Bachelor of Science in Mathematics–Probability and Statistics **2007**
University of California, San Diego *La Jolla, California*

RESEARCH EXPERIENCE

Graduate Research Assistant **2009–2014**
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant, Statistics **2009–2014**
University of California, Irvine *Irvine, California*

REFEREED CONFERENCE PUBLICATIONS

A statistical model for event sequence data **Apr 2014**
Seventeenth International Conference on Artificial Intelligence and Statistics (AISTATS)

ABSTRACT OF THE DISSERTATION

A Statistical Approach to Detecting Patterns in Behavioral Event Sequences

By

Kevin Andrew Heins

Doctor of Philosophy in Statistics

University of California, Irvine, 2014

Professor Hal Stern, Chair

The identification of recurring patterns within a sequence of events is an important task in behavior research. In this thesis, we develop a probabilistic framework for identifying such patterns from behavioral data. This framework allows us to distinguish between events that belong to a pattern and events that occur as part of background or unstructured behavior. Stochastic processes are introduced to describe the incidence of both background events and events that belong to recurring patterns. The behavioral events are modeled together using a competing risks framework combining the stochastic processes. We develop an inference procedure to detect the sequences present in observed data. The motivation for this work comes from a large scale longitudinal study to assess the impact of fragmented and unpredictable maternal behavior on emotional and cognitive development of children. We describe our results on both simulated data and the maternal data.

We also consider extensions to our model. We develop a model to study population level behavior, allowing us to compare separate populations as well as pool information across individuals. To perform this analysis, we describe how to extend both our model and inference procedure to a hierarchical setting. We describe results for both simulated data and the maternal data.

Finally, we explore the distributional assumptions inherent in our model. We consider a

variety of parametric forms for our model, as well as a nonparametric approach that is both flexible and computationally efficient. In addition to improved model fit, these approaches allow us to better describe background behaviors, such as behaviors that occur in bursts or with strict regularity. We also explore the effect of distributional assumptions on both simulated and real data.

Chapter 1

Introduction

The study of behavior plays an important role in a variety of fields, such as psychology, neuroscience, sociology, and zoology. However, most behavior studies are qualitative in nature, as quantitative constructs describing behavior have proven difficult to find. In particular, a great deal of subjectivity is often needed when determining what does or does not constitute an interesting or scientifically relevant behavior, and when attempting to differentiate between various types of behaviors. Researchers often want both to predict when an actor will exhibit various behaviors, and to use past behavior to predict other characteristics. Thus a quantitative construct of behavior would be beneficial to help drive these fields when investigating behavior.

One possible approach to the study of behavior that has proven popular is the identification of recurring behavioral patterns, which are repeated sets of events that occur in a specific order. These have proven useful for characterizing temporal streams of behavior (Eibl-Eibesfeldt, 1970). Most analyses identify patterns of interest a priori, and then attempt to locate these patterns in a behavioral record. Approaches to automatically detect relevant patterns are in their infancy. In this thesis, we present a probabilistic model which attempts to model some

of these intricacies of behavior by identifying repeated patterns within a sequences of events. We posit that such patterns describe characteristic elements of the behavior of an individual or individuals. Once identified, the patterns can be utilized in additional analyses.

To establish some terminology for our problem of interest, let a behavioral sequence be defined as a series of events performed by an actor or actors, with each event occurring at some point in continuous time. The data we consider can be represented as a sequence of time-event pairs, including the type of event and the time that it occurred. Event types are defined from a fixed set of possible behaviors identified by subject matter experts. Patterns are defined as an ordered set of specific events occurring relatively close together in time, which occur multiple times within a behavioral sequence.

Figure 1.1 provides a graphical representation of a behavioral sequence. The upper plot displays the initial data, with no patterns identified. Time in seconds is represented on the horizontal axis. Different events types are denoted by integers on the vertical axis. The point process for each value on the vertical axis represents the events for that particular event type. The lower plot displays the same data, but includes lines representing patterns. For example, the pattern consisting of event types 11, 14, and 15, which we denote $11 \rightarrow 14 \rightarrow 15$, occurs three times.

One might hope that patterns would be easy to detect, perhaps through visual inspection of the data. However, as noted in Magnusson (2000), it is relatively difficult to identify patterns from this type of data via visual inspection alone. Typically analysts resort to guessing which patterns they expect may exist, often based on past research or observation, and then note how often the expected patterns occur. However, this has obvious drawbacks, as proposed patterns may not actually exist, and more importantly, patterns may exist that are not hypothesized.

The aim of this thesis is to provide a principled approach to automatically identifying pat-

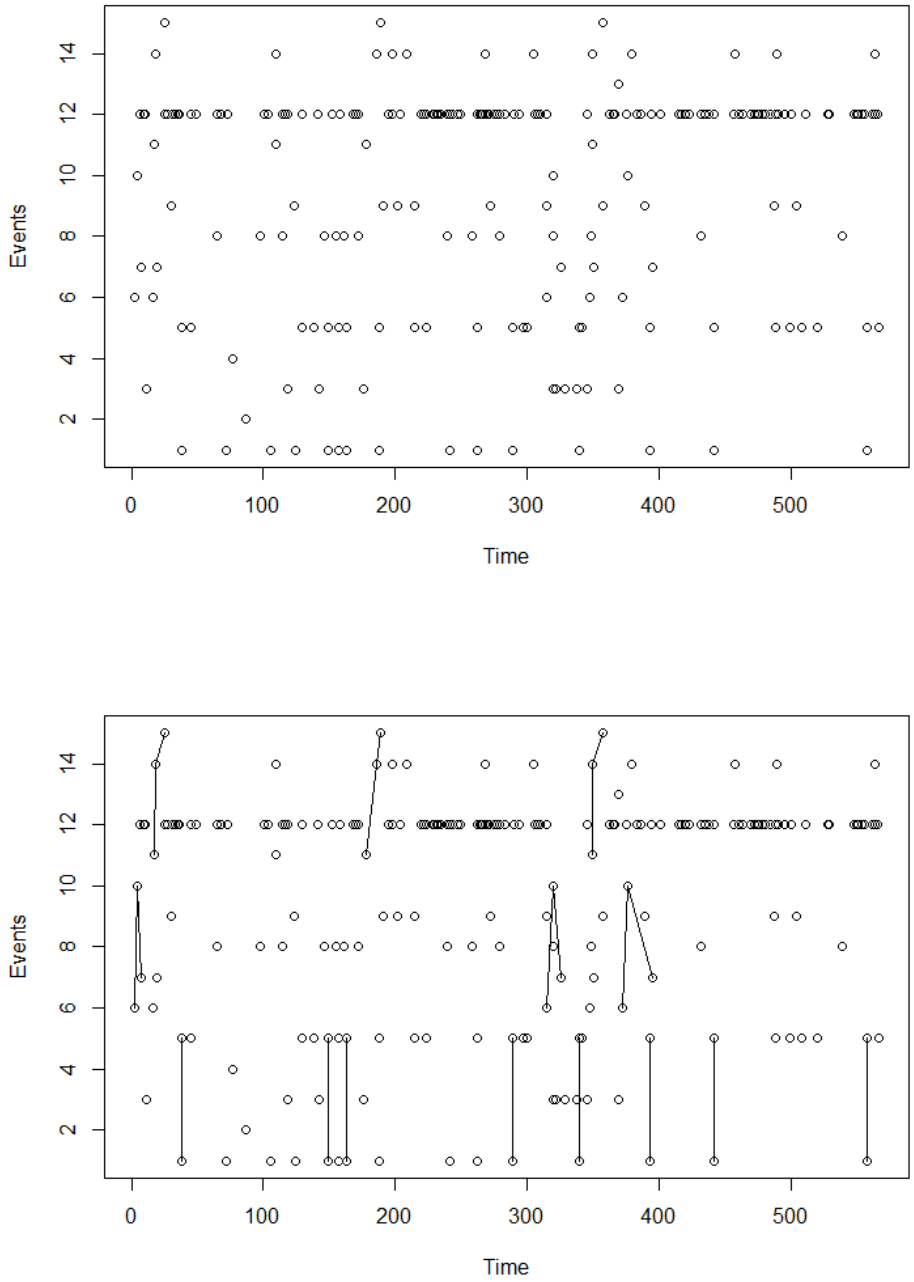

Figure 1.1: Displays of behavior data. There are 15 events types, here denoted by integers, which are indicated on the vertical axis. The horizontal axis gives time in seconds. Each dot represents a recorded event. Top and bottom panels are the same, with patterns identified in the lower plot.

terns from behavioral sequences without the need for prior definition of relevant patterns. We achieve this by introducing a probabilistic model that can automatically extract these repeated patterns in sequences of time-indexed events. The existence of such patterns, as well as their features, often carry scientific significance, such as in behavior studies.

1.1 Pattern Detection Models

There are a range of techniques for identifying patterns in sequences of observations. Many of the methods focus on identifying patterns in observed data from among a candidate set of patterns that are specified a priori. This allows a more focused analysis and enables the inclusion of more complex behaviors, but limits the ability to fully capture the behavior of an individual. If a pattern is not hypothesized a priori, then it will not be included in the analysis. We do not discuss finding known patterns further; see Marcum & Butts (2014) for additional details.

We want to explicitly learn all the patterns that exist, as well as detect their occurrence for certain individuals. Methods for this purpose often assume that time is discrete, which allows the analysis to focus solely on the order of events. A small but growing literature also exists for methods that identify patterns in continuous time, which is what we consider for our model. We describe a number of such techniques below, and evaluate their suitability for our applications and data.

1.1.1 Pattern Detection in Discrete Time

We first discuss a number of methods for detecting patterns in discrete time. For example, one could consider a set of behavioral events that occur in order, and discard the time the occurs between events. Then the focus is on a set of ordered events and not on the time at

which they occur. Indeed, in many such cases, time does not actually exist in the traditional sense. For example, DNA and protein sequences exist as an ordered set of events but contain no time component.

A rich literature exists on pattern finding within DNA and protein sequences. Repeated patterns, known as motifs, are important for understanding gene expression (Kellis et al., 2004; Liu et al., 1995). A motif finding algorithm typically operates by first establishing a background model, where each state (such as a base or residue) is assigned some probability from a zero or higher order Markov process. Motifs are then found using a variety of strategies. For instance, the expected frequency of all motifs of a certain length can be calculated, and motifs that occur more often than expected from the background model will be included in the final set of motifs. However, this requires one to enumerate all motifs of a given length, which is often impractical. An alternative strategy is to only enumerate motifs with shorter lengths, and then build up longer motifs using either an expectation maximization or Gibbs sampling algorithm.

Data compression methods can be applied to discrete time series to identify recurring patterns, such as the Lempel-Ziv algorithm and its derivatives (Welch, 1984; Ziv & Lempel, 1978). The goal of these methods is lossless compression, whereby the original sequence is represented with fewer bits without a loss of information. The algorithm first creates a dictionary of recurring patterns from the data. Then, whenever a given pattern occurs in the sequence, the algorithm replaces the pattern with a new event type to represent the pattern's occurrence. This ultimately decreases the number of bits required to represent a full sequence of events, as several events are subsumed into pattern events.

1.1.2 Pattern Detection in Continuous Time

The study that motivates our work attempts to search for patterns in sequences of events measured in continuous time. In cases like this one where continuous time points exist, discrete time series methods could be implemented by simply removing the time element, and modeling the order in which the event occur. Alternatively, time could be discretized into equally sized time bins, with several 'no action' events assigned to bins where no events are recorded. Both of these approaches have limitations. For example, the former method could fail when the time differences contain useful information, and different scales of discretization could have significant effects on the results obtained by the latter approach.

Time series pattern finding for continuous time data (e.g., looking for change points) usually occurs in the context of continuous measurements made at regular time points. This allows the use of standard time series models as the foundation for a model. However, our data are categorical measurements occurring at irregular time points. This type of data could instead be modeled as a series of states, such as a Markov process, but this would be insufficient for addressing our scientific questions. Instead, we would like to model each action as an instantaneous event, and then identify temporal relationships between several events.

There are a growing number of models that attempt to model events in this way, but most do not have a component that explicitly accounts for patterns. For example, a range of models have been developed from Hawkes processes, which are similar to Poisson processes, but have an excitatory feature (Hawkes, 1971). Each event increases the rate of future events, either of the same event type (self-exciting) or different event types (mutually-exciting). Hawkes process models have been used in many different application areas, such as models of financial transactions, earthquakes, and crime (Engle & Russell, 1998; Ogata, 1988; Mohler et al., 2011).

Simma & Jordan (2010) model events based on cascades of Poisson processes, which retains

the excitatory feature in an attempt to identify which events trigger other events. They consider a background Poisson process as a baseline, as well as numerous triggered Poisson processes, which are initialized after certain events and feature a decay function. They consider a number of applications, such as social networks where one event (e.g. a posted message) may trigger several resulting events. While similar in spirit to the model we describe in Chapter 2, where we model patterns as processes triggered by background events, their model focuses on improving model fit to the observed sequence of events rather than identifying specific recurring patterns.

We know of one existing method to specifically extract patterns from continuous data: the software package Theme (Magnusson, 2000). Theme partitions the data into time bins, and assumes that the events occur uniformly in those bins. Then, using a bottom up search procedure, it identifies patterns that appear more often than expected under the uniform model via repeated hypothesis testing. The number of patterns identified in this way depends of the significance level. Our experience is that for the default significance level, there are many false positives. To account for this, Theme includes a permutation procedure which provides a distribution for the number and length of patterns that would be expected under randomized data. We can then compare the distribution of the patterns in the original data to determine how many were due to random error, such as from the hypothesis testing. However, the assessment of which specific patterns are genuine and which are false positives is not explicit in Theme's output. Our model aims to be more explicit about which patterns are discovered, while not relying on hypothesis testing or discretization.

1.2 Maternal Data

The motivation for our model comes from a study on the effect of maternal sensory signals on child development. Previous experimental research has identified fragmented and unpre-

dictable maternal behavior in rodents as a risk factor for outcomes that are analogous to emotional and cognitive disorders in humans (Baram et al., 2012). However, how to best characterize fragmented and unpredictable behavior in humans in clinically meaningful ways remains an open research question. One proposed method is to identify the patterns that a mother follows in her behavior. Fragmented and unpredictable behavior would then be defined by statistics derived from each mother's set of patterns, such as the number and length of recurring patterns.

In this context, we think of consistent behavior as being characterized by long repeated patterns, whereas fragmented behavior likely has few, if any, long repeated patterns. Unpredictable behavior is characterized by the inability to determine the mother's next action after considering the previous actions performed by her and the child.

Our data comes from a longitudinal study conducted by the NIMH-funded Conte Center on Brain Programming in Adolescent Vulnerabilities at UC Irvine, which is attempting to examine the impact of early life interactions between a mother and her child on the child's development through adolescence. The center is collecting data from both humans and rodents, and using a wide array of data types, including behavioral data, genetic data, and brain imaging data. In this thesis, we only consider behavioral data. We will focus primarily on human data, but also consider rodent behavioral data in Chapter 3.

For the human studies, our collaborators invited the participant mothers and their children to the lab, where our collaborators video recorded the mothers playing with their children. The data derives from these short videos, each about 10 minutes long. Each video was annotated by the researchers to include the type of each event and the time when it occurs. Event types include changes in the mother's expression or emotional state (smiling, content, bored), where she is looking, and physical interactions (hug, play with toy, pick up child, speak).

The human data that we consider comes from an initial cohort of individuals that was collected using an earlier funded study, prior to the Conte Center. The initial data that is available to us is comprised of 121 mothers playing with their 12 month old child. Further data is currently being collected, including repeated observations from the same mothers and children. One of the goals of this thesis is to identify recurring patterns from this data, which can then be used to derive numerical summaries of maternal behavior. Ideally, these summary statistics would help characterize each mother for the duration of this study.

An additional goal of the Conte Center is to validate measures of fragmented behavior on rodents, where fragmented behavior can be experimentally induced. We will not focus on this issue in this thesis, but will explore the rodent data in Chapter 3 to help illustrate one of our techniques.

Data was collected from rats in a fashion similar to human mothers. Fragmented behavior was induced in experimental rats by providing them with limited bedding materials. This causes the mother rats (dams) to become extremely stressed, causing abnormal erratic behavior. Both control and experimental dams were observed, but not video recorded, for a total of two hours each day, for a total of eight days. A total of seven event types were recorded for the dams, such as eating, grooming her pups, or nursing. Observations were made from days two through nine after the birth of the pups, which are roughly analogous to early childhood in humans. We currently have data from a total of twelve dams, including six controls and six experimental rats. We will briefly explore the differences in behavior between the two populations of rats in Chapter 3.

1.3 Outline

In Chapter 2, we present a renewal process model for identifying patterns from behavioral data. We also describe the inference procedure that we use to detect patterns and estimate the parameters that characterize their behavior. The rest of this thesis considers various extensions of the basic model. Chapter 3 presents a hierarchical version of our model, to allow for a population level analysis of patterns. Chapter 4 evaluates the strengths and weaknesses of using various probability distributions to characterize interarrival times in our model, as well as propose a nonparametric alternative. Finally, in Chapter 5 we present concluding remarks and discussion.

Chapter 2

A Renewal Process Model for Pattern Detection

In this chapter, we develop a renewal process model (RPM) for the detection of patterns within behavioral time-event data. We develop the model by specifying a data generating process with the properties we expect in our data. The first component of this process assumes background (i.e. non-pattern) events occur randomly in time, with the occurrences of each type of event governed by an independent stochastic process. The second component of the data generating process incorporates stochastic processes that generate patterns of events. Essentially, our model seeks to detect pattern sequences by identifying a sequence of events that occur together more often than would be expected according to the background model.

First, we briefly review details from event history analysis that are important to our model. Next we describe the model in detail, followed by a description of our inference procedure. We then explore results from both simulation experiments and our maternal data. We finish this chapter with a discussion on possible future directions. A portion of the content of this

chapter follows Heins & Stern (2014).

2.1 Event History Analysis

Before we describe our model, we first describe concepts from event history analysis, one possible framework to model a discrete set of events that occur in continuous time. Event history analysis, or survival analysis, is a general statistical framework used in a multitude of areas, such as medical studies. We focus on a stochastic process formulation of survival analysis models (Aalen et al., 2008).

The stochastic process framework assumes that events occur according to a point process. The time between events is known as the interarrival time, and for the purposes of this thesis, we assume each interarrival time is independent of other interarrival times. Processes with independent time increments are known as renewal processes, and are uniquely characterized by the distribution of the time increments.

We denote the time that event i in the point process occurs as T_i. Because the times T_i are strictly increasing, such that $T_i > T_{i-1}$, we then define the random interarrival times as $t_i = T_i - T_{i-1}$.

Renewal processes are typically specified by the probability density of the interarrival times. Alternatively, they may be specified by transformations such as the hazard or survival functions, which are all unique to a given distribution. All three functions are instrumental to our model, so we next discuss them.

The distribution function is the probability that an event occurs before some time t. The distribution function of an interarrival time, t_i, can then be defined as $F(t) = \Pr(t_i < t)$. The survival function is the complement of the distribution function, and hence the probability

that an event occurs after some time t. The survival function is defined as $S(t) = 1 - F(t) = Pr(t_i > t)$.

If we know that t_0 time units has passed since the last event, then we can define the conditional survival function. This is the conditional probability that a time increment $t_i > t$, given that we also know $t_i > t_0$. We denote the conditional survival function as $\bar{S}(t|t_0)$, and define it as:

$$\bar{S}(t|t_0) = Pr(t_i > t \,|\, t_i > t_0) = \frac{S(t)}{S(t_0)} \tag{2.1}$$

The hazard function $\lambda(t)$ is defined as the instantaneous rate of events, given that an event has not yet occurred by time t. Hence it can be defined as $\lambda(t) = f(t)/S(t)$. It can also be derived via the survival function, such that $\lambda(t) = -\frac{d}{dt}\log(S(t))$. Given both this definition of the hazard function and the definition of the conditional survival function in Equation 2.1, it is straightforward to show that conditioning on $t_i > t_0$ does not affect the hazard function.

The probability density of a interarrival time $f(t) = F'(t)$ is equal to the product of the survival and hazard functions, which is clear from the first definition of the hazard function given above. Thus, the density of the interarrival times is $f(t) = \lambda(t)\, S(t)$.

2.2 Model Description

We describe our model and our approach to inference by focusing on the model's data generating process. Events are assumed to be generated by competing renewal processes, which we classify as either background or sequence processes. The model assumes most

observed events occur independently of other recent events; these are assumed to belong to the background processes. Other events are assumed to occur as part of a sequence; these events are generated by sequence processes.

First, we establish a model with only background processes. We then discuss sequence processes independent of background processes, and finally explain how to combine the two into a single model.

We consider data comprised of a set of n events that occur over some period of time. Events are observed as time-event pairs $\{E_i, T_i\}$ with occurrence times $T_i \in \mathbb{R}_+$ and event types $E_i \in Q$, where $Q = \{1, 2, ..., J\}$ is a set of possible event types. Occurrence times are strictly increasing, so $T_i > T_{i-1}$, and we define event interarrival times as $t_i = T_i - T_{i-1}$. While events are observed as $\{E_i, T_i\}$, our model is most easily described in terms of the equivalent interarrival time-event pairs $\{E_i, t_i\}$.

2.2.1 Background Processes

We first describe the background process corresponding to a single event type. For the event type j, we define its background process as a continuous time renewal process with independent and identically distributed interarrival times from some arbitrary strictly-positive distribution $t_i \sim F^j(t_i)$. For example, if $F^j(\cdot)$ is defined as an exponential distribution, the resulting process would be a Poisson process. For subsequent sections, we assume that $F^j(\cdot)$ is arbitrary, but only consider $F^j(\cdot)$ to be exponential in the simulation experiments and applications of this chapter. We explore more general distributions in Chapter 4.

Our data are assumed to include J possible event types. For each event type, we assume the events belong to an independent background process, each with its own unique interarrival time distribution. Thus we consider a multi-event model with several independent processes,

each one corresponding to an event in Q. We will distinguish the interarrival times of each of these processes by a superscript, so for event type $j \in Q$ we have $t_i^j \sim F^j(t_i)$.

We consider a competing risks framework to generate the event sequence $\{E_i, t_i\}_{i=1,\ldots,n}$ from the independent renewal processes. In this scenario, each of the possible actions in Q is a competing event, where the first to occur will be labeled as the next event to occur in our observed sequence. The competing risks framework will prove especially useful when we incorporate sequence processes in the next subsection.

Let the initial time be T_0. Then we want to know the first event-time pair $\{E_1, t_1\}$ to occur in the sequence. The data generating mechanism generates initial interarrival times from the J event processes, denoted $\left\{t_1^1, ..., t_1^J\right\}$. The events are sorted into a schedule according to increasing times, and we observe the first scheduled event next. Thus the time until the next event can be modeled as the minimum over the interarrival times $t_1 = \min_{j \in Q} \left\{t_1^j\right\}$ and the corresponding event type is labeled the next event $E_1 = \arg\min_{j \in Q} \left\{t_1^j\right\}$. Once the pair $\{E_i, T_i\}$ has been established, we need to determine the next event $\{E_{i+1}, T_{i+1}\}$.

An illustration of how this happens in the competing risks approach can be seen Figure 2.1. The left side of the figure shows three interarrival times active at time T_{i-1} corresponding to three different background processes. The smallest corresponds to event A, so the next event E_i will be event type A. The corresponding time T_i will occur at $T_{i-1} + t_i^A$.

After the time increment for event type A has been established as the next event to occur, we need to draw a new time increment for event type A, as seen on the right side of 2.1. The time increments for event types B and C remain unchanged. The next scheduled times for both events B and C remain unchanged, but our model must address the fact that these events did not occur between the times T_{i-1} and T_i. We next explain how this is done.

