9-10-2010

Generalizations of the statistical flowgraph model framework

Richard Warr

Richard L. Warr
Candidate

Mathematics and Statistics
Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

, Chairperson

Generalizations of the Statistical Flowgraph Model Framework

by

Richard Lyman Warr

B.S., Mathematics, Southern Utah University, 1996
M.A., Mathematics, University of Nebraska at Omaha, 2005
M.S., Statistics, University of New Mexico, 2009

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Statistics

The University of New Mexico

Albuquerque, New Mexico

July, 2010

The views expressed in this article are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

Acknowledgments

I would first like to thank James Cotts, my undergraduate statistics professor, for cultivating my interest in the field of statistics, and to master's thesis advisor Jack Heidel. Additionally, many thanks to the UNM statistics group that has provided me with an excellent foundation in the thoery and application of statistical methods. I would also like to thank the statistics group at Los Alamos National Laboratory for their hospitality and assistance, especially Mike Hamada and Dave Collins for their editorial comments. To my dissertation committee for their guidance and assistance with my schooling and research; to Donna George and Roxanne Littlefield for their advisement and patience with me. I am also appreciative of my fellow statistics students who significantly enhanced my educational experience and to the USAF for considering and funding me in this program. I would especially like to thank my advisor, Aparna Huzurbazar, for her time, encouragement, and mentoring in statistical research and academia. Above all, I thank my wife and family for their help, patience, and long-suffering through this challenging experience.

Generalizations of the Statistical Flowgraph Model Framework

by

Richard Lyman Warr

ABSTRACT OF DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
Statistics

The University of New Mexico

Albuquerque, New Mexico

July, 2010

Generalizations of the Statistical Flowgraph Model Framework

by

Richard Lyman Warr

B.S., Mathematics, Southern Utah University, 1996

M.A., Mathematics, University of Nebraska at Omaha, 2005
M.S., Statistics, University of New Mexico, 2009

Ph.D., Statistics, University of New Mexico, 2010

Abstract

Statistical flowgraphs model multistate semi-Markov processes and provide a way to perform inference for these processes. This methodology provides powerful results that significantly impact the study of multistate semi-Markov processes. This dissertation extends previous work in several ways. First, by demonstrating how any "smooth" transition distribution can be incorporated into a statistical flowgraph model (SFGM), we provide a method to use popular distributions, such as the log-normal, that have not been used in the past. Next, we propose an alternate way to consider Bayesian SFGMs by showing how computation can be accomplished when the traditional methods of SFGMs fail to be computationally feasible. We demonstrate this method with a Bayesian non-parametric example. We extend flowgraph models to handle time-varying covariates using an accelerated failure time model.

We also show how SFGMs can be used to make inference in multistate semi-Markov models to calculate exact likelihood functions when faced with incomplete data. Finally, we develop a goodness-of-fit criterion that is applicable to any continuous model and can be applied to SFGMs. This goodness-of-fit test criterion is general enough to be useful when dealing with censored and incomplete multistate data.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This dissertation focuses on some important problems in the area of statistical flowgraph models (SFGMs). Flowgraph models are a framework that can be used to develop multistate models. They have been used to compute the distribution of waiting times in complex stochastic networks with feedback loops. SFGMs have been used in both Bayesian and frequentist frameworks. The final result of a Bayesian SFGM is a posterior predictive density (PPD) of the first passage time from one state to another, and similarly for frequentist SFGMs the result is a probability density function (PDF).

SFGMs connect a vast number of areas that are of interest in statistics, mathematics, computer science, and engineering. Some application areas include survival analysis and disease progression in medical studies, reliability engineering, and queuing theory; all of these involve stochastic processes. The applications we tend to focus on in this thesis are in the areas of survival analysis and reliability; namely in the prediction of time until a specific event occurs. In survival analysis this is often some significant event such as death, or in reliability, the time until some type of failure occurs.

Statistical flowgraph models provide a way to do prediction and inference in multistate models. Multistate models often represent longitudinal data consisting of states and waiting times until events occur. Figure 1.1 gives an example of a SFGM. The diagram represents a patient's transition through a notional disease process for a recurrent illness. State 0 represents "good health". In state 1 a patient is "diseased" or ill. The patient is allowed to recover from the disease with a transition from state 1 to state 0. Eventually the patient dies from the disease and makes a transition to state 2, "death".

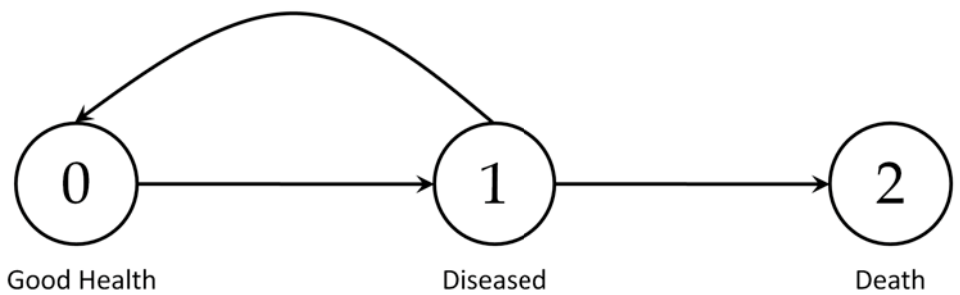


Figure 1.1: A diagram of a recurring illness process.

Our interest may be in the time until death, given the patient is in state 0, or state 1. SFGMs can answer such questions as: What is the mean predicted time until death occurs, given the patient is in good health, or given the patient is ill? How likely is it that an individual recovers from this illness? What is the predicted probability of survival beyond a certain time? Or in a reliability context, how long do we expect this engine to run without any downtime? How do these answers change when we have additional information in the form of patient covariates, or manufacturing and usage information about the engine? Although these scenarios fall under the general realm of stochastic processes, traditional methods from stochastic processes do not provide general solutions. SFGMs (using the semi-Markov assumption) provide a method to estimate and predict first passage times from one state to another.

Statistical flowgraphs have their limitations. Flowgraphs require that transitions be conditionally independent; this may be a reasonable assumption in some situations, but it is often violated. For example, in a medical study, if a patient is resistant to a treatment the first time, there may be some correlation if the patient is treated a second time. Another limitation is that SFGMs are restricted to distributions that have moment generating functions, which limits their modeling capabilities. Also, flowgraphs only model processes with a finite state space, so if the process has an infinite number of possible states, the state space must be redefined to be finite. Another shortcoming is that when a SFGM is developed there is no quantitative way to assess how well it models the process. In this dissertation we address some of these limitations to make SFGMs more attractive as a modeling tool. We demonstrate the techniques and methods with real and simulated data examples. We carry a simulated data example across chapters to provide continuity in illustration and discussion.

In this dissertation we propose generalizations and extensions of SFGMs. The first area we address is generalizing the distributions that can be used in SFGMs. We develop the methodology to incorporate any "smooth" distribution in a SFGM. We use the word *smooth* to denote a continuous and differentiable probability density function. We also introduce a Bayesian technique to the flowgraph framework that allows faster estimation of the posterior predictive distribution. In the Bayesian framework this technique also allows any distribution to be used in SFGMs whether smooth or otherwise. Using this technique we also introduce Bayesian non-parametric methods to SFGMs. In addition, with this added flexibility of modeling with any distribution, we introduce a method to handle time-varying covariates in SFGMs, an important problem in survival analysis. Next, we discuss how these advances affect the way incomplete data are handled in a semi-Markov model and propose a more general method of modeling incomplete data using SFGMs. We also propose a goodness-of-fit criterion that can be applied to complex models such as flowgraphs.

These are the primary contributions of this dissertation to the SFGM framework.

We give a summary of the chapters to follow. Chapter 2 provides a brief introduction into stochastic processes and how they relate to SFGMs. Chapter 2 defines a SFGM as a stochastic process and discusses its implementation. This is an important section that should be read if not familiar with the statistical flowgraph methodology.

In Chapter 3 we introduce how any smooth time-to-event distribution can be used in SFGMs. This allows any distributions that do not have moment generating functions (MGFs), such as the lognormal or certain parameterizations of the Weibull, to be used in flowgraphs. Incorporating these distributions into SFGMs is an important step for modeling in survival analysis and reliability. Examples demonstrating these techniques are provided.

Chapter 4 discusses Bayesian SFGMs and demonstrates how to efficiently predict in complex SFGMs. We specifically address an alternative way to estimate the posterior predictive distribution in especially complicated Bayesian SFGMs. Advanced Bayesian SFGM examples are provided. One illustrates a non-parametric flowgraph and another focuses on the important problem of time-dependant covariates in an accelerated failure time model.

Chapter 5 shows how SFGMs can be used to calculate the likelihood function of a semi-Markov model when incomplete data are present. Using the methods from Chapter 3, we show how SFGMs can calculate an exact likelihood function, whereas the previous literature addressing incomplete data only developed approximations for it. This provides better inference based on fewer assumptions.

In Chapter 6 we introduce a goodness-of-fit method that can be applied to any continuous model, but are specifically adapted to SFGMs. This is an important area that has not been addressed in SFGMs. The suggested method enables models to be appropriately appraised before implementation. This methodology also enhances

building models, by helping to determine the most appropriate distributions for a particular transition of a flowgraph.

In the final chapter we review the contributions of this paper to SGFMs and their impact. We also list some of the open problems and additional areas for promising research in SFGMs.

We include a simulated example at the end of each major chapter. The code is contained in the appendix, so that the reader can apply the techniques without undue investment in time. The computation was conducted using R, which is a free software environment for statistical computing and graphics; it runs on UNIX/Linux, Windows, and Mac operating systems (see Hornik (2009) or R Development Core Team (2009) for details). For all the distributions used in this dissertation, the parameterizations are the same as defined in R or the R code.

Chapter 2

Stochastic processes and statistical flowgraph models

A stochastic process is a random process which may change in state over time. Statistical flowgraphs are a type of stochastic process, where the state of a flowgraph varies over time. We review some definitions and areas of stochastic processes that are applicable to SFGMs. For a detailed introduction to stochastic processes see Taylor and Karlin (1998) or Ross (1996). We present only the material necessary to understand statistical flowgraph models.

2.1 Stochastic processes

Definition A *stochastic process* is a collection of random variables, $X(t)$, indexed by a parameter, t, regarded as time (see Billingsley (1995)).

In most instances this index parameter belongs to a set T, typically $T = \ 0, 1, 2, 3, \ldots$ or $T = [0, \quad)$. In our context for survival analysis and reliability we will use the

index set $T = [0, \quad)$ exclusively, where T represents time. The random variable $X(t)$ corresponds to the state of the process at time t. The *state space* is the set of all possible values that $X(t)$ can assume.

Recall the SFGM in Figure 1.1. Let $X(t)$ denote the process for a particular patient at time t. If $X(t) = 0$ then the patient would be in "good health". Likewise if $X(t) = 1$ or $X(t) = 2$ we could make a statement about the health of the patient.

The state space can be a finite set such as $\text{alive}, \text{dead}$ in survival analysis or $\text{fully-operational}, \text{degraded}, \text{failed}$ in reliability. The state space can also be infinite, either countable or uncountable (see Rudin (1976)). An example with an infinite uncountable state space is an environmental model where $X(t) \quad [-273.15, \quad)$, and $X(t)$ represents the temperature in degrees Celsius at time t. This same example could also be interpreted as an infinite countable state space if we are only able to measure the temperature to the nearest tenth of a degree.

SFGMs have been developed for applications that have a finite state space representing potential outcomes, and we confine our discussions to these. Stochastic processes have been used to successfully model numerous phenomena. Familiar examples are stock market prices, audio signals, medical data such as blood pressure, and random movements similar to Brownian motion (Taylor and Karlin (1998)). Inference and prediction in stochastic processes can be very difficult unless simplifying assumptions are made.

2.1.1 Markov processes

A Markov process is an example of a stochastic process with simplifying assumptions. The *Markov property* was introduced by A. A. Markov (1856-1922) while trying to model Brownian motion. The Markov property is essentially one of conditional independence; given the state of a process at a particular time, the future of the

process is independent of the past. For example: if the weather were a stochastic process that possessed the Markov property, then the weather yesterday would not help predict the weather tomorrow, if we know the weather today. The weather today would determine the probabilities of possible types of weather tomorrow. Clearly this property is quite restrictive and is often not completely accurate, nevertheless, it can be very useful in stochastic modeling.

The definition of a *Markov process* is simply a stochastic process that possesses the Markov property. More formally, if $X(t)$ is a stochastic process in state j, and i is any possible adjacent state, then

$$P(X(t+\varepsilon) = i|X(t), X(s)) = P(X(t+\varepsilon) = i|X(t)),$$

for all $\varepsilon > 0$ and all s such that $0 \leq s < t$, then we say $X(t)$ is a Markov process. State i is considered *adjacent* to state j if and only if the process can proceed directly from state i to state j without transitioning through any other intermediate state. Assuming a process has the Markov property greatly simplifies calculations and allows difficult problems to be solved more easily.

Markov processes have been used in a variety of applications. They have been used to model generic disease progression in Fix and Neyman (1951), and for progression of specific diseases such as cancer (Lagakos (1976)), kidney disease (Gross et al. (1971)), and HIV (Longini et al. (1989)). A complete literature review on Markov processes is too vast to include here, but Stroock (2005) and Grimmett and Stirzaker (2001) provide an introduction to the development of Markov process theory and its applications.

The exponential distribution has the memoryless property, which makes it the natural distribution to model Markov processes. The exponential distribution is the only continuous distribution with this memoryless property (the geometric is the only discrete distribution). Therefore if a finite state continuous time process is truly

Markovian, then the exponential distribution will perfectly model its transitions.

Although Markov models have been successful in modeling many processes, some processes are not appropriately modeled within this framework.

2.1.2 Semi-Markov processes

A semi-Markov process is a stochastic process that has fewer restrictions than a Markov process. The *semi-Markov property* relaxes the Markov property by allowing the probability of a future state to depend not only on the last observation, but also on the amount of time the process has been in the current state. This generalization greatly increases the model's flexibility, by allowing the duration of time in a particular state to "affect" the transition time. Therefore any distribution with positive support could be used to model the transitions of a semi-Markov process, in contrast to the Markov model where the exponential distribution is the only allowable transition distribution. A semi-Markov process has also been called a duration dependent Markov process.

We define a *semi-Markov process* to be a stochastic process that possesses the semi-Markov property. A process $X(t)$ in state j, is a semi-Markov process if and only if, for any possible adjacent state i, $P(X(t+\varepsilon) = i|X(t), X(s), t_j) = P(X(t+\varepsilon) = i|X(t), t_j)$, for every $\varepsilon > 0$, and all s such that $0 \leq s < t$, and t_j represents the time $X(t)$ has been in state j.

The concept of semi-Markov processes is generally agreed to have been simultaneously introduced by Lévy (1954), Takacs (1954), and Smith (1955). The theory was formalized soon after in Pyke (1961a) and Pyke (1961b). Further developments came from Takacs (1959), Pyke and Schaufele (1964), Pyke and Schaufele (1966), and Çinlar (1969).

Barbu and Liminios (2008) mention that semi-Markov processes are applied in queuing theory, reliability, survival analysis, performance evaluation, biology, DNA analysis, risk processes, insurance and finance, earthquake modeling, and more. For a formal introduction to semi-Markov processes in reliability see Limnios and Oprişan (2001).

We have introduced the concept of a semi-Markov process in the notation of stochastic processes. When dealing with SFGMs we do not usually use this notation. SFGMs were created to perform data analysis for semi-Markov processes without getting bogged down in the mathematics required for semi-Markov processes. We introduced semi-Markov processes to help the reader understand both the powerful modeling capabilities of statistical flowgraphs, but also their limitations given the semi-Markov assumption. Next, we define a SFGM and more convenient notation for dealing with SFGMs.

2.2 Statistical flowgraph models

Flowgraph models are robust enough to model any finite state semi-Markov process. This section introduces the basic concepts regarding SFGMs. For a comprehensive treatment see Huzurbazar (2005c).

We now formally define a SFGM. A *statistical flowgraph model* is a directed graphical depiction of a finite state stochastic process that is assumed to have the semi-Markov property. In this graph the nodes represent the states of the process. The time until a transition of the process occurs is characterized by one or more directed branches. Each branch has an associated waiting time distribution which represents the random time it takes for the transition to occur. If there are two or more paths leaving a node then we also include a probability of passage for each path.

Flowgraph models were first used in engineering and appeared in the literature of electrical engineering as "signal flow graphs" (Mason (1953)). Mason (1953) was primarily concerned with solving systems of linear equations. Signal flowgraphs are concerned with "transmitting" current with respect to inductance and capacitance. In reality, the branches of such a flowgraph can be labeled with anything. SFGMs began by labeling the branches with moment generating functions (MGFs) of waiting time distributions. This was convenient for flowgraph algebra to solve the system of linear equations. Using MGFs on the branches of a flowgraph makes SFGMs more accessible, but limits what distributions can be used. It is equally valid, and sometimes more appropriate, to label the branches with other functions that may represent the same distribution. For a branch connecting state i to state j, we use $M_{ij}(s)$ to represent the MGF, $L_{ij}(z)$ as the Laplace transform (LT), and $F_{ij}(t)$ as the cumulative distribution function (CDF); these functions are the various ways we represent the random waiting time the process resides in state i before a transition to state j. For a random variable T_{ij}, we say the MGF does not exist if $\varepsilon > 0, M_{ij}(\varepsilon) =$. Using $F_{ij}(t)$ on the branches is more general, since the CDF always exists.

Butler and Huzurbazar (1997) adapted flowgraph models for use in Bayesian stochastic models. Since then, the use and theory of statistical flowgraph models has continued to expand. Huzurbazar (1999a) used SFGMs to generalize phase-type distributions. Huzurbazar (2000) demonstrated a Bayesian application of SFGMs on a complex cellular telephone network. Butler and Huzurbazar (2000) improved on some of the techniques used in flowgraph modeling and demonstrated their use in Bayesian prediction of waiting times in queuing theory. Yau and Huzurbazar (2002) show how SFGMs can be used to model incomplete data in multistate systems. The theory linking semi-Markov processes with multistate models using SFGMs was explained in Huzurbazar (2004b). Huzurbazar (2005b) provides an excellent example of how Bayesian SFGMs can be applied in various fields, using an example in con-

struction project management. Huzurbazar (2005c) is a full length text devoted to SFGMs and its applications. Other applications of SFGMs can be found in Huzurbazar (2002), Huzurbazar and Williams (2005), and Huzurbazar (2004a). Williams and Huzurbazar (2006) provides a Bayesian approach on how to construct a likelihood function when faced with incomplete data. Collins (2009) formally introduces non-parametric methods to SFGMs and Collins and Huzurbazar (2008) use a simple non-parametric flowgraph to model cumulative earthquake damage to buildings. Huzurbazar and Williams (2010) is a significant publication, which provides the methodology to incorporate covariates into a Bayesian SFGM. However, they do not consider time-varying covariates. To date, these are the main developments of flowgraph models in statistics. There are still many open research problems in SFGM theory.

There is a systematic way to implement statistical flowgraphs. The first step is to propose the system diagram or graphical model. Often graphical models in statistics literature refer to models with the random variable modeled as the node in the graph (see Edwards (2000)); however, this is not the case in SFGMs. We design our graph by identifying the states the process can assume. These states are represented by the nodes of the graph. Next, we identify the possible transitions between states, which are the directed branches (or edges) of the graph. Hougaard (1999) provides an excellent introduction on developing multistate models. Once the graphical model is in place, we examine the data to suggest appropriate distributions for the branch transition times. This is usually accomplished by comparing a histogram of the data with several families of parametric distributions. Huzurbazar (2005a) suggests a method to construct a histogram for situations with censored data. The selected distributions model the time it takes to transition from one state to another. Next, we find the MGFs or LTs of the distributions assigned to the branches of the flowgraph.

Definition A *first passage distribution* from state i to state j is the distribution of

the random time it takes a process to transition from state i to state j, regardless of the path the process takes.

Mason's rule, described in Mason (1953), is a way to find the first passage MGF in a flowgraph. The general form of Mason's rule (as found in Huzurbazar (2005c, pp. 36)) provides the overall MGF from input to output as

$$M(s) = \frac{\sum_i P_i(s)[1 + \sum_j (-1)^j L_j^i(s)]}{1 + \sum_j (-1)^j L_j(s)}, \tag{2.1}$$

where:

- $P_i(s)$ is the transmittance for the i^{th} path.
- $L_j(s)$ in the denominator is the sum of the transmittances over the j^{th}-order loops.
- $L_j^i(s)$ is the sum of the transmittances over the j^{th}-order loops sharing no common nodes with the i^{th} path (i.e., loops not touching that path).

We apply Mason's rule to the flowgraph to find the MGF of the first passage from state i to state j. However, this first passage MGF is not of much practical use, until we transform it into a probability density function (PDF). Using the PDF of the overall flowgraph we can then proceed with inference and prediction.

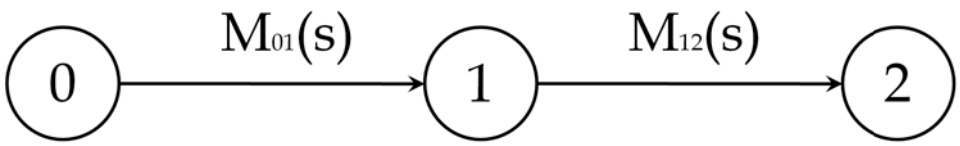


Figure 2.1: A SFGM for a simple series system.

Consider the simple SFGM in Figure 2.1 with states 0, 1, and 2 representing the states of a system. T_{01} represents the random waiting time to transition to state

1 starting in state 0. T_{12} represents the random waiting time to transition from state 1 to state 2. In SFGMs the random variables T_{01} and T_{12} are conditionally independent given the process transitioned to state 1. Now define $T_{02*} = T_{01} + T_{12}$, which represents the total waiting time starting in state 0 to reach state 2. We use the "*" as the notation to denote first passage distributions. Now consider the modified SFGM in Figure 2.2. We have a branch connecting state 0 to state 2 directly; this "*" notation distinguishes the first passage distribution, T_{02*}, from the direct transition from state 0 to state 2, T_{02}. So in Figures 2.1 and 2.2, T_{02*} represents the random waiting time to get from state 0 to state 2. In Figure 2.1 T_{02*} is the convolution of the PDFs of T_{01} and T_{12}. We can obtain the distribution of T_{02*} via its MGF, $M_{02*}(s) = M_{01}(s)M_{12}(s)$, which can be transformed into a density. For example, if we model T_{01} with a $gamma(a_1, b)$ and T_{12} with a $gamma(a_2, b)$, then $T_{02*} \sim gamma(a_1 + a_2, b)$. However, convenient models that convolve to closed form solutions are often overly simplistic and not justified by the data or *a priori* information of the true transition behavior. Now consider a less convenient model where $T_{01} \sim Weibull(a_1, b_1)$ and $T_{12} \sim Weibull(a_2, b_2)$. In this case we must now rely on a numerical solution to transform $M_{02*}(s)$ into a PDF, provided $M_{02*}(s)$ even exists. For the Weibull distribution $M_{02*}(s)$ will not exist if $a_1 < 1$ or $a_2 < 1$; however, in this case we would need to change the parameterization of our model, or use a transform that exists for all distributions. We introduce the latter option of complex LTs in the next chapter.