Suppose we are at time T_i and have just observed the pair $\{E_i, t_i\}$. Consider the process by which $\{E_{i+1}, t_{i+1}\}$ is generated. Recall that the next event in each background process

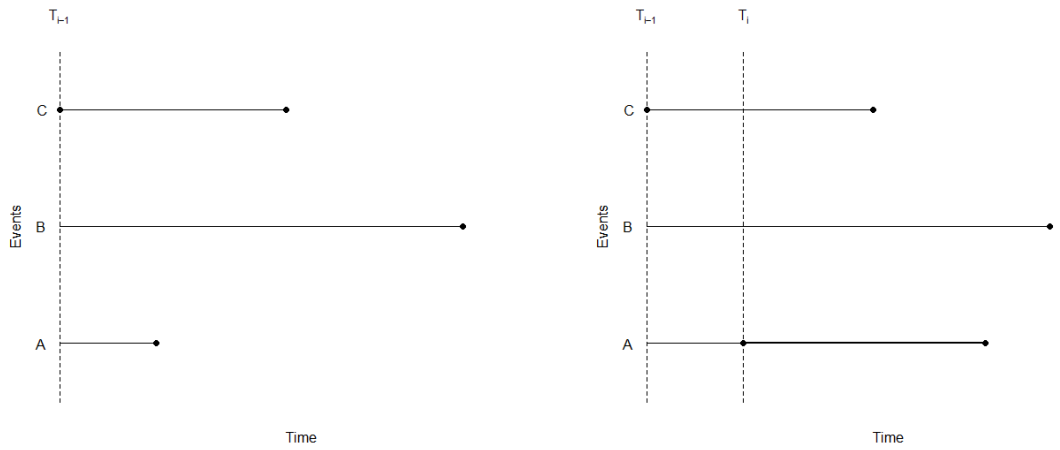

Figure 2.1: An example containing three background processes. The figure on the right shows the three competing events at time T_{i-1}. The figure on the left shows the three competing events at time T_i, including the newly drawn interarrival time for event A.

is already scheduled to occur at t_i^j for $j \neq E_i$. We keep these other events that were already scheduled, $\left\{t_i^j,\, j \neq E_i\right\}$ in the queue of possible events. However, two additional steps are required before determining the next time-event pair. First we need to account for the fact that these already scheduled events did not occur between times T_{i-1} and T_i. We update the corresponding interarrival times by decrementing each time by the value of t_i, $t_{i+1}^j = t_i^j - t_i^{E_i}$ for $j \neq E_i$. Second, we need to incorporate the next event from the process corresponding to E_i, so we draw a new interarrival time $t_{i+1}^{E_i} \sim F^{E_i}\left(t_{i+1}\right)$ and append it to the set of interarrival times. The new set of active interarrival times $\left\{t_{i+1}^{E_i},\, \left(t_i^j - t_i^{E_i}\right)_{j \neq E_i}\right\}$ can now be denoted as $\left\{t_{i+1}^1,\, ...,\, t_{i+1}^J\right\}$.

It is important to recognize that as a consequence of the decrement, $t_{i+1}^j = t_i^j - t_i^{E_i}$ is no longer distributed according to $F^j\left(t_{i+1}\right)$ for $j \neq E_i$. We can derive the correct distribution for the modified interarrival times using the survival function. Let the survival function for the original interarrival time for event type j be $S^j(t)$. Then, for event types $j \neq E_i$, the

conditional survival function of the decremented interarrival time, t_{i+1}^{j}, is

$$\bar{S}^{j}\left(t|t_{i}^{E_i}\right) = Pr\left(t_{i+1}^{j} > t \,|\, j \neq E_i\right) = \frac{S^{j}\left(t + t_{i}^{E_i}\right)}{S^{j}\left(t_{i}^{E_i}\right)} \tag{2.2}$$

We refer to the survival function of the next event E_{i+1} as the multi-event survival function $S\left(t\right)$. Because we assume that the J processes are independent, the multi-event survival function can be defined as the product of the relevant conditional survival functions and the survival function for the next event in the process corresponding to event E_i,

$$S\left(t\right) = S^{E_i}\left(t\right) \prod_{j \neq E_i} \bar{S}^{j}\left(t|t_{i}^{E_i}\right) \tag{2.3}$$

To perform inference for our model (below), we need to derive the density of the interarrival times. Recall that the density of the interarrival time, $f\left(t\right)$, is the product of the survival and hazard functions, $f\left(t\right) = \lambda\left(t\right) S\left(t\right)$. Following the notation above, we use the superscript j with the hazard function $\lambda^{j}\left(t\right)$ to denote the hazard function for the process of the j^{th} event type, and let $\lambda\left(t\right)$ denote the hazard function for the multi-event background process. Recall from above that the hazard functions corresponding to both the survival and conditional survival functions are identical. Then we can use the relationship of the hazard and survival functions to establish that the multi-event hazard is $\lambda\left(t\right) = \sum_{j=1}^{J} \lambda^{j}\left(t\right)$, and thus $f\left(t\right) = \sum_{j=1}^{J} \lambda^{j}\left(t\right) S\left(t\right)$ is the density of the interarrival times in our set of multiple background processes.

The previous paragraph defines the density function for the interarrival time. We are interested in the joint density of the event type and the interarrival time, or the event-time pair

$\{E_i, t_i\}$. Because the probability density for the interarrival times marginalizes the joint density over all possible events j, it follows that the background process time-event pairs have the following density, where $S(t)$ is as defined in (2.3):

$$f_{t_i,E_i}(t, j) = \lambda^j(t) S(t) \tag{2.4}$$

Equation 2.4 establishes the density for a single time-event pair four our model when it contains only background processes. To perform inference, we need the likelihood for all data points. Due to the generative point process construction of the density, the likelihood can simply be defined as the product over the densities of the time-event pairs $L_{BG}(\mathbf{t}, \mathbf{E}) = \prod_{i=1}^{n} f_{t_i,E_i}(t, j)$.

In this subsection, we only described the model when no sequences are present. In the next subsection we will describe the construction of sequence processes. This is followed by a subsection describing how to augment the likelihood for the model containing only background processes to accommodate the sequence processes.

2.2.2 Sequence Processes

In addition to background events, we also want to consider the possibility of recurring sequences, where sets of events tend to occur in sequence and close together in time. We model these sequences of events as arising from sequence processes, which are assumed independent of the background processes. Sequence processes will be identified by our model when events occur in sequence more often than expected if they belonged to the background processes. We require that all preceding events must have occurred to observe the next event in a sequence process.

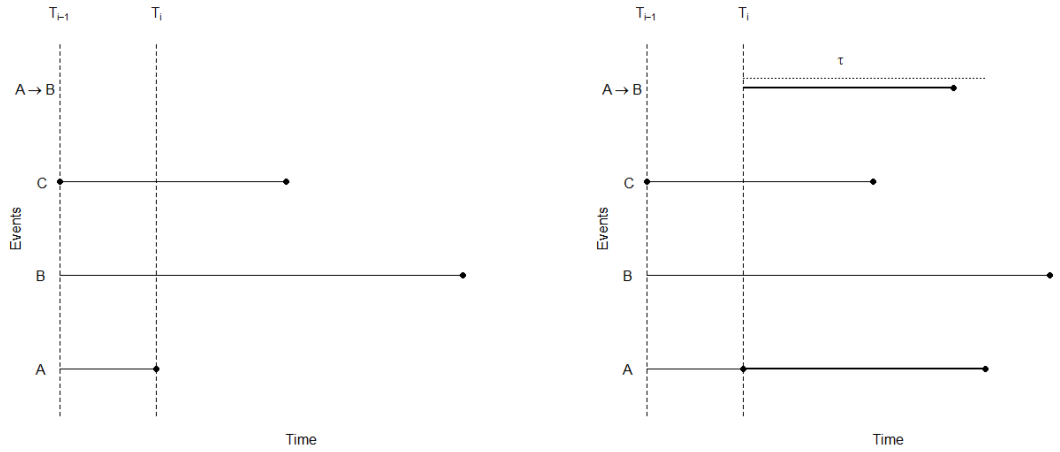

Figure 2.2: An example of the data generating process for a model that includes the sequence $A \to B$. On the left, the event type A just occurred at time T_i, which triggers the sequence $A \to B$. On the right, a new time increment is drawn for both event type A as well as event type B, which belongs to the sequence.

An example of how a sequence process can be interpreted as part of the data generating process is shown in Figure 2.2. In this example, event type A just occurred at time T_i, as seen on the left. If the sequence $A \to B$ exists as part of the model, then the occurrence of event type A would trigger the sequence, and hence an event B would likely occur soon after A. Thus we draw both a new time increment for A, as well as a new time increment for B, as part of the sequence $A \to B$. An example of both of these draws can be seen on the right side of Figure 2.2.

Developing the sequence processes requires additional notation. Let s index a specific sequence. For example, if a sequence s consists of behaviors $A \to B \to C$ occurring in that order, then we need a way to identify which events in our observed data $\{E_i, t_i\}$ correspond to events in the sequence. For sequence s, we define $s\{\ell\}$ to be the event index in our data corresponding to the ℓ^{th} event in that sequence, so that $E_{s\{1\}} = A$, $E_{s\{2\}} = B$, and $E_{s\{3\}} = C$.

We expect that the events corresponding to consecutive indexes $s\{\ell\}$ and $s\{\ell+1\}$ will frequently be consecutive events in the data, but this need not always be the case. We want to allow for the possibility that unrelated events may occur in the midst of a sequence by chance. We refer to these events as noise events, intervening events that occur during a sequence but do not belong to that sequence. These events can occur as part of either background processes or other sequence processes. Furthermore, our model allows a single event to belong to multiple sequence processes.

An important aspect of a sequence is that the events should occur relatively close together in time. To ensure this condition, we introduce parameters that govern the maximum time allowed between events in a sequence. We include these parameters both because it is well accepted that events occurring well in the past have much less direct influence on current behavior, and also for computational convenience. However, no limit is placed on either the number of events in a sequence or its overall duration, so it remains possible that events well in the past can have indirect influence on current events.

For event $s\{\ell\}$, we denote the corresponding time parameter as $\tau_{s\{\ell\}}$, a positive number that represents the maximum amount of time that can elapse between events $E_{s\{\ell-1\}}$ and $E_{s\{\ell\}}$ for each $\ell > 1$. Hence we require that the ℓ^{th} event in the sequence $s\{\cdot\}$ must occur within the time window $\left[T_{s\{\ell-1\}},\, T_{s\{\ell-1\}} + \tau_{s\{\ell\}}\right]$. If $E_{s\{l\}}$ does not occur by the end of this window, then the sequence does not occur. Given events $E_{s\{1\}}$, ..., $E_{s\{\ell-1\}}$ have occurred, there is positive probability that event $E_{s\{\ell\}}$ does not occur, and thus neither does the sequence $s\{\cdot\}$.

An example of how a maximum time parameter can be included is displayed as a dotted line on the right side of Figure 2.2. Had the dotted line been shorter than the time increment for sequence $A \rightarrow B$, then the sequence would not have occurred, as thus the time increment would be removed.

The probability of non-occurrence for $E_{s\{\ell\}}$ that an event does not occur is equivalent to the probability that the interarrival time is greater than $\tau_{s\{\ell\}}$. Hence it can be denoted by the survival function evaluated at $\tau_{s\{\ell\}}$. To distinguish the survival function of the sequence processes from the background processes, we denote $\dot{S}^{s\{\ell\}}\left(\tau_{s\{\ell\}}\right)$ as notation for the survival function of the next event in a sequence process. Then the probability that the ℓ^{th} event in a sequence process does not occur is

$$Pr\left(T_{s\{\ell\}} > T_{s\{\ell-1\}} + \tau_{s\{\ell\}}\right) = Pr\left(t_{s\{\ell\}} > \tau_{s\{\ell\}}\right) \equiv \dot{S}^{s\{\ell\}}\left(\tau_{s\{\ell\}}\right) \tag{2.5}$$

The censoring at $\tau_{s\{\ell\}}$ complicates the calculation of the survival distribution associated with the sequence event $s\{\ell\}$. The survival function is right censored at $\tau_{s\{\ell\}}$, with a nonzero probability that the event does not occur. We define a new censored survival function $\tilde{S}^{s\{\ell\}}(t)$ as

$$\tilde{S}^{s\{\ell\}}(t) = \dot{S}^{s\{\ell\}}(t)\, 1_{\left(t<\tau_{s\{\ell\}}\right)} + \dot{S}^{s\{\ell\}}(\tau)\, 1_{\left(t>\tau_{s\{\ell\}}\right)} \tag{2.6}$$

The latter half of the sum in (2.6) refers to a point mass giving the probability that the event does not occur. Let $\dot{\lambda}(t)$ denote the hazard function associated with $\dot{S}(t)$. Then by the properties of the density function described earlier, the probability density of the ℓ^{th} sequence event can be derived as

$$f_{t_{s\{\ell\}}}(t) = \dot{\lambda}^{s\{\ell\}}(t)\, \dot{S}^{s\{\ell\}}(t)\, 1_{\left(t<\tau_{s\{\ell\}}\right)} + \dot{S}^{s\{\ell\}}(\tau)\, 1_{\left(t>\tau_{s\{\ell\}}\right)} \tag{2.7}$$

To model the noise events defined earlier, assume the ℓ^{th} event in the sequence occurs at T_i, so $s\{\ell\} = i$ and $T_{s\{\ell\}} = T_i$. The previous event in the sequence, $E_{s\{\ell-1\}}$, occurs at some time $T_{s\{\ell-1\}}$ which is less than T_{i-1} if intervening events are present, but equal to T_{i-1} if not. Hence $T_{s\{\ell-1\}} \leq T_{i-1}$. Thus, the time of the intervening event can be represented as $T_{i-1} = T_{s\{\ell-1\}} + r$ where $r \geq 0$ is the time between the previous event in sequence and the noise event. If no noise event is present, then $r = 0$, but if a noise event is present, then $s\{\ell - 1\} < i - 1$ and $r > 0$. The interarrival time for the sequence can then be represented as $T_{s\{\ell\}} - T_{s\{\ell-1\}} = t + r$, where t is the time between T_{i-1} and $T_{s\{\ell\}}$, and r is the time between $T_{s\{\ell-1\}}$ and T_{i-1}. Here we considered the noise event occurring directly before $E_{s\{\ell\}}$, but the argument above can be easily generalized to handle any number of noise events that may occur between events $E_{s\{\ell-1\}}$ and $E_{s\{\ell\}}$.

The hazard function $\dot{\lambda}(t)$ remains unchanged, so the new density for event $s\{\ell\}$ can be expressed as

$$f_{t_{s\{\ell\}}}(t|r) = \dot{\lambda}^{s\{\ell\}}(t+r)\frac{\dot{S}^{s\{\ell\}}(t+r)}{\dot{S}^{s\{\ell\}}(r)}1_{\left(t+r<\tau_{s\{\ell\}}\right)} + \dot{S}^{s\{\ell\}}(\tau)\,1_{\left(t+r>\tau_{s\{\ell\}}\right)} \tag{2.8}$$

This defines the density for a single event in a particular sequence process. Using this density, we next describe now to add multiple sequence processes to the previously established likelihood function for background processes $L_{BG}(\mathbf{t}, \mathbf{E})$.

2.2.3 Combining Background and Sequence Processes

Our model assumes that background and sequence events occur independently, which allows us to combine them in a relatively straightforward manner. We assume that all sequence events are initialized by a background event as their initial event, thus $s\{1\}$ effectively

belongs to both a background process and a sequence process.

As before, assume the current time is T_i and we have observed $\{E_i, t_i\}$. Now the pool of possible next events is comprised both of the next events in all background processes as well as the next events for all live sequences. We define a live sequence as any sequence such that all prior events in the process have occurred, and a requisite amount of time since the last event in the sequence has not yet passed. Mathematically, a live sequence is any process for which $s\{1\}$ through $s\{\ell-1\}$ have occurred previously and $T_{s\{\ell-1\}} < T_i < T_{s\{\ell-1\}} + \tau_{s\{\ell\}}$. Let $\mathcal{B} = \{1, ..., J\}$ denote the background processes, and $\mathcal{S} = \{s_1, ..., s_K\}$ denote the K live sequence processes at time T_i, so the next event must occur from either $\mathcal{B}$ or S.

From our independence assumption, $S^{\mathcal{B}}(t) = \prod_{j=1}^{J} \bar{S}^j(t)$ and $S^{\mathcal{S}}(t) = \prod_{k=1}^{K} \tilde{S}^{s_k}(t)$. Furthermore, from the definition of the sequence process, we have

$$\begin{aligned} S(t) &= S^{\mathcal{B}}(t) S^{\mathcal{S}}(t) \\ &= S^{\mathcal{B}}(t) \dot{S}^{\mathcal{S}}(t) 1_{(t<\tau)} + S^{\mathcal{B}}(t) \dot{S}^{\mathcal{S}}(\tau) 1_{(t>\tau)} \end{aligned} \tag{2.9}$$

where $\dot{S}^{\mathcal{S}}(t) = \prod_{k=1}^{K} \dot{S}^{s_k}(t)$.

At time T_i, fix the current event to be $E_i = e$. If there are K live sequences, then there are $K^* \leq K$ sequences that are both live and the next event in the sequence is e, hence $s\{\ell\} = e$. Denote the set of these sequences as $\mathcal{S}'$. Finally, define $\dot{\lambda}^e$ to be the sum over all the hazard functions corresponding to the sequences in $\mathcal{S}'$, $\dot{\lambda}^e(t) = \sum_{s \in \mathcal{S}'} \dot{\lambda}^s(t)$.

Finally, we get the following density, inserting the modification for noise events from (2.8)

as needed:

$$
\begin{aligned}
& f_{t_i,\,E_i}\left(t,\,e\right) \\
& = \left(\lambda^e\left(t\right) + \dot{\lambda}^e\left(t\right)\right) \prod_{j=1}^{J} \bar{S}^j\left(t\right) \prod_{k=1}^{K} \dot{S}^{s_k}\left(t\right) 1_{\left(t<\tau_{s_k}\right)} \\
& + \lambda^e\left(t\right) \prod_{j=1}^{J} \bar{S}^j\left(t\right) \prod_{k=1}^{K} \dot{S}^{s_k}\left(\tau\right) 1_{\left(t>\tau_{s_k}\right)}
\end{aligned}
\tag{2.10}
$$

Finally, we can calculate the likelihood of the model by taking the product of the densities for all time-event pairs: $L\left(\mathbf{t},\,\mathbf{E}\right) = \prod_{i=1}^{n} f_{t_i,\,E_i}\left(t,\,e\right)$.

2.3 Inference

Section 2.2 derives the likelihood function for an observed series of events. The notation of that section refers only to the unspecified interarrival distribution of each process in the model. In practice it is natural to assume that these processes may depend on unknown parameters. We are now interested in performing inference on these parameters. To this point, we have made no assumptions on the distribution of the time increments, allowing the functional form of the survival and hazard functions to remain arbitrary. Because of this, we will consider a general inference procedure that can accommodate a wide variety of functional forms.

We assume that each background process j depends on a set of parameters θ_j, and each sequence process k depends on a set of distribution parameters θ_k and time parameters τ_k. Note that the sequence process parameters are each vectors of length $\ell - 1$, so each level of a sequence has its own distribution parameters and time parameter. Then both the background

and sequence processes have their respective supersets of parameters $\Theta_B = \{\theta_j,\, j = 1, ..., J\}$ and $\Theta_S = \{(\theta_k, \tau_k)\,,\, k = 1, ..., K\}$. We denote the entire parameter set as $\Theta = \{\Theta_B, \Theta_S\}$.

We take a Bayesian approach to learning the parameters in our model. In this context, all parameters are assumed to be random, and our goal is to describe a posterior distribution. Furthermore, in addition to the likelihood, we must place a prior distribution on all parameters. Then the posterior distribution is defined as $\Pr(\Theta|\mathbf{t},\, \mathbf{E}) \propto L(\mathbf{t},\, \mathbf{E}|\Theta) \Pr(\Theta)$, and the parameters can be sampled from the posterior distribution via a Markov Chain Monte Carlo (MCMC) procedure (Brooks et al., 2011).

While we do not consider the possibility of finding MAP estimates, such a procedure would be of practical interest. A conditional maximization algorithm is probably feasible, where parameters are obtained by maximizing the conditional probabilities rather than sampled from the conditional probabilities.

2.3.1 Prior Distributions

Prior to sampling parameters, it is necessary to describe prior distributions for each parameter. For the application later in this chapter, we only consider renewal processes with exponentially distributed increments. In this case, each process has a single rate parameter, so here we discuss prior distributions for the exponential rate parameter. This can clearly be modified for other assumptions. The higher the rate parameter, the more often that event occurs. In chapter 4, we consider more general distributions for the time increments, and provide further discussion on prior distributions.

In the case that some prior knowledge is available, informative priors can be straightforward to include. For example, consider the rate parameters λ_j for background processes, although a similar prior distribution could be used for sequence process parameters. An expert could

specify how often they expect event j to occur in a given amount of time. For instance, let $1/\phi_j$ be an estimate for the number of times that event j occurs, then we could assume the prior distribution follows from an exponential distribution with rate $\lambda_j \sim Exp(\phi_j)$. While informative prior distributions are appealing in some situations, we do not consider them further in this chapter. Instead, for the background processes, we simply place exponential prior distributions on all background rate parameters.

Our goal for this dissertation is to accurately identify patterns from behavioral data. In the case that some patterns are known to exist, then we consider a prior distribution for the pattern rates with low mass around zero. For example, we could consider a gamma distribution with parameters a and b, such that the mode $\frac{a-1}{b}$ is equivalent to our prior guess variance $\frac{a}{b^2}$ is relatively small. Otherwise, we assume the prior distributions have high mass around zero to ensure regularization. In this case we could consider a gamma distribution with shape parameter $a < 1$, which will ensure most of the mass is around zero. For the application we consider, we do not expect any particular patterns in our data, so we will not consider this prior structure either.

Instead, we consider a different type of prior distribution for the sequence process rate parameters. When an event only occurs once in the observation period, especially when following another rare event, the likelihood alone cannot clearly differentiate whether that event belongs to a background or sequence process. This could plausibly lead to spurious patterns being discovered. In cases such as this, we prefer to err on the side of the background processes. To account for this, we place shrinkage priors on any sequence process rates but not on background process rates, such that more mass is near zero for sequence processes (Bhattacharya et al., 2012). In our application, we consider the mixture of two half normals, both with a mode of 0, but one with much higher variance.

We also note that we consider weakly informative priors for the τ parameters. Because their purpose is to constrain the amount of time between events in a sequence, we want to place

most of the mass of the prior distribution greater than zero and centered around a reasonable guess. Placing too much mass around zero or limiting the variance around a poor prior guess will both impede pattern discovery, which we discuss in the experiments section. In practice, the exact location of the prior mass would be determined by the application area. For our particular application, we expect transitions between events in a pattern to take roughly 2-30 seconds on average. Hence we consider a gamma prior distribution with shape $a = 2$ and scale $b = 0.2$, which places most of the prior mass in the range 2-30 seconds.

2.3.2 MCMC

Our complete approach to inference uses the No-U-Turns Sampler (NUTS) for sampling the parameters, with occasional birth-death steps incorporated for model selection (Hoffman & Gelman, 2012; Stephens, 2000). First, we describe NUTS, implemented with the Stan software package (Stan Development Team, 2013), and then describe the birth-death steps.

NUTS is a form of Hamiltonian Monte Carlo (HMC), an MCMC algorithm that avoids the random walk behavior of several other popular MCMC methods (Neal, 2010). HMC methods use an analogy to Hamiltonian dynamics to sample from a distribution. We consider the joint distribution over the parameters of interest p ("position" variables) and auxiliary "momentum" variables q. We can then simulate the evolution over time of the Hamiltonian dynamics of the system via a leapfrog integrator. A total of L leapfrog steps are considered, each of step size ϵ. More details can be found in Neal (2010).

However, standard HMC is sensitive the number of leap frog steps L and the step size ϵ, which are user provided parameters. NUTS improves upon the standard HMC implementation by eliminating the need to set either parameter. However, our implementation requires that we set the step size, which we describe below. Furthermore, the Stan implementation uses reverse-mode algorithmic differentiation to calculate the gradient for the conditional

distributions, eliminating the need for us to calculate it by hand. Hence all that is needed to run NUTS via Stan is the probability model itself, and step size in our application.

2.3.3 Model Selection

While NUTS is very effective for sampling from the posterior distribution, it does not address model selection. Our model allows for a potentially unlimited number of sequence processes to exist, so we need to determine which sequences actually exist in the observed data. Including all possible sequences is impractical; when we consider sequences of increasing maximum length, the number of possible sequences grows exponentially. Instead, we only include a small subset of all possible sequences in our model at any given point in the sampler, and use a birth-death MCMC technique to vary the subset of sequences.

Sequences that are very unlikely to occur will have a hazard rate near zero, and thus a survival function constantly near one. When $\lambda(t) \approx 0$ and $S(t) \approx 1$ for all t, from (2.10) we see that the sequence is effectively not included in the model. In this case, we fix the sequence process parameters θ such that the hazard rate is fixed at zero. Furthermore, this is equivalent to simply removing the sequence process from the model.

To determine which sequence processes belong in the model, we apply the birth-death process of Stephens for model selection, an alternative to reversible jump steps. Stephen's birth-death MCMC views the parameters of the model as a point process, which allows the number of components to vary by allowing new ones to be 'born' and existing ones to 'die.' At various times within the overall NUTS chain, we perform a birth-death step by constructing an easily simulated process that changes the number of sequences in the model. The births and deaths both occur according to Poisson processes. Births occur at a constant rate, while deaths occur at a low rate for sequences critical for explaining the data, and a high rate for sequences that do not help explain the data.

When a birth-death step occurs, we fix some simulation time W and run through a series of birth or death steps until the simulated time exceeds W. Birth steps occur according to a Poisson process with rate δ_B. Each sequence s currently in the model receives an independent death process with rate δ_D^s, described in the next paragraph. The overall rate of death steps is $\delta_D = \sum_s \delta_D^s$. The time to the next step is exponentially distributed with rate $\delta_B + \delta_D$. Times are generated until they pass time W, with a birth or death step occurring at each time with probability proportional to the rate for the birth/death.

To ensure that important sequences are retained within the model, the death rate for a given sequence s will be equal to the birth rate multiplied by the ratio of the time-event likelihoods with and without sequence s included. Let Θ_s denote the parameter set with sequence s and Θ_{-s} the parameter set without sequence s. Then the death rate for sequence s would be $\delta_D^s = \delta_B \frac{L(t,E|\Theta_{-s})}{L(t,E|\Theta_s)}$. Thus sequences where the data suggests they belong in the model will have very low death rates.

For the birth step, we choose a sequence not currently in the model, and draw parameters from an appropriate distribution, e.g. the prior distribution for it. To choose the new sequence, we pool all existing sequences in the model, randomly sample two of them and combine them to form a new sequence. For these purposes, we include all single events as sequences in addition to the multi-event sequences in the model. For example, assume the model has background events A, B, and C, as well as the C→A pattern. We sample, with replacement, two of those four possible events, and combine them to form the new pattern. Possible new patterns include A→B, C→A→B, or C→A→C→A, among others.

A death step simply chooses an existing sequence process, and fixes its parameters such that the hazard function is zero for all possible times t.

The simulated birth-death process proceeds as follows:

Finally, we describe how our model selection technique is integrated within the NUTS algo-

Algorithm 1: Birth-Death Process

Fix δ_B, W ; `// Birth Rate, Birth Death Sim Time`
$w \leftarrow 0$;
while $w < W$**:**
  $\delta_D^s \leftarrow \delta_B \frac{L(t,E|\Theta_{-s})}{L(t,E|\Theta_s)}$; `// Sequence Death Rate`
  $\delta_D \leftarrow \sum_s \delta_D^s$; `// Overall Death Rate`
  Sample $u \sim unif\,(0,1)$;
  if $u < {}^{\delta_B}/_{\delta_B+\delta_D}$**:**
    Birth Step
  else:
    Death Step
  Sample $t \sim \exp\left({}^{1}/_{\delta_B+\delta_D}\right)$; `// Next Jump Time`
  $w \leftarrow w + t$;

rithm. We initially run the HMC sampler for the background model (no sequence processes included) for a large number of iterations, though not necessarily until convergence. After that period, the birth-death process is run once every m iterations of the sampler, allowing NUTS to explore the state space for every model. The standard implementation of NUTS within Stan determines the step size needed for HMC during the warmup phase. To ensure ergodicity, we estimate the step size via the results of the sampler for the initial background model, and then fix the step size at this value when running NUTS for the model including sequence processes.

2.3.4 Convergence

We monitor convergence of the MCMC by running multiple chains, and monitoring convergence statistics. This is somewhat problematic when the size of the model can change, which occurs for our model when a pattern is either added or removed. To handle this issue, we use the convergence criterion developed in Sisson & Fan (2007). This method assumes that the models change according to a point process, such as the birth death process that we consider for model selection.

At iteration i, the current set of model parameters is denoted $\Theta^{(i)}$. We assume this parameter set is comprised of parameters found in all possible models (background process parameters, denoted here as $\Theta_{BG}^{(i)}$) and those found only in certain samples (sequence process parameters, denoted here as $\Theta_S^{(i)} = \left(\Theta_{s_1}^{(i)}, ..., \Theta_{s_K}^{(i)}\right)$).

This method first sets several reference points ν. These reference points are instances drawn from the sample space of possible models and parameters, which can be used to compare against each iteration of each chain. In practice, the reference points can be drawn from the Markov chains themselves, so $\nu = \Theta^{(j)}$ for some random j. For each of the reference points, we can define the distance $x_\nu^{(i)}$ as the minimum distance between ν and one of the subsets of $\Theta_S^{(i)}$. A subset of $\Theta_S^{(i)}$ refers to one of the combinations of parameters that are found in it. For example, a set of one parameter would have two subsets (the parameter and the null set); a set of two parameters would have four subsets (the null set, each individual parameter, and both parameters); a set of three parameters would have eight subsets. In practice, we consider Euclidean distance between ν and each subset, where parameters not in a particular model are set as zero.

Let us consider a toy example. Our current sample consists of $\theta_1 = 1.3$ and $\theta_2 = 1.7$. Our reference point is set as $\nu = \{\theta_1 = 1.1\}$. The possible distances we consider are for the parameter sets $\{\emptyset\}$, $\{\theta_1\}$, $\{\theta_2\}$, and $\{\theta_1, \theta_2\}$. Respectively, the distances would be $\sqrt{(1.1-0)^2 + (0-0)^2} = 1.1$, $\sqrt{(1.1-1.3)^2 + (0-0)^2} = 0.2$, $\sqrt{(1.1-0)^2 + (0-1.7)^2} = 2.0$, and $\sqrt{(1.1-1.3)^2 + (0-1.7)^2} = 1.7$. In this case, we would set $x_\nu^{(i)} = 0.2$.

Using these distances, we can compute an analogue of the Rubin-Gelman statistic $\hat{R}_\nu$ for each of the reference points (Gelman & Rubin, 1992). We can then declare that the chain has converged if the value is sufficiently close to one for all of the reference points. By choosing a set of reference points, each with a potentially different set of patterns, this method allows us to examine various deviations in the model from both between and within chains to assess whether further iterations are needed.

2.4 Simulations

We consider a number of simulation experiments to explore the performance characteristics of the renewal process model. All experiments assume that background and sequence processes have a constant hazard function, and therefore the interarrival times are exponentially distributed. We explore more general hazard rates in Chapter 4.

Our simulation experiments will focus on several aspects of our model. First we establish that our model performs well on various types of simulated data. We then explore how capably our model performs in various more difficult conditions, such as when limited data exists and when especially long patterns are present.

We note that unless otherwise specified, we consider the following prior distributions for both our simulation experiments and the following data analysis: an exponential prior distribution with rate one for all background rate parameters; a mixture of two half normal distributions for all sequence process rate parameters, both with mode zero, one with variance 5 and the other with variance 0.01, and a mixing parameter of 0.5; and a gamma prior distribution for all τ parameters with shape 2 and scale 0.2.

2.4.1 Proof Of Concept

We first consider the results for simulated data under a simple scenario, one with only three possible event types which we label events 1, 2, and 3. We include only one sequence process, a simple 2-event sequence: $3 \rightarrow 1$. This will allow us to test our model on a simple data set, and compare the results directly with those obtained from Theme. We fix the background exponential process parameters $\lambda_1 = 0.16$, $\lambda_2 = 0.57$, and $\lambda_3 = 0.40$, the sequence exponential process parameter $\lambda_{3\rightarrow 1} = 0.25$ and the corresponding time window parameter as $\tau_{3\rightarrow 1} = 1.12$. Using these parameters, we sample $M = 100$ data sets of length

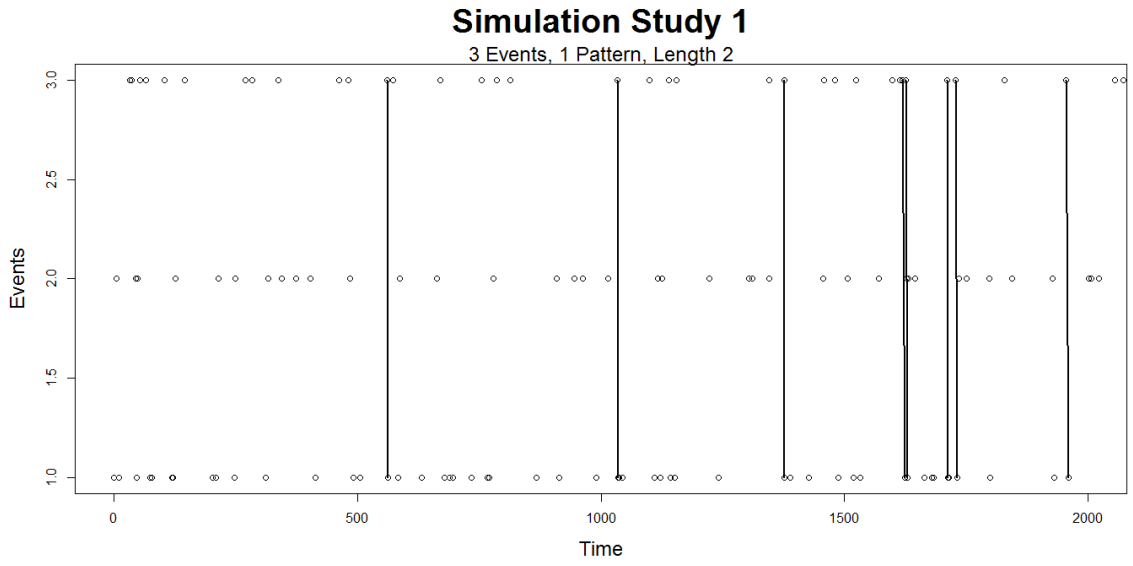

Figure 2.3: An example of simulated data for the proof of concept simulation. Three different events are present, with patterns identified by lines.

Parameter	True Value	Post. Mean	S. E.	Cov.
λ_1	0.16	0.16	0.03	0.97
λ_2	0.57	0.57	0.08	0.95
λ_3	0.40	0.40	0.06	0.97
$\lambda_{3\to1}$	0.25	0.23	0.08	0.97
$\tau_{3\to1}$	1.10	1.14	0.18	0.95

Table 2.1: The parameter values for each of the five parameters in our first simulation experiment. The estimated mean is the average over the posterior means across the simulations. Standard error (S. E.) and coverage probability (Cov.) are also included for each parameter.

$n = 500$ events, and fit the model to each data set. An example of the simulated data, with patterns identified, can be seen in Figure 2.3.

For each simulated data set, we run our inference procedure treating all parameters as unknown. For each simulation we record the posterior mean, and report the average and standard error of these values. Additionally, we calculate a 95% posterior credible interval for each parameter. Table 2.1 summarizes the simulation results. The parameters are estimated well and the coverage probabilities for the 95% posterior intervals for the parameters are all 0.95 or higher.

	Renewal Process Model	Theme
# of Correct Sequences	99/100	94/100
# of False Positives	57	384
# of False Negatives	1	6
Power/Recall	0.99	0.94
Precision	0.64	0.20

Table 2.2: Performance summaries for $M = 100$ simulated data sets

We also record which sequences were included by the model selection process for each data set. This is done by including the patterns corresponding to all sequence processes with a posterior probability greater than 0.5 of being nonzero. We record the proportion of trials in which the correct sequence was identified (power), the number of null sequences incorrectly identified as being present (false positives), and the proportion of declared sequences that are true sequences (precision). The Renewal Process Model's results for the first scenario are in the first column of Table 2.2.

We also ran Theme with its default settings on each of the simulated data sets for the first scenario, and recorded the same summaries. As described earlier, assessing significance in Theme is not always straightforward, and not all patterns are assumed real. After the patterns are identified, a permutation procedure can be used, which compares the results with the results using randomized data. However, it does not indicate which patterns are most important. Here we just summarize the patterns identified before the permutation test was applied, recognizing that practiced Theme users would not consider all of these as real. These results are also included in Table 2.2. Theme identified the correct sequence about as often as our model, though as expected it often identified several other sequences that do not actually exist.

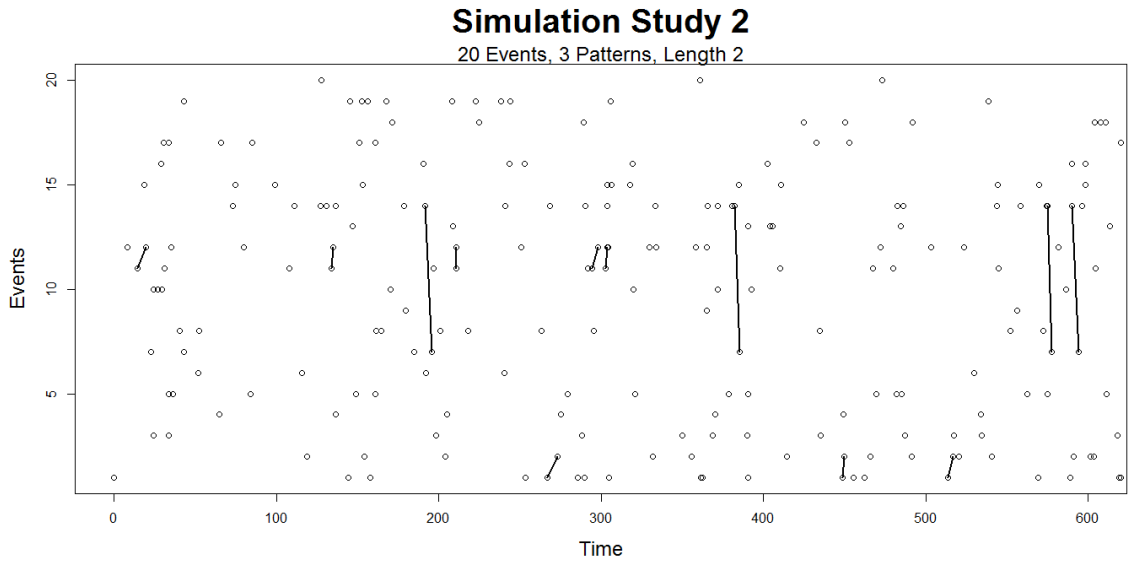

Figure 2.4: An example of simulated data consistent with a mother who is fragmented and unpredictable. There are 20 possible events, with three short patterns.

2.4.2 Realistic Scenarios

Our first simulation scenario is not consistent with data from our motivating example, so we would like to test our model on more complex data. Here we consider two simulations with data more consistent with our application. Both simulations are comprised of 20 different possible event types. The first has three 2-event sequences, which represents the type of observation expected from a fragmented and unpredictable mother in our motivating example. An example of the data simulated in this scenario is shown in Figure 2.4

The second simulates data with both more patterns (six) and more complex patterns (up to 4-events in one pattern). Again, we sample $M = 100$ data sets of length $n = 500$ events for both of the more complex scenarios. An example of data from this scenario, along with patterns identified, can be seen in Figure 2.5. Note that for this scenario, we consider sequences $1 \to 2$ and $1 \to 2 \to 3$ to be different patterns. However, the parameters for $1 \to 2$ are the same as the parameters for the first half of $1 \to 2 \to 3$. Thus the parameter values seen in Table 2.6 refer only to the parameters for the final part of the sequence process.

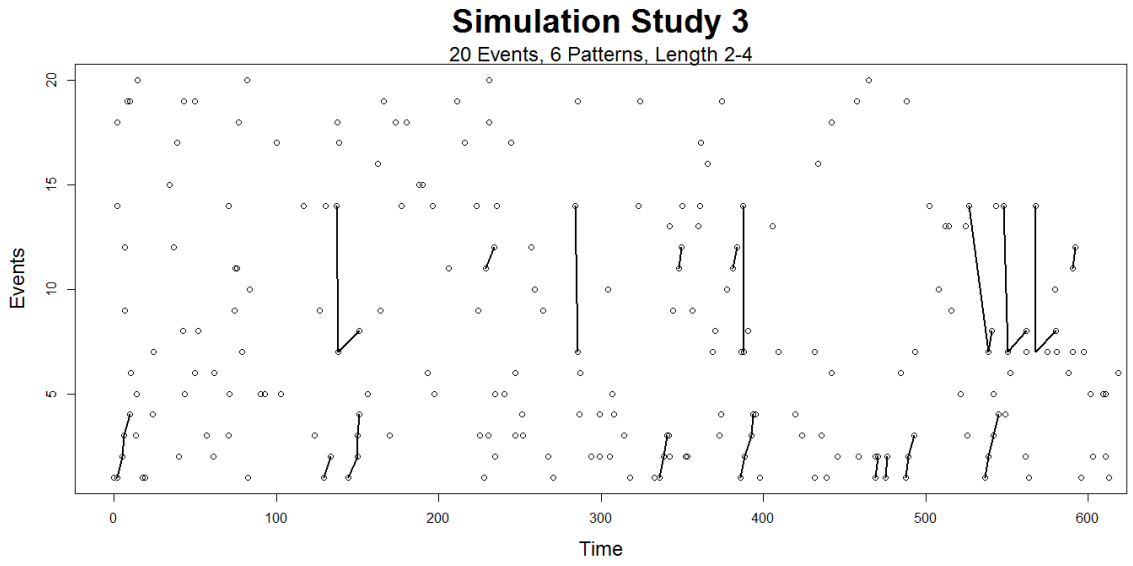

Figure 2.5: An example of simulated data consistent with a mother who is consistent and predictable. There are 20 possible events, with six patterns.

For these second and third scenarios, there are considerably more parameters. Results for the background processes parameters in the second simulation scenario can be found in Table 2.3, along with the true values of the parameters. Results include estimates of the rate parameter, as well as standard errors and coverage probabilities. Similarly, the same results for both the rate parameters and the maximum time parameters for the sequence processes from the second simulation scenario can be found in Table 2.4. The same results for both the background processes and sequence processes of the third simulation scenario can be found in Table 2.5 and Table 2.6.

The estimate for each parameter is the average over the posterior means from the different simulations. The standard error is calculated as the standard deviation of the posterior means. The coverage probability refers to the percentage of times that the true parameter value was found in the 95% credible interval.

The estimated rate parameter is similar to the true parameter value in essentially all cases. For sequence process parameters, when a simulation did not successfully identify the sequence process, we did not include that simulation when calculating the average over the posterior

Event	Truth	Est.	S.E.	Cov.	Event	Truth	Est.	S.E.	Cov.
1	0.2	0.20	0.08	0.94	11	0.2	0.19	0.07	0.95
2	0.2	0.34	0.15	0.94	12	0.1	0.15	0.07	0.93
3	0.2	0.25	0.12	0.92	13	0.1	0.10	0.03	0.96
4	0.2	0.24	0.12	0.97	14	0.2	0.18	0.08	0.93
5	0.2	0.20	0.05	0.94	15	0.1	0.10	0.03	0.94
6	0.2	0.16	0.04	0.95	16	0.1	0.10	0.03	0.93
7	0.1	0.14	0.07	0.96	17	0.1	0.10	0.03	0.95
8	0.1	0.10	0.06	0.92	18	0.1	0.10	0.03	0.96
9	0.1	0.10	0.03	0.96	19	0.1	0.10	0.02	0.95
10	0.1	0.10	0.03	0.98	20	0.1	0.10	0.03	0.93

Table 2.3: Results for background process rate parameters in the second simulation scenario (three simple patterns). The estimated mean is the average over the posterior means across the simulations. Standard error (S. E.) and coverage probability (Cov.) are also included for each parameter.

Pattern	Param.	Truth	Est.	S.E.	Cov.
$1 \rightarrow 2$	λ	0.38	0.33	0.04	0.92
	τ	4.41	4.38	0.11	0.91
$11 \rightarrow 12$	λ	0.22	0.19	0.03	0.96
	τ	3.05	3.00	0.09	0.95
$14 \rightarrow 7$	λ	0.35	0.36	0.05	0.91
	τ	2.19	2.20	0.04	0.90

Table 2.4: Results for sequence process rate and time parameters in the second simulation scenario (three simple patterns). The estimated mean is the average over the posterior means across the simulations. Standard error (S. E.) and coverage probability (Cov.) are also included for each parameter.

Event	Truth	Est.	S.E.	Cov.	Event	Truth	Est.	S.E.	Cov.
1	0.2	0.27	0.07	0.90	11	0.2	0.18	0.05	0.97
2	0.2	0.26	0.11	0.95	12	0.1	0.13	0.07	0.96
3	0.2	0.26	0.12	0.92	13	0.1	0.10	0.03	0.94
4	0.2	0.18	0.09	0.96	14	0.2	0.20	0.05	0.95
5	0.2	0.23	0.05	0.97	15	0.1	0.11	0.03	0.94
6	0.2	0.19	0.05	0.94	16	0.1	0.10	0.04	0.98
7	0.1	0.16	0.07	0.99	17	0.1	0.10	0.04	0.95
8	0.1	0.09	0.04	0.95	18	0.1	0.10	0.04	0.99
9	0.1	0.11	0.02	0.93	19	0.1	0.09	0.03	0.95
10	0.1	0.10	0.05	0.95	20	0.1	0.10	0.03	0.95

Table 2.5: Results for background process rate parameters in the third simulation scenario (six patterns, including longer ones). The estimated mean is the average over the posterior means across the simulations. Standard error (S. E.) and coverage probability (Cov.) are also included for each parameter.

Pattern	Param.	Truth	Est.	S.E.	Cov.
$1 \to 2$	λ	0.38	0.41	0.06	0.96
	τ	4.41	4.45	0.06	0.95
$1 \to 2 \to 3$	λ	0.39	0.45	0.07	0.92
	τ	4.42	4.16	0.08	0.93
$1 \to 2 \to 3 \to 4$	λ	0.41	0.36	0.11	0.88
	τ	4.49	4.70	0.16	0.91
$11 \to 12$	λ	0.22	0.14	0.10	0.93
	τ	3.05	3.27	0.18	0.92
$14 \to 7$	λ	0.35	0.31	0.08	0.94
	τ	2.19	2.36	0.11	0.95
$14 \to 7 \to 8$	λ	0.38	0.42	0.13	0.89
	τ	2.34	2.25	0.08	0.91

Table 2.6: Results for sequence process rate and time parameters in the third simulation scenario (six patterns, including longer ones). The estimated mean is the average over the posterior means across the simulations. Standard error (S. E.) and coverage probability (Cov.) are also included for each parameter.

means for the parameters. Thus we only averaged over simulations where the pattern was successfully identified. The same method was used when calculating standard errors.

Coverage probability was calculated by the percentage of times that the credible interval contained the true parameter values. For sequence processes, this means both that the model successfully identified the sequence process, and the true parameter value was contained in the credible interval. Because of this, the coverage probabilities of the sequence process parameters were fairly low, ranging from 0.88 to 0.96, which tend to be lower than 0.95 but are fairly close nonetheless. The coverage probabilities of the background processes were generally quite good, all were over 0.90.

For the second simulation scenario, all 3 patterns were found in 91% of the simulations using the RPM, with an average of 0.87 false positive patterns per simulation. In the final scenario, all 6 patterns were found in 87% of all simulations with an average of 1.6 false positives per simulation. Thus our method performs well for more complex data. Furthermore, these simulations were designed to be consistent with our application (see below), and it does

not appear that the complexity of the model, including the larger parameter space, has any negative effects on our model.

2.4.3 Effect of Sample Size

The data in our motivating example, the maternal behavior study, is very difficult to obtain because there is considerable labor involved in annotating the videotaped sessions. It is therefore important to consider how the identification of patterns is effected by the amount of collected data. For example, an individual who performs more events during their observation period will have more data. It is plausible that such individuals will display more patterns on average, even if the total time duration is the same for all actors (as it is in our motivating example below). A small n individual has fewer events observed overall, so the likelihood that a given pattern will occur often enough to be correctly identified is less than for a large n individual.

To study the impact of sample size, we develop a single simulation scenario and then vary the number of observations. As in the previous section, we use a set of 20 unique events. We set 10 patterns to be active, all of length two events. The patterns have decreasing rates of occurrence, such that some patterns are very rare and others are very common. These can be seen in Table 2.7. Most of the sequences have equivalent time parameters τ (see Table 2.7), allowing a better comparison of the rate parameters. Each of the background rates is fixed at 0.1. We consider four different sample sizes, each with data generated from the same model, but the number of events varied from 100 to 700. Data is simulated a total of 100 times for each sample size.

The results for this experiment can be seen in Table 2.7. As the amount of data increases, the ability to detect relatively rare patterns increases significantly, with only a slight increase in the number of false positives. With at least 300 events, over half of the patterns were

Seq	True Rate	True τ	100	300	500	700
$1 \rightarrow 2$	0.40	2.5	0.57	0.82	0.84	0.96
$3 \rightarrow 4$	0.40	1.0	0.30	0.70	0.92	0.83
$5 \rightarrow 6$	0.35	2.5	0.41	0.92	0.88	0.85
$7 \rightarrow 8$	0.30	2.5	0.33	0.85	0.95	0.87
$9 \rightarrow 10$	0.30	1.0	0.20	0.51	0.72	0.78
$11 \rightarrow 12$	0.25	2.5	0.28	0.74	0.96	0.86
$13 \rightarrow 14$	0.20	2.5	0.28	0.63	0.72	0.67
$15 \rightarrow 16$	0.15	2.5	0.42	0.51	0.52	0.65
$17 \rightarrow 18$	0.10	2.5	0.24	0.29	0.48	0.65
$19 \rightarrow 20$	0.05	2.5	0.16	0.36	0.44	0.61
Number of Detected Patterns			3.20	6.33	7.43	7.73
False Positives:			0.87	1.20	1.60	2.27

Table 2.7: The results of our sample size experiment. The true parameter values of the rates for the 10 sequence processes are given. For each sample size, the percentage of times that a pattern is identified, out of 100 simulations each, is also given. At the bottom, the total number of detected patterns and the number of false positives for each sample size are given.

detected on average. As the amount of data increased, the rate of improvement in pattern detection started to decrease. With 700 events, which is more events than any individual has in our maternal data, our model still had difficulty detecting the most infrequent patterns for a third of the simulations.

2.4.4 Detecting Longer Sequences

Our renewal process model works well for shorter patterns, as seen above. Fortunately, these comprise the bulk of the patterns that we expect to detect in behavior data similar to our maternal data. When the number of events is fairly small, finding longer patterns is difficult. Nonetheless, we would like to show that our model is capable of finding such patterns when they do exist.