There are three main characteristics or features in a SFGM. The first is a *series* structure shown in Figure 2.1. This is solved by finding the convolution of the random variables in the series. The second feature of a SFGM is a *parallel* structure, an example of this is shown in Figure 2.2. A parallel structure exists in a graphical model if there are two different paths from state i to state j. The first passage distribution in a parallel structure is the mixture distribution of the paths. For example in Figure 2.2 the MGF of T_{02*} is $pM_{01}(s)M_{12}(s) + (1-p)M_{02}(s)$. The third

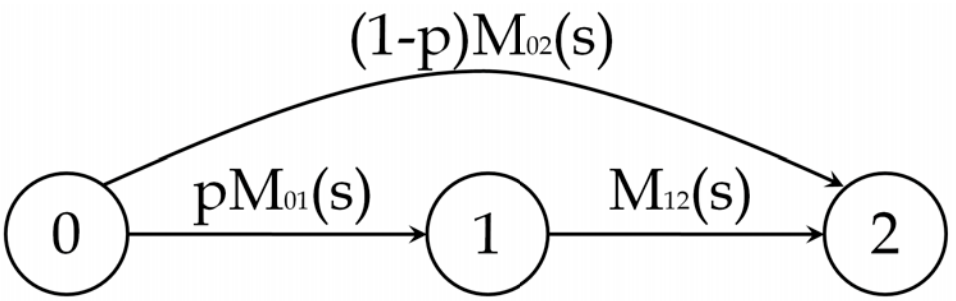


Figure 2.2: An example SFGM.

feature of a SFGM is the *loop*. If there is a possibility that once in state i the process can return to state i at a later time, then the SFGM has a loop. The SFGM in Figure 2.3 is a fairly simple graph that includes all three features.

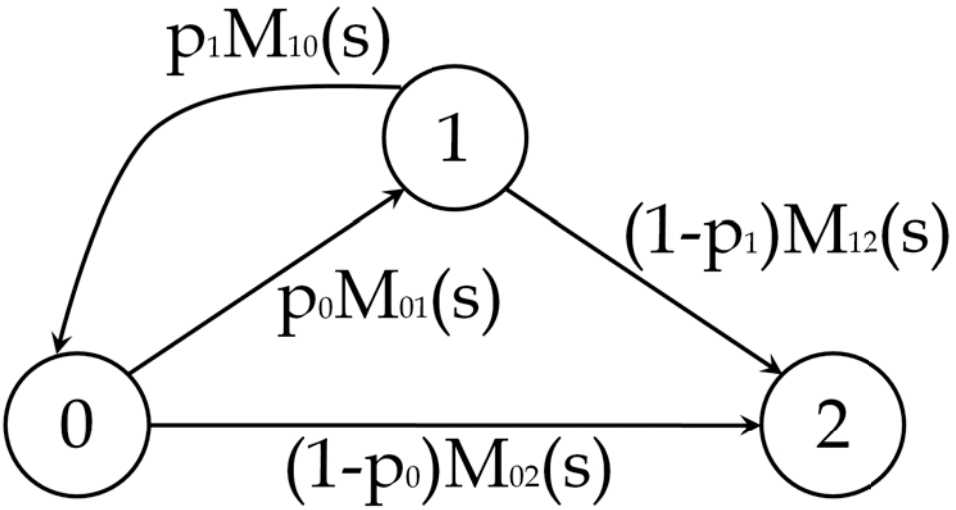


Figure 2.3: A SFGM that includes a series, parallel path, and a loop.

Without loss of generality, we restrict our consideration to SFGMs that have the following regularity properties:

1. There can be only be one directed branch from node i to node j (this does not

exclude a directed branch from node j to node i).

2. There are no isolated states (state i is isolated if $P(X(t) = i) = 0$ for all $t > 0$).

3. There are no states that have instantaneous transition(s) of probability 1.

4. There are no irrelevant states (if considering the first passage time from state i to state j, state k is irrelevant if $P(X(t) = j|X(t-s) = k) = 0$ for all $0 < s < t$).

5. The process is stationary (i.e., the transition distributions (and their weighting probabilities) do not change over time).

6. There must be more than one state.

When formality requires, we will refer to SFGMs that satisfy the above conditions as *regular SFGMs*.

2.2.1 "Solving" statistical flowgraphs

The term "solving" a flowgraph refers to finding the MGF of the overall waiting time distribution from a beginning node to an ending node. This requires using flowgraph algebra or Mason's rule. Consider the semi-Markov process in Figure 2.4. It has three states, where state 2 is absorbing. We model each branch of the graph with a distribution $f_{ij}(t)$ and corresponding MGF, $M_{ij}(s)$. Applying Mason's rule we find that the MGF of the first passage from state 0 to state 2 is

$$M_{02*}(s) = \frac{(1-p)M_{01}(s)M_{12}(s)}{1 - pM_{01}(s)M_{10}(s)}, \tag{2.2}$$

where p is the probability that the process will proceed to state 0 before state 2, given the process is in state 1. We could replace Figure 2.4 with an equivalent model found in Figure 2.5. This equivalent model has only two states 0 and 2, with one

connecting branch. The MGF assigned to the branch is $M_{02*}(s)$ as defined in (2.2). This flowgraph is referred to as the *solved* or *reduced* flowgraph.

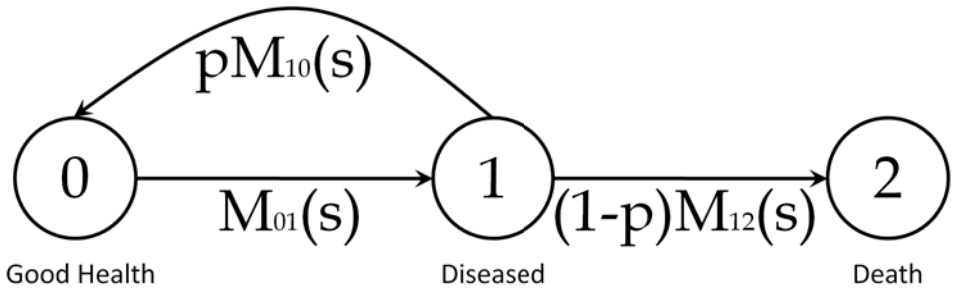


Figure 2.4: The recurring illness process.

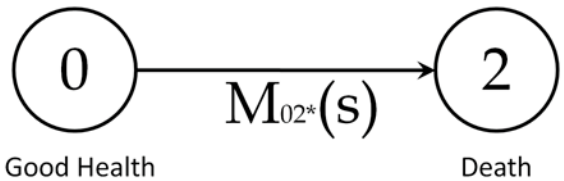


Figure 2.5: The "solved" flowgraph of Figure 2.4.

The following steps are a simple way to use Mason's rule and find the MGF of a first passage distribution. To find the first passage MGF from state i to state j, do the following:

1. Remove any branches in the SFGM departing from state j
2. Remove any node k (and its incoming and departing branches) if it is irrelevant to the first passage from state i to state j
3. Renormalize the branch probabilities so that the probabilities on the departing branches of each node sum to one
4. Renumber the remaining nodes so state i is 1 and state j is m

5. Create an m x m matrix T such that the entry in the q^{th} row and r^{th} column of T is comprised of the branch MGF that directly connects state q with state r, this is multiplied by the corresponding probability p.

6. Then $M_{1m*}(s) = (1, 0, \ldots, 0)(I - T)^{-1}(1, 0, \ldots, 0)$

When creating the matrix T if there is no direct connection between state q and state r, the entry on the q^{th} row and r^{th} column of T is 0. It is important to note that the branches are directed, so state q may be connected to state r but the reverse is not necessarily true. $(I - T)^{-1}$ exists if certain conditions are imposed on T, see Collins (2009).

If we follow these steps to find the MGF of the first passage from state 0 to state 2 for the semi-Markov process in Figure 2.4 we find that

$$T = \begin{matrix} 0 & M_{01}(s) & 0 \\ pM_{10}(s) & 0 & (1-p)M_{12}(s) \\ 0 & 0 & 0 \end{matrix} .$$

Therefore,

$$M_{02*}(s) = \frac{(1-p)M_{01}(s)M_{12}(s)}{1 - pM_{01}(s)M_{10}(s)}.$$

There are times when the graphical model of a flowgraph must be redefined to obtain the first passage MGFs. An example of this is finding the first passage MGF from state 0 to state 0. To do this we must divide state 0 into two states, $0a$ and $0b$. We begin in state $0a$, which retains the departing branches, and find the first passage MGF to state $0b$, which keeps the incoming branches. We can also creatively find second passage MGFs and other items of interest in a similar manner.

2.2.2 Inversion via saddlepoint method

Once we obtain an MGF of the first passage, we use an inversion technique to convert the MGF into a density. There are several methods to accomplish this. The most frequently used in SFGM literature is the saddlepoint method. Although we mostly use this method for comparison of other proposed methods, we provide a short introduction.

The saddlepoint method was introduced into statistics by Daniels (1954). Its primary use is not for MGF inversion, but has been used for this purpose in SFGMs. Reid (1988) and Huzurbazar (1999b) provide introductory discussions on saddlepoint methods in statistics. Huzurbazar (2005c, chap. 3) gives an excellent introduction of saddlepoint methods for SFGMs.

The saddlepoint method uses a function of the MGF called the cumulant generating function, which is defined as $K(s) = log[M(s)]$, where $M(s)$ is the MGF. The saddlepoint density approximation for a single random variable (n=1) is

$$f(t) \quad \frac{1}{\sqrt{2\pi K\ (s)}} \exp\ K(s) - st\ , \tag{2.3}$$

where $K\ (s) = d^2K(s)\ ds^2$ and $K\ (s) = t$. This approximation is only valid if the MGF exists.

This approximation can be fast and accurate, but in some cases it can be very imprecise. We give an example of two cases.

Consider the convolution of two gamma distributions. Let $T_1 \sim gamma(\alpha_1, \beta)$, $T_2 \sim gamma(\alpha_2, \beta)$, and $\alpha_3 = \alpha_1 + \alpha_2$. We know the MGF of $T_3 = T_1 + T_2$ is $M(s) = (1 - s\ \beta))^{-\alpha_3}$, then $K(s) = -(\alpha_3)\log(1 - s\ \beta)$,

$$K\ (s) = \frac{\alpha_3}{\beta - s} \text{ and } K\ (s) = \frac{\alpha_3}{\beta(\beta - s)^2}.$$

Solving $K(s) = t$ gives that $s = \beta - \alpha\ t$. Putting these into Equation 2.3 we find

$$f(t) \quad t^{\alpha_3 - 1} \exp\ -\beta t\ .$$

This is just a $gamma(\alpha_3, \beta)$, which is what we expect to get from the convolution $T_3 = T_1 + T_2$. So in this case, if $f(t)$ is properly normalized, the saddlepoint method is exact. This illustrates one of the drawbacks of the saddlepoint approximation: we must compute it at many points on the support of t to get an accurate normalizing constant.

For the second example consider the mixture of two gamma distributions. Let $X \sim gamma(5, 5)$, $Y \sim gamma(50, 10)$, and define the PDF of Z to be $f(z) = f(x)\ 2 + f(y)\ 2$. We can obtain a closed form MGF for Z, however the calculations to find $f(z)$ must be done numerically. In Figure 2.6 we show both the exact mixture and the saddlepoint approximation. Clearly the saddlepoint method smoothes out the density we are trying to approximate. In fact, Collins (2009) proves that the error of the saddlepoint method can be arbitrarily large. In both examples the saddlepoint method performs well in the tails of the distribution, but there is no guarantee how it will perform in the center of the distribution. We see later that there are other methods that do better than the saddlepoint; however, there may be instances when the saddlepoint method is preferable. In several instances we find that the theory of SFGMs has been limited by using the saddlepoint method to invert MGFs. In later chapters we show how some aspects of SFGMs are improved by using alternate inversion methods.

2.2.3 Markov SFGMs

There has been confusion in the past if a particular SFGM is a Markov process. Intuitively, for a SFGM to be Markovian, all waiting times must be exponentially distributed. Having exponential distributions in a Markov SFGM is a necessary

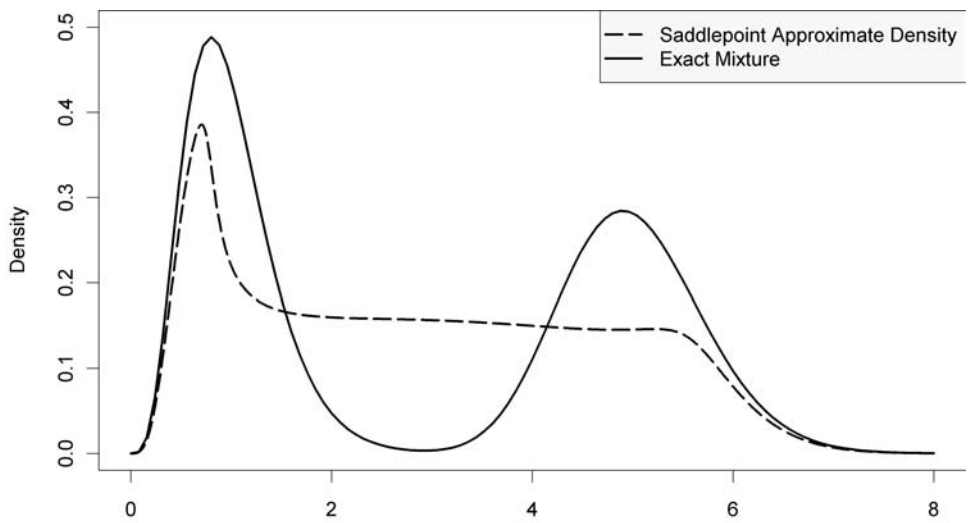


Figure 2.6: The saddlepoint approximation of a mixture of gammas.

but not a sufficient condition. Therefore, if a process has exponential waiting time distributions it may or may not be a Markov process. This begs the question, what does a branch on a SFGM really represent? Previously, we stated that it represents the random waiting time until a transition occurs. This is true, but when there are two or more branches leaving a node then how do these two branches interact or compete with each other? If there are two or more paths leaving a node in a SFGM, then each branch represents the random waiting time before a transition occurs given that this transition occurs before any others. Therefore, it is a conditional distribution that this transition occurs first. This makes sense from a practical point of view, since these conditional distributions are the ones we can actually observe. Consider Figure 2.2; let the $0 \to 1$ transition, $T_{01} \sim Exp(\lambda)$ and the $0 \to 2$ transition, $T_{02} \sim Exp(\theta)$, with $p = \lambda\ (\lambda+\theta)$ ($Exp(\lambda)$ denotes the exponential distribution with rate parameter λ). Then the expected time until departure from state 0 is not $1\ (\lambda + \theta)$ as in a competing risks model. In this SFGM, the expected time until departure from state 0 is $pE[T_{01}] + (1-p)E[T_{02}] = 2\ (\lambda+\theta)$. Therefore, a necessary

condition for a SFGM to be Markov is that the transitions are all exponential, unlike in a true competing risks model where it is a necessary and sufficient condition.

Lemma 2.2.1 *A regular SFGM must have all exponentially distributed waiting times to be a Markov process.*

Proof: Consider the SFGM (in stochastic process notation) $X(t)$. Assume the waiting time from state i to state j is not exponentially distributed. Then $P(X(t+\varepsilon) = j|X(t) = i, X(s) = i) = P(X(t+\varepsilon) = j|X(t) = i)$ for all $0 < s < t$ and $\varepsilon > 0$ (since the exponential distribution is the only continuous distribution with the memoryless property). Therefore this SFGM, $X(t)$, is not Markovian. ∎

The fact is, many SFGMs with all exponentially distributed transitions are not Markov. If we again look at the last example the only possible way for this to be a Markov process is if $\theta = \lambda$. Then the expected time until departure is $pE[T_{01}] + (1-p)E[T_{02}] = 2\ (\lambda + \theta) = 1\ \theta$. This leads us to our next lemma.

Lemma 2.2.2 *A regular SFGM is a Markov process if and only if the random time to depart any node (with an exit branch) has a unique exponential distribution.*

Proof: (Sufficiency) If the random time to depart from any node i has the exponential distribution, then $P(X(t+\varepsilon) = j|X(t) = i, X(s) = i) = P(X(t+\varepsilon) = j|X(t) = i)$ for all $0 < s < t$, $\varepsilon > 0$, and any adjacent state j, therefore $X(t)$ is a Markov process.

(Necessity) If an SFGM is Markov then $P(X(t+\varepsilon) = j|X(t) = i, X(s) = i) = P(X(t+\varepsilon) = j|X(t) = i)$ for all $0 < s < t$ and $\varepsilon > 0$ and all branches have exponential distributions. If all nodes have less than two exit branches then we are done. For any node j with k exit branches (where $k \geq 2$), the probability of departing

to an adjacent state i during the interval $(s,t]$ is

$$p_i \frac{\int_s^t \sum_{l=1}^k p_l \theta_l \exp\ -\theta_l x \quad dx}{\int_s \quad \sum_{l=1}^k p_l \theta_l \exp\ -\theta_l x \quad dx}.$$

In general, this probability of departure changes as s varies. This violates the Markov property unless $\theta_1 = \theta_2 = \ldots = \theta_k$. Therefore we must have $\theta_1 = \theta_2 = \ldots = \theta_k$, which implies $\sum_{l=1}^k p_l \theta_l \exp\ -\theta_l t \ = \theta_1 \exp\ -\theta_1 t\ $, and we have a unique exponential distribution for the random departure time for each node with an exit branch. ∎

This lemma forces Markov SFGMs that have two or more branches leaving one node to have the same exponential distribution. Therefore we can easily construct Markov SFGMs by choosing the parameterization to meet these requirements.

2.2.4 Bayesian statistical flowgraph models

Bayesian SFGMs use the same principles as frequentist SFGMs. The primary difference is that the Bayesian framework provides a posterior predictive distribution (PPD) as the final result. This is very useful since prediction is the primary focus of SFGMs. As with any Bayesian parametric model, we begin by assigning parametric distributions to the transitions, then incorporating any *a priori* information into the prior distributions of the parameters. Once the model is parameterized with appropriate prior distributions, the posterior distribution is defined using Bayes' Theorem,

$$\pi(\theta|x) \quad f(x|\theta)\pi(\theta). \tag{2.4}$$

Letting x be a future observation, such that $x|\theta \quad x|\theta$, then the PPD $f(x|x)$ is

$$f(x|x) = \quad f(x|\theta)\pi(\theta|x)\ d\theta \qquad f(x|\theta)f(x|\theta)\pi(\theta)\ d\theta. \tag{2.5}$$

Except in simple situations, this integral is usually computed via Monte Carlo integration. Often in a Bayesian analysis we do not have a closed analytic form for

the posterior distribution. However, we can obtain samples from the posterior using Markov chain Monte Carlo (MCMC) techniques (see Marin and Robert (2007) or Givens and Hoeting (2005) for an introduction). With a sufficiently large sample from the posterior we can use it to approximate the PPD. Since $x|\theta \quad x|\theta$, the PPD can be written as

$$\begin{aligned} f(x|x) &= f(x,\theta|x)\ d\theta \\ &= f(x|\theta,x)\pi(\theta|x)\ d\theta \\ &= f(x|\theta)\pi(\theta|x)\ d\theta \\ &= E_{\theta\, x}\left[f(x|\theta)\right]. \end{aligned} \tag{2.6}$$

This characterization of the PPD gives us an approximate of the PPD using the n posterior samples (denoted as θ_i). The approximation is

$$f(x|x) \approx \frac{1}{n} \sum_{i=1}^{n} f(x|\theta_i). \tag{2.7}$$

Unfortunately, with this approximation we must find each $f(x|\theta_i)$ by inverting an MGF. If our posterior sample is too large, the computations become overwhelming. We address this further in Chapter 4.

This is a basic introduction of how to use the Bayesian methodology in SFGMs. We expand upon this introduction with additional examples in the following chapters.

2.3 Recurring illness process example

Recall the flowgraph of the recurring illness process in Figure 2.4. Assume we are researching a chronic illness which can be fatal. Once the illness is contracted it can go into remission, where an individual is considered healthy, but the symptoms will

eventually recur. We are interested in modeling the time until death of an individual who has contracted this disease. If the patient is ill, the time until remission or death is relatively short, however, the time until recurrence is usually much longer. Our simulated data consists of 20 patients.

In this chapter we conduct an initial analysis of the simulated data contained in Table 2.1. We do not include observations 1, 14, and 15 until we discuss incomplete data in Chapter 5.

Table 2.1: Simulated data from the process in Figure 2.4

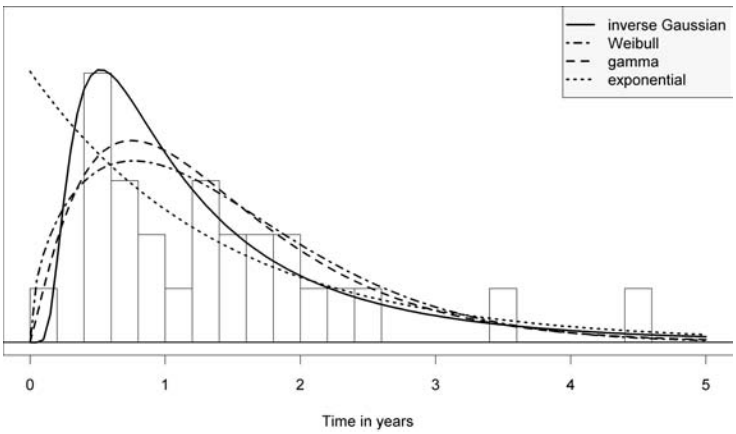
Obs	$0 \to 1$	$1 \to 0$	$0 \to 1$	$1 \to 0$	$0 \to 1$	$1 \to 2$	Total
1	0.53					3.04	3.58
2	1.20	0.02	0.14			3.16	4.52
3	0.43					2.59	3.03
4	4.93	0.01	1.05	0.04	0.90	2.24	9.16
5	1.39					1.80	3.19
6	0.44	0.03	0.65			3.78	4.90
7	1.63					4.61	6.23
8	2.09					2.52	4.61
9	1.78					2.12	3.90
10	0.74	0.01	0.49			3.85	5.09
11	4.53					2.63	7.16
12	1.48					2.94	4.42
13	0.54	0.01	1.41	0.02	1.28	2.43	5.69
14	0.11	0.01	0.28			5.59	5.99
15	3.08	0.01	0.89	0.01			
	1.21	0.02	1.28	0.05			
	1.61	0.01	0.89			1.87	10.94
16	0.96					2.84	3.79
17	0.98	0.02	1.08	0.01	3.50	3.87	9.47
18	2.57	0.01	0.66			3.74	6.98
19	2.27	0.03	1.91			2.31	6.53
20	1.81					2.73	4.54

Suppose we want to predict the time it takes an individual, who has contracted this disease, to transition from state 0 (good health) to state 2 (dead) (beginning at $t = 0$). It may be that the individual contracts the disease then goes into remission

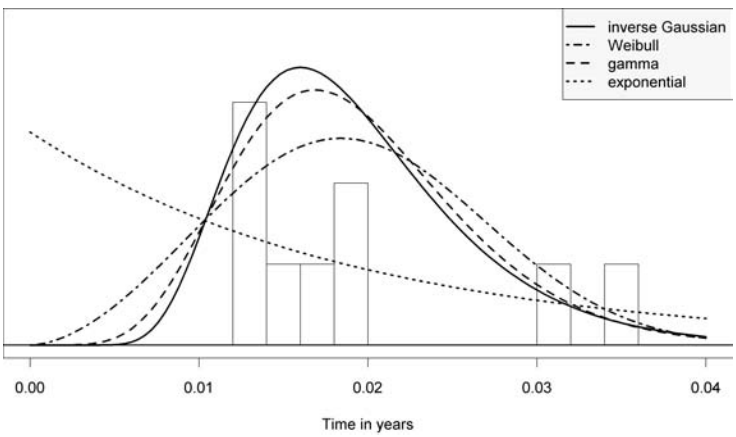
any number of times before death, as suggested by Figure 2.4.