For this simulation, we consider simulated data sets with varying numbers of events: 200, 500, and 1000 events per data set. We again have 20 unique event types. This time the

Pattern Length	200 Events	500 Events	1000 Events
2	0.96	1.00	1.00
4	0.90	0.96	0.98
6	0.49	0.69	0.78
8	0.18	0.25	0.26
10	0.08	0.18	0.24

Table 2.8: Percentage of simulated data sets where the model correctly identified the pattern. While longer patterns can be difficult to detect, there is some dependence on the size of the data set.

data generating process contains a single pattern, but we vary the length of the pattern from 2 to 10 events in different simulation runs. The pattern rates will be set high to ensure that the pattern occurs multiple times despite the small size of the data. Data for each set of parameters is simulated 100 times. To ensure consistency, the smaller data sets are comprised of the first 200 or 500 events of the respective 1000-event data sets.

The results are summarized in Table 2.8. The model does an admirable job detecting patterns up to length four, before becoming much less reliable. However, there is a clear tendency to better detect patterns when more data exists. This is a clear sign that longer patterns will be difficult to detect when the total number of events is small, as hypothesized.

Furthermore, even in cases when the full pattern was not discovered, subpatterns were always discovered. For example, consider a data set in which the pattern $1 \to 2 \to 3 \to 4 \to 5 \to 6$ is present. Our model often detects subpatterns such as $1 \to 2 \to 3$ and $4 \to 5 \to 6$, but cannot successfully join the two together in the model selection phase. This is because the proposed parameters for the full sequence are often sub-optimal compared to the previously sampled parameters of the partial patterns. The posterior distributions for both the model with the full sequence and the model with the partial sequences are often comparable, which increases the importance of optimal parameter proposal. This is one of the outstanding issues that is unresolved in this thesis, and a possible avenue for future work.

	Shape	Scale	P. Mode	Cred Int
Gamma	0.1	1	$\emptyset$	$\emptyset$
	2	0.2	2.37	(2.31, 3.52)
	1	1	2.36	(2.31, 3.20)
	30	10	2.37	(2.32, 3.32)
	100	10	$\emptyset$	$\emptyset$
Half-Cauchy	Scale = 10		2.37	(2.32, 3.49)

Table 2.9: On the left: The parameters for each of the six prior distributions that we consider for the τ parameter. On the right: the posterior mode and 95% credible interval for each of the posterior distributions.

2.4.5 Effect of Prior Distribution on Time Parameter

The posterior distribution of the τ parameter that dictates the maximum amount of time between two events in a sequence may be especially susceptible to its prior distribution. In particular, if the mass of the prior distribution is far from the true value, then the posterior distribution may not correctly contain mass around the truth. Poorly specified prior distributions could have negative effects for discovering patterns. In this section, we will explore the effect of various prior distributions for the τ parameter, and their affect on the analysis.

We consider a scenario with 20 possible events, all with background rates of 0.1. One sequence process is also included, with a rate of 0.3 and a maximum time parameter of 2.4. We consider a total of six possible prior distributions: five gamma distributions and a Cauchy distributions. The parameters that we consider can be found on the left side of Table 2.9.

For the gamma prior distributions, we consider two prior distributions with high mass around zero, one with low variance (shape 0.1, scale 1), and one with relatively higher variance (shape 1, scale 1). We also consider two distributions with mass well above the true value $\tau = 2.4$: one with low variance (shape 100, scale 10) and one with high variance (shape 2, scale 0.2). We also consider a scenario where the mass is concentrated close to the true value, but with

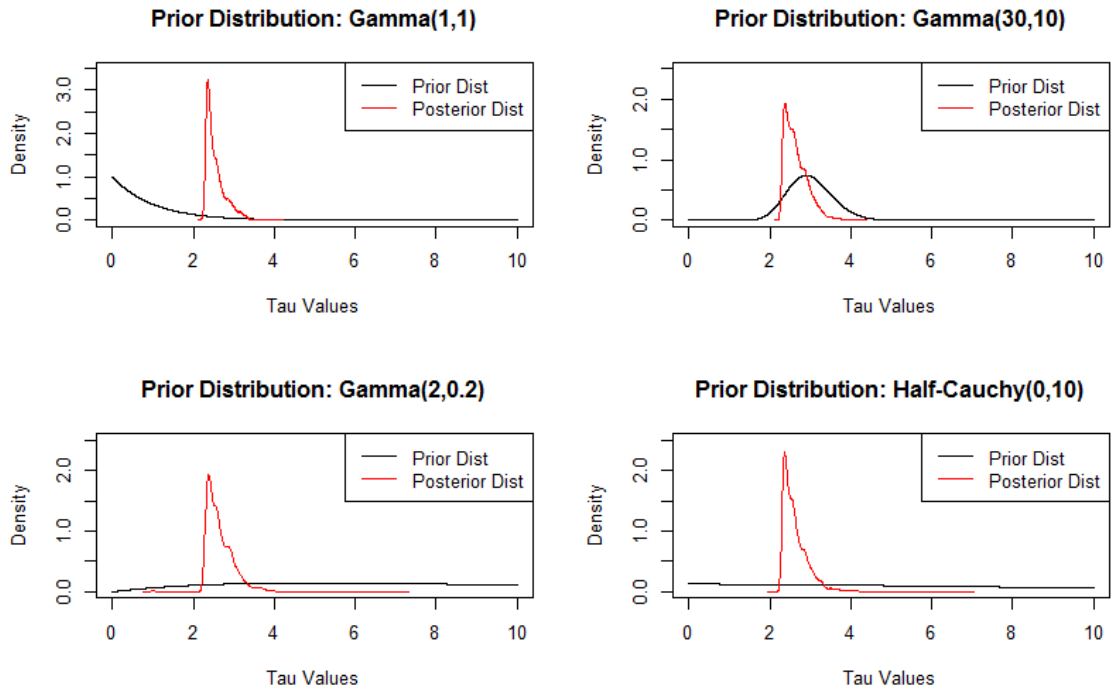

Figure 2.6: Density plots of the prior and posterior distributions corresponding to four choices of prior distributions for the maximum time parameter τ. Note that the prior distribution for both plots on the bottom have been scaled for visibility.

low variance (shape 30, scale 10). Finally, we consider a half-Cauchy with scale parameter equal to 10 as a purely non-informative prior.

For each of the six prior distributions, we fit the same simulated data set. Four of them successfully identified the pattern, and both the prior and posterior densities can be seen in Figure 2.6. The posterior mode and the credible interval can both be seen on the right side of Table 2.9 for each of these four prior distributions. The other two prior distributions, the gamma(0.1, 1) and the gamma(100, 10), were unsuccessful at identifying the patterns. We briefly examine why this happens below.

As seen in Figure 2.6, while the prior distributions vary widely, the posterior distributions are all fairly similar. Both the exponential and half-Cauchy prior distributions place the most mass at zero, and then the mass decreases as τ increases. This is somewhat undesirable as we

expect τ to be greater than zero, though both distributions have heavy tails. The gamma(2, 0.2) prior distribution also has a heavy tail, but puts very little mass at zero. Even though it places most of the mass well above the true value of 2.4 (the mean is 10 and mode is 5), the wide variance allows us to consider a wide range of τ values and is likely the most desirable in practice. The last successful prior distribution, the gamma(30, 10), places most of the prior mass around the true value, and while it is successful in this case, the small variance could be a problem when centered away from the true parameter.

In particular, the gamma(100, 10) prior distribution has similarly small variance, but is centered far away from the true value. Finally, the gamma(0.1, 1) prior distribution, which placed most of the mass at zero with fairly light tails, restricts the possible range for τ too close to zero, making pattern identification difficult. For both of these cases, we ran 100 simulations and assessed how often we discovered the pattern of interest. For the gamma(0.1, 1), the pattern was identified occasionally, 14 out of 100 times, with no false positives. Even with some success, placing too much prior mass at zero is clearly a bad idea. For the gamma(100, 10), the pattern of interest was also found occasionally, in 9 out of 100, but false positives were also quite prolific, with 3.4 false positives occurring on average. In this case, restricting the range caused spurious patterns to appear, where even one occurrence of an event following another within 10 seconds would signal a pattern given the low variance of the prior distribution.

In the following section, where we discuss the results on our maternal data, we consider the gamma(2, 0.2) prior distributions. This places 95% of the prior mass between 1 and 28 seconds, which is a reasonable transition time between events in patterns in the context. This also provides a lot of variability with limited mass at zero, both desirable features.

2.5 Maternal Behavior Analysis

The motivation for this research comes from a study on the role of fragmented and unpredictable maternal behavior on various childhood outcomes. Rodent studies have suggested that fragmented behavior, characterized by frequent shifts in attention, leads to poor outcomes for the rodent pups. A key translational step is developing methods that can identify fragmented behavior in humans. We believe that different behavior patterns may be relevant. Our first step is to detect patterns in the maternal behavior. Then, summary statistics can be derived from the patterns, which can then be used to characterize the behavior as either fragmented and unpredictable, or not. For example, we might consider the longest pattern for an individual or the number of patterns longer than two events as possible measures of fragmentation.

To obtain the data, each mother was invited to the research lab and observed playing with their child. There are a total of 121 mothers, with each recorded session lasting for 10 minutes. For this analysis, we consider data that was collected when the child was 12 months old. Video data was also collected at other times during the child's development (6 months, 2 years, etc.), but this data has not yet been annotated.

The number of events in any particular session ranges from about 100 to 400 events. Table 2.10 lists the 24 different event types that were observed. An example of the data recorded from a single mother-child session can be seen in Figure 2.7, where time is on the horizontal axis, and events are on the vertical axis.

Exploratory Data Analysis

One possible method for verifying the plausibility of the patterns that the RPM detected is to do an exhaustive search for patterns, and see if they occur more often than expected.

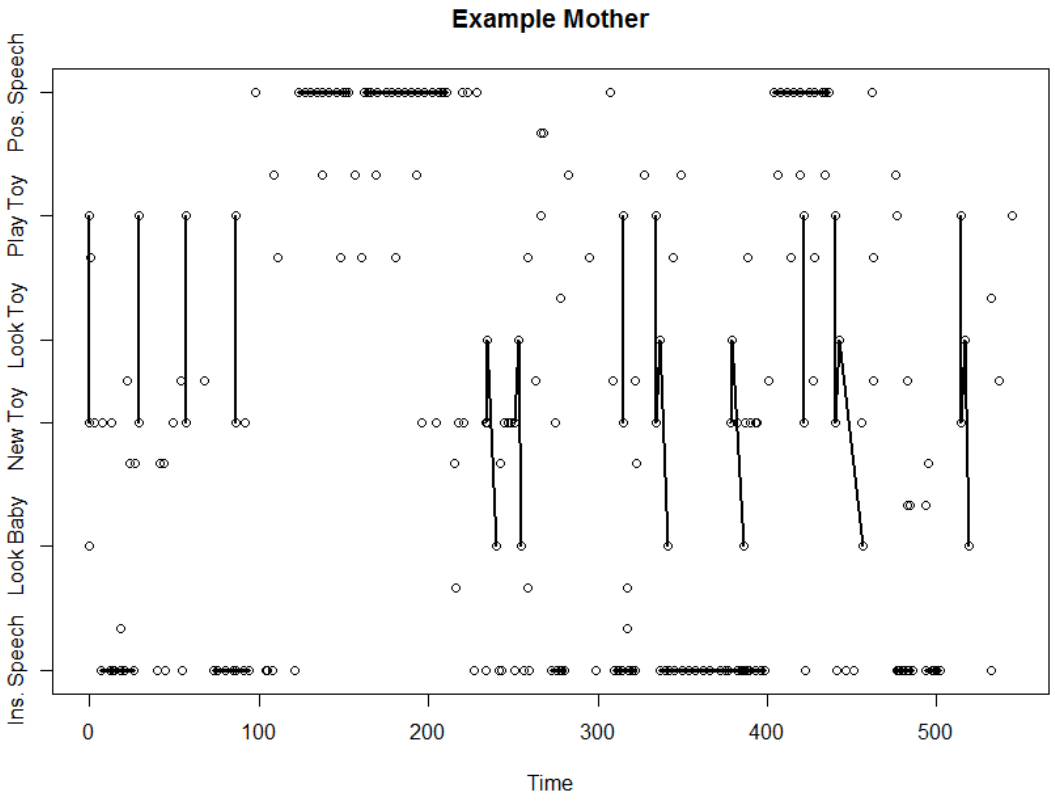

Figure 2.7: An example of maternal behavior data. The three patterns seen in this example include New Toy → Manipulating Toy, Instructive Speech → Instructive Speech, Positive Speech → Positive Speech, and New Toy → Look at Toy → Look at Baby.

Event Types		
Positive Speech	New Toy	Laugh
Negative Speech	Manipulate Toy	Smile
Instructive Speech	No Toy in Hand	Point
Looking at Baby	Give Toy to Child	Frown
Looking at Toy	Remove Toy	Bored
Not Looking	Hold Baby	Content
Affectionate Touch	Restrain Child	Hold Child in Lap
Functional Touch	Pickup Child	Support Child

Table 2.10: The possible event types in the maternal behavior data set. There are a total of 24 different event types.

This can also be done as an initial exploratory step, which can help inform the direction of the analysis. We will only briefly discuss this idea here, but we believe it is an important area for future work.

For example, consider the pattern $A \to B$. We assume that events A and B occur according to Poisson processes with rates λ_A and λ_B respectively. If the pattern $A \to B$ exists, then event B will occur more often than expected in the τ seconds following A, for some fixed τ. Assume there are n_A instances of event A. If $A \to B$ does not exist, then the probability of observing a B event in the τ seconds following an A event is $p_B = 1 - \exp(-\lambda_B \tau)$. Of the n_A of these periods, we would expect to see a B in $E_{A\to B} = p_B n_A$ of them, with a standard deviation of $S_{A\to B} = \sqrt{n_A p_B (1 - p_B)}$.

If we observe B following A considerably more often than expected, then we can assume that the pattern $A \to B$ exists. We denote the number of times that B follows A within a fixed τ seconds to be $O_{A\to B}$. Then we can consider the normalized quantity $\frac{O_{A\to B} - E_{A\to B}}{S_{A\to B}}$. If this number is very large, then we can assume that the pattern exists. Furthermore, if this quantity is very small, then we can say that A appears to inhibit B.

Let us consider the mothers as an example. We consider three different values of τ, 3, 10, and 30 seconds, to examine how it affects pattern identification. We say a pattern exists if the normalized statistic is greater than three. We only consider patterns of length two in this exercise. For each mother, we identify a set of patterns, and then count the number of times, out of 121, that each pattern was identified for a mother. Some of the more common patterns are shown in Table 2.11. These patterns generally align well with the results from the RPM analysis of the maternal data that follows in the next section. Note that τ seems to play an important role in the discovery of patterns by this method. For $\tau = 3$, the most common patterns are New Toy $\to$ Manipulating Toy and Smiling $\to$ Laughing, which become increasingly less common as τ increases. Conversely, the pattern Smiling $\to$ Content never occurs when τ is small, but is quite common when τ is large. Hence one clear advantage of

Pattern	$\tau = 3$	$\tau = 10$	$\tau = 30$
New Toy → Manipulating Toy	107	58	7
Look at Toy → Look at Baby	42	99	105
Smiling → Content	0	78	73
No Toy → Manipulating Toy	27	72	47
Smiling → Laughing	50	25	7
Positive Speech → Positive Speech	38	40	3

Table 2.11: The number of mothers that display various patterns as identified with the descriptive statistics method. Results are shown for three values of τ.

Number of Patterns	1	2	3	4	5	6	7	8	9	10	11
Number of Mothers	7	6	23	22	18	18	9	7	5	3	3

Table 2.12: Distribution of mothers who exhibited a given number of patterns.

our model over this method is that our model is capable of detecting a pattern regardless of the optimal τ.

Fitting the RPM Model

For each of the 121 mothers, we fit the Renewal Process Model with exponentially distributed interarrival times, and using the prior distributions presented in the beginning of Section 4. For each mother, we identify a set of patterns than she performs as a product of our model. The degree to which a mother's behavior is deemed fragmented or unpredictable will be determined from each mother's set of pattern. In particular, we calculate several summary statistics to describe fragmentation, which include the length of the longest repeated sequence, the percentage of sequences longer than two events, and the number of different sequences identified. Note that our definition of fragmented and unpredictable behavior does not consider whether the actions are positive or negative, but rather only considers the structure of the behavior through these summary statistics.

The distribution of the three summary statistics we consider can be found in the following tables: the number of patterns in Table 2.12, the longest pattern in Table 2.13, and the

Maximum Pattern Length	2	3	4	5
Number of Mothers	35	75	10	1

Table 2.13: Distribution of mothers who had a given maximum pattern length.

Percentage	0%	1-20%	21-40%	41-60%
Number of Mothers	35	38	42	6

Table 2.14: Distribution of mothers who had a given percentage of patterns that were longer than two events. The highest percentage was 60%.

percentage of patterns greater than length two in Table 2.14. All three statistics are correlated. In particular, the maximum pattern length and percentage of patterns greater than length two have a correlation of 0.83. The number of patterns has a correlation of 0.05 with maximum pattern length, and a correlation of 0.21 with the percentage of patterns greater than length two.

The simulations of Section 2.4 suggest that individuals with more recorded events are likely to have more patterns identified. Thus there may be a concern that the number of patterns that a mother displays may simply be an artifact of how active she was during the ten minute observation window. However, we did not find that to be the case, as there was no significant relationship between the number of events and the number of patterns ($R = 0.02$). Thus, there is reason to believe that we have been able to identify most of the true patterns.

Pattern	Number of Mothers
New Toy $\rightarrow$ Manipulating Toy	95
Look at Toy $\rightarrow$ Look at Baby	54
Smiling $\rightarrow$ Content	36
Instructive Speech $\rightarrow$ Instructive Speech	25
Smiling $\rightarrow$ Laughing	23
Positive Speech $\rightarrow$ Positive Speech	21
Affectionate Touch $\rightarrow$ Affectionate Touch	17
New Toy $\rightarrow$ Look at Toy $\rightarrow$ Look at Baby	7

Table 2.15: The most common patterns discovered in the maternal behavior data using the Renewal Process Model. The number of mothers who exhibited a given pattern is also provided (out of a total of 121 mothers).

The sequences that our model detects are largely quite sensible. Examples of detected

patterns, with the number of mothers (out of 121 total) demonstrating each pattern, can be found in Table 2.15. The most common patterns are not particularly surprising, and are similar to the patterns identified with our exploratory data analysis. One might actually expect these patterns to occur even more often. For example, all the mothers probably play with toys, or manipulate them in some way, once they select a new toy. However, only about 80% of the mothers actually exhibited this pattern. Most likely, this occurs because the remaining 20% of mothers did not pick up a new toy often enough during the ten minute session for our model to detect the pattern. We will return to this point in the next chapter.

Another interesting point is that several of the most common patterns are characterized by a mother performing the same action repeatedly. For example, an incident of instructive speech is often followed by another incident of instructive speech and an affectionate touch is often followed by another affectionate touch. Such bursts of action are interesting to behavior researchers, though researchers may not characterize them as patterns in the same sense as other patterns involving different event types. We return to this issue in Chapter 4.

There is still a lot of work to be done with the maternal behavior analysis. For instance, as mentioned above, we only consider data recorded for children at 12 months old. At the moment, only the videos for the 12 month old data have been annotated. Once the rest of the videos have been annotated, considerably more work can be done using RPM to identify, and validate, fragmented and unpredictable maternal behavior.

The mother-child interaction changes significantly as the child ages, which would significantly alter the patterns. For example, at 6 months the child cannot walk, so mother-holding-child events are far more common than later in the child's life. While the patterns should change, ideally we expect the characterization of the mother to remain consistent. To this end, we expect significant correlation across the child's age for the summary statistics that we considered above.

Furthermore, we only consider data for the mother's actions. Our collaborators are also recording data for the baby's actions, which was not available in time for this thesis. The baby's actions will often dictate what the mother does, so seemingly random maternal behavior may actually be in response to her child. Modeling patterns for the mother while accounting for the child's actions will be a future direction for the Conte Center.

2.6 Discussion

In this chapter, we proposed a generative model, the renewal process model, and an inference procedure for identifying recurring patterns from a stream of events in continuous time. Our experiments show that our model correctly identifies patterns in simulated data, performs admirably with increasing complexity, and finds fewer false positives than the current state of the art algorithm. Furthermore our model identifies both interesting and intuitive patterns in real data.

However, our current model as discussed in this chapter has several limitations. As witnessed by our motivating example, behavior data can be very difficult to obtain and thus only exists in limited quantity. This negatively impacts both inference and pattern detection. In settings where multiple exchangeable observation sets occur, one possible solution is to pool the data as part of a hierarchical model, which we discuss in Chapter 3.

While our model makes no distributional assumptions, we only considered Poisson processes for modeling both the background and sequence processes in our experiments. This may be a poor model in some applications. In Chapter 4, we consider both parametric and nonparametric alternatives, and discuss the advantages and disadvantages of the alternatives.

While our model works well in practice, it does suffer from being fairly slow. MCMC is notoriously slow due to the poor computational cost of sampling and the potentially large

time required for convergence. While the HMC implementation that we used is relatively fast for a sampling procedure, an optimization procedure may be preferable. We are currently in the early stages of designing a mode finding algorithm to identify the most likely patterns, but will not consider it further in this thesis.

Finally, our model assumes that all background processes are independent, and whatever dependency might exist is completely accounted by the sequence processes. At the moment, we have not examined the potential ramifications of this assumption. If this assumption leads to issues in practice, then alternative models will need to be considered. We do not discuss this further, however we believe that it is an important future direction for our model.

Chapter 3

Population Level Pattern Analysis

Chapter 2 develops a model for mother-child behavioral data collected from a single observation period. However, data is often collected from a set of unrelated actors observed in similar settings or via repeated observations of a particular actor. One approach in this case would be to model each series of behavioral events independently. We adopted this approach when examining the maternal behavior data at the end of the previous chapter. However, modeling data independently like this can lead to poor inference when there are population effects that may not be observed in a single observation period.

Patterns that are common across the population may not be evident for certain actors with limited data, even if the patterns would be detected over a longer observation period. By grouping the observation sets into a single model, we can pool across the actors to help identify patterns common across the population, even when they occur insufficiently often to be identified for a single actor (Gelman et al., 2013). This issue is particularly important with data such as ours, where converting a video to data is expensive and time consuming, limiting the amount of available data. Furthermore, the results of Chapter 2 suggest that it is reasonable to expect similar patterns across the population of mothers, and this assumption

will likely be valid for other behavioral situations as well.

In this chapter, we propose a hierarchical model to improve inferences by borrowing strength across actors. The Renewal Process Model of Chapter 2 is applied to each observation period, with each observation period having unique parameters. Those parameters are then modeled as elements from a common population distribution. This has a number of advantages, most notably that we may better identify sequences that tend to exist in the population, but are not evident in each and every observation period.

Hierarchical models also provide a principled approach to assessing differences between populations. The behavior of individuals from different populations often have qualitative differences, but such differences may be difficult to describe quantitatively. A hierarchical model will provide population parameters for both populations, which can be compared to assess a difference in behavior. Populations of interest might include demographic differences (such as comparing younger and older mothers), or with experimental differences (for instance, assessing the behavior of individuals after different stimuli).

In Section 3.1, we describe a hierarchical renewal process model for modeling behavioral data. Section 3.2 develops a specific population model that allows for individuals to exhibit both unique patterns and patterns common across the population. Our approach to inferences is described in Section 3.3. Simulated examples are presented in Section 3.4 to illustrate the power of a hierarchical model. We then apply the hierarchical model to examples from our motivating study on fragmented maternal behavior in Section 3.5.

3.1 A Hierarchical Renewal Process Model

We consider a hierarchical renewal process model as an alternative to fitting the RPM model of Chapter 2 independently to each observation period. We consider M different records of

behavioral observations, and use the subscript m to note quantities relevant to the mth observation period. For example, we assume that observation m contains of n_m behavioral events.

Denote an observed sequence of events for the m^{th} record as $\{t_{i,m}, E_{i,m}\}$, $i = 1, \ldots, n_m$. Then the RPM of Chapter 2 can be applied to the data with $\Theta_m^{\mathcal{B}}$ denoting the background process parameters and $\Theta_m^{\mathcal{S}}$ denoting the sequence process parameters. As a shorthand for the model, we write

$$\{t_{i,m},\, E_{i,m}\}\,|\Theta_m^{\mathcal{B}},\, \Theta_m^{\mathcal{S}} \sim RPM\left(\Theta_m^{\mathcal{B}},\, \Theta_m^{\mathcal{S}}\right) \tag{3.1}$$

The hierarchical model assumes that the parameters $\Theta_m^{\mathcal{B}}$ and $\Theta_m^{\mathcal{S}}$ are drawn from population distributions. We will denote these population distributions as $g\,(\cdot)$, which are assumed to depend on population parameters $\Theta^{\mathcal{B}}$ and $\Theta^{\mathcal{S}}$ for background and sequence processes respectively. These parameters represent process parameters for the population, and are thus indicative of the population level effects. The individual-level parameters are drawn from the population distributions as such:

$$\Theta_m^{\mathcal{B}}|\Theta^{\mathcal{B}} \sim g^{\mathcal{B}}\left(\Theta_m^{\mathcal{B}}|\Theta^{\mathcal{B}}\right) \qquad \Theta_m^{\mathcal{S}}|\Theta^{\mathcal{S}} \sim g^{\mathcal{S}}\left(\Theta_m^{\mathcal{S}}|\Theta^{\mathbf{S}}\right) \tag{3.2}$$

Finally, we assume that the population level parameters are drawn from prior distributions $h\,(\cdot)$ that depend on hyperparameters $\omega^{\mathcal{B}}$ and $\omega^{\mathcal{S}}$ for background and sequence processes

respectively. We represent these draws as follows:

$$\Theta^{\mathcal{B}}|\omega^{\mathcal{B}} \sim h^{\mathcal{B}}\left(\Theta^{\mathcal{B}}|\omega^{\mathcal{B}}\right) \qquad \Theta^{\mathcal{S}}|\omega^{\mathcal{S}} \sim h^{\mathcal{S}}\left(\Theta^{\mathcal{S}}|\omega^{\mathcal{S}}\right) \tag{3.3}$$

To make things more concrete, let us consider a single parameter in more detail. For example, consider the time parameter τ, which must exist no matter the assumed distribution for interarrival times in $RPM\left(\Theta_m^{\mathcal{B}}, \Theta_m^{\mathcal{S}}\right)$. For the ℓ^{th} event in sequence s, we are interested in the individual level parameters $\tau_{s\{\ell\},m}$. For individual M, this gives the maximum amount of time for the ℓ^{th} event to occur after event $\ell - 1$ in sequence s. From 3.2, the set of parameters $\tau_{s\{\ell\},m}$, $m = 1, ..., M$ are drawn from some distribution $\tau_{s\{\ell\},m}|\tau_{s\{\ell\}} \sim g\left(\cdot|\tau_{s\{\ell\}}\right)$. For example, we can fix $f\left(\cdot|\tau_{s\{\ell\}}\right)$ to be a gamma distributions with parameters as follows:

$$\tau_{s\{\ell\},m}|\tau_{s\{\ell\}} \sim gamma\left(c\left(\tau_{s\{\ell\}}+1\right), c\right). \tag{3.4}$$

This is a gamma distribution with a mode at $\tau_{s\{\ell\}}$, so the population level parameter represents the most likely outcome for an individual from the population. The parameter c controls the amount that the τ parameters may vary in the population. In particular, the variance is $\frac{\tau_{s\{\ell\}}+1}{c}$, so a large c corresponds to small variance, while a small c corresponds to large variance. Note that we can use an alternative parameterization with gamma shape parameter $c\tau_{s\{\ell\}}$ to fix the population mean, rather than the mode, at the population level parameter $\tau_{s\{\ell\}}$.

Finally, we must also consider a prior distribution on the population parameters $\tau_{s\{\ell\}}$ and c. As both of these are strictly positive quantities, we can simply assume they are drawn

from positive distributions, such as $gamma\,(a_\tau, b_\tau)$ and $gamma\,(a_c, b_c)$ respectively. The parameters for both these distributions are determined by the researcher in practice to reflect whatever prior information is known.

3.2 Population Distributions for Sequence Processes

The population model described in the previous section assumes that the parameters for all individuals in the population are drawn from the distribution $g\,(\cdot)$. The distribution $g\,(\cdot)$ can be very different for background and sequence processes. It is natural to assume background processes are always present so $g\,(\cdot)$ would describe the distribution of the rate parameter. However, for sequence parameters, a given pattern may only be observed in a fraction of the population, so any specific sequence process parameter would only be present in that fraction of the population. Thus we want to consider a population distribution for the sequence process parameters that conforms to this assumption. The results of our study of maternal behavior in Chapter 2 supports this approach. Patterns such as smiling→content were common throughout the mothers, indicating that these are widespread in the population. However, other patterns were exhibited by only a few mothers, and these may be difficult to identify in a population model.

Our expectation is that a given sequence is present in some of the population, but not in the rest of the population. We accommodate this by assuming the population distribution $g\,(\cdot)$ for the sequence process parameters is a mixture of two distributions: the first corresponding to the portion of the population where the sequence is present, and the second corresponding to the other portion.

In this chapter, we again assume that the renewal processes for all individuals have exponentially distributed time increments. Hence each sequence process is characterized by a

rate parameter, which controls how often the corresponding sequence process occurs. For a given sequence process s and individual m, denote the rate parameter for the ℓ^{th} event in the sequence as $\theta_{s\{\ell\},m}$. For individuals who display the sequence, $\theta_{s\{\ell\},m}$ is drawn from a population distribution $g^s(\cdot)$ that depends on population parameters $\Theta_{s\{\ell\}}$. Otherwise, we assume that the pattern may be present, but rare, for individual m and thus assign a second population distribution $g_0^s(\cdot)$ with heavy mass near zero.

We denote the weight for the nonzero portion as p_s, where $0 < p_s < 1$, and thus the weight for the zero portion is $1 - p_s$. We refer to p_s as the prevalence parameter, indicating the percentage of the population expected to display pattern s during a sequence of events. The resulting population distribution is then

$$\theta_{s\{\ell\},m}|\Theta_{s\{\ell\}} \sim p_s g^s\left(\cdot|\Theta_{s\{\ell\}}\right) + (1 - p_s)\, g_0^s(\cdot) \tag{3.5}$$

For example, we can consider a mixture model where one component is a gamma distribution with mean $\theta_{s\{\ell\}}$ and the other is a gamma distribution with mode equal to zero and small variance. We note that a gamma distribution with shape parameter α and scale parameter β will have a mode or asymptote of 0 if $\frac{\alpha-1}{\beta} < 0$. An example of this distribution can be seen in Figure 3.1. The population distribution for this example would then be

$$\begin{aligned} \theta_{s\{\ell\},m} \sim \quad & p_s\, gamma\left(c\theta_{s\{\ell\}},\, c\right) + (1 - p_s)\, gamma\left(\alpha,\, \beta\right) \\ \theta_{s\{\ell\}} \sim \quad & gamma\left(a,\, b\right) \end{aligned} \tag{3.6}$$

where $\frac{\alpha-1}{\beta} < 0$, and a and b are hyperparameters chosen by the researcher to reflect whatever prior knowledge exists.

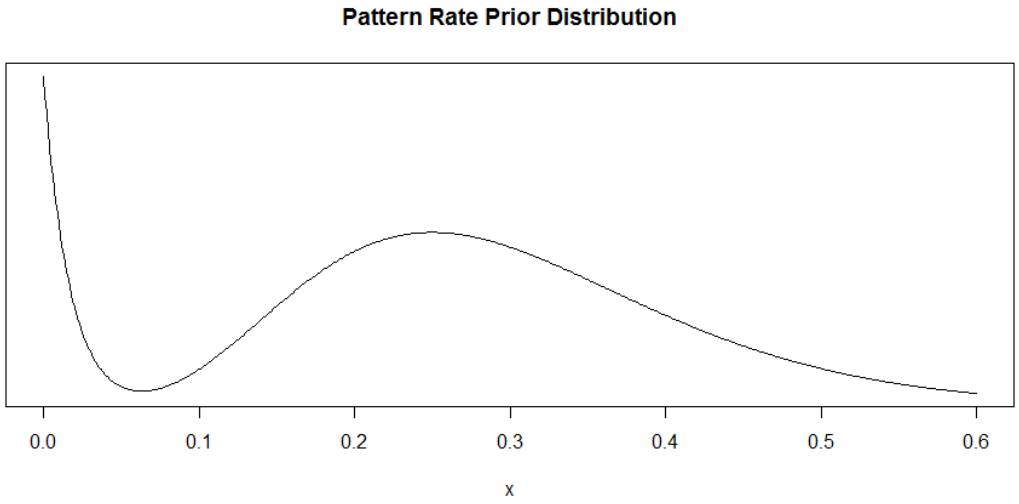

Figure 3.1: The population level prior distribution that we consider for the sequence process rates. It is a mixture model with two components, one near zero indicating no pattern is present, and one centered around a nonzero rate.

In addition to providing the population prevalence, we can also determine the likelihood that a sequence exhibits a given pattern by examining the posterior distribution of $\theta_{s\{\ell\},m}$. When the posterior distribution of $\theta_{s\{\ell\},m}$ is concentrated close to 0, then we can conclude that the actor either does not exhibit pattern s or exhibits it only rarely.

3.3 Inference

Our inference procedure for the hierarchical renewal process model is similar to the inference procedure described in the previous chapter. To build the posterior distribution, first recall the form of the likelihood for a single individual m: $L\left(\mathbf{t},\, \mathbf{E}|\Theta\right) = \prod_{i=1}^{n_m} f_{t_{i,m},\, E_{i,m}}\left(t,\, e\right)$. Hence the likelihood for the entire sample of mothers is $L\left(\mathbf{t},\, \mathbf{E}|\Theta\right) = \prod_{m=1}^{M} \prod_{i=1}^{n_m} f_{t_{i,m},\, E_{i,m}}\left(t,\, e\right)$. Finally, define the entire parameter space Θ as $\Theta = \left\{\Theta_m^{\mathcal{B}}, \Theta_m^{\mathcal{S}}, m = 1, ..., M, \Theta^{\mathcal{B}}, \Theta^{\mathcal{S}}, \omega^{\mathcal{B}}, \omega^{\mathcal{S}}\right\}$.

Then the posterior distribution is proportional to the following quantity:

$$P\left(\Theta|\mathbf{t},\,\mathbf{E}\right) = \prod_{m=1}^{M}\prod_{i=1}^{n_m} f_{t_{i,m},\,E_{i,m}}\left(t,\,e\right) g^{\mathcal{B}}\left(\Theta_m^{\mathcal{B}}|\Theta^{\mathcal{B}}\right) g^{\mathcal{S}}\left(\Theta_m^{\mathcal{S}}|\Theta^{\mathcal{S}}\right) h^{\mathcal{B}}\left(\Theta^{\mathcal{B}}|\omega^{\mathcal{B}}\right) h^{\mathcal{S}}\left(\Theta^{\mathcal{S}}|\omega^{\mathcal{S}}\right) \quad (3.7)$$

We again consider the No-U-Turns Sampler (NUTS) for sampling from the posterior distribution, with birth-death process steps occurring every several iterations of sampling to explore the space of all possible patterns.

The birth-death process we consider for the hierarchical model has some slight modifications from what we considered in the previous chapter. Most notably, when adding a pattern during the birth step, we add the pattern for all individual observation periods. However, this creates a problem when the pattern is only exhibited by a small number of individuals, as the negative impact for the other individuals is likely to overshadow whatever positive impact there is for the posterior distribution. One approach to addressing this issue is to propose a small rate parameter (near zero) when adding a new sequence. This provides an opportunity for the data to provide information on prevalence. For example, if we expect the rate of a pattern to be around 0.1 to 1.0 (this is a typical range for rate parameters in the mother-child data), we could propose a rate about a magnitude lower, such as 0.01. For individuals who exhibit the pattern, this is generally provides enough signal not to remove the pattern, while not significantly negatively affecting the likelihood for the individuals who do not exhibit the pattern.

3.4 Experiments

In this section we consider several simulation studies to experiment with the accuracy and applicability of the hierarchical RPM model.

3.4.1 Efficacy of the Hierarchical Renewal Process Model

We first explore the hierarchical model with a simulated data set of ten individuals and four sequence processes. We vary the number of events per individual to assess whether the hierarchical model affects small data individuals differently than large data individuals. The total number of events per individual is drawn from a negative binomial distribution with mean $4.5 * m$ for $m = 1, ..., 10$. This ensures that the individuals range from very few events to over 500, which is consistent with our maternal data. In this initial experiment, all ten individuals perform all four possible sequences, and hence each sequence process has a prevalence of one. All patterns contain only two events. All renewal processes have exponentially distributed time increments.

Throughout our experiments and the data analysis in the following section, we will fix the prior distribution for the sequence rate parameters to follow Equation 3.6 with fixed c = 1. The prior distribution for the population rate parameters will be an exponential with rate one. The prior distribution for the individual level time parameters τ_m will be a gamma distribution centered at the population level τ, which will have a gamma(2, 0.2) prior distribution, following from the results of Chapter 2.

In this scenario, we consider a total of 20 possible event types. The population rates for the corresponding background processes have rates of either 0.1 or 0.2. The background rates for individuals are sampled from a normal distribution centered around the population rate, with a standard deviation of 0.02. This means that in addition to the 20 background processes,

	Rate λ_s		Max Time τ_s	
	Truth	CI	Truth	CI
1→2	0.16	(0.15, 0.22)	2.01	(1.94, 2.81)
2→6	0.23	(0.17, 0.24)	2.49	(2.14, 2.95)
3→4	0.18	(0.18, 0.30)	2.49	(2.46, 3.12)
11→12	0.25	(0.25, 0.40)	3.05	(2.54, 3.08)

Table 3.1: Population parameters for the four patterns. All four patterns are assumed to occur for each simulated mother, hence no prevalence parameter.

we consider a total of four sequence processes. The population rates and population τ parameters for each of these processes can be found in Table 3.1. The individual level sequence rates are drawn from a normal distribution centered at the population rate with a standard deviation of 0.02. The maximum time parameter for each individual is also drawn from a normal distribution centered at the population parameter, but with standard deviation 0.5.

Our model performs well in this instance, with all patterns successfully identified. We will first consider the results of the hierarchical model, and then compare them against the individual-level results at the end of this section. Using the MCMC samples, we calculated 95% credible intervals for the population parameters, which appear in Table 3.1. The results are very promising, indicating successful estimation of the population parameters.

As a second example, we vary this simulation to include prevalence values under one. Each of the four patterns is given a prevalence between 0.5 and 0.8, indicated in Table 3.2, so that each individual does not exhibit all four patterns. All model parameters are determined exactly as in the previous simulation experiment.

The model performs well in this scenario as well, and successfully identifies all four patterns. We again calculated 95% credible intervals for the population parameters, which can be found in Table 3.2. Again, the credible intervals look good, and perhaps most importantly, the model appears to accurately model the prevalence of each pattern.

	Rate λ_s		Max Time τ_s		Prevalence p_s	
	Truth	CI	Truth	CI	Truth	CI
1→2	0.16	(0.13, 0.17)	2.01	(3.27, 4.04)	0.6	(0.57, 0.63)
2→6	0.23	(0.21, 0.26)	2.49	(3.09, 4.02)	0.5	(0.46, 0.54)
3→4	0.18	(0.15, 0.21)	2.49	(2.04, 2.74)	0.7	(0.71, 0.79)
11→12	0.25	(0.29, 0.34)	3.05	(3.07, 4.16)	0.8	(0.79, 0.86)

Table 3.2: Population parameters for the four patterns. All four patterns are no longer expected to occur for each mother, thus we estimate the prevalence of each pattern.

In addition to the population parameters, we want to see the effect of the hierarchical model on the individual level parameters. In particular, we generated the data such that individuals would range from almost no data to over 500 data points. For these simulations, the lowest amount of data for an individual was 25 data points, ranging to 500 at the maximum. For individuals with little data, we expect that the estimates, for both background and sequence processes, would be poorly estimated for an individual level model as described in Chapter 2.

Because of this, we demonstrate the effect of partial pooling across the population on a selection of individual level rates. In particular, we consider the background rate for event type 1, and the sequence rate for the sequence $3 \rightarrow 4$.

For each of the ten simulated individuals, we recorded parameter estimates for the background and sequence processes. We then compared these estimates with the estimates obtained using the hierarchical model. The results for the background rate of type 1 events can be seen in Figure 3.2. In addition to the individual and hierarchical model results, the true parameter values were also plotted. In most cases, there was clear evidence of shrinkage towards the true values. The number of total events, and the number of type 1 events, are both indicated. Some level of shrinkage is displayed, even for the simulated individuals with relatively more data. Data containing about 500 events is comparable to the mothers with the most data in our application, so we can assume that shrinkage is expected for most mothers in our data set.

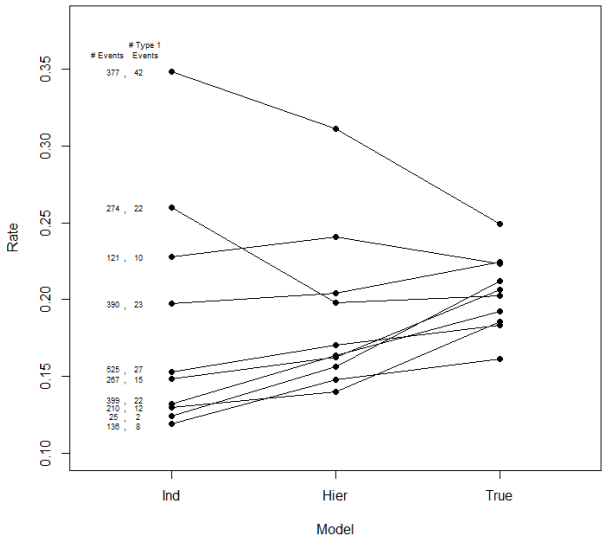

Figure 3.2: Estimates (posterior mean) for the background rate parameter of the type 1 event. The figure includes results from the individual level analysis (left), results from the hierarchical model (middle), and the true parameters (right). Most parameters display some level of shrinkage under the hierarchical model. The number of total events, as well as the number of type 1 events, are indicated on the far left.

Similarly, the results for the sequence process rate can be found in Figure 3.3. Out of the ten simulated individuals, three of them never displayed the pattern. Unsurprisingly, the individual-level RPM was unable to identify the pattern. For the hierarchical model, these individuals had low rates, though due to limited data and a pooling effect, the rates were much larger than the true value of 0.001. However, there is still clear separation between these three and the rest of the simulated individuals, as we would expect.

In addition, two individuals displayed the sequence rarely enough that the individual model was unable to identify it. However, for both of these individuals, the hierarchical model was able to estimate rates much higher than the three no-pattern individuals, displaying a successful pooling effect. These two individuals, who had a limited amount of data but displayed the pattern nonetheless, clearly benefited from the hierarchical model.

Of the five simulated individuals where the standard individual-level RPM successfully iden-

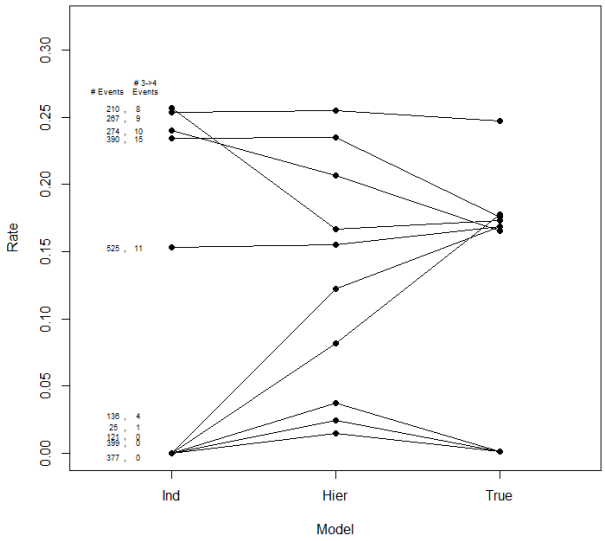

Figure 3.3: Estimates (posterior mean) for the sequence rate parameter for the $3 \rightarrow 4$ sequences. The figure includes results from the individual level analysis (left), results from the hierarchical model (middle), and the true parameters (right). Most parameters display some level of shrinkage under the hierarchical model. The number of total events, as well as the number of patterns, are indicated on the far left.

tified the sequence, the hierarchical RPM displayed clear shrinkage for two of them, and a similar value for the other three. For these individuals, who had more data and displayed the pattern more often, the hierarchical model had less of an effect, though it was still positive.

3.4.2 Detecting Sequences with Varying Levels of Prevalence

While the model performed well with varying prevalence, we would like to get a better idea of how prevalence affects model selection. For this task, we consider a simpler model, with only 10 events types and only one pattern, but with a larger sample of 100 individuals. To gauge the effect of prevalence, we vary the prevalence parameter for the patterns from 0.1 to 0.9 in increments of 0.1. The population-level background rates ($\lambda_{BG} = 0.1$) and the population-level sequence rate ($\lambda_S = 0.2$) were held constant for each of the different prevalence scenarios. These values were specifically chosen so that the pattern would eas-

Prevalence	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Percentage	0.02	0.05	0.19	0.42	0.60	0.89	0.97	0.99	0.98

Table 3.3: For each of nine different simulation scenarios, each with a different prevalence level, 100 simulations were run. The percentage of times a pattern was discovered out of total of 100 simulations is displayed. with varying levels of prevalence.

ily be discovered with high prevalence, thus limiting the effect of rates on the simulation experiment.

For each prevalence value, we simulate 100 data sets, each comprised of 100 simulated individuals. The percentage of time that the pattern was discovered is included in Table 3.3. For scenarios where the prevalence of a pattern was over 0.6, the patterns were successfully detected most of the time. However, pattern detection degrades quickly as prevalence decreases. For prevalence values between 0.4 and 0.6, the pattern was detected roughly half of the time, and for prevalence values below 0.4, the pattern was rarely, if ever, detected.

This could be due to a number of reasons. The hierarchical model may not give much support to rare patterns, making detection difficult. The more likely cause is due to poor parameter proposals in the birth step. For rare patterns, if the birth step gives a large parameter value to a sequence that never occurs for a mother, it has a significant impact on the likelihood. When there are few mothers who actually display this pattern, a few poor proposals could negatively affect the posterior distribution, so the pattern quickly is removed in a death step. We are currently working on alternative proposal distributions for the birth step.

3.5 Maternal Behavior Analysis Revisited

In Chapter 2 we focused on data from human mothers interacting with their children. The Conte Center research program includes both human and animal studies. We can apply the hierarchical model in humans, and do so below. However, this is challenging because we do

not have information about population heterogeneity with respect to fragmentation. Thus the appropriate population distribution is not obvious. In animal studies it is possible to induce fragmented maternal behaviors. Thus we have two clear populations, so we can fit the hierarchical RPM to both and compare results.

3.5.1 Rodent Maternal Behavior Analysis

In addition to the human behavioral data that we examined in Chapter 2, we also have rodent behavioral data. This data comes from observations of mother rats (dams) caring for their offspring (pups) over a series of days. Rats were observed from days two through nine after the birth of the pups, which roughly corresponds to the period of early childhood in humans. Each rat was observed for two hours each day, for a total of 16 hours of observation per rat. A total of seven possible behavioral events were recorded for the dams: carrying pups (C), eating (E), nursing (N), self grooming (SG), licking or grooming the pups (LG), nest building (NB), and off pups (O). Off pups indicates the dam is not attending to the pups and not otherwise engaged in a behavior of interest.

To induce fragmentation, select dams were only given a single sheet of paper for bedding. This provides the dam with inadequate bedding, causing her behavior to become erratic. Ultimately, pups raised in a fragmented environment display negative outcomes. The control dams are given adequate bedding, and thus display normal maternal behavior. Furthermore, the offspring for the control mothers have been shown to display better outcomes.

The sample that we consider has a total of 12 rats: six raised their pups in a control environment, and six raised their pups in a fragmented environment. All have 16 hours of observation. For each of the two populations (control and fragmented), we fit the hierarchical model assuming exponentially distributed time increments. We consider the same prior distributions as those outlined in the simulation experiments, with one exception. Because

	Control		Fragmented	
Event	Rate	95% Cred. Int.	Rate	95% Cred. Int.
C	0.18	(0.13, 0.25)	0.29	(0.25, 0.32)
E	0.35	(0.29, 0.46)	0.34	(0.30, 0.40)
LG	0.37	(0.32, 0.43)	0.36	(0.24, 0.43)
N	0.35	(0.22, 0.48)	0.50	(0.44, 0.58)
NB	0.24	(0.21, 0.29)	0.36	(0.32, 0.41)
O	0.69	(0.55, 0.89)	0.64	(0.57, 0.74)
SG	0.49	(0.38, 0.66)	0.60	(0.48, 0.79)

Table 3.4: The population-level background rate parameters for all events in both control and fragmented rats. Both the posterior mean and 95% credible interval are included. There is overlap in the credible intervals for all events but nest building, indicating that the rates are relatively similar.

the rats have much larger time between events than the humans do, the population level prior distribution is set as a gamma distribution with shape 2 and scale 0.02.

Parameter estimates (posterior mean and 95% credible interval) for the background process rates can be found in Table 3.4. The background rates for each of the seven event types were similar for the two groups, though a number of events (nursing, carrying pups, and nest building) appear to occur at a higher frequency in the fragmented rats.

The same six patterns were discovered for both the control and fragmented rats. For all six parameters, population prevalence was nearly one, indicating that all the rats displayed each pattern. The estimates (posterior mean) for the population level sequence process rates are displayed in Table 3.5, and the estimates for the maximum time parameters τ can be found in Table 3.6. The sequences themselves are displayed in both tables.