After obtaining the data we must find appropriate parametric distributions for each of the three transitions, along with the probability of recovering from the disease. To find a good parametric fit, we plot a histogram of the data with some distributions at their maximum likelihood estimates (MLEs). We consider an exponential, gamma, inverse Gaussian, and the Weibull. The Weibull distribution only has an MGF if the shape parameter is greater than or equal to 1. In Figure 2.7 we have plotted the candidate distributions at their MLEs with a histogram of the transition data. The $0 \to 1$ transition can be seen in Figure 2.7(a). It appears that the $gamma(\alpha_{01}, \beta_{01})$ distribution provides the best fit. For the $1 \to 0$ transition we choose the $inverse\ Gaussian(\mu_{10}, \lambda_{10})$ as seen in Figure 2.7(b), and for the final transition, $1 \to 2$ we choose the $gamma(\alpha_{12}, \beta_{12})$ as seen in Figure 2.7(c). Therefore our model parameterization is

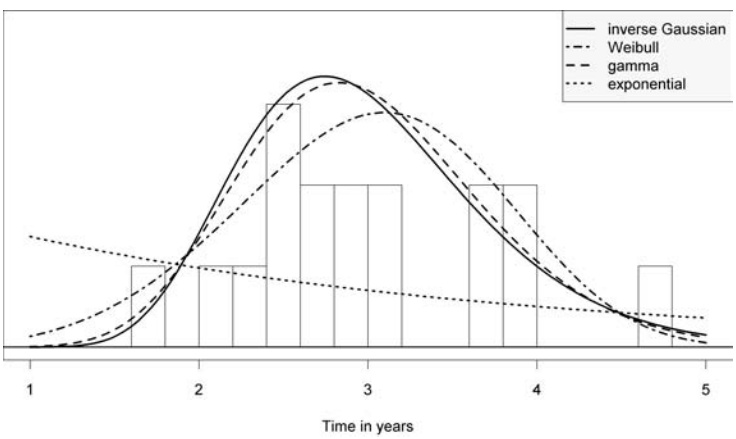
$$\begin{aligned} T_{01} &\sim gamma(\alpha_{01}, \beta_{01}), \\ T_{10} &\sim inverse\ Gaussian(\mu_{10}, \lambda_{10}), \text{ and} \\ T_{12} &\sim gamma(\alpha_{12}, \beta_{12}). \end{aligned} \tag{2.8}$$



(a)



(b)



(c)

Figure 2.7: Histograms of data from Table 2.1 with possible parameterizations, (a) shows the data and some possible fitted distributions for the $0 \rightarrow 1$ transition, (b) is the same, but using the data from the $1 \rightarrow 0$ transition, and (c) also shows the data and some possible fitted distributions, but for the $1 \rightarrow 2$ transition.

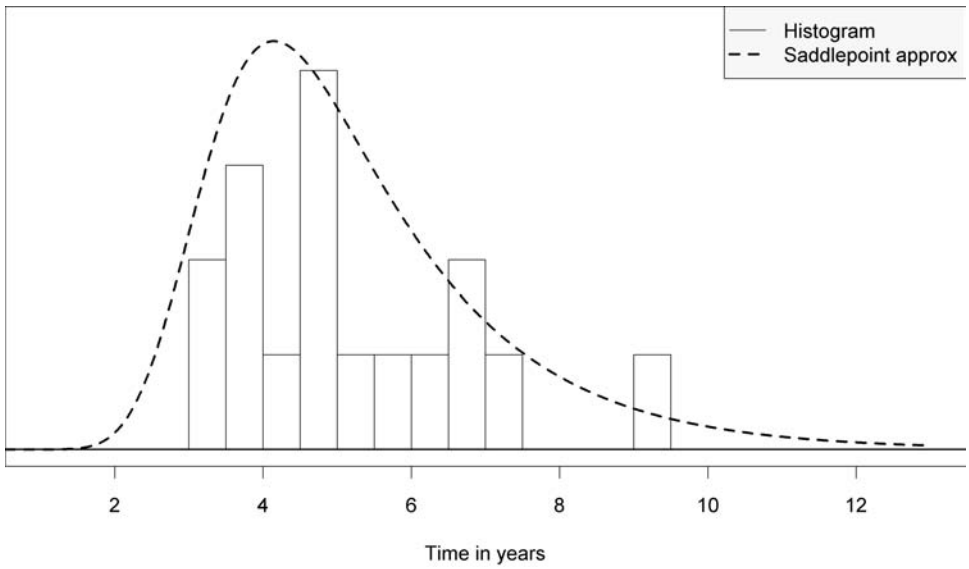


Figure 2.8: The saddlepoint approximation of the first passage distribution from state 0 to state 2 for the simulated recurring illness process.

To find the MGF of the first passage from state 0 to state 2, denoted $M_{02*}(s)$, we create the transition matrix T and find $(I-T)^{-1}$. The entry in the first row and last column of $(I-T)^{-1}$ contains this MGF, which we have previously found in Equation 2.2. The MGFs for the inverse Gaussian and gamma are in closed form, so we can find a closed form expression for $M_{02*}(s)$. After substituting our MLE estimates we get

$$M_{02*}(s) = \frac{(1-p)\left(1-\frac{s}{\hat{\beta}_{12}}\right)^{-\hat{\alpha}_{12}}}{\left(1-\frac{s}{\hat{\beta}_{01}}\right)^{\hat{\alpha}_{01}} - p\exp\left\{\frac{\hat{\lambda}_{10}}{\hat{\mu}_{10}} - \sqrt{\frac{\hat{\lambda}_{10}^2}{\hat{\mu}_{10}^2} - 2s\lambda_{10}}\right\}}.$$

From $M_{02*}(s)$ we find the cumulant generating function, $K_{02*}(s)$, and its first and second derivatives. Once we find s for each desired time point t, we have everything we need to invert $M_{02*}(s)$ to $f_{02*}(t)$ using the saddlepoint method. Figure 2.8 shows the approximated density overlaid on a histogram of the data. We can see that this

model appears to be reasonable given the data. In the following chapters we continue this example and show how we can improve this model with our new techniques. Now that we have provided background on the statistical flowgraph methodology, we introduce novel material and demonstrate why it is useful.

Chapter 3

Extending SFGMs to use any smooth time-to-event branch distributions

Parametric SFGMs have used the moment generating function (MGF) to represent the probability of transition between states. This has greatly limited the number of distributions available for SFGMs, since not all distributions have MGFs. Popular distributions such as the lognormal or certain Weibulls do not have MGFs and have not been used in SFGMs. By using complex Laplace transforms (LTs) in lieu of MGFs we can use all continuous and differentiable parametric distributions in SFGMs. Complex LTs are a generalization of MGFs and characteristic functions and exist for all lifetime distributions. In mathematics, the term "Laplace transform" includes both real and complex variables; however, this is usually not the case in statistics literature. Therefore, we use the term complex LT to avoid ambiguity. This chapter is organized as follows: First, we introduce complex LTs and an efficient method to invert (or transform) these into probability density functions (PDFs). Then we show how this inversion technique is implemented in some illustrative examples and a real

data example with interval-censoring. Finally, we return to the simulated recurring illness process example.

Figure 3.1 represents a SFGM for a portion of a construction engineering application discussed in Huzurbazar (2005b). The states 0, 1, 2, and 3 are outcomes that represent stages of a construction engineering project for a given month. The branches are labeled with transition probabilities and complex LTs of waiting time distributions. The original analysis of these data in Huzurbazar (2005b) was constrained to the use of distributions with MGFs. We will return to this example to demonstrate the extension of SFGMs with any smooth distribution.

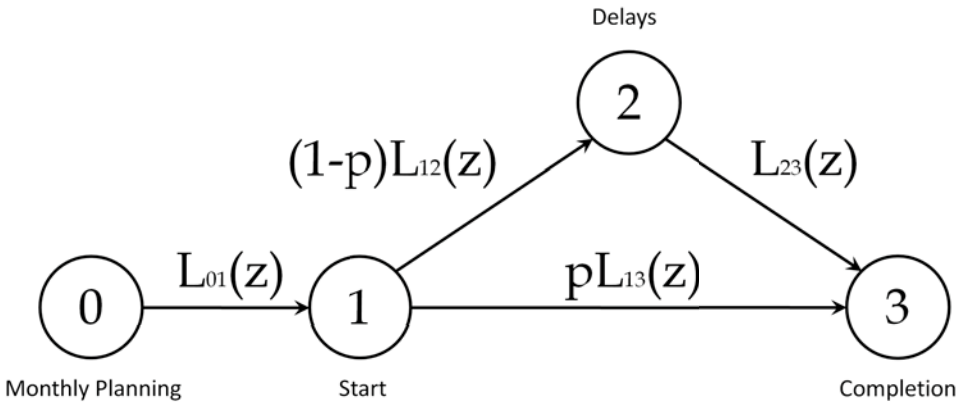


Figure 3.1: A SFGM for construction engineering data.

3.1 Transforming complex LTs to PDFs

As mentioned in the previous chapter, the most common method for transforming the overall MGF of the SFGM into a PDF is the saddlepoint method. The saddlepoint method transforms the MGF to an approximate density. This method is cumbersome when solving complicated flowgraphs and requires the use of a symbolic algebra software package. It cannot handle distributions that do not have MGFs and also

fails in multimodal situations, as seen in Figure 2.6. Another possible method of inversion is the EULER method, developed by Abate and Whitt (1992). This method has been used in past SFGM literature but only with MGFs that have a closed analytic form. It is fast and accurate, but similar to the saddlepoint method, it does not have a means of bounding the error of the approximation. However, if used for smooth PDFs the EULER method provides an approximate error bound that can be controlled and reduced if necessary. In this chapter, we implement an additional capability of the EULER method to transform complex LTs to PDFs. Complex LTs exist for all lifetime distributions and can be used in the EULER method even if they do not exist in a closed analytic form. This greatly extends the parametric flexibility of SFGMs.

The Laplace transform, as defined in statistics, is a simplification of the LT as defined in mathematics. In mathematics, the Laplace transform, $L(z)$, is defined for all real and complex z, while in statistics it is usually only defined for real z. A complex LT for a positive random variable T, defined on $[0, \quad)$ with $z = x + iy$ and $x \geq 0$, is

$$L_T(z) = \int_0 e^{-zt} f_T(t) dt. \tag{3.1}$$

The complex LT is a generalization of the characteristic function. If we transform the variable z to let $x = 0$ and $u = -y$ then $L_T(u)$ is the characteristic function for the random variable T. The complex LT is $E[\exp\ -x - iy\ ]$, whereas the MGF is $E[\exp\ x\ ]$ and the characteristic function is $E[\exp\ iy\ ]$.

If the MGF has a closed analytic form, then so does the complex LT, and finding the complex LT for a random variable is a transformation similar to finding the MGF. However, for distributions that do not have closed form MGFs, such as some of the Weibull family, we must find the complex LT or MGF by numerical integration. Once we have all the complex LTs for the SFGM, we can use it to obtain the complex LT for the overall SFGM. For example in Figure 2.1, if we use complex LTs in lieu

of MGFs, we would have $L_{02*}(z) = L_{01}(z)L_{12}(z)$. Using the complex LT does not change any theory we introduced in Chapter 2. Mason's rule still applies, all we do is replace the MGFs with their associated complex LTs.

Given that we have a first passage complex LT, $L_{ij*}(z)$, for a SGFM, we can use the EULER method to obtain $f_{ij*}(t)$.

3.1.1 Overview of the EULER method

The EULER method uses the Bromwich integral and Euler summation (and hence its name). For an applied introduction to this method that includes code for the algorithm, refer to Abate and Whitt (1995). The Bromwich contour inversion integral is

$$f(t) = \frac{1}{2\pi i} \int_{a-i}^{a+i} e^{zt} L(z)\, dz,$$

where $i = \sqrt{-1}$, $L(z)$ is the complex LT, and the contour is any vertical line $z = a$ such that $L(a)$ has no singularities on or to the right of it. With a change of variable and some manipulation, $f(t)$ can be rewritten as

$$f(t) = \frac{2e^{at}}{\pi} \int_0 Re\left[L(a+iu)\right]\cos(ut)\, du,$$

where $Re[z]$ is the real part of a complex number. Using the trapezoidal rule with step size π $(2t)$ and letting $a = A$ $(2t)$ gives the approximation

$$f(t) \approx \frac{e^{A\ 2}}{2t} Re\left[L\left(\frac{A}{2t}\right)\right] + \frac{e^{A\ 2}}{t} \sum_{k=1} (-1)^k Re\left[L\left(\frac{A+2k\pi i}{2t}\right)\right].$$

This is a nearly alternating series so Abate and Whitt (1992) use Euler summation as an acceleration method. Combining these, we have our approximation

$$f(t) \approx \sum_{j=0}^{m} \frac{e^{A\ 2}}{t} \frac{2^{-m} m!}{j!(m-j)!} \left[\frac{Re\left[L\left(\frac{A}{2t}\right)\right]}{2} + \sum_{k=1}^{n+j} (-1)^k Re\left[L\left(\frac{A+2k\pi i}{2t}\right)\right] \right],$$

or,

$$f(t) \approx \frac{e^{A\ 2}}{t} \sum_{k=0}^{m+n} (-1)^k w_k Re\left[L\left(\frac{A + 2k\pi i}{2t}\right)\right], \tag{3.2}$$

where $w_0 = 1\ \ 2$, $w_k = 1$ for $k = 1 \ldots n$, and

$$w_k = w_{k-1} - \frac{2^{-m} m!}{(k-n-1)!(m+n+1-k)!} \text{ for } k = n+1 \ldots n+m.$$

Abate and Whitt (1995) recommend setting $m = 11$, $n = 15$, and increasing n if better accuracy is required. This approximation contains two different errors. First is the error introduced by the trapezoidal approximation, and the second by the truncated sum and Euler acceleration. Abate and Whitt (1995) show how to bound the first type of error by choosing A (often they choose $A = 18.4$). However, the error introduced by the truncated sum and Euler acceleration cannot be bounded, only estimated.

3.1.2 Using the EULER method

For $L(z)$, the EULER method restricts $z = x + iy$ to have non-negative real part (i.e., $Re(z) = x \geq 0$). The transformation formula of complex LTs depends only on the real portion of $L(z)$, which we denote as $Re[L(z)]$. For example, in the flowgraph of Figure 2.1, we can find the real portion of $L_{02*}(z)$ by substituting $z = x + iy$ in

(3.1) and obtain $Re[L_{02*}(z)]$ as follows:

$$\begin{aligned} Re\left[L_{02*}(z)\right] &= Re\left[L_{02*}(x,y)\right] = Re\left[L_{01}(x,y)L_{12}(x,y)\right] \\ &= Re\left[\int_0^{} e^{-(x+iy)u} f_{T_{01}}(u)\, du \int_0^{} e^{-(x+iy)v} f_{T_{12}}(v)\, dv\right] \\ &= Re\left[\int_0^{} e^{-xu}e^{-iyu} f_{T_{01}}(u)\, du \int_0^{} e^{-xv}e^{-iyv} f_{T_{12}}(v)\, dv\right] \qquad (3.3) \\ &= \left(\int_0^{} e^{-xu}\cos(yu) f_{T_{01}}(u)\, du\right)\left(\int_0^{} e^{-xv}\cos(yv) f_{T_{12}}(v)\, dv\right) + \\ &\quad \left(\int_0^{} e^{-xu}\sin(yu) f_{T_{01}}(u)\, du\right)\left(\int_0^{} e^{-xv}\sin(yv) f_{T_{12}}(v)\, dv\right). \end{aligned}$$

We can apply the EULER method to transform $Re[L_{02*}(z)]$, in (3.3), to an approximate $f_{T_{02*}}(t)$. In practice, it is usually easier to compute each complex LT, include it in the first passage complex LT formula, after which we find the real part, and then transform it into a density using the EULER algorithm. In this way, we do not need to concern ourselves with the many cross-products of the imaginary parts which become real.

We have mentioned that the complex LT exists for all distributions with support $[0, \quad)$. The following argument demonstrates this.

Proposition 2.1 *$L(z)$ is finite for any density $f_T(t)$ such that $0 \le t < \quad$, $z = x+iy$ and $x \ge 0$.*

Proof:

$$L(z) = \int_0^{} e^{-(x+iy)t} f_T(t)\, dt = \int_0^{} e^{-xt}\cos(yt) f_T(t)\, dt - i \int_0^{} e^{-xt}\sin(yt) f_T(t)\, dt$$

However, $|e^{-xt}\cos(yt)| \le 1$ and $|e^{-xt}\sin(yt)| \le 1$ therefore $|L(z)| \le 2$. ∎

By Proposition 2.1, any random variable with a density defined on $[0, \quad)$ has a complex LT. We have also shown that the complex LT is a generalization of the characteristic function. Since a characteristic function uniquely identifies a density,

this guarantees the same for complex LTs. The EULER method requires that the densities involved be continuous and differentiable. Thus by using complex LTs in the place of MGFs on the branches of a SFGM, we can use any relatively smooth lifetime distribution in statistical flowgraph modeling. This is a big step forward for SFGMs. A drawback of the EULER method is that it cannot guarantee error bounds. In addition, it is only defined for random variables with support [0,). However, this is not problematic when modeling semi-Markov processes that begin at time $t = 0$.

Since the EULER transformation method cannot quantify the error bound, we must ensure that we are getting the right answer. One rudimentary check for reasonable results is to compare it with a histogram of a Monte Carlo sample from the model. If our approximation closely matches this histogram, we can be confident that the amount of error in the transformation is acceptable. Brute force Monte Carlo simulation of a network can be performed when information on network transitions and waiting times are completely known, but if this is not the case, Monte Carlo simulation may not be possible. In the next sections, we validate our calculations with Monte Carlo simulations.

3.2 Illustrative examples

We demonstrate how the EULER method can be applied to a variety of SFGMs using two examples. The first is a series system with two non-identical Weibull waiting times; this is basically the convolution of two Weibull random variables. The other example is also a series system but with three waiting time distributions that do not have MGFs.

3.2.1 A series system with two Weibull waiting times

We continue with our simple flowgraph model example from Figure 2.1. Let $T_{01} \sim Weibull(0.5, 1)$ and $T_{12} \sim Weibull(1.9, 2.2)$. We know that $M_{01}(s) =$ when $s > 0$, therefore the MGF does not exist and we are unable to use the saddlepoint method. However, the EULER method is well suited for this situation. We can calculate $L_{02*}(s)$ with numerical integration and then invert it to a PDF. Figure 3.2 shows the approximate density using the EULER method along with a histogram of a sample of 1,000,000 points from $f_{02*}(t)$. This figure indicates that the EULER method performs well in approximating the PDF $f_{02*}(t)$.

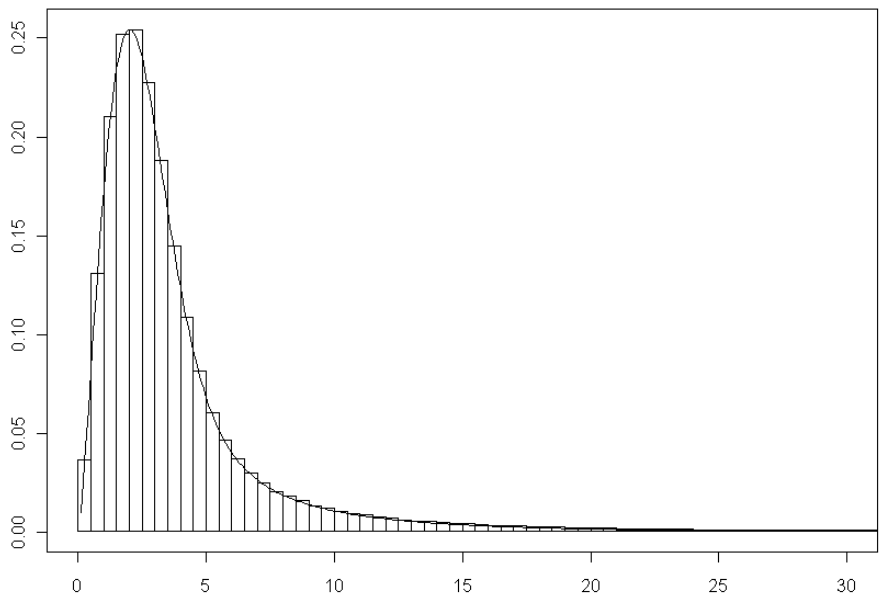


Figure 3.2: The line represents the approximate convolution of a Weibull(0.5,1) and Weibull(1.9,2.2) and the histogram is a Monte Carlo estimate of 1,000,000 samples of the same distribution.

3.2.2 Production repair model

This example illustrates the implementation of the EULER method in a flowgraph specifically choosing a few distributions that do not have MGFs.

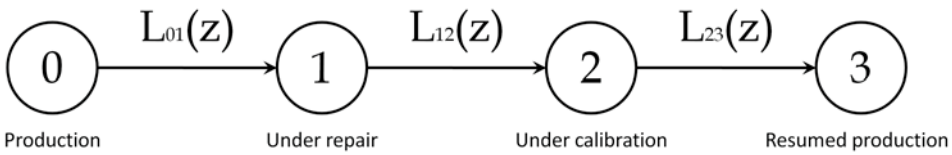


Figure 3.3: The SFGM of a machine with four states.

Consider a machine at a manufacturing plant. The machine is set into production at time $t = 0$ and functions until an unusually large part is introduced, which causes failure. The machine must be repaired and calibrated before it can be put back into production. Figure 3.3 represents the possible states of the machine where state 0 is "production" (the machine is working), state 1 is "under repair", state 2 is "under calibration", and state 3 is "resumed production". We are interested in finding the distribution of the time it takes to start at state 0, proceed through states 1 and 2, and arrive at state 3. We assume that the time-to-failure of the machine should be modeled as an extreme value type distribution, specifically the Fréchet. The PDF of a Fréchet distribution is

$$f(t|\xi,\sigma) = \frac{\xi t^{-(\xi+1)}}{\sigma} \exp\left\{-\frac{1}{\sigma}t^{-\xi}\right\}.$$

Additionally, repair times are commonly modeled by the lognormal distribution, so we will model the machine repair time as such. Finally, we model the calibration time with a Weibull distribution, which is a common distribution used in reliability. Let T_{ij} be the random variable that represents the transition density from state i to state j. Assume that $T_{01} \sim Fréchet(0.1, 3.5)$, $T_{12} \sim lognormal(0, 1)$, and $T_{23} \sim Weibull(0.9, 1.5)$. None of these particular distributions have MGFs because

$M_{01}(s) =$, $M_{12}(s) =$, and $M_{23}(s) =$ if $s > 0$. However, the EULER method can find an approximation for $f_{03*}(t)$. Letting $z = x + iy$, we obtain

$$L_{ij}(z) = E[e^{-(x+iy)T_{ij}}] = \int_0 e^{-xt}\cos(yt) f_{T_{ij}}(t)\,dt - i\int_0 e^{-xt}\sin(yt) f_{T_{ij}}(t)\,dt. \quad (3.4)$$

We find the three complex LTs numerically, because they have no simple closed-form expression. We then take the real part of $L_{03*}(z) = L_{01}(z)L_{12}(z)L_{23}(z)$ and use the EULER method to transform it into a PDF.

Figure 3.4 shows the EULER approximated PDF $f_{03*}(t)$ along with a histogram of a sample of 1,000,000 points from $f_{03*}(t)$. From the plot we can see that the approximation is very good.

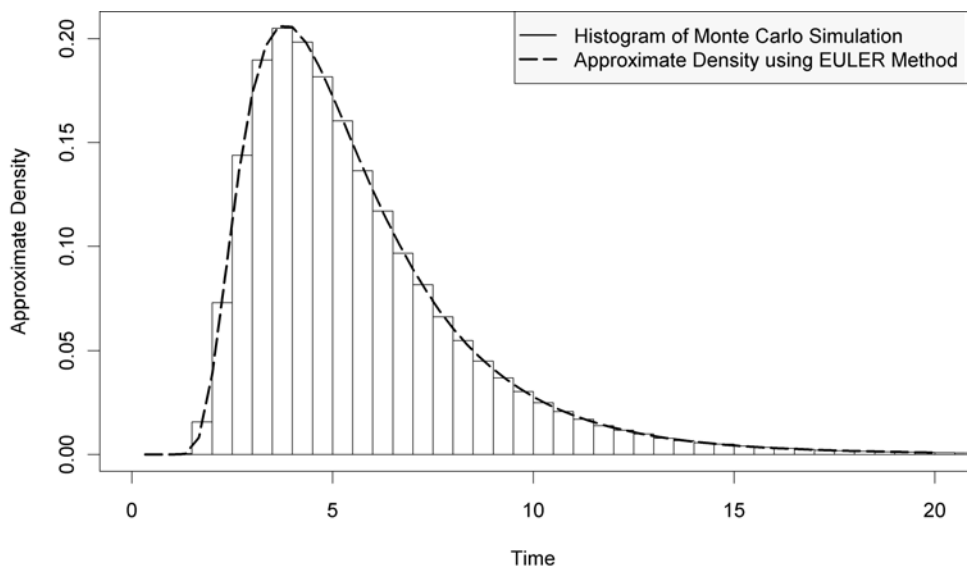


Figure 3.4: Comparison of the EULER approximated convolution of a Fréchet, log-normal, and Weibull distribution over the histogram of a Monte Carlo sample of the same distribution.

Table 3.1: Construction engineering transition times (in days)

0→1	4.5	4.0	3.5	3.5	4.0	4.5	4.0	4.0	3.5	4.0	4.5	4.5	4.0	4.0
	4.0	3.5	4.0	4.5	4.0	4.0								
1→2	2	5	5	5	3	3	15	5	6	20	34	28	5	5
1→3	5	40	15	5	25	20								
2→3	21	18	18	18	20	20	18	17	18	18				

3.3 Construction engineering application

3.3.1 Non-censored data

We demonstrate the EULER method using data from construction engineering projects considered in Huzurbazar (2005b). We focus on the monthly planning phase of this process, as shown in the flowgraph of Figure 3.1. State 0 represents "Monthly Planning", and state 1 is the "Start" of the project. From state 1 we transition to state 2 if there are "Delays", and go to state 3 when the project is "Complete". The data are reproduced in Table 3.1 from Huzurbazar (2005c, pp. 184).

Let Y_1 be the waiting time from state 0 to state 1, Y_2 be the waiting time from state 1 to state 2, Y_3 be the waiting time from state 1 to state 3, Y_4 be the waiting time from state 2 to state 3, and X the random variable where $X = 1$ if the process goes from state 1 directly to state 3, $X = 0$ otherwise. Let $Y_i \sim Weibull(\nu_i, \ _i)$ for $i = 1 \ldots 4$ and $X \sim Bernoulli(p)$ and n_i be the number of *iid* observations we have for each Y_i. The analysis of these data in Huzurbazar (2005b) parameterized the SFGM differently, with a point mass at 4 for the state 0 to state 1 transition, a gamma for the state 1 to state 2 transition, a point mass at 18 for the state 2 to state 3 transition, and an inverse gamma for the state 1 to state 3 transition. We use the current parameterization to demonstrate the powerful capability of the EULER method.

Table 3.2: MLEs of the construction engineering data

ν_1	$_1$	ν_2	$_2$	ν_3	$_3$	ν_4	$_4$	p
13.438	4.179	1.142	10.638	1.538	20.416	15.003	19.196	0.300

The maximum likelihood estimates (MLEs) of the data are given in Table 3.2. The likelihood function for this SFGM is

$$L(\theta|x, y_i) = p^6(1-p)^{14} \prod_{i=1}^{4} \prod_{j=1}^{n_i} \frac{\nu_i}{\frac{\nu_i}{i}} y_{ij}^{\nu_i - 1} \exp\left\{ -\left(\frac{y_{ij}}{i} \right)^{\nu_i} \right\} \quad , \tag{3.5}$$

where θ is the generic vector representing the parameters in the model.

For comparison, we calculate the PDF using the saddlepoint approximation and then the EULER method. We use a Bayesian approach using flat priors and find the posterior predictive density (PPD) for the first passage time from state 0 to state 3.

Even though the SFGM for this example is fairly straightforward, the saddlepoint approximation is complicated to program. We must adapt the formulas to keep the numerical calculations within R's capacity. The default integration routine in R is not able to handle many of the integrals, therefore we use an LAPACK routine DGAUS8 (see Anderson et al. (1999)) and transform the integrals to be on the support $(0, 1]$ as opposed to $[0, \quad)$. This method provides a solution, but Figure 3.5 suggests that the saddlepoint approximation does not perform well in this situation. The saddlepoint approximation tends to smooth out the multimodality of the density and truncates the right tail.

For comparison, we use the EULER method on the same data with the MLEs. The R code for the EULER method is only slightly more complicated from the illustrative example. In fact, the computations and formulas are still straightforward. We find $Re(L_{03*}(z)) = Re\left[pL_{01}(z)L_{13}(z) + (1-p)L_{01}(z)L_{12}(z)L_{23}(z)\right]$ using numerical integration and the EULER algorithm to transform $Re(L_{03*}(z))$ into a PDF. Referring again to Figure 3.5, we can see how the approximation of the two meth-

ods differ. The Monte Carlo simulation indicates the EULER method is much more accurate than the saddlepoint method.

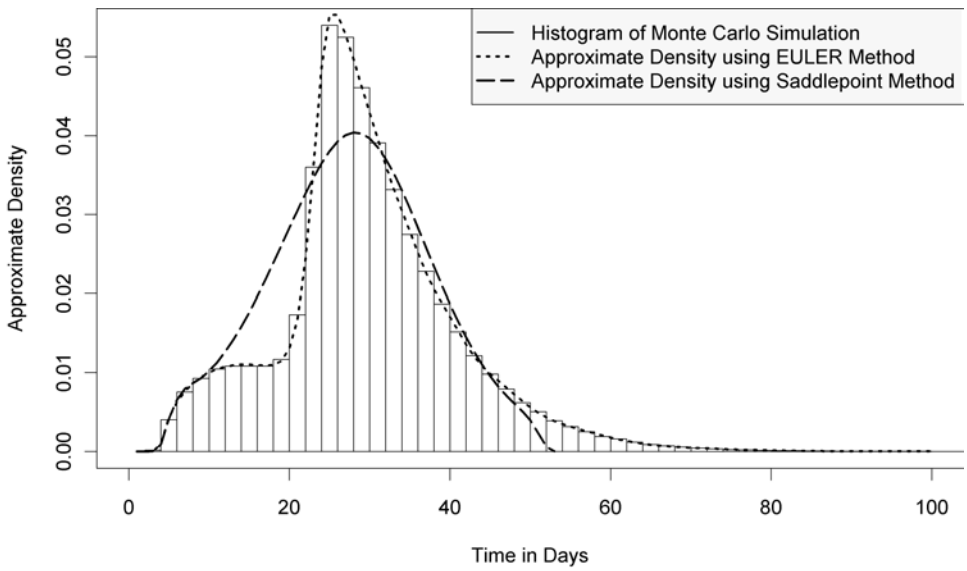


Figure 3.5: A comparison of the EULER approximation, saddlepoint approximation, and a Monte Carlo sample of the first passage distribution from state 0 to state 3 for the construction engineering example.

It is more computationally intensive to perform a Bayesian analysis on these data. For convenience only, we use flat priors to make our results fairly comparable to the frequentist approach. A benefit of the Bayesian method is that it provides a predictive density. With flat priors, we expect similar results to those from the EULER method, but with larger variance, since we will be obtaining a predictive density. Prediction is an important goal in complex systems such as these stochastic networks. In reality, all the parameters we consider are nuisance parameters. What we really want to know is, given another engineering project, what is the probability it will be finished within a certain amount of time. This information is readily available from the PPD and all our predictive inference can be based on it.

We use Gibbs sampling for p and include a Metropolis (random-walk) step for

the parameters ν_1, $_1, \nu_2$, $_2, \nu_3$, $_3, \nu_4$, and $_4$. A convenient place to initialize the Markov chain Monte Carlo (MCMC) is at the MLEs.

To find the PPD, we run the Gibbs sampler and save the posterior samples. We take each sample from the posterior, use them in the first passage complex LT, and transform it using the EULER method to get a density. Then the PPD is the average of all the densities we obtained from our posterior samples (Huzurbazar (2005c, pp. 90)). This is very time consuming if the posterior sample is even moderately sized. An additional complexity of the Bayesian approach is that we must ensure our Markov chains have converged and that they have explored the posterior space adequately. In-depth details of MCMC techniques can be found in Robert and Casella (2004).

After running our MCMC on these data, the convergence diagnostics are acceptable. We have suitable mixing, with Metropolis acceptance rates around 40% and proper decay in the autocorrelation of the Markov chains. For a more formal check we use the Heidelberger-Welch diagnostic included in the **boa** package for R (see Smith (2005) for more information). This diagnostic recommends discarding the first 1,100 samples for burn-in. After removing the first 1,100 samples, we have 9,900 samples from which we obtain the PPD. The PPD is shown in Figure 3.6. As expected, the EULER method and Bayesian results are quite similar with the Bayesian having slightly higher variance, due to the fact that it is a predictive density.

3.3.2 Interval-censored data

A powerful result is that this method can handle interval-censored data. In survival analysis and reliability many data sets are censored, so for any technique to be generally effective it must be able to cope with censored data. Time-to-event censored data can be defined as introducing uncertainty regarding when the event of interest

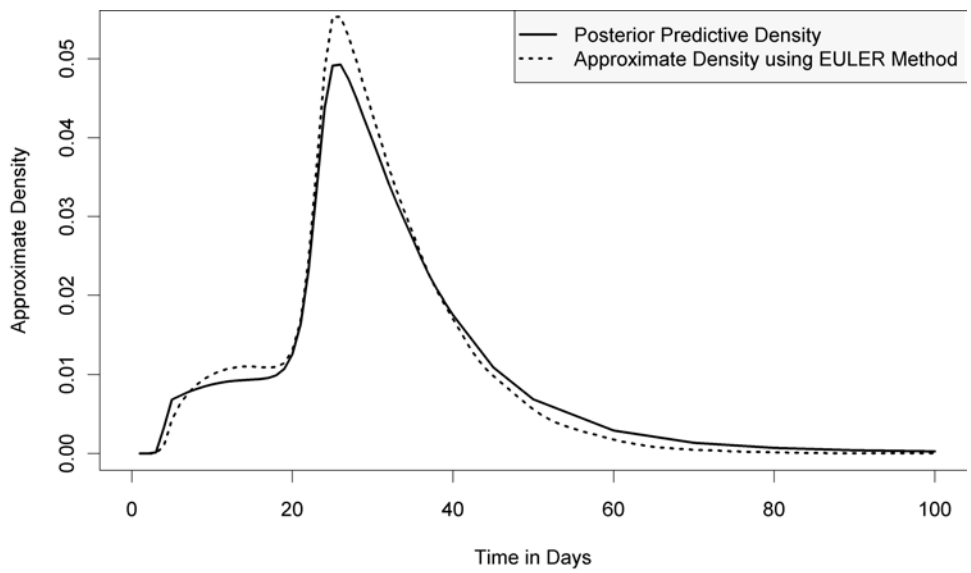


Figure 3.6: A comparison of the EULER estimated density and the posterior predictive density with flat priors of the first passage distribution from state 0 to state 3 for the construction engineering example.

occurred. Therefore, *interval-censoring* can be defined as only knowing the event occurred in some known interval $[t_0, t_1)$, *right-censoring* is knowing the event did not occur before some known time t, and *left-censoring* is knowing that the event occurred before some known time t. Clearly, right- or left-censoring is a special case of interval-censoring when $t_1 =$ or $t_0 = 0$ respectively.

Wolstenholme (1999, pp. 39) argues that all time-to-event measurement data is interval-censored due to the rounding accuracy of nearly all measurements. Since our data, in this example, are all whole or half integers, we know the data up to the nearest 1 2 or 1 4 day. If we have a transition that took 12 days, we know the true transition time is in $[11.5, 12.5)$ or if it is 4.5 days then our known time can be found in $[4.25, 4.75)$. Now define $c_1 = 0.25$ and $c_i = 0.5$ for $i = 2, 3, 4$, which are half the width of our censoring intervals.

Given the assumption of interval-censored data, the likelihood function is now

$$L(\theta|x, y_i) = p^6(1-p)^{14} \prod_{i=1}^{4} \prod_{j=1}^{n_i} \left(\exp\left\{ -\left(\frac{y_{ij} - c_i}{i} \right)^{\nu_i} \right\} - \exp\left\{ -\left(\frac{y_{ij} + c_i}{i} \right)^{\nu_i} \right\} \right) . \tag{3.6}$$

Observing the likelihood in (3.6), it is fairly intuitive for the special case of right-censoring, the last term in the product is 0, or if left-censored the first term in the product is 1. Clearly, if we can handle interval-censoring, right- or left-censoring uses the exact same machinery.

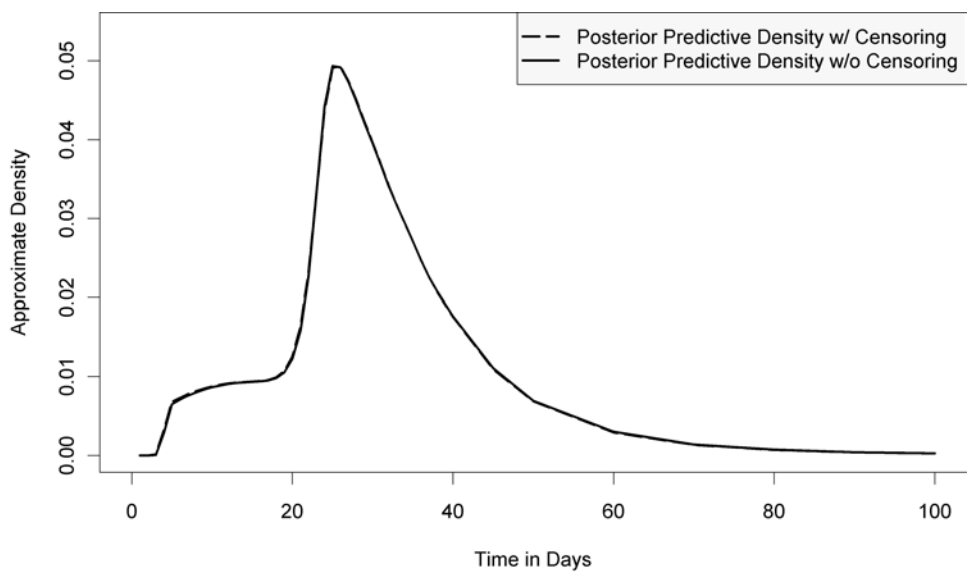


Figure 3.7: A comparison of the two Bayesian posterior predictive densities of the first passage time from state 0 to state 3 for the construction engineering example. One PPD is found assuming interval-censored data, the other assumes transition times are completely known.

Again, our MCMC mixing is acceptable with Metropolis acceptance rates at approximately 40%, and the autocorrelations of the chains decay quickly. The

Heidelberger-Welch diagnostic recommends not discarding any samples for burn-in. However, we discard the first 1,000 because we were tuning the MCMC during those iterations. So we are left with 10,000 samples from the posterior to obtain the PPD. With the small amount of censoring we introduced, we would not expect to see a large change in the PPD. Figure 3.7 confirms this, and we see our results are barely distinguishable from the Bayesian results without censoring.

3.4 Simulated recurring illness process example (continued)

In this chapter we have extended SFGMs to allow the use of any smooth distribution for branch modeling. We have done this by using the complex LT and the EULER method for its transformation into a PDF. This allows much richer modeling possibilities in SFGMs. Recall the recurring illness process example from Chapter 2. We were constrained to use parametric distributions that had MGFs; we are no longer under that limitation. The only limitation we now have is to use distributions that are continuous and differentiable, which is a very large class of distributions.

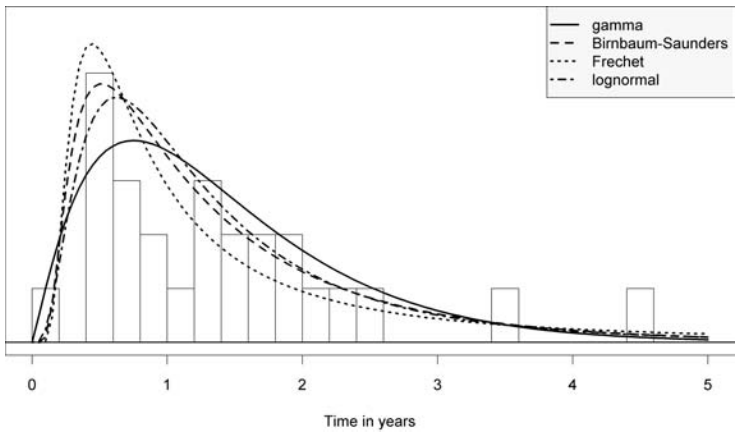
We return again to our simulated example in Chapter 2. Now that we have a larger selection of distributions to choose from, we reconsider finding a suitable fit for each transition distribution. The selected distributions were gammas for the $0 \to 1$ and $1 \to 2$ transitions and an inverse Gaussian for the $1 \to 0$ transition. We compare these with a few other common lifetime distributions, the lognormal, Fréchet, and the Birnbaum-Saunders (see Birnbaum and Saunders (1969)). The Birnbaum-Saunders distribution is an adaption of Miner's rule (Miner (1945)), to model fatigue cracking. The PDF of a Birnbaum-Saunders distribution is

$$f(t|\alpha,\lambda) = \frac{\overline{\lambda t}+1 \quad \overline{\lambda t}}{\overline{2\pi}2\alpha t} \exp\left\{-\frac{1}{2\alpha^2}\left(\quad \overline{\lambda t}-1 \quad \overline{\lambda t}\right)^2\right\}.$$

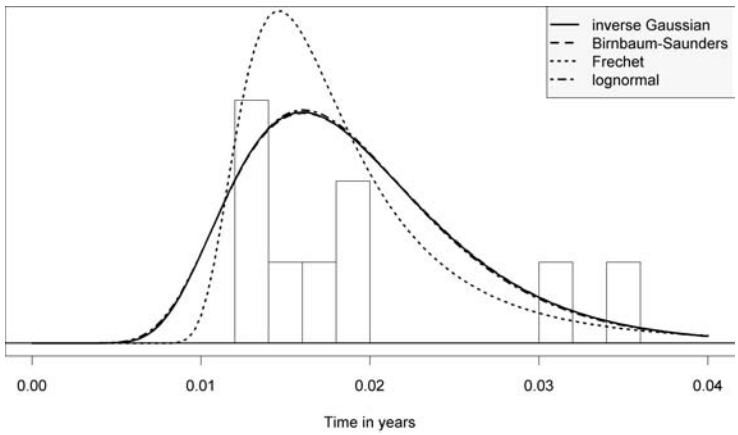
Previously, we could not use the lognormal or the Fréchet because neither of these have a MGF that exists, however, the Birnbaum-Saunders distribution does have an MGF.

We again find the MLEs of these distributions and compare these with a histogram of the data. Figure 3.8 shows each transition. Deciding which would be most appropriate is subjective and each of these distributions have theoretical properties that might make them more appealing even if the data do not support them quite as much as another. For now, we will put these issues aside and look strictly at the data. For the $0 \to 1$ transition in Figure 3.8(a) the Birnbaum-Saunders seems to fit a little better than the gamma. The $1 \to 0$ transition in Figure 3.8(b) is a challenge, but the Fréchet captures the mode well, and is adequate in the tails. Three of the four distributions in the $1 \to 2$ transition in Figure 3.8(c) look reasonable, the best being the Birnbaum-Saunders and the lognormal distributions. Both these distributions look almost identical, so we choose the lognormal for some variety. Now our model parameterization is

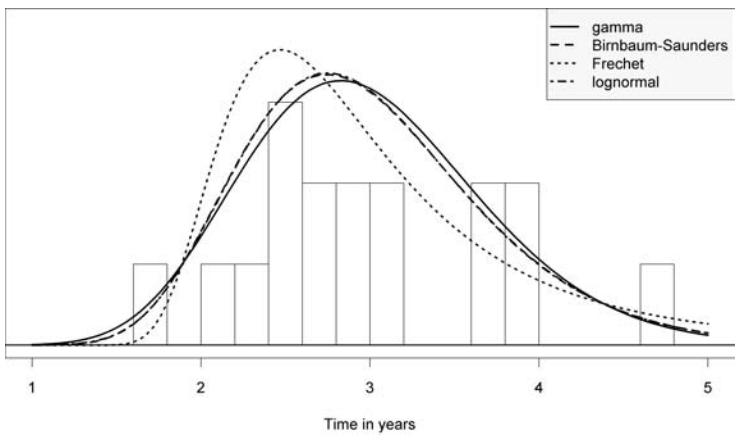
$$\begin{aligned} T_{01} &\sim \textit{Birnbaum-Saunders}(\alpha_{01}, \lambda_{01}), \\ T_{10} &\sim \textit{Fréchet}(\xi_{10}, \sigma_{10}), \text{ and} \\ T_{12} &\sim \textit{lognormal}(\mu_{12}, \sigma_{12}). \end{aligned} \tag{3.7}$$



(a)



(b)



(c)

Figure 3.8: Histograms of data from Table 2.1 with possible parameterizations, (a) shows the data and some possible fitted distributions for the $0 \rightarrow 1$ transition, (b) is the same, but using the data from the $1 \rightarrow 0$ transition, and (c) also shows the data and some possible fitted distributions, but for the $1 \rightarrow 2$ transition.