The maximum times for each pattern are comparable between control and fragmented rats. However, the population sequence process rates are uniformly larger for the control rats than the fragmented rats. Because the maximum time parameters are comparable, higher sequence process rates indicate that the patterns occur more frequently. Thus the control and fragmented rats display the same patterns, but the patterns occur more often in control

	Control		Fragmented	
Patterns	Rate	95% Cred. Int.	Rate	95% Cred. Int.
SG→O	0.33	(0.25, 0.42)	0.05	(0.03, 0.06)
O→E	0.21	(0.17, 0.24)	0.15	(0.12, 0.19)
LG→N	0.25	(0.16, 0.29)	0.09	(0.07, 0.10)
O→SG	0.17	(0.14, 0.21)	0.05	(0.04, 0.05)
NB→NB	0.10	(0.07, 0.13)	0.08	(0.07, 0.09)
SG→LG	0.10	(0.08, 0.13)	0.07	(0.06, 0.08)

Table 3.5: The population-level sequence process rate parameters for all patterns in both control and fragmented rats. Both the posterior mean and 95% credible interval are included. The rate parameters for the control rats are uniformly larger, indicating that the patterns occur more often.

	Control		Fragmented	
Patterns	Time	95% Cred. Int.	Time	95% Cred. Int.
SG→O	223	(177, 357)	307	(272, 355)
O→E	328	(234, 430)	233	(200, 263)
LG→N	320	(268, 370)	192	(154, 237)
O→SG	371	(300, 449)	450	(322, 543)
NB→NB	428	(311, 542)	412	(346, 511)
SG→LG	244	(185, 310)	342	(271, 402)

Table 3.6: The population-level sequence process time parameters for all patterns in both control and fragmented rats. Time indicates the maximum amount of time between events in a sequence in seconds. Both the posterior mean and 95% credible interval are included.

Pattern	Prevalence	Ind. Model Percentage
New Toy → Manipulating Toy	0.86	0.79
Look at Toy → Look at Baby	0.79	0.45
Smiling → Content	0.68	0.30
Smiling → Laughing	0.65	0.19
Positive Speech → Positive Speech	0.57	0.17
Pointing → New Toy	0.42	0.11

Table 3.7: Each of the six patterns discovered by the hierarchical model. The prevalence column is the posterior mean of the prevalence parameter. The individual model percentage refers to the percentage of mothers, out of 121, who displayed that pattern according to the individual level RPM from Chapter 2.

rats. This allows us to consider a new hypothesis; that pattern frequency, in addition to complexity, may be an important indicator of the level of fragmented behavior.

3.5.2 Human Maternal Behavior Analysis

We next apply the hierarchical renewal process model to the human mother-child behavior data. As a first step we apply the hierarchical model to all 121 mothers in our data set, treating them as coming from a single population. The prior distributions we chose for this analysis follow directly from the ones used in Section 4. A total of six patterns were discovered, which are displayed in Table 3.7. All of the detected patterns are fairly common patterns from the individual level model. In addition to the patterns themselves, the table also displays hierarchical model prevalence and the percentage of mothers who exhibited the pattern under the individual level model.

Prevalence was higher than the individual level percentage for all patterns. This indicates that for each pattern, there seem to be mothers who display the pattern, but not often enough for it to be detected from the individual level model alone. It is possible that other patterns exist, but as shown in our simulation results, the ability for our model to detect them degrades as the prevalence decreases.

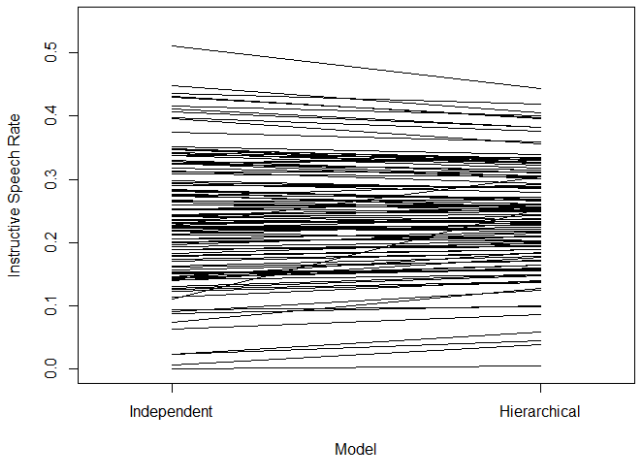

Figure 3.4: Estimates (posterior mean) of the background rate parameter for Instructive Speech. The figure includes results from the individual level analysis (left) and results from the hierarchical model (right). Most parameters display some level of shrinkage under the hierarchical model.

In addition to better pattern detection, the hierarchical RPM is also capable of improved inference. The estimated rate parameters (posterior mean) for the Instructive Speech background event are plotted in Figure 3.4. Estimates from both the individual level RPM and the hierarchical RPM are included. As expected, some level of shrinkage occurs across the sample. This particular event is fairly common, hence the relatively modest shrinkage effect.

Shrinkage results for one of the sequence process rate parameters can be found in Figure 3.5. In particular, we consider the pattern New Toy $\rightarrow$ Manipulate/Play with Toy. The individual level model discovered this pattern for 95 out of 121 mothers. Of the 26 mothers where the pattern was not detected, the rate is increased from zero but remains small. However, the distinction between mothers who display the pattern and those who do not, according to the individual model, is blurred by the hierarchical model. The pattern does seem to occur in the data for some of these mothers, but not often enough for the individual model to properly detect.

Unlike the rodents, there is no natural way to split the human mother-child pairs into two or

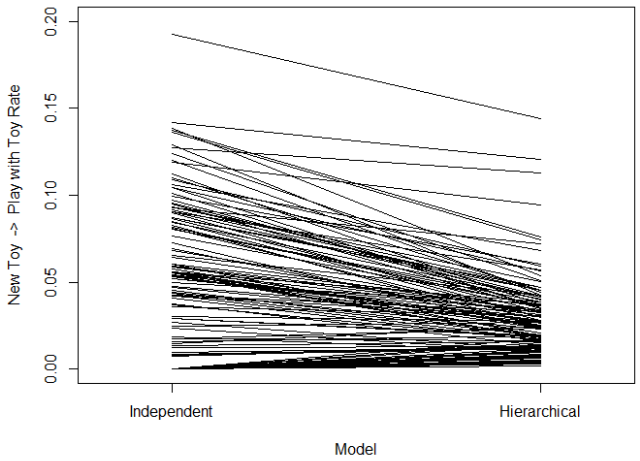

Figure 3.5: Estimates (posterior mean) of the sequence rate parameter for the pattern New Toy $\rightarrow$ Play with Toy. The figure includes results from the individual level analysis (left) and results from the hierarchical model (right). The rates are generally smaller for the hierarchical model than the individual model. The mothers who did not appear to display the sequence from the individual model results have estimated rates that indicate the pattern might actually occur, just not as often.

more subpopulations. We have split the mothers into mothers of boys and mothers of girls, and run the hierarchical RPM on both groups. However, there were no apparent differences between the two groups. Our collaborators have suggested using depression (high and low depressive symptoms) as well as parity (nulliparous and multiparous mothers, or mothers with no previous children and mothers with previous children). This is currently an active area of research.

3.6 Discussion

In this chapter, we presented a hierarchical renewal process model, allowing us to generalize our model to obtain population level inferences. We consider a specialized population distribution for sequence process rates, which incorporates pattern prevalence into our model. This allows patterns that exist widely across the population, but not in all individuals, to

be found by our model. The model performed well on simulation data, and helped to reveal interesting information about different populations of rodents. Our work with the human mothers is ongoing.

However, the hierarchical model does suffer from issues similar to those for the individual level model. In particular, efficient exploration of the posterior distribution is critical. As the number of mothers included in the model increases, the MCMC scales poorly. Our current inference procedure works well, but room for improvement certainly exists.

Chapter 4

Renewal Processes with General Probability Distributions

A renewal process is completely characterized by the distribution of the interarrival times. Different distributions will result in data with different properties, which is important to consider when attempting to model behavior. The probability distribution governing interarrival times can be specified via the density, distribution, survival or hazard function. We focus primarily on the hazard function because the likelihood of the renewal process model is easiest to describe using the hazard functions of the underlying renewal processes, as seen in Equation 2.10.

The Poisson process models that have been used in the analyses for Chapters 2 and 3 assume exponentially distributed interarrival times. The exponential distribution has a constant hazard rate, which means the likelihood of an event occurring, given that is has not yet occurred, remains constant at any point in time. This is known as the memoryless property, and an example of data generated from a Poisson process can be seen in the bottom row of Figure 4.1.

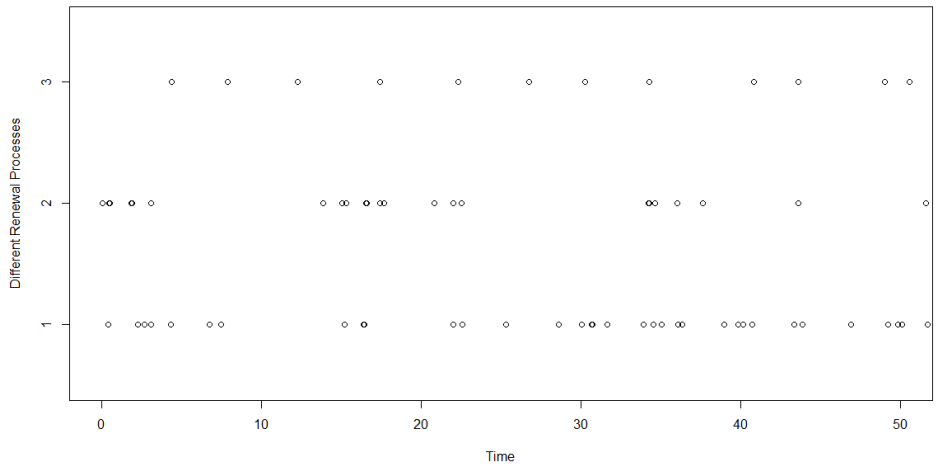

Figure 4.1: The plot shows simulated data from three distributions:

1. Data occurs haphazardly, with little structure (constant hazard).
2. Data is subject to bursts, with long periods between bursts.
3. Data occurs regularly, with little variance.

However, this may be a poor assumption in some cases. For example, some behaviors may occur in bursts, where the same action is performed repeatedly in a short time window, and then not performed again for a longer period of time. A renewal process with an interarrival distribution that has this patten is illustrated in the middle row of Figure 4.1, to compare against the Poisson process in the bottom row.

Another possible example of non-exponential data are behaviors that are performed cyclically. Some behaviors, such as eating, often occur at similar times throughout the day, so the interarrival times are relatively consistent. In such instances, once the behavior is performed, it will be performed again predictably after some lag period. Again, this is poorly modeled with the exponential distribution. An example of data illustrating this type of behavior is included as the top row of Figure 4.1.

Other distributions for interarrival times are also possible. We would like our model to

adapt to such variation to correctly model the behaviors of the background processes, and to ensure patterns are successfully discovered. In the following section, we discuss a number of parametric interarrival distributions, and the shapes of the corresponding hazards. In Section 4.2, we discuss a nonparametric alternative that allows for a more flexible class of hazard shapes. We explore the use of different interarrival distributions for simulated data in Section 4.3, and for our maternal data in Section 4.4. Finally, we conclude with a discussion of the use of different interarrival distributions. Note that in this chapter, we only consider alternative distributions for the time increments of background processes, though we hope to consider such alternatives for sequence processes in the future.

4.1 Parametric Renewal Processes

We first generalize from the Poisson process model of Chapters 2 and 3 by considering a number of families of parametric hazard functions. While these models are not as general as the nonparametric models that are considered in the following section, the parametric models are simple to implement, well understood, and can conform to a variety of shapes of hazard functions. There are several possible parametric distributions to consider in this setting, including the Weibull, gamma, and lognormal distributions (Klein & Moeschberger, 2003). In this section, we discuss the advantages and disadvantages of these distributions.

The Weibull distribution is widely used for time to event models due to its flexibility and simple parametric form. The Weibull distribution has two parameters, a shape and a scale parameter, and includes the exponential distribution as a special case when the shape parameter is equal to one.

The hazard function of the Weibull distribution can be either decreasing, flat, or increasing, see Figure 4.2. When the shape parameter is less than one, the hazard is monotonically

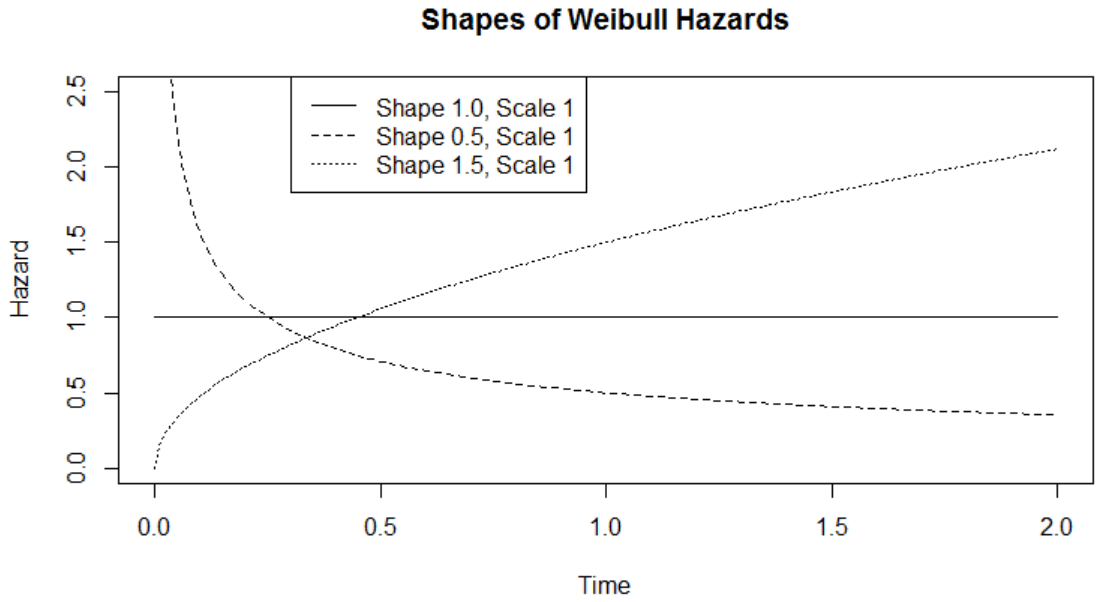

Figure 4.2: The different shapes of the Weibull hazard function: Constant, Decreasing, and Increasing.

decreasing. This indicates a higher likelihood for events soon after an event has occurred, similar to the bursts example described above. When the shape parameter is greater than one, the hazard is monotonically increasing. This indicates that the likelihood of a successor event increases as more times passes without the event occurring. We explore the use of the Weibull distribution as an alternative for the exponential distribution in the experiments section of this chapter.

The gamma distribution can also support a monotonically increasing, decreasing, or constant hazard function, as seen in Figure 4.3. However, both the survival and hazard function of the gamma distribution are analytically intractable. This means they need to be solved numerically, decreasing their practicality for our approach.

Parametric distributions that have hazard functions with non-monotonic shapes, unlike the Weibull and gamma distributions, are also interesting for our purposes. For instance, the hazard function for the lognormal distribution increases from 0 to a maximum, before de-

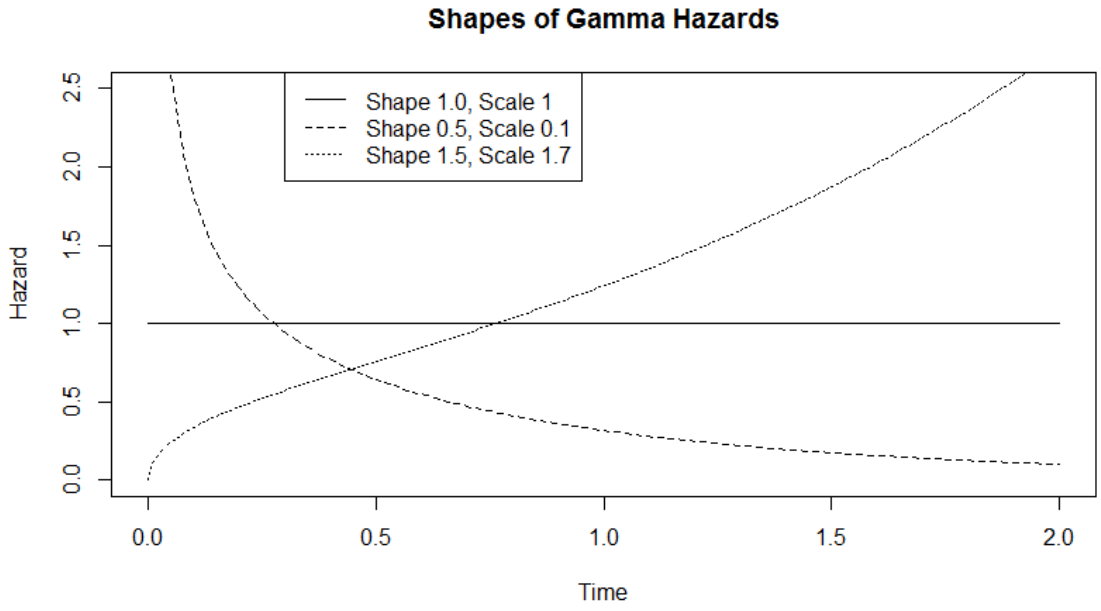

Figure 4.3: The different shapes of the gamma hazard function: Constant, Decreasing, and Increasing.

creasing. Some sample hazard shapes for the lognormal distribution can be seen in Figure 4.4. Such hazard functions may be of great interest in behavior, such as in the cyclic behavior example. We refer to these hazard shapes as upside down bathtub shapes. However, using the lognormal for behavior modeling is restrictive because its hazard function does not generalize well beyond the upside down bathtub hazard shapes. We will not explore modeling with either the gamma or the lognormal distribution in the following sections.

If we want maximum flexibility in a parametric family, then there are two options. One is to use even more general parametric distributions. These generally have intractable hazard functions or are difficult to model. A second possibility is to consider mixture models, such as a finite mixture of Weibull distributions, which allow a general form for the hazard function. Furthermore, we could consider a Dirichlet process mixture process for a fully nonparametric approach (Kottas, 2006). While this approach is viable when considering the survival distribution, modeling the hazard of a mixture model analytically is more difficult.

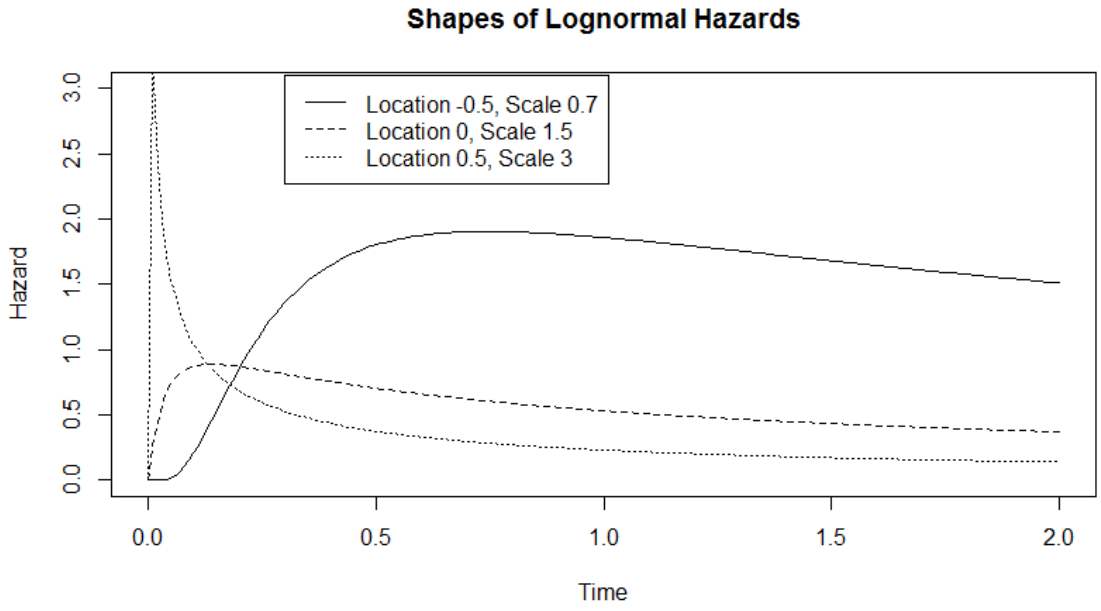

Figure 4.4: The different shapes of the lognormal hazard function: Constant, Decreasing, and Increasing.

4.2 Non Parametric Renewal Processes

Determining the most appropriate probability distribution a priori for modeling renewal processes is difficult. To avoid the need to choose between several potential distributions, such as the examples described in the previous section, we now consider one possible nonparametric approach for constructing hazard functions.

To achieve this goal, we consider modeling the hazard function as a step function. Step functions were chosen due to their simplicity and the fact that they can easily be implemented in our modeling framework. Step functions can model commonly considered shapes such as decreasing, flat, and increasing hazards, as well as bathtub and upside down bathtub hazard functions.

We next develop the notation for a general step-based hazard function. The interarrival times t_i occur according to a distribution on the positive real numbers. We partition the

time axis into K intervals at jump times $0 < \xi_1 < \xi_2 < ... < \xi_K$, and for the interval $[\xi_{k-1}, \xi_k)$, we set the hazard rate to be a constant λ_k. The jump times can either be fixed a priori or treated as unknowns. The hazard function is thus given by

$$\lambda(t) = \sum_{k=1}^{K} \lambda_k I_{[\xi_{k-1}, \xi_k)}(t) \tag{4.1}$$

To define the cumulative hazard function at t requires that we identify the step in which t lies, which we denote as step ϕ. Suppose that $\xi_{\phi-1} < t < \xi_\phi$ so that t in in step ϕ. We can define the cumulative hazard function as

$$\begin{aligned} \Lambda(t) &= \int_0^t \lambda(u)\, du \\ &= \lambda_\phi (t - \xi_{\phi-1}) + \sum_{k:\, \xi_k < t} \lambda_k (\xi_k - \xi_{k-1}) \end{aligned} \tag{4.2}$$

Furthermore, from this we can define the survival function as $S(t) = \exp(-\Lambda(t))$ and the density function as $\lambda(t) \exp(-\Lambda(t))$.

This model is especially convenient when calculating the density of interarrival times and events for our model. Recall the density that formulates the Renewal Process Model likeli-

hood, as presented in Chapter 2, is:

$$
\begin{aligned}
& f_{t_i,\,E_i}\left(t,\,e\right) \\
& = \left(\lambda^e\left(t\right) + \lambda^e\left(t\right)\right) \prod_{j=1}^{J} \bar{S}^j\left(t\right) \prod_{k=1}^{K} \dot{S}^{s_k}\left(t\right) 1_{\left(t<\tau_{s_k}\right)} \\
& + \lambda^e\left(t\right) \prod_{j=1}^{J} \bar{S}^j\left(t\right) \prod_{k=1}^{K} \dot{S}^{s_k}\left(\tau\right) 1_{\left(t>\tau_{s_k}\right)}.
\end{aligned}
\tag{4.3}
$$

Now we can simply replace all the survival functions in (4.3) with $S\left(t\right) = \exp\left(-\Lambda\left(t\right)\right)$, where $\Lambda\left(t\right)$ is defined as in (4.1). Furthermore, we note that $\lambda\left(t\right)$ is equal to exactly one of the values of λ_k in (4.1): the one corresponding to step ϕ, λ_ϕ.

Such a model has other advantages as well. With a large number of steps, any smooth function can be approximated to a desired precision. Using a large number of steps is computationally expensive. However, even with a small number of steps, a wide variety of hazard shapes can be estimated. Both increasing and decreasing hazard functions can be modeled with only two steps. The bathtub and upside down bathtub hazard functions can easily be modeled with three steps.

The shape of the step function is dictated both by the number of steps, as well as the height of the steps λ_k. We next consider a probability model for the step heights. One approach is to model the step heights as independent random variables a priori, such as in Walker & Mallick (1997),where the prior distribution for each λ_k is an independent gamma random variable. However, the independence assumption implies the height of a step is unrelated to the height of neighboring steps. This is clearly a poor assumption if we assume that the hazard functions are relatively smooth. While no step function will be able to model a smooth function perfectly, we can consider an alternative prior distribution for λ_k that

allows for a notion of smoothness between the steps.

A Markov gamma process for the prior distribution on the hazard rates λ_k proves useful for smoothing the step function (Nieto-Barajas & Walker, 2002). The following description of the prior distribution follows almost entirely from Nieto-Barajas & Walker (2002). We no longer assume the hazard levels λ_k are independent, but instead assume the levels are related according to a Markov process. First, we set the prior distribution on the initial hazard rate as $\lambda_1 \sim gamma\,(a_1,\, b_1)$. Then, the approach we use introduces a latent process $\{\mu_k : k = 2, \ldots, K\}$ that depends on hyperparameters $\{a_k, b_k, c_k : k = 2, \ldots, K\}$. Hyperparameters a_k and b_k can be interpreted as providing prior information for the location of λ_k, with $E\,(\lambda_k) = \frac{a_k}{b_k}$. The c_k hyperparameters dictate the smoothness of the hazard functions, which we describe below. The process $\{\mu_k : k = 2, \ldots, K\}$ is defined as:

$$\begin{aligned} \mu_{k-1}|\lambda_{k-1} \quad &\sim Poisson\,(c_k\lambda_k) \\ \lambda_k|\mu_{k-1} \quad &\sim gamma\,(a_k + \mu_{k-1},\, b_k + c_k) \\ k = 2, \ldots, K & \end{aligned} \tag{4.4}$$

When $c_k = 0$, the prior process reduces to independent gamma distributions. For large c_k, $E\,(\lambda_k|\lambda_{k-1}) \approx \lambda_{k-1}$, and $Var\,(\lambda_k|\lambda_{k-1}) \approx \frac{2\lambda_k}{c_k}$which means that the λ_ks tend to a constant, forcing an exponential model.