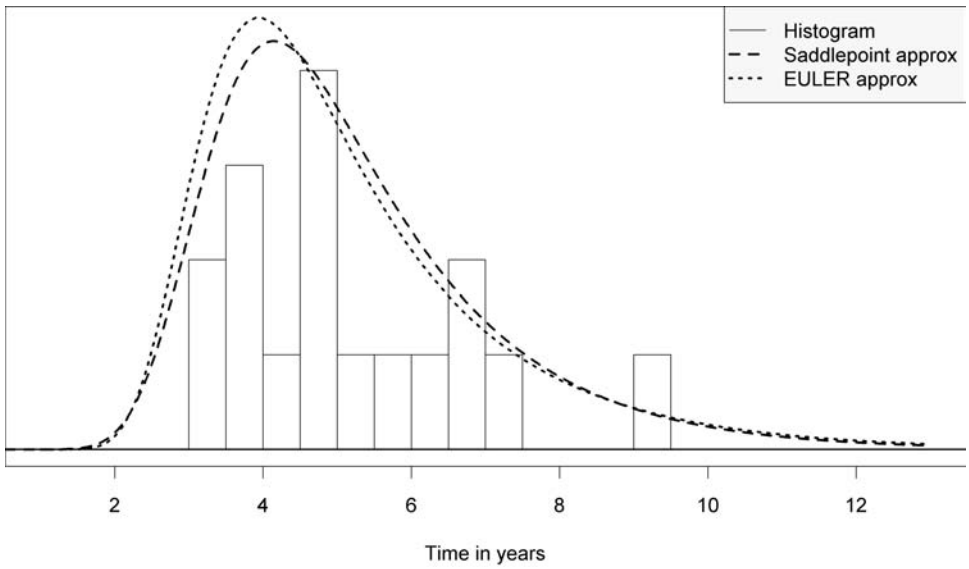


Figure 3.9: A histogram of the first passage data from state 0 to state 2 for the simulated recurring illness process, along with plots of the models from (2.8), estimated with the saddlepoint approximation, and (3.7), estimated with the EULER method.

We again use Mason's rule to find the first passage complex LT which is

$$L_{02*}(z) = \frac{(1-p)L_{01}(z)L_{12}(z)}{1-pL_{01}(z)L_{10}(z)}. \tag{3.8}$$

Since we have selected the distributions for each transition, we must find their complex LTs by numerical integration and then use the EULER method to invert $L_{02*}(z)$ to $f_{02*}(t)$. Figure 3.9 shows a histogram of the data with the two models found in (2.8) and (3.7). Even though for each transition it had appeared we have improved our fit, the fit for first passage time appears to have become worse. This leads us to question, if we find the optimal models for each transition will the combination of these provide the optimal model for the first passage in a SFGM? How should we determine which one of these models is better, and is the one we select satisfactory? It is difficult to determine these questions just by looking at a histogram of the data and a plot of the modeled transition. In Chapter 6 we introduce a quantitative measure to assist in choosing appropriate transition distributions and assessing if our

model is an adequate representation of the data.

Chapter 4

New methods and models in Bayesian SFGMs

SFGMs can be used in either a Bayesian or frequentist framework, and in a parametric or non-parametric setting. Up to this point we have only briefly discussed Bayesian SFGMs. The Bayesian methodology naturally lends itself to prediction; using predictive distributions are very convenient. In this chapter we describe some advances in SFGMs using the Bayesian framework. First, we demonstrate a few ways to calculate the posterior predictive density of a first passage time. Then, we use one method in a Bayesian non-parametric model and use it again by introducing time-varying covariates in an flowgraph of an accelerated failure time model.

Techniques to handle Bayesian SFGMs are computationally intensive. SFGM literature suggests a way to calculate the posterior predictive density (PPD), but it is quite time consuming. The current state-of-the-art method to find the PPD uses the posterior sample. Then, for each of these fixed sample values, finds a density using flowgraph algebra and numerical inversion. The PPD is the average of all these densities we have calculated over the posterior sample. If the posterior sample

is large, this requires an inordinate amount of computation. We present a more efficient way to estimate the first passage PPD in a SFGM. This method makes Bayesian SFGMs faster computationally and allows for additional flexibility in our modeling choices.

4.1 Improved methodology for calculating the posterior predictive density

This chapter radically alters the way we think about the Bayesian SFGM framework by suggesting an alternative method for computing the PPD. Our approach to Bayesian SFGMs reduces the amount of computation required, and makes Bayesian SFGMs an attractive modeling choice.

We review three common methods to estimate the PPD in a Bayesian analysis, identify their strengths and weaknesses, and recommend one that works best with SFGMs. The first way to obtain a PPD is to calculate it exactly with an analytic solution. Obviously if this was always possible it would be the preferred method. In some simple parameterizations we can analytically find the PPDs of each transition, and then use Mason's rule and numerical methods to obtain the PPD for a first passage time. In this case, our computation time would be equivalent to what we would expect from a frequentist analysis of a SFGM. However, the Bayesian analysis has a major advantage over the frequentist method because we have a predictive distribution for the same amount of computation. The primary drawback of this method for finding the PPD is that it is applicable only in a very limited number of situations.

The next method also provides a way to estimate the PPD. This method is what has been used in SFGMs up to this point and is explained in Huzurbazar (2005c).

Given a sampling distribution $f(x|\theta)$, the prior $p(\theta)$, and the data, the density of a predicted future observation X is

$$f(x|x) = \quad f(x,\theta|x)\ d\theta, \tag{4.1}$$

if $x|\theta \quad x|\theta$, and after some algebra we find that

$$f(x|x) = E_{\theta\, x}[f(x|\theta)]. \tag{4.2}$$

In words, this means that the PPD is the average of $f(x|\theta)$ over the posterior density of θ. This method to find the PPD is straightforward, and as in the exact computation method we end up with a functional estimate of the PPD, $f(x|x)$, at the desired points on the support. The primary drawback to this method is that it is very computationally intensive. For even moderately sized posterior samples all of these inversions could take days. Like the exact method, this method requires a "smooth" parameterization of the SFGM.

The last method we discuss finds an estimate of the PPD by obtaining a Monte Carlo sample from it. The fact is, for a Bayesian analysis, all we need is the likelihood function, a loss function, and the priors to make inference. For prediction we only need the likelihood function and the prior distribution of the parameters. In most cases, we have all these components "independent" of the SFGM framework. To find an estimate of the PPD, for each given sample θ_i from our posterior we also sample from $f(x|\theta)$ (see Albert (2007)). This provides us with a Monte Carlo sample from the PPD. Instead of having an estimated PPD, we have a sample from the PPD. Even though we do not have a numeric estimate of $f(x|x)$, having a sample from the PPD is very convenient, since it readily provides us with prediction intervals and point estimates. Sampling from the PPD is not as accurate when calculating tail probabilities. If we are interested in means, medians or $100\%(1-\alpha)$ prediction intervals (where α is not too near 0) this method is fast and accurate. But, if we wanted to find the threshold c where $P(T < c) = 0.000001$, this method may not be as accurate as the others, unless a very large number of samples are obtained.

We compare the outcomes of the three methods with an introductory example. Consider the SFGM of the recurring illness process in Figure 1.1. We choose to model the transition from state 0 to state 1 with an $Exp(\theta_1)$, the transition from state 1 to state 0 with an $Exp(\theta_2)$, and the transition from state 1 to state 2 with an $Exp(\theta_3)$. We use Jeffrey's priors, $p(\theta_i) \quad 1\ \theta_i^2$ and $p \sim beta(1\ 2, 1\ 2)$.

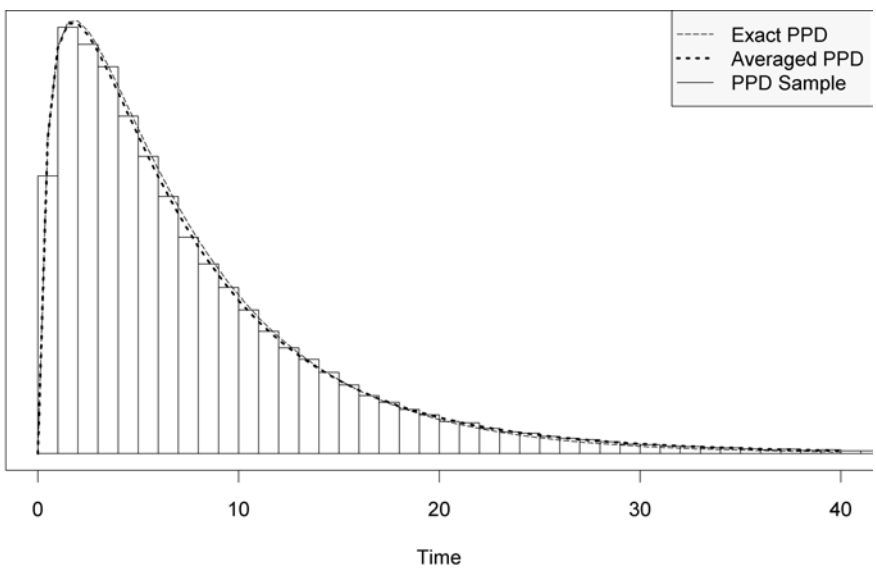


Figure 4.1: Approximations of the first passage posterior predictive density from state 0 to state 2 for the SFGM in Figure 1.1.

As you can see from Figure 4.1, in this example all three methods appear to be estimating the same thing. The time to compute the first is about 60 seconds, which is the numerical calculations for the combination of the three PPDs. (We are using a laptop with average computational power. We give the computational time not as a benchmark, but to make relative comparisons between the different methods.) The second method, which averages $f(x|\theta_i)$ for each sweep i of the MCMC, took about 3 hours to invert the 100,000 posterior samples. The third method only took about 15 seconds to obtain 100,000 samples from the PPD. Clearly this third method is computationally the quickest and very easy to implement in conjunction with an

MCMC method. This is a very simple example, and for more complex models the time to obtain an estimate of the PPD can take much longer.

By using this third method of sampling from the PPD, with SFGMs we obtain several benefits. Obviously, it dramatically reduces the computational time. In addition, we do not need to incorporate an inversion routine to find the PPD. This reduces the complexity of finding a solution and enables us to use any distribution in our SFGM, whether "smooth" or not. Basically, it allows us to model with any distribution that we can effectively sample from. Some may argue that this is a "brute force simulation" approach; this may be true, but most complex Bayesian models apply the exact same principles and it is just another avenue of accurately predicting the same things.

If we choose not to work with MGFs, we can slightly modify the graphical representation of SFGMs. Since we no longer need to represent the edges of the graph with an MGF we can use cumulative distribution functions (CDFs) or PDFs. Theoretically, we could model SFGMs using any distribution with non-negative support, if we can find a likelihood function for it. The primary practical limitation is that we need to be able to sample from the distribution. Figure 4.2 is a SFGM of the recurring illness process, but with the branches relabeled with CDFs in lieu of MGFs, and again for nodes with two or more exit paths we include an associated probability of taking that path. These probabilities can be interpreted as the parameters of a binomial or multinomial trial.

For the SFGM in Figure 4.2 after a sweep of our MCMC we have a sample from the posterior. With this posterior sample we can get a sample from the PPD. The following pseudo-code can be used to get a draw from the PPD of the SFGM in Figure 4.2.

1. Set $Time = 0$

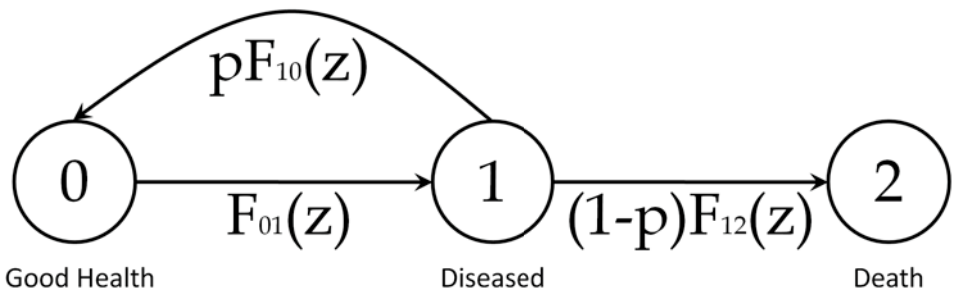


Figure 4.2: A flowgraph of the recurring illness process where the branches are labeled with CDFs.

2. Sample T_{01} and set $Time = Time + T_{01}$

3. Sample $X \sim Bern(p)$

4. If $X = 0$ sample T_{12} and set $Time = Time + T_{12}$ goto END

5. Draw T_{10} and set $Time = Time + T_{10}$

6. Goto step 2

7. END

Once we have completed our MCMC, checked for convergence and discarded our burn-in samples, we also can get samples from the PPD, as illustrated above. If the distributions we have parameterized our SFGM with are easy to sample from, it will be relatively easy to get a sample from the PPD. Now we demonstrate an example of the flexibility we can employ by estimating the PPD in this more efficient manner.

4.2 A Bayesian non-parametric model

We demonstrate our suggested method with data from 131 patients that received a bone marrow transplant. These data are found in Klein and Moeschberger (2003) (for simplicity we propose a slightly different model and ignore six observations from the analysis in Klein and Moeschberger (2003)). There are several states that describe the stages a patient may follow after a transplant. The patient begins in the transplant state (state 0), where the time-in-state is the amount of time since the transplant. From there, the patient will either proceed to state 1, which is platelet recovery, or to state 3 which is relapse or death. From state 1, the individual goes to state 2, which is chronic graft versus host disease (CGVHD), or to state 3. From state 2, the patient will eventually have a relapse or die. Figure 4.3 is a graphical representation of the SFGM.

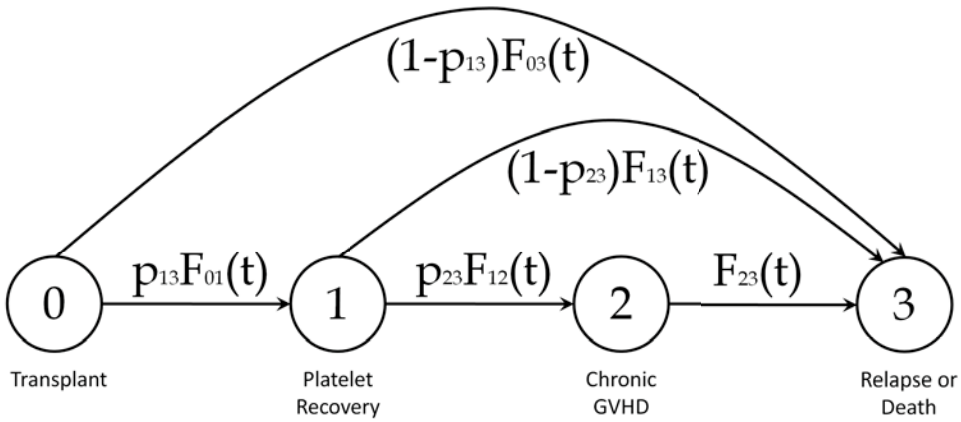


Figure 4.3: SFGM for bone marrow transplant patients.

Our goal is to predict the time to relapse or death for a new bone marrow transplant patient. There are six sets of data, one for each transition, plus a set of censored observations in state 1 that may or may not have transited through state 2 before reaching state 3. Three of the transitions, $1 \to 2$, $1 \to 3$ and $2 \to 3$, contain right-

censored data. To find an adequate parameterization of the model we consider the lognormal, gamma, Weibull, and exponential distributions and compare these with a histogram of the data. From these, we choose the distribution that seems to fit the data best. For the censored data sets we use the censored data histogram as described in Huzurbazar (2005a).

Trying to find a good parameterization for this data is challenging. The transition from state $0 \to 3$ does not fit any of the usual lifetime distributions well, as seen in Figure 4.4. The lognormal looks the best, but does not perform well when fitting the complete model. Therefore we consider a mixture of finite Polya trees (MPTs) to "parameterize" this transition distribution. For the other transitions the lognormal also seems to fit the histograms best, but once implemented this parameterization fails to adequately capture key features of the the data, such as the median. Therefore, we also use MPTs for the other transitions. Christensen et al. (2008) provides an excellent introduction to MPTs, and for convenience we use the same notation and methodology to construct the finite Polya trees. MPTs are a fairly common Bayesian method that can be considered semi-parametric or non-parametric depending how they are applied.

To understand MPTs, we must first define finite Polya Trees (FPTs). An informal way to look at FPTs is first to consider any distribution with continuous support, F. The support of F, which we call Ω, is divided into n partitions, such that $\bigcup_{i=1}^{n} A_i = \Omega$, and for all $i = j$, $A_i \cap A_j = \emptyset$. These partitions naturally have an assigned probability, which is defined by the measure F assigns to A_i. FPTs generalize distributions by assigning a new (possibly random) probability to these partitions. We show an example of a FPT, where F is an exponential distribution with four partitions in Figure 4.5. The more partitions defined in a FPT the more flexible they become. Most likely, FPTs are not "smooth" at a finite number of points even when the underlying distribution is. Therefore, we cannot use the standard methods to invert

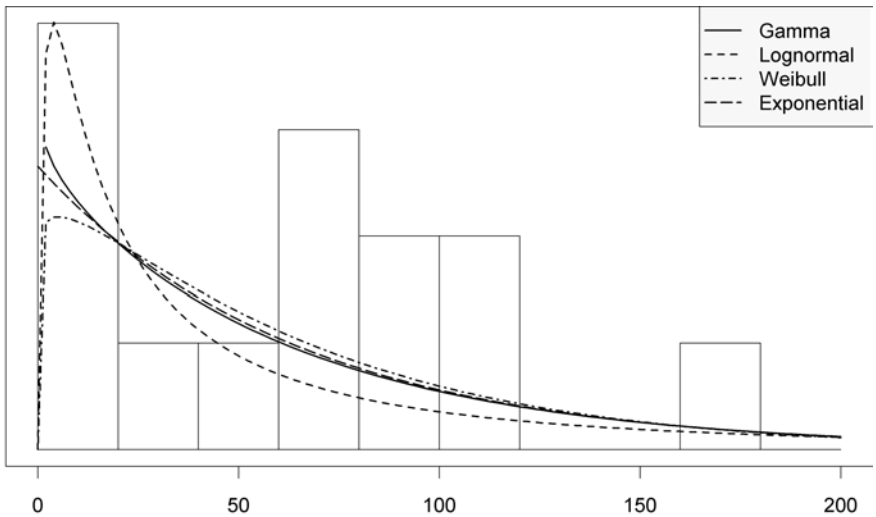


Figure 4.4: Parametric lifetime distributions fitted to the direct $0 \rightarrow 3$ transition data.

an MGF or complex LT, so we use our suggested method to obtain an estimate of the PPD by sampling from it.

A mixture of finite Polya trees is a generalization of a FPT. If we allow the underlying distribution of a FPT to have a parameter, such that the parameter itself a random variable, then for every possible parameter value we have a FPT. If we multiply the underlying distribution with the distribution of the parameter and integrate out the parameter, we are left with an MPT distribution.

For our finite Polya trees we will use the $Exp(\lambda)$ as the underlying parametric distribution. The support of the distribution is then divided into q partitions. For our application we choose $q = 8$, where each partition has probability 1 8. Next, a new probability is assigned to each partition using a rule that depends on the data. One realization of a finite Polya tree with an underlying distribution of an $Exp(1)$ is given in Figure 4.5. We use the notation $PT_k(Exp(\lambda))$ to denote the finite Polya tree

that has an $Exp(\lambda)$ underlying distribution, with 2^k equal partitions of the support. If for $PT_k(Exp(\lambda))$ we set a prior on λ, multiply it by the PDF of $PT_k(Exp(\lambda))$ and then integrate λ out, we have a mixture of finite Polya trees. MPTs are very flexible and can range from very parametric to non-parametric.

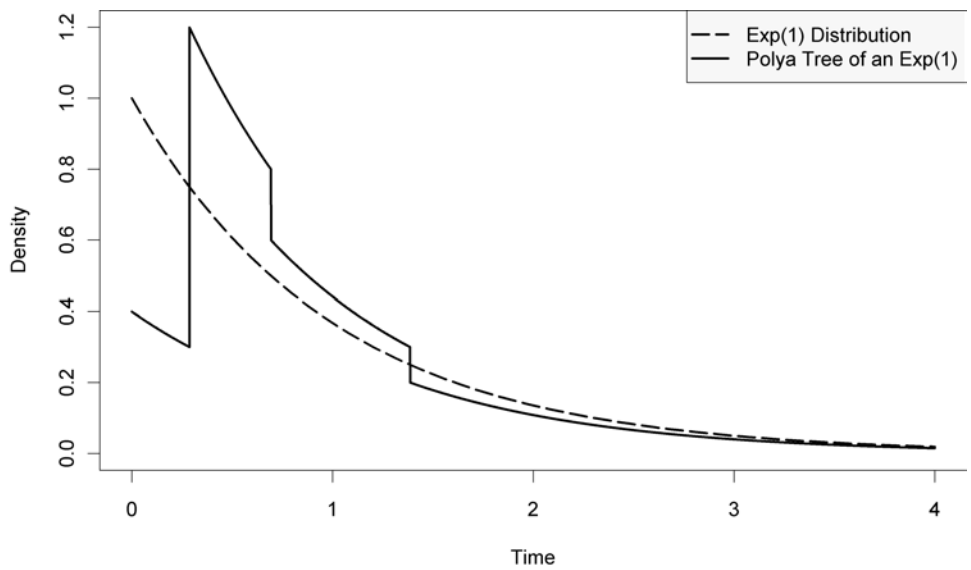


Figure 4.5: An example finite Polya tree $PT_2(Exp(1))$ overlaid on an $Exp(1)$ distribution.

Figure 4.6 shows the PPD of the MPT distribution of the direct $0 \to 3$ transition, along with a histogram of the data. Comparing this with Figure 4.4, we see that this looks to be more appropriate than any of our parametric options considered earlier.

The parameterization of our SFGM is as follows. Let

$$G_{ij} \sim \quad PT_3(Exp(\lambda_{ij}))p(\lambda_{ij})\ d\lambda_{ij}.$$

Then

$$T_{ij1}, \ldots, T_{ijk_{ij}} \Big| G_{ij} \overset{\text{iid}}{\sim} G_{ij} \quad \text{and} \quad G_{ij} \Big| \lambda_{ij} \sim PT_3(Exp(\lambda_{ij})).$$

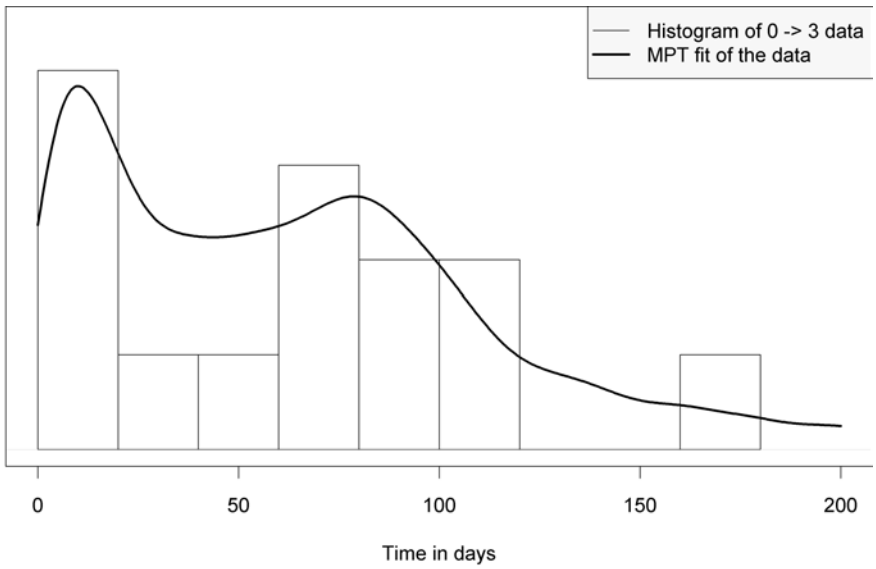


Figure 4.6: An MPT fit of the direct $0 \to 3$ transition with a histogram of the data.

T_{ij} is the random waiting time to proceed from state i to state j, and λ_{ij} is the rate parameter of an exponential distribution. The PDF of a $PT_3(Exp(\lambda_{ij}))$ is

$$\begin{aligned} f_{ij}(t|\lambda_{ij}, \boldsymbol{\theta}_{ij}) = \lambda_{ij} \exp\ -\lambda_{ij} t\ \ 2^3 * \\ \big[\theta_{ij11}\theta_{ij21}\theta_{ij31} I_{(0\ r_1]}(t) + \theta_{ij11}\theta_{ij21}\theta_{ij32} I_{(r_1\ r_2]}(t) + \\ \theta_{ij11}\theta_{ij22}\theta_{ij33} I_{(r_2\ r_3]}(t) + \theta_{ij11}\theta_{ij22}\theta_{ij34} I_{(r_3\ r_4]}(t) + \\ \theta_{ij12}\theta_{ij23}\theta_{ij35} I_{(r_4\ r_5]}(t) + \theta_{ij12}\theta_{ij23}\theta_{ij36} I_{(r_5\ r_6]}(t) + \\ \theta_{ij12}\theta_{ij24}\theta_{ij37} I_{(r_6\ r_7]}(t) + \theta_{ij12}\theta_{ij24}\theta_{ij38} I_{(r_7\quad]}(t)\big]\,, \end{aligned} \tag{4.3}$$

where the θ_{ijkl}'s are the parameters that control the probabilities of the finite Polya trees and the r_k's represent the k^{th} octile of the underlying $Exp(\lambda_{ij})$ distribution. Clearly, for this to be a valid PDF $\theta_{ij12} = 1 - \theta_{ij11}$, $\theta_{ij22} = 1 - \theta_{ij21}$, $\theta_{ij24} = 1 - \theta_{ij23}$, and so on, we use this extra variable for notational convenience. We also define $(X_{13}|p_{13}) \sim Bernoulli(p_{13})$, where $X_{13} = 1$ if the $0 \to 1$ transition is realized, otherwise we observe a $0 \to 3$ transition. Similarly, $(X_{23}|p_{23}) \sim Bernoulli(p_{23})$.

For each of the transitions we have the following prior information: we expect the mean of T_{ij} to be in the interval (a_{ij}, b_{ij}) with roughly 98% certainty. We will want our prior distribution for λ_{ij} to have 98% of its "mass" to be in the interval $(1\ \ b_{ij}, 1\ \ a_{ij})$. If we use the gamma conjugate prior, it notably influences the posterior distribution towards the mode of the prior distribution. This poses a problem because we do not necessarily favor any value in this interval over another. We therefore opt to use a prior for the mean that is flat over (a_{ij}, b_{ij}) and tapers in the tails. We must transform this prior to work with our parameterization of the gamma. The prior distribution of λ is

$$p(\lambda|a,b,\alpha) \quad \left[I_{[a\ b]}(\lambda) + I_{(0\ a)}(\lambda) \exp\left\{ \frac{2(1-\alpha)}{\alpha(b-a)}(\lambda - a) \right\} + I_{(b\quad)}(\lambda) \exp\left\{ \frac{2(1-\alpha)}{\alpha(b-a)}(b - \lambda) \right\} \right]. \tag{4.4}$$

The hyper-parameters (a, b) define the interval we believe contains $1\ \ \lambda$, and α controls the amount of area in the tampered tails. The shape of this prior and the transformed prior can be seen in Figure 4.7, where $a = 1, b = 2$, and $\alpha = 0.05$.

We choose $beta(1, 1)$ priors for p_{ij}. The priors for θ_{ijkl} (the finite Polya tree probabilities) are $beta(c\rho(k), c\rho(k))$ reference priors, where c is a positive constant that controls how non-parametric we want our model. If c is small then the model acts more non-parametric and the reverse is true for large values of c. Also, $\rho(k) = 1, 4$, or 9 for $k = 1, 2$, or 3 respectively; this weights the importance of the data at each k level, where data at level $k = 1$ has the most influence. We assume all λ_{ij}, θ_{ijkl}, and p_{ij} are mutually independent (except where $\theta_{ijkl} = 1 - \theta_{ijk(l-1)}$).

There are 22 observations that are censored in state 1 so we do not know if they would eventually proceed to state 3 directly, or through state 2; we designate these

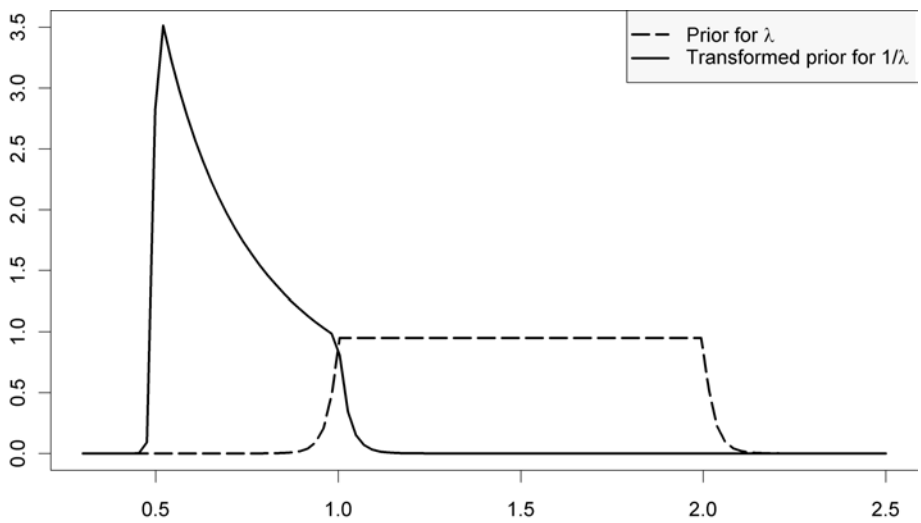


Figure 4.7: An example prior for λ and 1 λ.

as $t_{1(23)k}$ for $k = 1 \ldots 22$. The likelihood function for fixed λ_{ij}'s is:

$$
\begin{aligned}
&L(\boldsymbol{\theta}, \boldsymbol{p} | \boldsymbol{t}_{ij}, \boldsymbol{\lambda}) = \\
&\prod_{k=1}^{14} [f_{03}(t_{03k}|\lambda_{03}, \boldsymbol{\theta}_{03})] \prod_{k=1}^{117} [f_{01}(t_{01k}|\lambda_{01}, \boldsymbol{\theta}_{01})] * \\
&\prod_{k=1}^{39} [f_{13}(t_{13k}|\lambda_{13}, \boldsymbol{\theta}_{13})] \prod_{k=1}^{56} [f_{12}(t_{12k}|\lambda_{12}, \boldsymbol{\theta}_{12})] * \\
&\prod_{k=1}^{22} \left[(1 - p_{23} F_{12}(t_{1(23)k}|\lambda_{12}, \boldsymbol{\theta}_{12}) - (1 - p_{23}) F_{13}(t_{1(23)k}|\lambda_{13}, \boldsymbol{\theta}_{13})\right] * \\
&\prod_{k=1}^{25} [f_{23}(t_{12k}|\lambda_{23}, \boldsymbol{\theta}_{23})] \prod_{k=1}^{31} [1 - F_{23}(t_{23k}|\lambda_{23}, \boldsymbol{\theta}_{23})] * \\
&(p_{13})^{117} (1 - p_{13})^{14} (p_{23})^{56} (1 - p_{23})^{39} .
\end{aligned}
\tag{4.5}
$$

We use Gibbs sampling as our MCMC method. We can easily sample from the full conditionals of the $\boldsymbol{\theta}$ and $\boldsymbol{p}$ variables. We use random walk Metropolis

steps for the $\boldsymbol{\lambda}$ variables, and the Heidelberger-Welch method for our convergence diagnostic included in the **boa** package for R (see Smith (2005) for more information). The calculation of the 30,000 posterior samples took slightly less than an hour and we discarded 6,000 samples for burn-in. This time includes generating the 24,000 samples from the PPD of the first passage time from state 0 to state 3. This is a vast improvement in speed over previous Bayesian SFGM methods; if it were possible, it would have taken about 46 hours to invert the 24,000 posterior samples to obtain an estimate of the PPD. A kernel smoothed density function of the PPD sample is overlaid on a censored data histogram of the data in Figure 4.8. This PPD is the predicted amount of time a new bone marrow transplant patient has before a relapse or death. This sample from the PPD readily provides predictive intervals. A 95% prediction interval of time from transplant to relapse or death is $[17, 4605]$ days, with the median at 513 days.

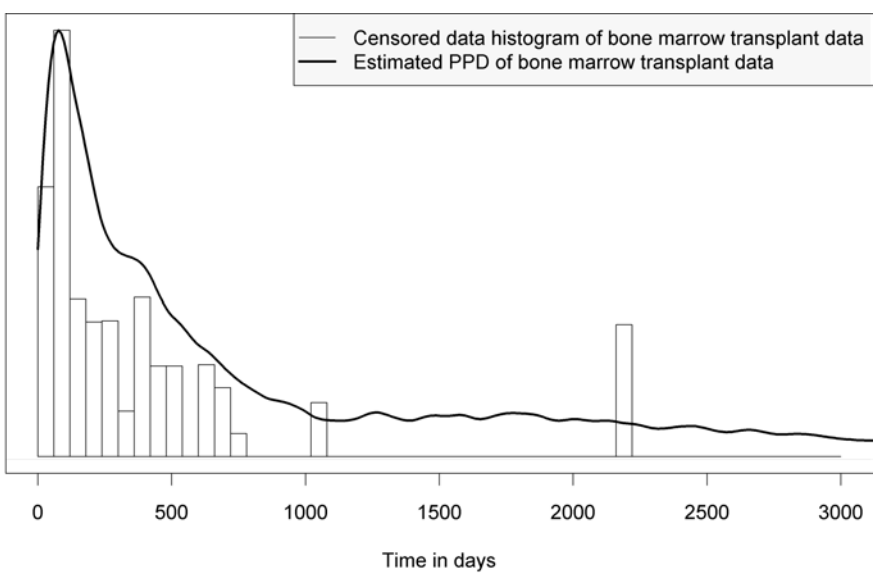


Figure 4.8: The MPT fit of the first passage PPD from transplant until time of death or recurrence. The MPT fit is overlaid on a censored data histogram.

The model we developed for this data provides a much better fit than a parametric

SFGM and is computationally much faster than traditional Bayesian SFGM methods. This example demonstrates the computational advantage of sampling from the PPD rather than finding a functional form for it. We implemented a Bayesian semi-parametric SFGM with relatively fast computational speed. This and other methods that do not use smooth PDFs would not be possible without utilizing the method of sampling from the PPD. This example demonstrates how to increase speed and flexibility in Bayesian SFGMs.

4.3 Accelerated failure time models with time-dependent covariates

Only recently have covariates been introduced into SFGMs. Huzurbazar and Williams (2010) show how covariates are included in a SFGM using the generalized linear model framework. We use a different approach by using accelerated failure time models. For an introduction to covariates and linear model theory see Christensen (2002).

Accelerated failure time (AFT) models are an alternative to the popular proportional hazards model. When we have fixed covariates (with respect to time) the PH model assumes

$$h(t|z) = g(z)h_0(t), \tag{4.6}$$

where h is the hazard function, and $g(z)$ usually depends on $x\ \beta$. In this context we let x be a vector of predictor variables and β are the coefficients of x. This assumption is not always met, therefore another alternative which is easier to interpret can be used. The AFT model assumes

$$h(t|z) = g(z)h_0(g(z)t). \tag{4.7}$$

In an AFT model a parametric distribution must be assumed. Christensen et al. (2010) expresses the AFT model in a less general but more usable form. For event times $T_1, \ldots, T_n$ the model is

$$log(T_i) = x_i\beta + \sigma e_i, \tag{4.8}$$

where x_i is a vector of predictor variables, β the vector of coefficients for x_i, e_i is the random error with a fixed distribution, and σ is a parameter that controls the variance of the error term. Usually the error terms are assumed to be *iid* and have a mean or median of 0. An alternative way to write this model is

$$log(T_i) = \beta_0 + \beta_1 x_{i\,1} + \ldots + \beta_{(p-1)} x_{i\,(p-1)} + \sigma e_i, \tag{4.9}$$

where we have $p - 1$ predictor variables.

In an AFT model there are numerous possible distributions for e, but some popular choices are the normal, Gumbel, and the logistic. Other options include the gamma, inverse Gaussian, and the generalized Pareto. These last distributions are not as popular due to the fact that they are more difficult to handle analytically.

This framework for AFT models must be generalized if the predictor variables $x_{i\,j}$ are allowed to vary with time. In survival analysis time-varying covariates are usually some type of measurements that are repeatedly taken on a patient over time. Classic examples would be weight and blood pressure; at each visit it is common practice to have these measurements recorded. Therefore, a patient's risk of some illness or disease may increase or decrease depending on one or more of these time-varying covariates.

In most situations time-varying covariates are measured at certain points over time, but there is no information about the covariate values between these time points. For example, with a patient's weight it would be safe to assume it is continually varying over time, but we only observe it when a patient is actually weighed.

Therefore we do not know what it was in between the measurements. Collett (2003) offers some ways to deal with this uncertainty. We adopt the practice of holding the covariate fixed until the next measurement. If the patient was weighed at 165 lbs., we use that as the value of the covariate until we obtain another measurement. In many situations this makes computation easier than other methods of handling time-varying covariates.

The most difficult part of modeling time-varying covariates is obtaining a likelihood function. Petersen (1986) provides a straightforward way to obtain the likelihood for this type of model. He argues that

$$S(t|X(t_0), X(t_1), \ldots, X(t_n), \theta) = \exp\left(-\sum_{i=1}^{n} \int_{t_{i-1}}^{t_i} h(s|X(t_{i-1}), \theta)\, ds\right), \quad (4.10)$$

where $t_0 = 0$, S is the survivor function, h is the hazard function, θ is a generic vector of parameters, and $X(t_i)$ is the time-varying covariate at an observed time. We can use this formula as the contribution to the likelihood for a right-censored observation. If we have a complete observation we use

$$f(t|X(t_0), X(t_1), \ldots, X(t_n), \theta) = h(t|X(t_n), \theta) S(t|X(t_0), X(t_1), \ldots, X(t_n), \theta) \quad (4.11)$$

as the contribution to the likelihood function. Similarly, we could also find the contribution of left or interval-censored data.

Therefore, if we let x_i be the constant vector of predictor variables during the time interval $[t_i, t_{i+1})$ and the model error term, e, have the standard logistic distribution, then the hazard given x_i is

$$h(t|x_i\beta, \sigma) = \frac{\exp\left(\frac{\log(t) - x_i'\beta}{\sigma}\right)}{\sigma t \left(1 + \exp\left(\frac{\log(t) - x_i'\beta}{\sigma}\right)\right)}, \quad (4.12)$$

and the integral of this hazard over the time interval $[t_1, t_2)$ is

$$\int_{t_1}^{t_2} h(s|x_1\beta, \sigma)\, ds = -\log\left(\frac{1 + \exp\left(\frac{\log(t_1) - x_1'\beta}{\sigma}\right)}{1 + \exp\left(\frac{\log(t_2) - x_1'\beta}{\sigma}\right)}\right). \quad (4.13)$$

Using (4.13) and evaluating (4.10) gives us the contribution for the likelihood of right-censored observations. This in turn gives us everything that we need to evaluate (4.11) for a complete observation.

Using this AFT model with the standard logistic distribution as the error, we consider the diabetic retinopathy data discussed in Marshall and Jones (1995). Every patient has two or more observations over time, and each observation has several covariates. Using the results from Marshall et al. (1995), which suggests that glycohemoglobin ($\mathbf{H}\mathrm{b}\mathbf{A}_1$) is one of the most important factors related to changes in retinopathy. The glycohemoglobin level is found from a blood test that measures the amount of sugar bound to hemoglobin, the value of $\mathbf{H}\mathrm{b}\mathbf{A}_1$ is reported in percentages. We use this predictor as the time-varying covariate in the model. For simplicity $\mathbf{H}\mathrm{b}\mathbf{A}_1$ is the only covariate we include.

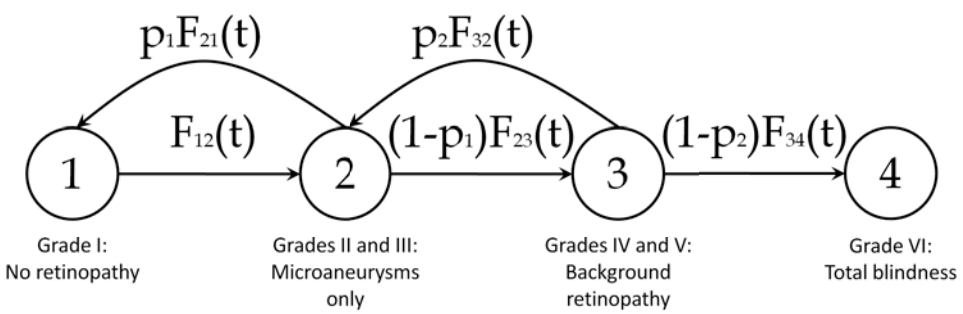


Figure 4.9: Flowgraph model for the diabetic retinopathy data.

Figure 4.9 is the multistate Markov model for this data suggested by Marshall and Jones (1995). We also use this as a SFGM of the data. Yau and Huzurbazar (2002) and Huzurbazar (2005c) have also analyzed this data using SFGMs but not with covariates. We continue the assumption made in earlier analyses; we treat the transition times as known. In Chapter 5 we relax this assumption and treat the transition times as unknown. Individuals with severe diabetes eventually develop retinopathy which leads to blindness. In state 1 no retinopathy has developed. In

Table 4.1: A few observations of the diabetic retinopathy data.

Subject	State	Time in study (months)	$\mathbf{H}\mathrm{b}\mathbf{A}_1$
1	2	0	10.2
1	2	12	
2	1	0	7.9
2	1	15	9.6
2	1	27	7.9
2	1	39	7.2
2	1	51	
3	1	0	8.6
3	1	17	12.5
3	1	30	8.9
3	1	40	

state 2 there is some damage to the retina which is categorized as an intermediate stage of retinopathy, but is reversible. State 3 indicates that prolonged high sugar levels have damaged the retina; this stage is also recoverable. State 4 indicates blindness due to diabetic retinopathy. In this study there were a total of 277 patients. Table 4.1 displays the first few observations. The final observation for each patient does not include covariate information. Therefore, we use the covariate information at a specific visit as the constant value of the covariate until the time of the next visit.

We have five transitions for this model, each with three parameters, and parameterized as follows:

$$\begin{aligned}
log(t_{i\,1}) &= \beta_{0\,1} + \beta_{1\,1}X(t_{i\,1}) + \sigma_1 e_{i\,1} \text{ for the } 1 \to 2 \text{ transition,} \\
log(t_{i\,2}) &= \beta_{0\,2} + \beta_{1\,2}X(t_{i\,2}) + \sigma_2 e_{i\,2} \text{ for the } 2 \to 1 \text{ transition,} \\
log(t_{i\,3}) &= \beta_{0\,3} + \beta_{1\,3}X(t_{i\,3}) + \sigma_3 e_{i\,3} \text{ for the } 2 \to 3 \text{ transition,} \\
log(t_{i\,4}) &= \beta_{0\,4} + \beta_{1\,4}X(t_{i\,4}) + \sigma_4 e_{i\,4} \text{ for the } 3 \to 2 \text{ transition, and} \\
log(t_{i\,5}) &= \beta_{0\,5} + \beta_{1\,5}X(t_{i\,5}) + \sigma_5 e_{i\,5} \text{ for the } 3 \to 4 \text{ transition.}
\end{aligned} \tag{4.14}$$

As seen in Figure 4.9 we define p_1 as the probability that an individual in state 1

proceeds to state 0 before state 2. Similarly we define p_2 as the probability that an individual in state 2 proceeds to state 1 before state 3. Because p_1 and p_2 also depend on $\mathbf{H}\text{b}\mathbf{A}_1$, we use logistic regression (see Kutner et al. (2005) or Wasserman (2004)) to determine their values for a given $\mathbf{H}\text{b}\mathbf{A}_1$ level. Therefore $logit(p_1) = \gamma_{01} + \gamma_{11}X(t)$ and $logit(p_2) = \gamma_{02} + \gamma_{12}X(t)$, where $logit(p) = log(p\ \ (1-p))$.

With this information we now express the log likelihood function:

$$
\begin{aligned}
l(\beta, \sigma, p|t, x) = & \\
& \sum_{i=1}^{5}\sum_{j=1}^{n_i}\sum_{k=1}^{m_{ij}} \log \frac{1+\exp\ \frac{\log(t_{ij(k-1)})-x'_{ijk}\beta_i}{\sigma_i}}{1+\exp\ \frac{\log(t_{ijk})-x'_{ijk}\beta_i}{\sigma_i}} \ + \\
& \sum_{i=1}^{5}\sum_{j=1}^{q_i}\sum_{k=1}^{r_{ij}} \log \frac{\exp\ \frac{\log(t_{ijk})-x'_{ijk}\beta_i}{\sigma_i}}{\sigma_i t_{ijk}\left(1+\exp\ \frac{\log(t_{ijk})-x'_{ijk}\beta_i}{\sigma_i}\right)} \ - \\
& \sum_{i=1}^{2}\sum_{j=1}^{N_i} \left(C_i \log\left[1+\exp(-x_i\gamma_i)\right] + (N_i - C_i)\log\left[1+\exp(x_i\gamma_i)\right]\right).
\end{aligned}
\tag{4.15}
$$

The first line of the log likelihood function is the log of all the survivor functions where n_i is number of patients that have an observation for the i^{th} transition, and m_{ij} is the number of $\mathbf{H}\text{b}\mathbf{A}_1$ measurements taken on the j^{th} patient for the i^{th} transition. The second line is the log of all the hazards for the complete observations where q_i is the number of patients with a complete observation for the i^{th} transition, and r_{ij} is the number of complete observations the j^{th} patient had for the i^{th} transition. The last line is the log of the logistic regression contributions, where N_i is the number of observed transitions from state $i+1$, and C_i is the number of observed transitions from state $i+1$ to state i. In the log likelihood function we denote the appropriate predictors and coefficients as $x\ \beta_i$ or $x\ \gamma_i$.