Thus the larger the value of c_k, the more the distribution of λ_k depends upon λ_{k-1}. We refer to c_k as a smoothness parameter which dictates how close λ_k and λ_{k-1} will be to each other. The parameter c_k can also be thought of a prior sample size, though in this case it is a sample size that dictates the strength of the influence of λ_{k-1} on the distribution of λ_k.

Figure 4.5 displays several examples of the step function for different values of c_k. In this

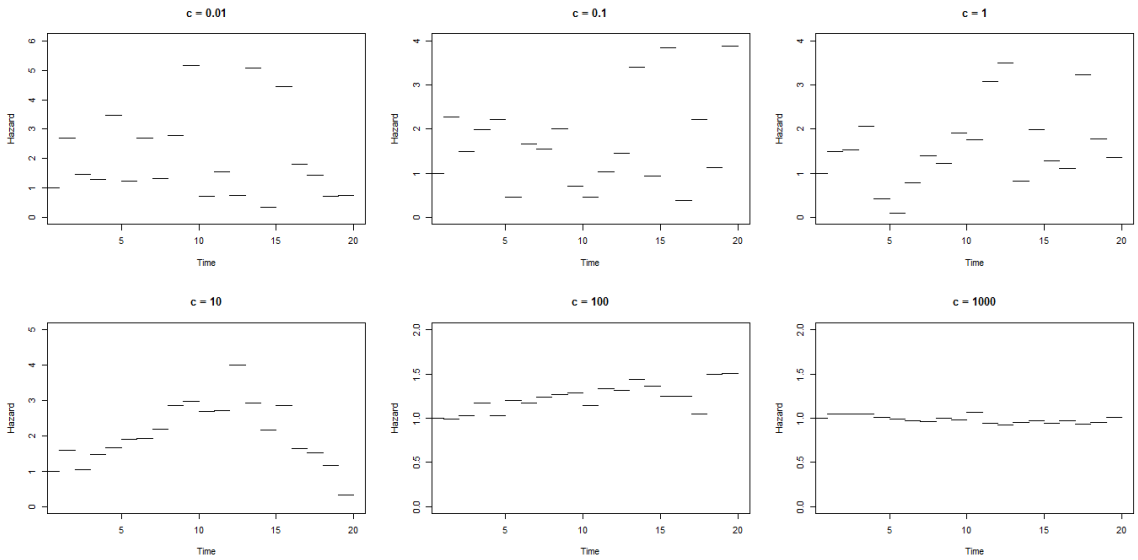

Figure 4.5: Simulated step functions plotted for various values of the smoothness parameter c.

example, the values of c_k are equivalent for all k, so we will denote the parameter as c. In each plot, we simulated the step function using the Markov gamma process. On the top left, $c = 0.1$, and the steps appear roughly independent. For $c = 1000$, on the bottom right, the steps are all roughly equivalent, implying a nearly constant hazard function. For $c = 10$, shown on the bottom left, there appears to be an interesting shape to the hazard function, with some notion of smoothness. In our experiments that follow, we will fix $c = 10$, though allowing c to be random may be an interesting future direction with model selection implications.

An advantage to this Markov process prior is that it is possible to describe the posterior distribution for a single renewal process. First, we can describe the prior distribution of λ_k dependent on both the preceding and succeeding values in the latent process, which is also a gamma distribution:

$$\lambda_k|\mu_k, \mu_{k-1} \sim gamma\left(a_k + \mu_{k-1} + \mu_k, b_k + c_{k-1} + c_k\right). \tag{4.5}$$

Now we can describe the posterior distribution for a single renewal process, such as the background process for a single event type. First, we need to introduce some statistics that depend on the data. Let n_k denote the number of observations in the interval $[\xi_{k-1}, \xi_k]$. Additionally, define $m_k = \sum_i r_{ki}$ where

$$r_{ki} = \begin{cases} \xi_k - \xi_{k-1} & t_i > \xi_k \\ t_i - \xi_{k-1} & \xi_{k-1} < t_i < \xi_k \\ 0 & else \end{cases} \tag{4.6}$$

Given these statistics, we can now derive the posterior distribution as yet another gamma distribution:

$$\lambda_k | \mu_k, \mu_{k-1}, data \sim gamma\left(a_k + \mu_{k-1} + \mu_k + n_k, b_k + c_{k-1} + c_k + m_k\right). \tag{4.7}$$

If no data exists in the interval $[\xi_{k-1}, \xi_k]$, then the distribution of λ_k is determined by the prior distribution and the rates of the neighboring partitions. Thus even with an arbitrarily large number of partitions, all the steps will still exhibit some smoothness even with no data, which is a significant improvement over the independent steps model.

Once applied to our full model with sequence processes included, the posterior is not represented as easily, though it retains a similar form.

We also note that by letting the time intervals $[\xi_{k-1}, \xi_k]$ collapse to 0, the discrete Markov gamma process described above converges to a continuous variant, specifically a Levy process kernel mixture model (Nieto-Barajas & Walker, 2004). While such a model is more

generalizable than our discrete approximation, it is much more difficult to sample from that model. In particular, it is necessary to simulate the Levy process every iteration. We do not consider this alternative, though it is possible to consider it as a future direction.

4.3 Experiments

We consider the results of a variety of simulations to determine the effect of using different probability distributions for the time increments. The primary aims of this section include to demonstrate the efficacy of our nonparametric model, to demonstrate how well the various parametric and nonparametric distributions perform, and to examine the robustness of the RPM model results against a change in the assumed interarrival distribution.

For our experiments, we consider data that contains only background processes. We consider four different hazard function types: exponential, Weibull with increasing hazard, Weibull with decreasing hazard, and lognormal (upside down bathtub hazard). Throughout this section we fit our model assuming that the interarrival times for the sequence processes are exponential. For all experiments, the prior distributions for the sequence process parameters follow directly from those used in Chapter 2.

4.3.1 Effectiveness of the Nonparametric Model for Point Process Data

We first demonstrate that our nonparametric model is capable of approximating various hazard functions. For each hazard type, we simulate a renewal process whose interarrival times are distributed according to the respective distribution. In this initial study, we do not consider multiple event types, but just a single renewal process. We fit each parametric renewal

process with a nonparametric renewal process, to demonstrate how well the nonparametric hazard estimates the parametric one.

We simulate a total of $n = 500$ data points per renewal process. The parametric distributions that we consider are an exponential (scale/rate 1), a decreasing hazard Weibull (shape 0.5 and scale 1), an increasing hazard Weibull (shape 2 and scale 1), and a lognormal (log mean 3.9 and log standard deviation 0.55). We approximate each hazard shape with a step function hazard that has a fixed number of steps. In this case, we fix the number of steps at eight.

The results can be seen in Figure 4.6. The posterior distribution for each hazard function is plotted along with the true value. As seen in the figure, this method does a good job of approximating the true hazard function, as the steps generally contain the true hazard function. For the exponential case, the samples for each step have overlapping ranges that seem to approximate the constant hazard well. For the decreasing hazard, the sample for the leftmost step have a large range but a short step length, thus well approximating a steep shape, and the following steps generally decrease. For the increasing hazard, the step heights are generally increasing, with increasing variance in the range reflecting the decrease in the amount of observed data. Finally, the lognormal demonstrates a steep increase, followed by a roughly flat region, before decreasing again.

4.3.2 Fitting Different Distributions

In this section, we explore fitting the Renewal Process Model to data generated from a range of distributions for the background process time increments. In this section, we return to the scenario of several background processes, where our goal is pattern discovery. We will discuss fitting the exponential, Weibull, and nonparametric model in cases where they are both properly specified, and misspecified.

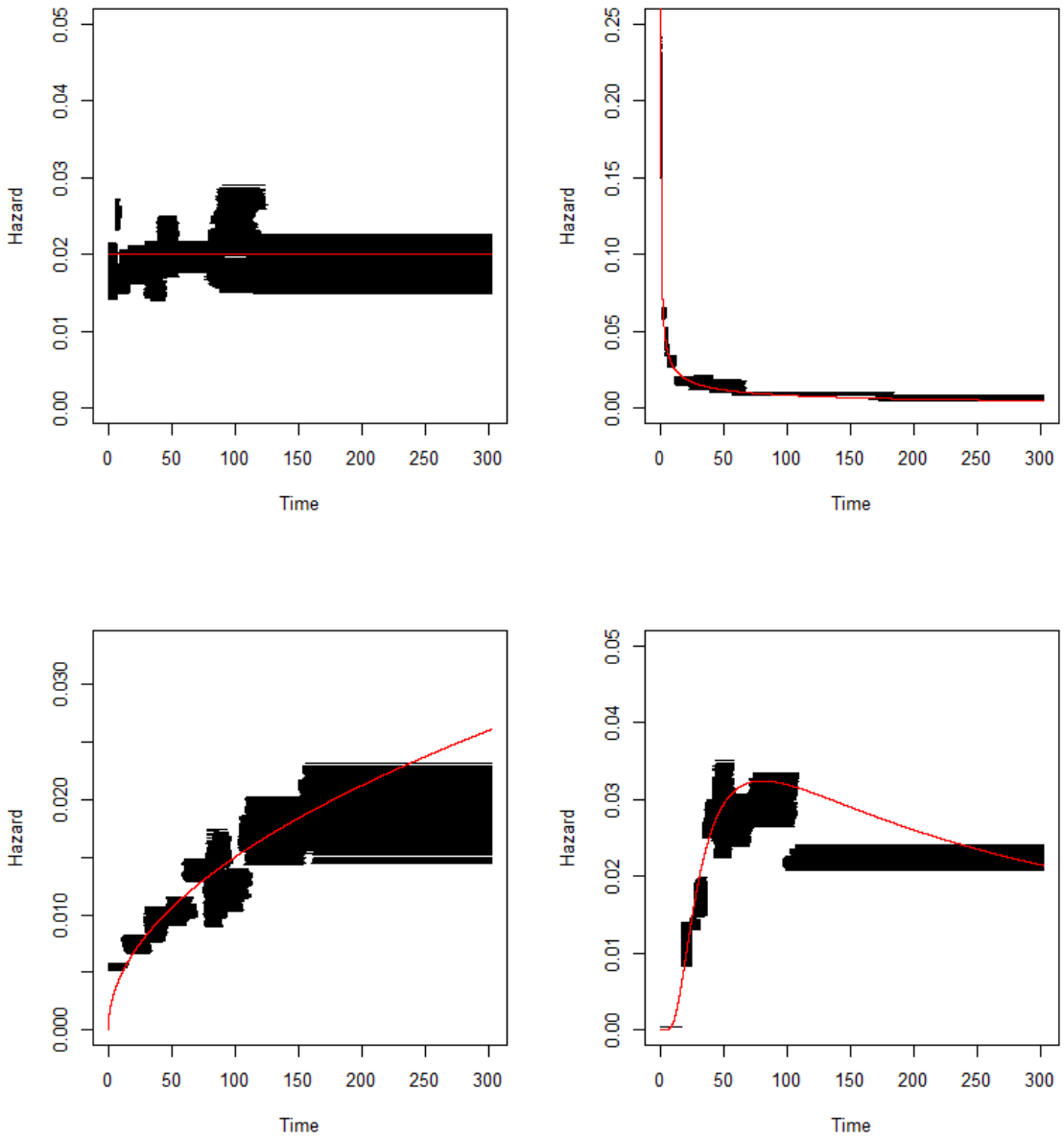

Figure 4.6: The results of fitting our nonparametric model for varying hazard functions. The red line indicates the true hazard. The black lines are sampled draws from the posterior distribution of the hazard function.

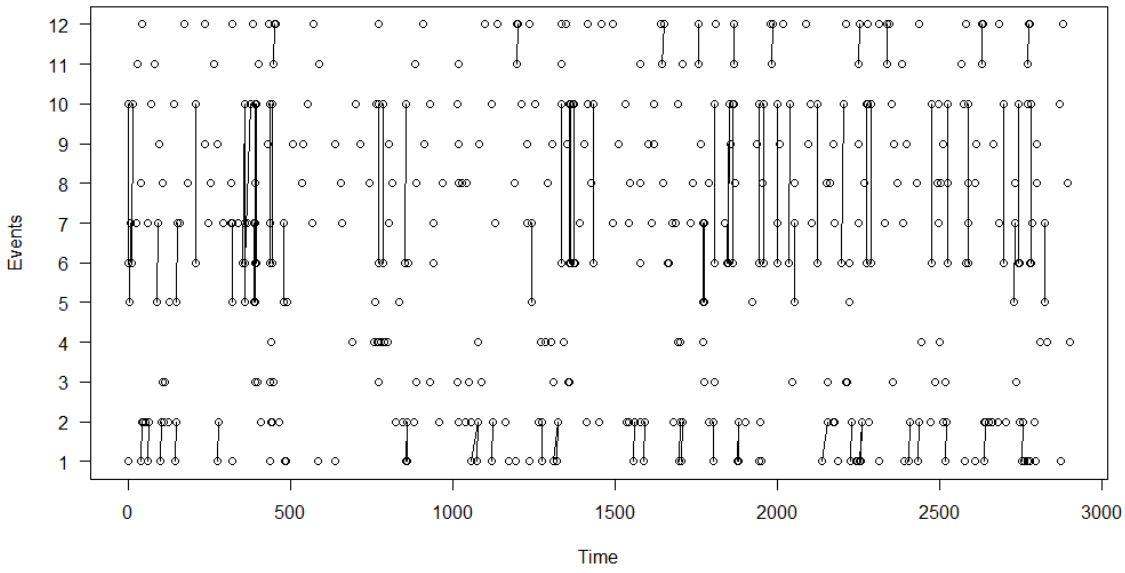

Figure 4.7: Simulated data that incorporates a variety of background processes with differing hazard functions, along with four different sequence processes.

For each case, we consider the same type of simulated data. The simulated data contains 12 different background processes, three for each of the hazard types described above. This includes three exponential distributions (rates 0.1, 0.2, 0.3), three Weibull distributions for both decreasing and increasing hazard functions (shape parameters 0.5 and 2, rates of 0.1, 0.2, 0.3 each), and three lognormal distributions (parameters chosen such that the mean and variance were equivalent to the mean and variance of an exponential with rates 0.1, 0.2, and 0.3). We also include four different sequence processes, each of length two and exponentially distributed (all four have rates of 0.2, and $\tau = 4$). An example of simulated data of this type can be seen in Figure 4.7.

Exponential Distribution

In Chapter 2, we discussed fitting the Renewal Process Model with exponentially distributed time increments for experiments with Poisson processes. Here we will discuss in greater detail

Pattern	Exponential	Weibull	Nonparametric
$1 \rightarrow 2$	0.98	0.94	0.94
$11 \rightarrow 12$	0.94	0.96	0.95
$5 \rightarrow 7$	0.95	0.98	0.90
$6 \rightarrow 10$	0.97	0.93	0.91

Table 4.1: For each of the exponential, Weibull, and nonparametric models, the percentage of 100 simulations where a given pattern was successfully detected.

fitting the exponential model when the true underlying distribution is not exponential. All prior distributions for this section are the same as those in Chapter 2.

For each of 100 simulations of the above data, we fit the exponential model. Despite the misspecification of the background processes (i.e. nine of the background processes are nonexponential), the model is quite successful at discovering the patterns, as seen in the left side of Table 4.1. From this we can assert that discovering patterns in time-event data is fairly robust to the misspecification of background processes.

However, because of the misspecification, the exponential model is prone to discovering additional patterns that correspond to bursts in behavior from the Weibull data. Events D, E, and F in Figure 4.7 all have time increments distributed according to a Weibull distributions with a decreasing hazard function. For these events, the exponential model identified burst patterns D $\rightarrow$ D, E $\rightarrow$ E, F $\rightarrow$ F at proportions of 0.33, 0.49, and 0.54 of the time, respectively. This does provide a powerful diagnostic for proper model specification.

Weibull Distribution

Again, for each of the 100 simulations of the above data, we fit the Weibull RPM to the data. For the background process parameters, we place an exponential prior distribution with rate one on both shape and scale parameters. Similar to the exponential model, the Weibull model successfully detected all the patterns most of the time, see the middle column

	Exponential		Decreasing Hazard		Increasing Hazard	
	Shape	Scale	Shape	Scale	Shape	Scale
True Value	1	0.2	0.5	0.2	2	0.2
Post. Mean	1.03	0.19	0.52	0.21	1.94	0.20
S. E.	0.13	0.02	0.11	0.03	0.17	0.02
Coverage	0.98	0.95	0.96	0.97	0.94	0.95

Table 4.2: The estimated shape and scale parameter determined using a Weibull RPM for each of three background processes. Posterior mean refers to the average of the posterior means over the 100 simulations. Standard error was calculated directly from the posterior means, and coverage was calculated using the 95% credible intervals.

of Table 4.1.

Of the four different distributions that we consider in this data set, three of them can be correctly modeled with a Weibull distribution. Thus we wanted to know how well the Weibull model performed in these cases, as proof that our model was working correctly. For each of the 100 simulations, we recorded the posterior mean for each of the scale and shape parameters, as well as a 95% credible interval. For each of these, we calculated the average over the means as our parameter estimate, and used them to calculate a standard error. Additionally, we computed coverage probabilities using the credible intervals. Results for three background processes (one for each a constant, decreasing, and increasing hazard function, and each with equivalent scale parameters) can be found in Table 4.2. The coverage probabilities are all relatively close to 0.95, and the estimated parameters are reasonably close to the true parameters.

The fourth distribution that we consider is the lognormal distribution. While the Weibull model is clearly inappropriate for correctly modeling this distribution, it has no apparent effect on pattern discovery. The estimated rate parameters (posterior means of 0.11, 0.21, and 0.30) are roughly 0.1, 0.2, and 0.3, as we chose the lognormal parameters to roughly align to these cases. The shape parameters (posterior means of 1.23, 1.87, 1.46 respectively) are all greater than one, so the Weibull model clearly attempts to conform to the initially increasing region of the lognormal hazard.

Nonparametric Distribution

We now want to examine the performance of the nonparametric model in this case. For this experiment, we consider a step function with eight jumps to model the hazard function. For the gamma Markov process prior distribution, we set the hyperparameters as $a = 0.2$, $b = 1$, and $c = 10$. The prior distribution for the jump locations ξ_k was set as a uniform between the previous and following jump locations. Thus the prior distribution is $\xi_k \sim Uniform\,(\xi_{k-1}, \xi_{k+1})$ where $\xi_0 = 0$ and ξ_{K+1} is a reasonable large value (for example, the total duration of the observation period).

Much like the other cases, the nonparametric model is successful at pattern detection, as seen on the right side of Table 4.1. The nonparametric model with eight jumps was successfully able to capture the increasing and decreasing nature of the processes, as well as the upside down bathtub pattern of the lognormal processes.

4.3.3 Robustness Against Misspecified Sequence Processes

Thus far we have only considered models that contain sequence processes with exponentially distributed interarrival times. The shape of the hazard for sequence processes is necessarily decreasing around time τ due to the truncation feature of the RPM, and it is not clear whether the shape of the hazard prior to τ will negatively affect our model's ability to detect patterns. As we have seen above, incorrectly specifying the distribution for background processes has relatively little effect when searching for patterns. We now want to investigate if the same is true for the sequence processes.

To explore this, we simulated data with a background structure identical to the previous section (12 processes, three for each of the four different types of hazard functions). However, the sequence processes had varying Weibull distributions. The scale parameter was held

Pattern	Shape	Exponential	Weibull	Nonparametric
$1 \rightarrow 2$	1.7	0.56	0.52	0.47
$11 \rightarrow 12$	1.3	0.41	0.55	0.61
$5 \rightarrow 7$	0.7	0.97	0.94	0.99
$6 \rightarrow 10$	0.3	0.96	0.98	0.93

Table 4.3: The proportion of simulations where each model successfully identified each pattern.

constant at 0.2 for each of the four sequence processes, but the shape parameter was allowed to vary (0.3, 0.7, 1.3, 1.7) to allow for varying hazard structures. Data was simulated 100 times, as before. Data was then fit with models where the background interarrival time distributions were each of the three variants that we have discussed (exponential, Weibull, and nonparametric). Results of this analysis can be seen in Table 4.3. In particular, we again calculate the proportion of simulations where the pattern was successfully identified.

The three models for the background process interarrival times performed similarly, indicating that the distribution of the background processes has little effect on the identification of non-exponential sequence processes. As shown in Chapter 2, patterns were successfully identified when their sequence processes have exponentially distributed time increments. Here we also see that sequence processes with decreasing hazards are successfully identified most of the time.

However, when the hazard is increasing (the first two rows of Table 4.3), the results show that the sequence patterns were only discovered half of the time. In such cases the data is clustered relatively close to the maximum time τ, which makes it difficult to discover patterns.

Overall, it appears that the exponential model for sequence processes is relatively robust in the case where the next event in the sequence often occurs well before the maximum allowed time (τ) has passed. However, when the next event in the sequence regularly occurs around the maximum time parameter, pattern detection starts to break down. Clearly, this issue

could negatively impact the results of an analysis, and should be explored in the future. Due to complications arising from the τ parameter, we do not consider any distributions other than the exponential for the sequence process interarrival times. In particular, for small τ, it is difficult to correctly distinguish between possible distributions.

4.4 Maternal Behavior Analysis Revisited

We now explore the effect of applying the RPM model based on alternative hazard functions to our maternal data. We will consider results from a Weibull model and a nonparametric model, and compare these results to the results obtained from the exponential model in Chapter 2. In the sections for both the Weibull and nonparametric model, we will again describe the distributions of the summary statistics derived from the patterns, and compare them against the summaries obtained from the exponential model. We also explore the effect of these models on the patterns themselves. Note that prior distributions for this section will be the same as those used in the simulation experiments.

4.4.1 Weibull Model Results

First, we consider the results of the Weibull model. For each mother, we modeled all of the background processes with a Weibull renewal process, each with its own shape and scale parameter. We calculate the same summary statistics as we did in Chapter 2: the longest pattern, the percentage of patterns longer than two events, and the number of sequences identified. Ultimately, we find considerably fewer patterns, but much of this decrease can be explained by the fact that we no longer identify patterns that demonstrate bursts (patterns of the type A $\rightarrow$ A).

The distribution for the number of patterns per mother under the Weibull model can be

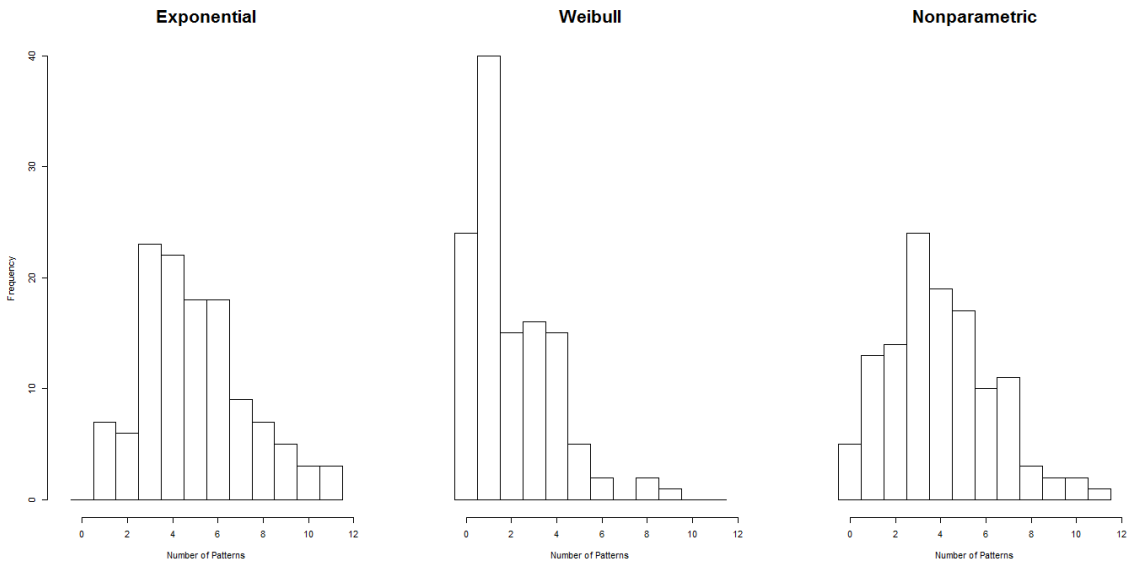

Figure 4.8: The distribution of the number of patterns per mother for each of the three models we consider in this section: the exponential, Weibull, and nonparametric RPM.

found in the middle of Figure 4.8. The most striking feature of this distribution is the significant decrease in the number of patterns. Under the exponential model (on the left side of Figure 4.8), mothers exhibited an average of 4.97 patterns. Under the Weibull model, the average number of patterns has decreased to only 2.14 per mother. This is at least partially attributed to the removed of bursts from the pattern set, which we discuss further below.