We set diffuse priors on the regression coefficients, so $p(\beta_{ij}) \sim N(0, 100)$ and $p(\gamma_{ij}) \sim N(0, 100)$. We also use vague priors for σ_i, where $p(\sigma_i) \sim Exp(1\ \ 100)$. We use Gibbs sampling with random walk Metropolis steps for each parameter.

The MCMC mixing looks good except for the final transition (state 3 to state 4), because there are few complete observations. To "remedy" this we use an informative prior, $p(\sigma_5) \sim gamma(2, 1)$. We now have 50,000 sample from the posterior after accounting for burn-in.

Once we have a posterior sample, we can begin to predict. However, since there are covariates, we need values for $\mathbf{H}\text{b}\mathbf{A}_1$ to be able to predict. If the value of glycohemoglobin is fixed (with respect to time) then it is straightforward to find a PPD for that fixed value. We proceed as in the previous sections; for each sample from the posterior we simulate an observation. For each transition we sample from a logistic random variable, multiply it by σ_i, add $x\,\beta_i$ and exponentiate it. This gives the time for one transition; we also need to simulate $Bernoulli(p_1)$ or $Bernoulli(p_2)$ random variables for each visit to state 2 or 3 respectively. Once we have simulated observations starting in state 1 and reaching state 4 for each of the 50,000 posterior samples we have a sample from the PPD of the first passage time from contraction of diabetes to blindness. We find a PPD for the mean value of all glycohemoglobin measurements in the study, which is approximately 11.26. A histogram of this PPD is found in Figure 4.10.

For other fixed values of $\mathbf{H}\text{b}\mathbf{A}_1$ we calculate PPDs. In Table 4.2 we compare the predicted probability of blindness due to diabetic retinopathy for fixed values of $\mathbf{H}\text{b}\mathbf{A}_1$. Clearly, the model distinguishes between given levels of $\mathbf{H}\text{b}\mathbf{A}_1$. The higher the level of $\mathbf{H}\text{b}\mathbf{A}_1$ the higher the risk of blindness due to diabetic retinopathy.

However, prediction when $\mathbf{H}\text{b}\mathbf{A}_1$ levels are time-dependent is more difficult. Suppose we have two individuals, one begins with a high level of $\mathbf{H}\text{b}\mathbf{A}_1$ which decreases over a 20 year period. The other individual begins with a low $\mathbf{H}\text{b}\mathbf{A}_1$ level, which increases over time. How do we get the PPD for these individuals?

Assume we can break the 20 year period into disjoint but adjacent time intervals,

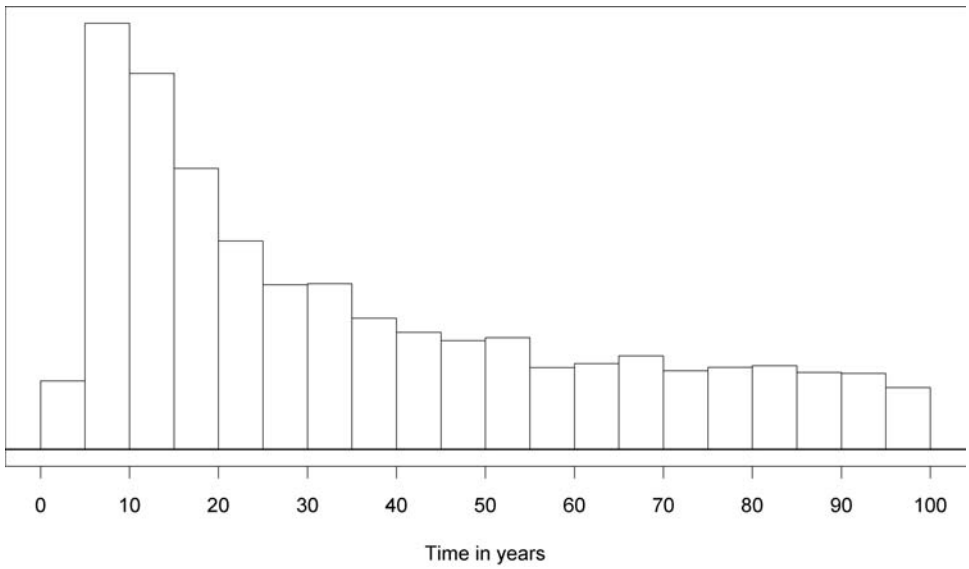


Figure 4.10: A histogram of the PPD of the first passage from contraction of diabetes (state 1) to blindness (state 4), using the mean level of $\mathbf{H}\mathrm{b}\mathbf{A}_1$.

where the $\mathbf{H}\mathrm{b}\mathbf{A}_1$ level for each individual is constant in each interval. Using Equation 4.10 we can find the survivor function. Since we have survivor function, S, we

Table 4.2: Predicted probabilities of blindness due to diabetic retinopathy before various times.

	$\mathbf{H}\mathrm{b}\mathbf{A}_1$ level					
Years	8	11	14	17	Individual 1	Individual 2
5	0.002	0.003	0.005	0.013	0.002	0.014
10	0.014	0.019	0.030	0.049	0.020	0.054
15	0.025	0.034	0.053	0.074	0.062	0.097
20	0.031	0.045	0.069	0.093	0.102	0.146
25	0.036	0.053	0.083	0.107	0.131	0.184
30	0.039	0.058	0.093	0.119	0.151	0.212
35	0.041	0.064	0.103	0.129	0.167	0.235
40	0.043	0.068	0.111	0.138	0.180	0.253
50	0.047	0.076	0.125	0.154	0.201	0.280
75	0.055	0.093	0.153	0.183	0.235	0.319

Table 4.3: Glycohemoglobin levels for two hypothetical individuals

	time in years										
	0	2	4	6	8	10	12	14	16	18	20
Individual	$\mathbf{H}\text{b}\mathbf{A}_1$ levels										
1	8	9	10	11	12	13	14	15	16	17	18
2	18	17	16	15	14	13	12	11	10	9	8

can sample directly from each transition distribution using the probability integral transform. For given values of the parameters, θ_i, from the posterior, we draw a $U(0,1)$ random variable, and then find $S^{-1}(U|\theta)$. $S^{-1}(U|\theta)$ must be found numerically, i.e. we find the value $0 < t <$ such that $S(t|\theta) - U = 0$. If we simulate an observation using these techniques for each sample from the posterior distribution, we obtain a sample from the predicted distribution of the time until blindness for these given glycohemoglobin levels.

For the two individuals introduced above, we assume their $\mathbf{H}\text{b}\mathbf{A}_1$ levels are the same as given in Table 4.3. Finding the PPDs for each of these hypothetical individuals is time consuming and takes as long as finding the posterior samples. We show histograms of the samples from each of these PPDs in Figure 4.11. The predicted time until blindness for these two individuals is in Table 4.2. Clearly, individual 1 has higher levels of $\mathbf{H}\text{b}\mathbf{A}_1$ after 12 years than individual 2; but after 75 years the probability that individual 1 goes blind is still much less than individual 2. This seems to indicate that the initial levels of glycohemoglobin are the most influential in predicting the probability of blindness due to diabetic retinopathy. Using glycohemoglobin as a covariate enhances the analysis of this data and in the predictive risk of blindness due to diabetic retinopathy.

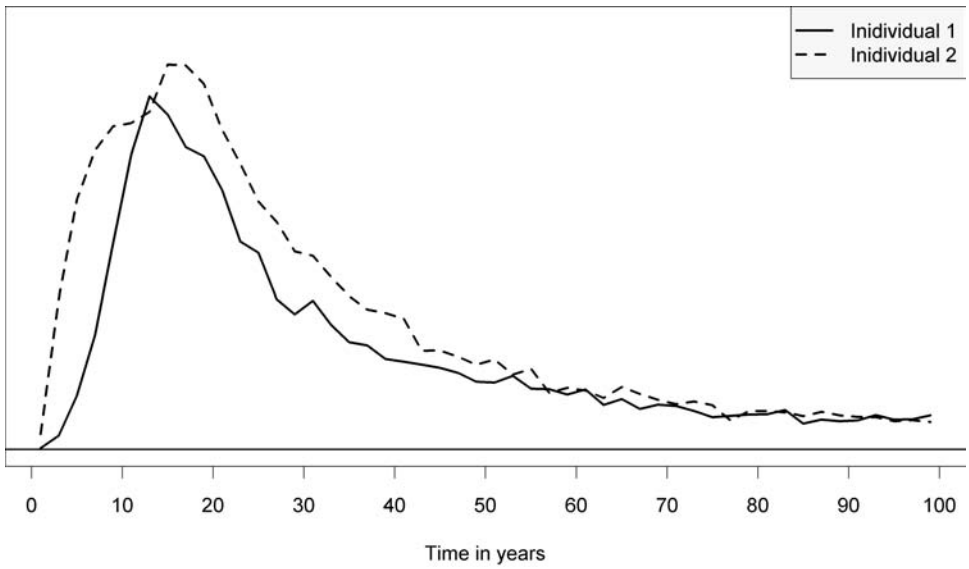


Figure 4.11: The PPDs of time from diagnosis until blindness for the two individuals with $\mathbf{H}\text{b}\mathbf{A}_1$ levels given in Table 4.3.

4.4 Simulated recurring illness process example (continued)

We have demonstrated that the PPD of the first passage time from one state to another can be estimated quickly by simulating from the SFGM for each sample from the posterior. We demonstrate this in our recurring illness process example.

We return again to the recurring illness process in Figures 2.4 and 4.2. In the past finding the PPD of even a simple Bayesian SFGM such as in Figure 4.2 has been time consuming to compute. Suppose we choose to model the branches with members from the exponential family with conjugate priors. With only a small amount of computational time we can obtain a substantial sample from the PPD of the first passage from state 0 to state 2.

We model the $0 \to 1$ transition with a $lognormal(\mu_1, \sigma_1^2)$, the $1 \to 0$ transition

Table 4.4: Predictive quantiles from the recurring illness process

Time in years								
0.5%	1%	2.5%	5%	50%	95%	97.5%	99%	99.5%
1.64	1.82	2.14	2.46	5.74	18.73	22.93	29.02	34.41

with a $lognormal(\mu_2, \sigma_2^2)$ and the $1 \to 2$ transition with a $lognormal(\mu_3, \sigma_3^2)$. The conjugate priors we choose are $p(\mu_i) \sim N(0, 100)$, $p(\sigma_i^2) \sim inverse\,gamma(1\ \ 20, 1)$, and $p \sim beta(1, 1)$. Using the Gibbs sampler on a PC laptop we find a sample of 100,000 from the posterior in 22 seconds and 100,000 samples from the PPD in 8 seconds. This is extremely fast, compared with the methods introduced in Chapter 3 that allows us to use the lognormal distribution in a SFGM. Figure 4.12 is a histogram of the 100,000 samples from the PPD, and Table 4.4 shows the quantiles from the posterior predictive distribution that can be used to find predictive intervals.

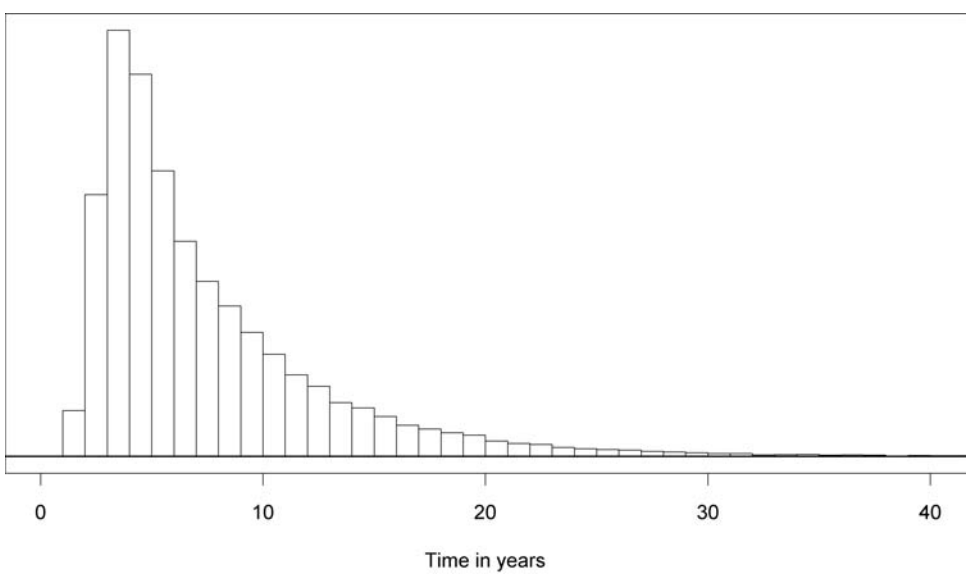


Figure 4.12: A histogram of the estimated first passage PPD from state 0 to state 2 for the recurring illness process.

In this chapter we have demonstrated that efficient Bayesian computation can

be extremely beneficial when working with SFGMs. This becomes invaluable when working with complicated SFGMs. Models with non-smooth distributions or difficult models with time-varying covariates are not easily handled using traditional Bayesian SFGM methods. This technique of sampling from the PPD also allows for fast Bayesian computation in some situations.

Chapter 5

General definition of incomplete data and model extensions

One advantage of SFGMs over other models for survival data is the ability to handle incomplete data. Incomplete data are distinct from censored data in that: "Incomplete data consists of data that have complete information on observed waiting times but incomplete information on the associated transitions" Huzurbazar (2005c). Incomplete data for SFGMs are considered by both Yau and Huzurbazar (2002) and Williams and Huzurbazar (2006). Their approach involves construction of the likelihood for missing transitions. Likelihood construction is computationally intensive even for simple flowgraphs. In addition, these previous approaches do not consider covariates. We propose an alternative method of likelihood construction which also allows for the inclusion of covariates.

The term "incomplete data" has often been mistaken for censored data. To avoid further confusion, we provide a general definition of incomplete data.

Definition *Incomplete data* are observations such that there is positive probability that one or more transition times are not completely known.

Note that this is more general than the definition of incomplete data given in Huzurbazar (2005c). The expanded generality of this new definition allows for transition times to be unknown. In other words, incomplete data can be and usually are censored. This also suggests that any type of censored data should be included in the broader definition of incomplete data. An example can help illustrate this definition.

Consider a stochastic process $X(t)$ displayed in Figure 5.1. We know at time $t = 0$ the process is in state i (i.e., $X(0) = i$), and that the process must transition through state j before reaching state k. We find that the process remains in state i until time $t = 10$. However, we do not receive any other information about the observation until $t = 30$ where we see that $X(30) = k$. Therefore, the observation is incomplete because we do not know when the transition from state i to state j and the transition from state j to state k occurred. The observation is also censored, by the fact we do not know how long the observation remained in state i until the first transition occurred and how long it had been in state k before $t = 30$.

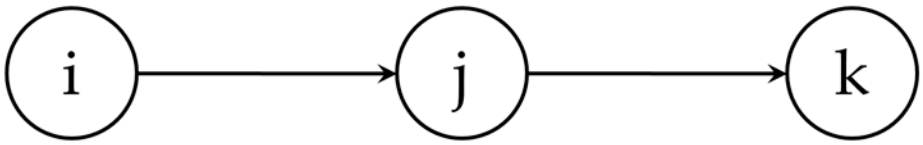


Figure 5.1: An example stochastic process.

In essence, incomplete data occurs when we "lose communication" with $X(t)$ for an interval of time. If this loss of communication introduces any uncertainty regarding the state of the process, then we say the observation is incomplete. The uncertainty is introduced if there is positive probability that a transition may have occurred, but we do not know either if or when one or more transitions actually took place. Imagine if we lose communication with a process, if we never again regain communication then we have the common case of right-censored data. If we

regain communication, the process may have transited between none or many states (depending on how the process is defined). The problem is we do not know how long it was before the process transitioned to the next state (if it did at all). If we regain communication with the process, unless we do at the instant it transitions, we do not know how long it has been in the current state. These are just some of the situations that we encounter with incomplete data.

It is helpful to consider some examples of what is and is not incomplete data when dealing with SFGMs. Consider the SFGM from Chapter 4 which we repeat in Figure 5.2. Suppose we know $X(0) = 0$ and $X(5) = 0$. If we lose communication with the process at time $t = 5$ and then regain communication at time $t = 10$ and see that $X(10) = 0$, then we would not have an incomplete observation, because we are certain that the process cannot return state 0 therefore it never left. Conversely, if we use this same scenario on our familiar example in Figure 5.3, then we would have an incomplete observation because the process could have left state 0 and then returned during the time we were not observing it. It is common practice to ignore this as incomplete if we are quite certain that the process did not transition out of state 0 (i.e., the probability that it transitioned and then returned to state 0 is small). Still, there are more blatantly incomplete observations that cannot be ignored. For example, we know the process in Figure 5.3 starts with $X(0) = 0$. If we observe the process again at time $t = 150$ where $X(150) = 2$ we do not know when the process transitioned through state 1 and how many times this may have occurred.

Yau and Huzurbazar (2002) and Williams and Huzurbazar (2006) construct an approximate likelihood where the contribution of this observation to the likelihood function would be the first passage distribution $f_{02*}(150)$. Using this as an approximate likelihood contribution assumes that the actual transition to state 2 occurred shortly before $t = 150$, otherwise this is a poor approximation. With the advances we have introduced we can provide an extension and treat this as a left-censored

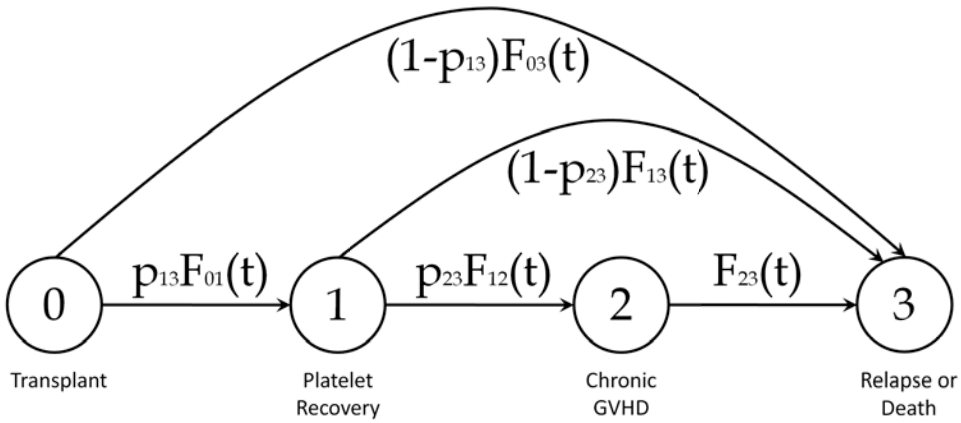


Figure 5.2: SFGM for bone marrow transplant patients.

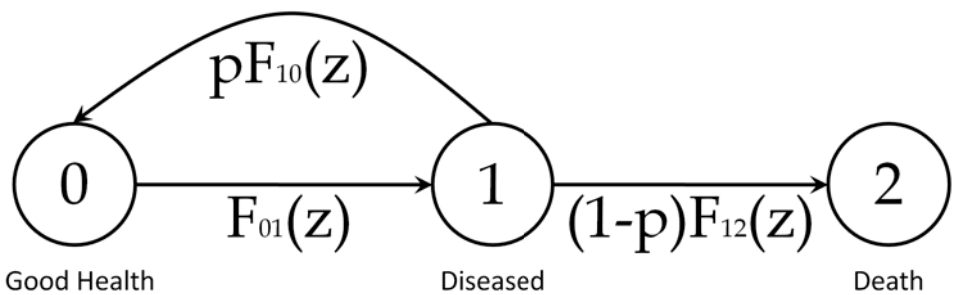


Figure 5.3: A flowgraph of the recurring illness process.

observation from the same distribution. The contribution to the likelihood function should be $F_{02*}(150)$, this is the exact contribution to the likelihood and is actually as easy to compute as $f_{02*}(150)$. We develop these ideas further in the next section.

Now that we have introduced a more general definition of incomplete data we need to show how SFGMs can help. Our aim is to introduce a method to handle and compute the exact likelihood in multistate models, given that we have incomplete data.

5.1 Constructing an exact likelihood for incomplete data

When using approximate likelihood functions with incomplete data, we must assume that the observed time is near to the actual transition time. In our last example when we only knew $X(0) = 0$ and $X(150) = 2$ for the SFGM in Figure 5.3, using $f_{02*}(150)$ as a contribution to the likelihood, it is only reasonable to use this value if we can assume that the actual transition to state 2 occurred shortly before or at time $t = 150$.

Often this assumption used to build an approximate likelihood function is not met. The likelihood construction approach of Yau and Huzurbazar (2002) and Williams and Huzurbazar (2006) cannot handle this case. We introduce a method that can incorporate information for almost any type of incomplete data. The key to modeling incomplete data is incorporating it into the likelihood function. Once we have done that, our analysis can proceed in the usual manner whether it be frequentist or Bayesian.

Suppose we have a process as described in Figure 5.3. We saw a patient that we knew was in state 0 at time $t = 0$ and then was later discovered at time t_1 to be in state 2 but with no other information; it would be unreasonable to make additional assumptions. In this scenario, we have some definite information that would be beneficial to include in the model. Using SFGMs, this subject gives us a left-censored observation from the distribution of the first passage from state 0 to state 2 or $f_{02*}(t)$. Therefore this observation's contribution to the likelihood would be $F_{02*}(t_1)$. Generally, the value $F_{02*}(t_1)$ is not trivial to find, but using flowgraphs it is not too difficult.

If we have an incomplete observation that we can glean some information from, we

can incorporate it into the likelihood. In our previous example, the contribution to the likelihood was $F_{02*}(t_1)$. This seems simple enough, but as seen in the introduction to SFGMs finding $F_{02*}(t_1)$ involves some numerical calculations. In some situations the numerical calculations are fast and in some situations it may be slower. It primarily depends on the complex LTs of our parameterization. If the complex LTs of the distributions we have chosen are in closed form, then the calculations are fast; if the complex LTs must be found numerically, then this dramatically slows down the calculations. Examples of distributions that have closed form complex LTs are the gamma and inverse Gaussian distributions, and examples that do not are the lognormal and Weibull distributions. In the next section we demonstrate how to compute $F_{02*}(t_1)$.

In the last example we were able to use all of the information the data contained in the likelihood, but we are not always able to to so. Consider the two state recurrent process in Figure 5.4. If we observe a subject at time t in state 1 and then observe the same subject again in state 1 at time $t + \varepsilon$ we know nothing about the process. The reason for this is the process may have transitioned many times or not at all in the time interval $[t, t + \varepsilon)$. One might argue that this observation contains no information, and therefore it does not contribute to the likelihood. In either case we have no way to use this information (or lack there of).

Figure 5.4: A two state recurrent process.

Even in simple situations, knowing what information is available in an incomplete observation can be challenging. We use the SFGM in Figure 5.3 to illustrate how we could use fragments of information in several scenarios. In each case we determine

that the incomplete information is just some sort of censored observation which could be from a first passage distribution.

Table 5.1: Examples of incomplete data in Figure 5.3.