The distribution for the maximum pattern length for each mother under the Weibull model can be found in the middle of Figure 4.9. Again, we can compare this distribution against the distribution of maximum pattern length under the exponential model, found on the left side of Figure 4.9. Several mothers (24 out of 121 total) simply did not have a pattern detected. Of the mothers who did have a pattern detected, the maximum pattern length was generally comparable to the results from the exponential model.

Finally, the distribution of the percentage of patterns greater than length two can be seen in Table 4.4. Again, several mothers simply did not have any patterns detected. Beyond that, the distribution differs significantly from the exponential model. The longer patterns are relatively more common than in the exponential model. This is likely due to the fact that

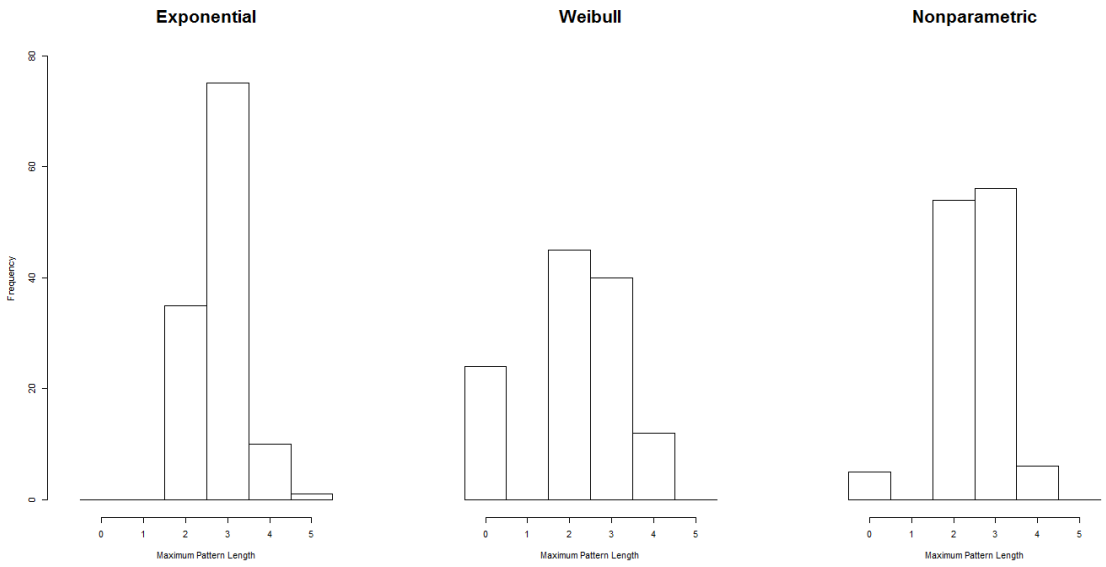

Figure 4.9: The distribution of the maximum pattern length per mother for each of the three models we consider in this section: the exponential, Weibull, and nonparametric RPM.

Percentage	0%	1-20%	21-40%	41-60%
Number of Mothers	45	2	27	23

Table 4.4: Distribution of mothers who had a given percentage of patterns that were longer than two events under the Weibull model. The highest percentage was 60%.

the event bursts that were detected under the exponential model as patterns were generally two event patterns. Once they are removed, the percentage of longer patterns is increased.

The most common patterns were very similar to those observed under the exponential model, as seen in Table 4.5. The three most common patterns in the exponential model were the most common patterns in the Weibull model, although they occurred far less often in the Weibull case. The more complex background process modeled much of what had been considered patterns in the simpler model, making patterns far less common. In particular, note that no patterns of repeating events, such as instructive speech $\rightarrow$ instructive speech, are present in the list of the most common patterns.

As we noted in Chapter 2, one of the most predominant pattern types found was a pattern comprised of repeated events of the same type. In our maternal data, this included repeated

Pattern	Number of Mothers
New Toy → Manipulating Toy	71
Smiling → Content	14
Look at Toy → Look at Baby	10
Set Down Toy → New Toy	4
Smiling → Laughing	2

Table 4.5: The most common patterns discovered in the mothers using the Weibull model. On the right is the number of mothers who exhibited a given pattern, out of 121, which are noticeably smaller than in the exponential model.

instances of speech (instructive, positive, and negative), as well as pointing and affectionate touching. These are potentially relevant patterns, but they are conceptually different from the patterns we are attempting to find. They describe the nature of a particular event (or in our model, a single background process) rather than a particular behavioral pattern intrinsic to the actor. As such, the Weibull model may actually be a superior model relative to the exponential one, though clearly this information can be obtained from the exponential model as well by removing such patterns manually.

Considering various different forms for the hazard functions provides an alternative approach to the issue of burst-type patterns. While the burst patterns are no longer detected, we can still identify similar structures in the data by examining the fitted Weibull shape parameters. As mentioned earlier in this chapter, the Weibull distribution is characterized by a monotonically increasing, monotonically decreasing, or constant hazard function depending on the Weibull shape parameter. By taking advantage of this fact, we can sort each background process into one of the three shapes.

For each background process, we calculated a 95% credible interval for the shape parameter, which controls the shape of the hazard function, using the MCMC samples. If the interval contained one, then we deemed there was insufficient evidence to conclude that the hazard was non-constant. If the interval was entirely greater than one, we concluded the corresponding process had an increasing hazard, and if it was entirely less than zero, we concluded that

the process had a decreasing hazard.

A decreasing hazard would be characterized by bursts of a behavior, where the same event would occur several times in succession. These bursts would be separated by an occasional longer gap between. Several event types were characterized by a decreasing hazard. The most prominent, with the number of mothers in parentheses, were as follows:

- Positive Speech (60)
- Affectionate Touch (49)
- Pointing (38)
- Functional Touch (26)
- Negative Speech (25)
- New Toy (20)

These occurrences are far more common than their respective patterns in the exponential model. For example, Positive Speech $\rightarrow$ Positive Speech appeared in 21 mothers according to the exponential model, and Affectionate Touch $\rightarrow$ Affectionate Touch appeared in 17 mothers.

Conversely, only two event types regularly had increasing hazards. Behaviors with an increasing hazard are characterized by more regular occurrences with few if any bursts. The two event types were Set Toy Down (40 mothers) and Instructive Speech (46 mothers).

These results are fairly intuitive. For example, positive and negative speech are likely linked to the child's behavior directly, and thus occur in bursts. Instructive speech will occur at the mother's discretion, and will likely be fairly regular without many bursts. Similarly, touching (affectionate and functional) and pointing occur in bursts, likely to increase the impact of the child's reaction to the behavior.

Furthermore, if each instance of a decreasing or increasing hazard was considered a pattern, the average number of patterns per mother would increase to 5.61. This is much closer to the 4.97 patterns on average per mother for the exponential model, further confirming that this is a primary reason for the significant difference in number of patterns.

4.4.2 Nonparametric Model Results

Next, we consider the results of applying the nonparametric model. For the nonparametric model, we fit a step function with a total of eight jumps to model the shape of the hazard function. For each background process, this means we estimate eight jump times ξ_k, $k = 1, ..., 8$, and the heights of the nine steps λ_k, $k = 0, ..., 8$, in addition to the hyperparameters μ_k, $k = 1, ..., 8$. We again calculate the summary statistics from applying the model to the 121 mother-child videos and compare them to the exponential.

The distribution for number of patterns can be seen in the right side of Figure 4.8, where it is compared to the distribution under the Weibull and exponential models. Perhaps the most noteworthy feature of this distribution is a tendency to discover more patterns than the Weibull model. Overall, under this model, mothers averaged 4.00 patterns each, which is considerably more than the Weibull model. While less than the exponential model, it is worth noting that A $\rightarrow$ A type patterns were largely absent from the nonparametric model, but made up a significant portion of the exponential model patterns.

The distribution for the maximum pattern length per mother can be seen in 4.9, and compared against the other two models. According to the nonparametric model, fewer mothers have no patterns than in the Weibull model, though the distribution of maximum pattern length remains relatively unchanged. None of the models could detect patterns greater than five events, and both four and five event patterns were relatively rare throughout.

Percentage Patterns > 2 Events	0%	1-20%	21-40%	41-60%
Number of Mothers	49	31	25	6

Table 4.6: Distribution of mothers who had a given percentage of patterns that were longer than two events under the nonparametric model. The highest percentage was 60%.

However, the increase in patterns is comprised almost entirely of simple patterns (those containing two events), as revealed by the distribution of the percentage of patterns greater than two events, seen in Table 4.6. These results directly mirror the results from the exponential model. This hints at the fact that while the Weibull model removed the burst patterns, it also suppressed other patterns found in the exponential model.

The patterns themselves are fairly similar to the previous models. Some of the most common patterns from the nonparametric model can be found in Table 4.7. The top several patterns are identical to those from the exponential and Weibull models, though the number of occurrences for each pattern seem to be closer to the Weibull model.

Pattern	Number of Mothers
New Toy → Manipulating Toy	93
Look at Toy → Look at Baby	33
Smiling → Content	23
Smiling → Laughing	24
Pointing → New Toy	10
Manipulating Toy → Instructive Speech	9
New Toy → Instructive Speech	8

Table 4.7: The most common patterns discovered from the mothers using the nonparametric model. On the right is the number of mothers who exhibited each pattern, out of a total of 121 mothers.

Unfortunately, one drawback of the nonparametric model is the inability to easily discern simple hazard shapes. The Weibull model is not nearly as flexible, but the ability of the Weibull to extract information such as increasing or decreasing hazard, by examining the shape parameter, is a significant advantage.

While it is difficult to extract such information, we can still explore the posterior distribution of the hazard function to obtain useful information. Let us first consider a specific mother,

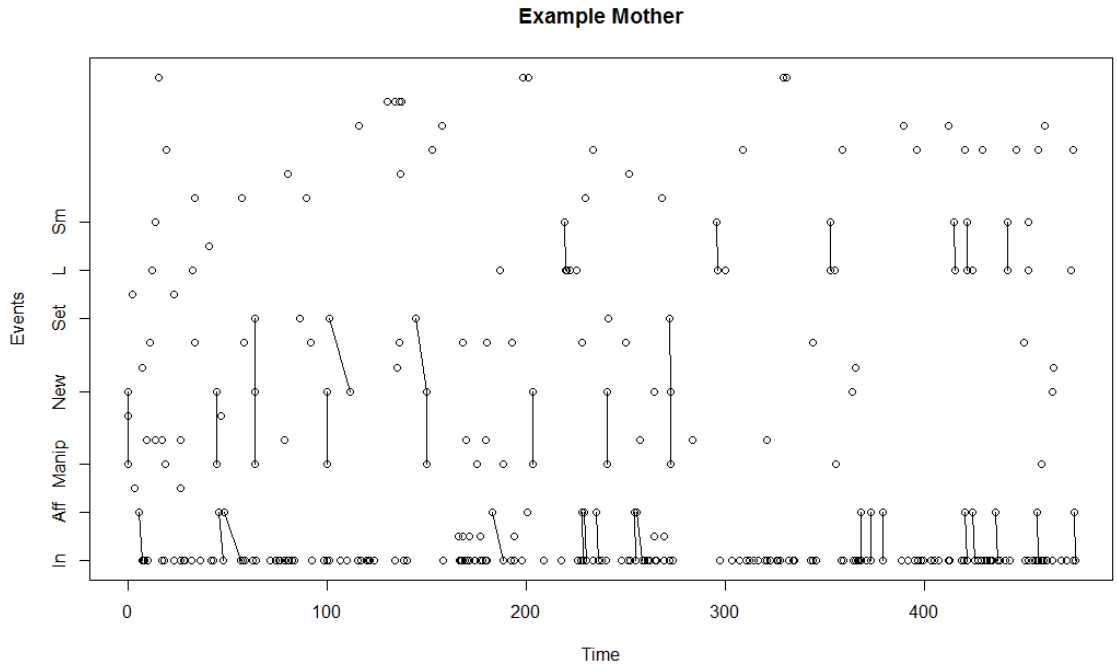

Figure 4.10: The data for an example mother, including lines indicating patterns found with the nonparametric model. The patterns include : New Toy (New) → Manipulating Toy (Manip), Smiling (Sm) → Laughing (L), Set Down Baby (Set) → New Toy (New), and Affectionate Touch (Aff) → Instructive Speech (Ins).

and compare her to the previous models. The data for our example mother can be seen in Figure 4.10

Our example mother displayed four patterns under the nonparametric model: New Toy → Manipulating Toy, Smiling → Laughing, Set Down Baby → New Toy, and Affectionate Touch → Instructive Speech. Under the exponential model, she displayed the first two patterns, as well as Negative Speech→ Negative Speech. For the Weibull model, she only displayed the first two patterns. However, Negative Speech, Positive Speech, and Looking at Baby all had decreasing hazards, indicating they were subject to bursts of activity. Instructive Speech had an increasing hazard.

Thus the nonparametric model displayed the same basic patterns, and found an additional two patterns that neither of the other models were able to detect. Furthermore, the shape

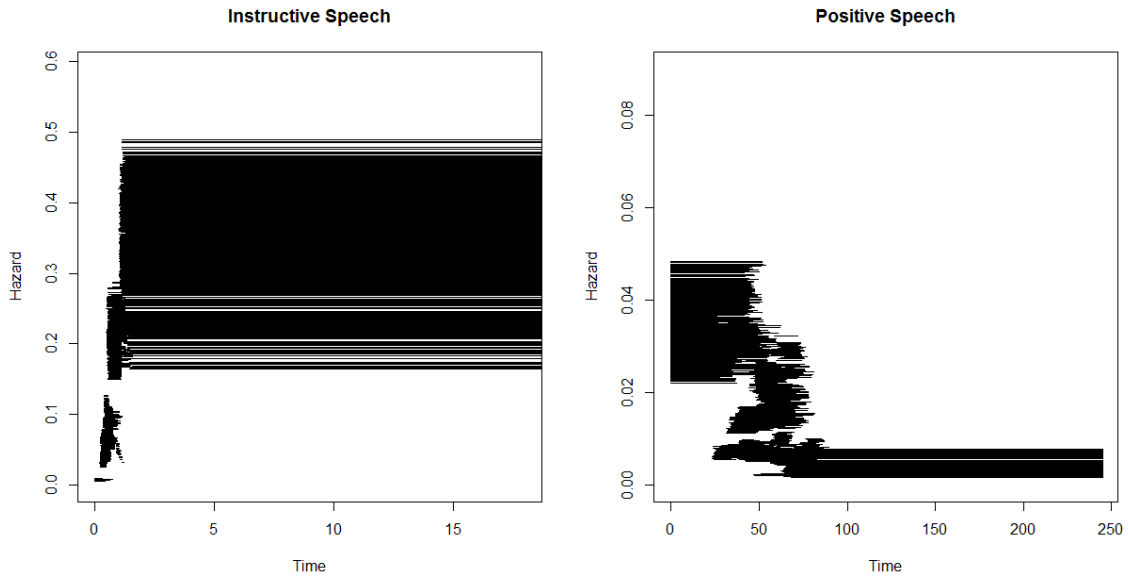

Figure 4.11: Samples from the posterior distributions of the hazard function for two background processes. Instructive speech shows an increasing hazard, while positive speech shows a decreasing hazard.

of the hazard functions for all three of the speech categories resembled the Weibull hazard function. The posterior distribution for the hazard functions of Instructive Speech and Positive Speech can be seen in Figure 4.11. Both plots show several samples of the step function from the posterior distribution. From these, we believe that the Weibull model would be sufficient for many events.

However, the hazard functions for some other events appear to deviate from the Weibull model. Laughing occurred 17 times, generally coming in bursts with little time between instances of Laughing. However, the minimum distance between two Laughing events was 5 seconds, with several occurrences (9 of the 17) occurring between 5 and 10 seconds of the previous Laughing event. The posterior distribution can be seen in Table 4.12. Because none of the Laughing data occurred between 0 and 5 seconds, the hazard is likely increasing in this duration, and so a decreasing hazard does not fit well. Because of this, the Weibull model failed to identify it as a decreasing hazard.

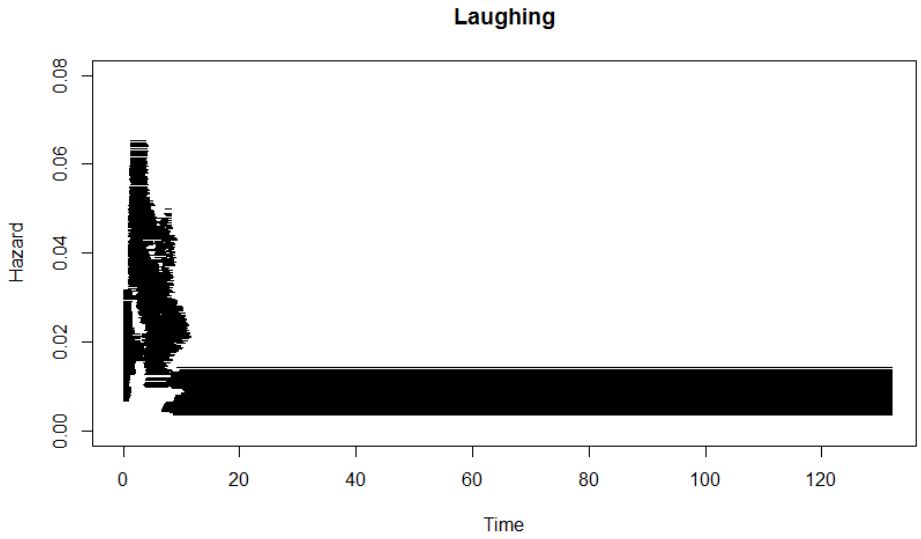

Figure 4.12: Samples from the posterior distributions of the hazard function for the background process corresponding to the event Laughing. The event Laughing generally comes in bursts, but always with a small gap between the individual events, causing a spike in the hazard function.

While such information can be acquired from the nonparametric model, it is still not clear how to obtain it in an automated fashion. We are currently working on methods to easily summarize information from the nonparametric hazard function, similar to the shape parameter for the Weibull distribution.

4.5 Discussion

In this chapter, we explored the implementation of different time increment distributions as part of our renewal process model. In particular, we explored parametric families that are more flexible than the exponential distribution, as well as a nonparametric alternative. These alternative distributions provide further information about the behavior of the actors, and should prove beneficial to behavioral research. Furthermore, we explored the robustness of our model, and in many cases, we determined that patterns will be successfully discovered even if the interarrival time is misspecified.

In this thesis, we only considered modeling the hazard function with step functions as a nonparametric alternative. Other alternatives to standard parametric approaches, such as Gaussian processes and splines, may be one possible area of future work (Rosenberg, 1995; Teh & Rao, 2011).

At the moment, we have only explored the use of alternative distributions for background processes. Extending this framework to accommodate nonexponential sequence processes may be beneficial to modeling the patterns, and thus may be a good direction for future exploration.

In this chapter, we considered several possible potential models for the distribution of the time increments, but we did not focus on model selection. This is another area ripe for future research. In particular, the exponential model of Chapter 2 is a special case of both the Weibull distribution (when the shape parameter is one) and the nonparametric model (as the smoothing parameter $c \to \infty$). This may be useful for exploring different models, potentially as part of a reversible jump procedure.

Chapter 5

Conclusion

In this thesis, our primary goal was to be able to identify recurring patterns in behavior, even when they are not known a priori. To accomplish this, we proposed a generative model, the Renewal Process Model, and an inference procedure for identifying recurring patterns from a stream of events in continuous time. Our approach considers a competing renewal process framework, which allows the patterns to be identified as distinct from background occurrences of the different events.

We expanded the model to allow for numerous observation periods via a hierarchical model. Such a model allows us to investigate how behavior and patterns differ across a population, while simultaneously improving individual level inference by borrowing information from other individuals in the population. Furthermore, we can investigate how subgroups differ across the population. Our model explicitly allows some actors to exhibit a pattern while others do not through our prevalence parametrization.

Our initial model assumes all interarrival times are exponentially distributed. We expanded the model to permit alternative parametric families of interarrival time distributions and nonparametric interarrival time distributions. These alternative distributions allow improved

inference in a variety of settings, and also allow us to better describe certain aspects of behavior.

The models we considered were fit using a fully Bayesian approach via an MCMC procedure that interweaves sampling steps and model exploration steps. This approach proved invaluable, as it allowed us to consider only a subspace of possible patterns, increasing the efficiency of sampling considerably. However, scalability issues persist, as do convergence concerns. Improvements to the model exploration procedure would likely help immensely, and comprise one area of possible research. Improving the model exploration component is especially important for our hierarchical model, which we believe has a difficult time picking up patterns belonging to a small proportion of the individuals in a population.

The models described in this thesis are motivated by a research project to determine the effect of fragmented and unpredictable maternal behavior on various childhood outcomes. While this is an interesting application area, a multitude of other possibilities exist. For example, researchers may be interested in patterns present during social interaction, whether it occurs in person, online, or across some other medium. Another area of interest is user behavior in online settings, where we want to model a user's patterns when web surfing, making purchases, taking online classes, or any number of other online activities.

Bibliography

Aalen, O., Borgan, O. & Gjessing, H. (2008). *Survival and Event History Analysis: A Process Point of View*. Statistics for Biology and Health. Springer.

Baram, T. Z., Davis, E., Obenaus, A., Sandman, C., Small, S., Solodkin, A. & Stern, H. (2012). Fragmentation and unpredictability of early-life experience in mental disorders. *Am J Psychiatry* **169**, 907–915.

Bhattacharya, A., Pati, D., Pillai, N. & Dunson, D. (2012). Bayesian shrinkage. *arXiv Preprint* **arxiv:1212.6088**, 1–42.

Brooks, S., Gelman, A., Jones, G. & Meng, X. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Taylor & Francis.

Eibl-Eibesfeldt, I. (1970). *Ethology: The Biology of Behavior*. Holt, Rinehart and Winston.

Engle, R. F. & Russell, J. R. (1998). Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica* **66**, pp. 1127–1162.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. & Rubin, D. (2013). *Bayesian Data Analysis*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 3rd ed.

Gelman, A. & Rubin, D. B. (1992). Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science* **7**, 457–472.

Hawkes, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika* **58**, 83–90.

Heins, K. & Stern, H. (2014). A statistical model for event sequence data. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Hoffman, M. D. & Gelman, A. (2012). The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* .

Kellis, M., Patterson, N., Birren, B., Berger, B. & Lander, E. S. (2004). Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery. *Journal of Computational Biology* **11**, 319–55.

Klein, J. & Moeschberger, M. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Statistics for Biology and Health. Springer.

Kottas, A. (2006). Nonparametric bayesian survival analysis using mixtures of weibull distributions. *Journal of Statistical Planning and Inference* **136**, 578 – 596.

Liu, J. S., Neuwald, A. F. & Lawrence, C. E. (1995). Bayesian models for multiple local sequence alignment and gibbs sampling strategies. *Journal of the American Statistical Association* **90**, 1156–1170.

Magnusson, M. S. (2000). Discovering hidden time patterns in behavior: T-patterns and their detection. *Behavior Research Methods, Instruments, and Computers: a Journal of the Psychonomic Society, Inc* **32**, 93–110.

Marcum, C. & Butts, C. (2014). Constructing and modifying sequence statistics for relevent using informr in r.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P. & Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association* **106**, 100–108.

Neal, R. M. (2010). MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo* **54**, 113–162.

Nieto-Barajas, L. E. & Walker, S. G. (2002). Markov beta and gamma processes for modelling hazard rates. *Scandinavian Journal of Statistics* **29**, 413–424.

Nieto-Barajas, L. E. & Walker, S. G. (2004). Bayesian nonparametric survival analysis via levy driven Markov processes. *Statistica Sinica* **14**, 1127–1146.

Ogata, Y. (1988). Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes. *Journal of the American Statistical Association* **83**, 9+.

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

Simma, A. & Jordan, M. (2010). Modeling events with cascades of poisson processes. In *Proceedings of the Twenty-Sixth Conference Annual Conference on Uncertainty in Artificial Intelligence*.

Sisson, S. A. & Fan, Y. (2007). A distance-based diagnostic for trans-dimensional markov chains. *Statistics and Computing* **17**, 357–367.

Stan Development Team (2013). Stan: A c++ library for probability and sampling, version 1.3.

Stephens, M. (2000). Bayesian analysis of mixture models with an unknown number of components- an alternative to reversible jump methods. *The Annals of Statistics* **28**, pp. 40–74.

Teh, Y. & Rao, V. (2011). Gaussian process modulated renewal processes. In *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira & K. Weinberger, eds. pp. 2474–2482.

Walker, S. G. & Mallick, B. K. (1997). Hierarchical generalized linear models and frailty models with bayesian nonparametric mixing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **59**, 845–860.

Welch, T. A. (1984). A technique for high-performance data compression. *Computer* **17**, 8–19.

Ziv, J. & Lempel, A. (1978). Compression of individual sequences via variable-rate coding. *IEEE Trans. Inf. Theor.* **24**, 530–536.