Case	Observation	A possible likelihood contribution
1	$X(t) = 2$	$F_{02*}(t)$
2	$X(t) = 1$	$1 - F_{02*}(t)$ or $F_{01}(t)$
3	$X(t) = 0$	$1 - F_{02*}(t)$
4	$X(t) = 1$ and $X(t+\varepsilon) = 1$	$1 - F_{02*}(t+\varepsilon)$ or $F_{01}(t)$
5	$X(t) = 0$ and $X(t+\varepsilon) = 0$	$1 - F_{02*}(t+\varepsilon)$
6	$X(t) = 1$ and $X(t+\varepsilon) = 0$	$F_{10}(t+\varepsilon)$
7	$X(t) = 0$ and $X(t+\varepsilon) = 1$	$F_{01}(t+\varepsilon)$
8	$X(t) = 0$ and $X(t+\varepsilon) = 2$	$F_{12*}(\varepsilon)$ or $F_{02*}(t+\varepsilon)$
9	$X(t) = 1$ and $X(t+\varepsilon) = 2$	$F_{01}(t)$

Consider the following incomplete data in Table 5.1 (where we assume $X(0) = 0$ for all cases). These scenarios show that if we know a subject was definitely in one or more states for a positive interval of time, we can use at least a piece of that information in the likelihood. We can even use very sparse information, such as in case 1, because the SGFM has an absorbing state. However, if the process described in Figure 5.3 had no absorbing state, we could not have obtained any information in some of these cases.

In Table 5.1 it is clear the same pieces of information can be interpreted differently, such as in case 8. We could use $F_{12*}(\varepsilon)$ or $F_{02*}(t+\varepsilon)$ in the likelihood, but we cannot use both. It may be tempting to glean more information out of an observation than is possible. At first glance we might try to use both $F_{12*}(\varepsilon)$ and $F_{02*}(t+\varepsilon)$ as contributions to the likelihood. However, we cannot use both pieces of information, because they are not independent; they are very dependent. We recommend caution when determining which pieces of information to include in the likelihood. This raises the question, if we have two or more pieces of dependent information from an incomplete observation, which one should we include in the likelihood? This question varies from sample to sample, but we recommend weighing which piece of

information provides the most information against the computation time it would take to use the information in the likelihood function.

To find the contribution to the likelihood of an incomplete data observation, we treat the observation as if it were one or more censored observations but possibly from a first passage distribution. For example, Figure 5.3 is a multistate semi-Markov model. We know $X(0) = 0$ and the process remains in state 0 until time $t = 10$. From then until $t = 30$, we do not receive any other information about the observation. We find that $X(30) = 1$, where $X(t)$ remains until at $t = 35$ it transitions from $1 \to 2$. The actual path of $X(t)$ may have been $0 \to 1 \to 2$, $0 \to 1 \to 0 \to 1 \to 2$, or $0 \to 1 \to 0 \to 1 \to 0 \to 1 \to 2$ et cetera.

To use this information we break this up into what we do know about the observation; it then becomes a straightforward problem. We know that the transition $0 \to 1$ occurred after $t = 10$. Using the censored contribution of this information we could use $1 - F_{01}(10)$ in the likelihood. The next piece of information is that the transition from $1 \to 2$ occurred in under 25 time units. So either $F_{12}(25)$ or $F_{12*}(25)$ would be valid contributions to the likelihood. Another option would be to use $f_{02*}(35)$. We reiterate the point that without the context of the situation, it is difficult to determine which bits of information would be most beneficial in the likelihood. Regardless of which one is chosen it will be beneficial to use as much of the information in the data as possible.

From the above examples we showed that there is often information to be gained by using incomplete information without as many assumptions. Now we suggest some techniques for computing the likelihood function of a SFGM with incomplete data.

5.2 Computation

Now that we know what to include in the likelihood function, we need to be able to calculate the actual quantities. We have already mentioned how to calculate $f_{ij*}(t)$ in Chapter 3. This can be done for $F_{ij*}(t)$ (the left-censored case), which is very similar to finding $f_{ij*}(t)$. First, we find the complex LT of $f_{ij*}(t)$ using Mason's rule, which we denote as $L_{ij*}(z)$. Then, using the properties of LTs for positive random variables defined on $[0, \quad)$ we find that the complex LT of $F_{ij*}(t)$ is $L_{ij*}(z) \;\; z$. From this formula we see that the time to compute $f_{ij*}(t)$ or $F_{ij*}(t)$ is essentially the same. Once we have $L_{ij*}(z) \;\; z$ we feed the real portion of this into the EULER algorithm to calculate the value of $F_{ij*}(t)$, again the details of the inversion can be found in Abate and Whitt (1995).

Until now, we have only considered the left-censored case. This easily extends to both the right- and interval-censored cases. For the right-censored we have $1 - F_{ij*}(t)$ as the contribution to the likelihood and $F_{ij*}(t_2) - F_{ij*}(t_1)$ as the contribution in the interval-censored case. These quantities can be approximated in the same fashion as shown above. Now that we have introduced how to handle incomplete data, we apply these techniques in an example.

5.3 Incomplete data application to diabetic retinopathy

We again consider the diabetic retinopathy data from Chapter 4. This is a longitudinal study of 277 diabetic patients. At each visit we have the following information: the time since diagnosis of diabetes, the time since the last visit, the glycohemoglobin percentage ($\mathbf{H}\mathrm{b}\mathbf{A}_1$), and the current state (as defined in Figure 4.9). We again in-

clude $\mathbf{H}\mathrm{b}\mathbf{A}_1$ as a covariate, but use the first measurement as a fixed value until we observe a change in state where $\mathbf{H}\mathrm{b}\mathbf{A}_1$ can then assume a new value. This is not as rigorous as using time-varying covariates but we still allow the value of $\mathbf{H}\mathrm{b}\mathbf{A}_1$ to vary when the patient changes states. This restriction to use the first observed measurement of $\mathbf{H}\mathrm{b}\mathbf{A}_1$ in the state is not unreasonable from what we observed in the PPDs of the two hypothetical individuals in the last chapter. The analysis indicated the first values had the strongest effect on the predictive probabilities of time until blindness.

Applying the techniques in this chapter we can make fewer assumptions about these data and still predict. We no longer assume that we know anything about the transition times, except the following: each patient began in state 1 at their time of diagnosis to diabetes. Next, if for two consecutive visits the patient was in the same state we assume no transitions were made during the time between appointments. The last assumption is that we observed the shortest transition path, so if we observe a patient in state 1 and then state 2 we assume that only the $1 \to 2$ transition took place, not $1 \to 2 \to 1 \to 2$ or other possibilities. The only information we really have in this study is the state of the patient at each visit. With this information we no longer have any complete information. Therefore we have right-, left-, and interval-censored observations for several possible transitions.

It is sensible to assume as little as possible about the transition times. One criticism of the earlier analyses of this data is that all of them assumed the transitions occurred at the time of the visit. For visits close in time this may be reasonable, but if the visits are spaced far apart this assumption breaks down. When dealing with decisions of human health it is good practice to make as few assumptions as possible, to give a more conservative analysis, and ensure the decisions are based on sound analysis of the data.

To parameterize our model we now look at a few different factors. In the previ-

ous chapters we were fairly unconcerned which distributions had closed form complex LTs; we simply selected the one that seemed to fit best. We must be cautious in choosing distributions with no closed form complex LT, since the likelihood function requires significant computation for incomplete data situations. If the data clearly indicate a distribution with no closed form complex LT is superior to other alternatives we can use it, but with a significant computational penalty. Thus if there are other reasonable alternatives with closed form complex LTs we should use them to save computation time.

For a model that incorporates incomplete data, if we allow the error term in (4.8) to have the Gumbel, normal or logistic distributions, then none of these models will have closed a form complex LT. This is a computational problem; since we have incomplete data we need to compute and invert complex LTs to calculate the likelihood. If the likelihood function takes too long to compute, our analysis will be overly difficult. So for the sake of illustration we conveniently choose a different parameterization of an accelerated failure time model. For each transition i, let

$$T_i = \exp \quad x\,\beta_i + e_i \quad , \tag{5.1}$$

where $e_i \sim gamma(\alpha_i, g(\alpha_i))$. We choose the function $g(\alpha_i)$ such that the median of e_i is 1. This forces $\log(T_i)$ to have a median of $x\,\beta$. With this parameterization we know $T \sim gamma(\alpha, g(\alpha)\exp(-x\,\beta))$. This model is equivalent to an AFT, since we control the median through $x\,\beta$ and our variance though α (instead of σ). We also have a closed form complex LT for T_i which is necessary to compute this model in a reasonable amount time. The priors we choose are $p(\alpha_i) \sim gamma(2, 1\ \ 2)$ and $p(\beta_{ij}) \sim normal(0, 16)$ for $i = 1, \ldots, 5$ and $j = 0, 1$.

Of the 277 patients we have 41 unique sequences of transitions. For example, 17 patients were observed in state 2 on their first visit and in state 2 on their second, and state 3 at the final appointment, after which we have no additional information. Under our assumptions, the first observation tells us the $1 \to 2$ transition took less

than the time from diagnosis to the first study visit. We know that the observation stayed in state 2 at least as long as the time between the first and second visits. We also know that we transitioned to state 3 before the time of the third visit. One of the strongest ways we could apply this information is to use it as a left-censored observation of the first passage from state 1 to state 3. However, to save computation time, we use this information as an interval-censored observation from the direct transition from state 2 to state 3. One must weigh the computational burden versus the information the observation provides when choosing how to incorporate each observation into the likelihood function.

Table 5.2: Incorporation of the diabetic retinopathy data for the model in (5.1). RXY, represents a right-censored observation from state X to Y, "*" indicates a first passage distribution, L or IC a left- or interval-censored observation respectively.

State sequence	Number of observations	Likelihood contributions			
11	51	R12			
12, 23	21, 6	L12			
21	5	L21			
22	75	L12	R23		
24	1	L12	L34		
32, 33	2, 7	L12	R34		
34	1	IC14*			
112	11	IC12			
121, 122	4, 16	L12	R23		
133	1	L23	R34		
211	7	L21	R12		
212	7	L21	L21		
213	3	L12	L23		
221	2	IC21			
223	17	IC23			
224	1	L12	R23	L34	
232	1	L12	L32		
233	2	L12	R34		
322	2	L32	R23		
334	1	IC34			
1121, 1122	1, 5	IC12	R23		
1123	1	R12	L23		
1133	2	IC12	R34		
2122	1	L21	R23		
2212	2	L12	R23	L12	
2213	2	IC21	L23		
2232, 12232	1, 1	L12	R23	L32	
2233	7	IC23	R34		
2234	1	L12	R23	L34	
3233	1	L23	L23	R34	
12123	1	L12	L12	R23	
21221	1	L21	IC21		
21321	1	L21	L23	L21	
22323	1	L12	R23	L32	R23
22332	1	L12	R23	IC32	
223233	2	IC23	L23	RC34	
233233	1	L12	IC32	R23	R34

In Table 5.2 we show what information we use for each patient. We must determine which transition(s) each of the 41 distinct sequences contributes to. It would not take the reader long to discover that we did not treat all the information in the most optimal way but also took computation time into consideration. We took special care not to use any of the information twice; however, we used the fact that we are modeling a semi-Markov process so sojourn times are independent. In our analysis we tend to err on the side of quick computation and sacrifice some information. However, if a practitioner needs all the data that is possible, where computation time is not a problem, this could be modified to keep as much of the data as possible.

For the MCMC we again use Gibbs sampling with random walk Metropolis steps. Tuning the Metropolis step is time consuming because the likelihood is slow to compute. The mixing is improved by subtracting the mean value of $\mathbf{H}\text{b}\mathbf{A}_1$ from the respective covariate for each observation. We have suitable mixing and obtain 10,000 posterior samples after burn-in.

For this model, prediction is again clear-cut for fixed values of $\mathbf{H}\text{b}\mathbf{A}_1$. We calculate four PPDs at set levels of $\mathbf{H}\text{b}\mathbf{A}_1$; we choose values of 8, 11, 14, and 17. We can use the quantiles from these PPDs to compare with our previous analysis with time-varying covariates. We would expect the PPDs in the current analysis to have larger variances because we are making fewer assumptions and therefore get less information from the data. Table 5.3 displays the PPD quantiles which can be compared to Table 4.2. The comparison between the two tables gives us some interesting information. We see that the predicted probability of blindness for low levels of $\mathbf{H}\text{b}\mathbf{A}_1$ is much smaller in this analysis than the first. The opposite is true but only slightly greater for high levels of glycohemoglobin. This makes sense since in the first analysis we were assuming an individual had just transitioned into their current state at the beginning of the study. In the second analysis we only assumed everyone started in state 1 at the time of diagnosis of diabetes. This analysis indicates that $\mathbf{H}\text{b}\mathbf{A}_1$

Table 5.3: Predicted probabilities of blindness due to diabetic retinopathy before various points in time (years).

	$\mathbf{H}\text{b}\mathbf{A}_1$ level			
Years	8	11	14	17
5	0.0000	0.0004	0.0066	0.0312
10	0.0004	0.0036	0.0415	0.1029
15	0.0010	0.0083	0.0727	0.1408
20	0.0014	0.0121	0.0997	0.1596
25	0.0020	0.0165	0.1151	0.1677
30	0.0024	0.0195	0.1273	0.1734
35	0.0029	0.0242	0.1351	0.1761
40	0.0037	0.0278	0.1423	0.1773
50	0.0043	0.0340	0.1499	0.1810
75	0.0064	0.0479	0.1575	0.1828

levels have a stronger connection to diabetic retinopathy than the first analysis. We can also see from the comparison of the two analyses at the 75 year mark that the predicted probabilities of blindness closely agree for higher levels of $\mathbf{H}\text{b}\mathbf{A}_1$.

In this analysis we have demonstrated that we can successfully make fewer assumptions about the transition times and still get meaningful predictive information. This method of using SFGMs to find the contributions of incomplete data for the likelihood function can be used in semi-Markov models. We again resume the discussion of the simulated recurring illness process example, but now including incomplete data.

5.4 Simulated recurring illness process example (continued)

We demonstrate the techniques for handling incomplete data with our simulated example. Refer again to the semi-Markov process in Figure 5.3. We use the model

as found in Equation 2.8, with a gamma as the distribution for the $0 \rightarrow 1$ transition, the inverse Gaussian for the $1 \rightarrow 0$ transition and another gamma for the $1 \rightarrow 2$ transition. We use a vague prior, $Exp(1\ \ 20)$, for all parameters, except for p (the probability of transitioning to state 0 given the process is in state 1) for which we use a flat prior.

We use the data included in Table 2.1, but now we include the three observations that have been left out of the previous analyses and treat them as incomplete data. The three randomly selected observations to be incomplete are observation 1, 14, and 15. We assume observation 1 reached state 2 before time $t = 5$ so the likelihood contribution is $F_{02*}(5)$. We saw observation 14 in state 1 at time $t = 4$, and have no other information, so the likelihood contribution is $(1 - F_{02*}(4))$. Observation 15 was in state 0 at time $t = 3$ and then in state 2 at time $t = 14$, so the likelihood contribution is $(F_{02*}(14) - F_{02*}(3))$.

We use Gibbs sampling with random-walk Metropolis steps as our MCMC method. To improve the MCMC mixing we reparameterize both gamma distributions to parameters that reflect the mean and variance (we keep the same vague priors for these new parameters). This reduces the correlation between the two parameters. We obtain 20,000 samples from which we use to get a PPD. In Figure 5.5 we can see a histogram of the data, the estimated PPD, and the true distribution that the data were generated from. It appears the PPD is performing well. We were able to incorporate three incomplete observations and complete the MCMC in about an hour. These techniques seem quite effective in this example.

5.5 Summary

We have demonstrated that SFGMs provide a excellent tool in incorporating incomplete data into multistate semi-Markov models. Although this increases the model

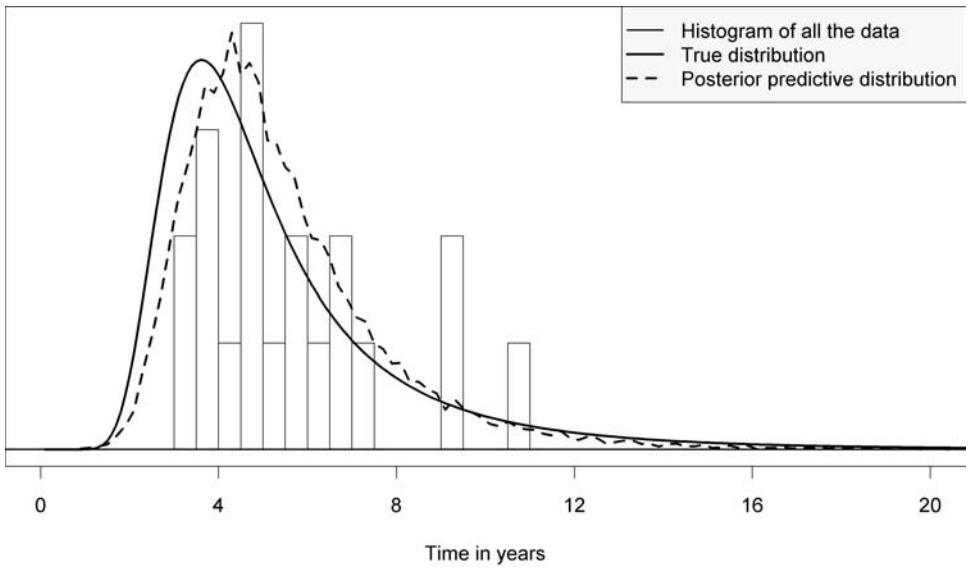


Figure 5.5: A histogram of the first passage from state 0 to state 2 (or total) data from Table 2.1 with a plot of the true distribution, and the estimated PPD.

computation time, it removes some restrictive assumptions. Because the cost of obtaining observations usually is much greater than computation time, the trade-off is justified to more fully use all the data collected. In our examples we demonstrated the basics of how to use SFGMs to incorporate incomplete data into the likelihood function, but there are situations in which most if not all of the data are incomplete. Most medical studies have some type of incomplete data. For example, when a patient sees a medical practitioner about a condition, the practitioner usually does not continuously observe the patient and only gets to see the patient at several snapshots in time. The transition times from one state to another are often treated as known on the date/time they were documented although this is not the case. This treats the inherently incomplete data as complete. Because of the sparseness of options to model incomplete data, most often incomplete data are ignored or handled with unrealistic assumptions. We provide a method to incorporate incomplete data into a model with fewer assumptions. This is a major step in modeling medical and

reliability data.

Chapter 6

Assessing model goodness-of-fit

In complicated statistical models it is often difficult to assess the model assumptions. Numerous methods attempt to solve the problem of variable selection, but these do not address the equally important issue of model adequacy. Naturally, with only one model we should determine if the model is adequate. Similarly, if we have 10 competing models, and use some rule to determine which is best, we should still determine if this "best" model is an adequate representation of the process of interest. The most popular variable selection methods such as Akaike information criterion (AIC) (see Akaike (1974)), deviance information criterion (DIC) (see Spiegelhalter et al. (2002)), and other similar criteria help determine a "best" model relative to other models, but none of them indicate if the model is reasonable with respect to the data. In this chapter we derive some goodness-of-fit results that can be applied to many statistical models, not just SFGMs.

Checking model adequacy is a difficult task. A variety of methods are available but none of them are completely satisfactory. The Pearson's chi-square statistic is a very popular method, however, this method is only valid if there is a sufficient amount of data. Using it with too small a sample produces inaccurate results. Lehmann

and Romano (2005) states that using the chi-square statistic on continuous data is hampered by the loss of information by grouping the data. Others such as the Kolmogorov-Smirnov (K-S), Anderson-Darling (A-D), or Cramér-von Mises (C-vM) tests are versatile, but difficult to use in complex models. In this chapter we suggest an A-D type criterion which is easier to interpret for a variety of models. Our primary goal is to develop a model selection and validation criterion that can be used for SFGMs.

This chapter introduces a method that helps determine if our model adequately represents the data. As with any stochastic model we know with almost certainty that the data were not generated by our model. We are trying to evaluate if our model is reasonable or useful, even though we know it is wrong (see Box and Draper (1987, pp. 424)). The proposed model assessment criterion is valid for small or large samples but is limited to continuous-time models. This is not a problem for SFGMs since they are continuous time models. This criterion is also applicable in both Bayesian and frequentist frameworks.

We start by developing the rationale for our proposed fit criterion. Then we expand it to handle more complex situations. Throughout each section we give examples of how the method can be applied.

6.1 Derivation of a new goodness-of-fit criterion

We propose a statistic that can be used as a fit criterion, but could also be adapted as a statistical test. We use the word *criterion* versus *test* to emphasize that the primary method we are advocating is not a formal statistical test, but a rule of thumb which allows us to gauge if the proposed model is reasonable. An obvious goal is to make model assessment as simple as possible. This is difficult because of complications in the data such as missing information and multiple states. We admit

that some models will be difficult to assess regardless of the goodness-of-fit method that is used.

The method we advocate is not a statistical test. Our focus is on model adequacy. We know we will not propose the correct model, but can we determine if a model is good enough to use for making predictions? A statistical test gives us a p-value, but it does not really tell us if our model is reasonable; it only tells us if we have enough evidence to reject our proposed model. If we do not have enough evidence to reject our model is it reasonable? Statistical testing does not answer this question.

As with many tests our criterion measures the distance between the observed data and the expected value of that observation given the model. A convenient way to put all models in the same framework is by transforming our data to $U(0,1)$ random variables through the probability integral transform. As mentioned before, any proposed model is almost certainly wrong, so if F is the proposed model, we know $F(X_i)$ will not truly be a $U(0,1)$ random variable. But, if we determine that $F(X_i)$ is approximately $U(0,1)$ we may be able to say our model is reasonable. We use similar concepts from the A-D and C-vM tests. First we apply the probability integral transform and then order the transformed sample. Assuming our model is correct then $F(X_i) \sim U(0,1)$ and the distribution of the i^{th} order statistic of a $U(0,1)$ is $F(X_{(i)}) \sim beta(i, n-i+1)$ (see Casella and Berger (2002, pp. 230)).

With this information we can assess the distance of the data, $F(X_{(i)})$, from its expected value, $i\ \ (n+1)$. A simple way of doing this for a sample of size n is to use the squared distance, where

$$Q^* = \frac{1}{n} \sum_{i=1}^{n} \left(F(X_{(i)}) - i\ \ (n+1) \right)^2 .$$

One problem is the distribution of Q^* is difficult to find exactly. However, when using the squared distance we find that the expectation of $\left(F(X_{(i)}) - i\ \ (n+1) \right)^2$ is just the variance of $F(X_{(i)})$, when $X_i \sim F$.

Therefore if we redefine our statistic to be

$$Q^* = \frac{1}{n} \sum_{i=1}^{n} \frac{\left(F(X_{(i)}) - i \ \ (n+1)\right)^2}{Var\left(F(X_{(i)})\right)},$$

then $E[Q^*] = 1$. This is convenient because if we get a value of Q^* that is less than one, then we can say the distance between our proposed model and the data is smaller than the average distance between the true model and a sample from it (as measured by Q^*).

This Q^* seems reasonable, but compared with a form of the A-D statistic it actually performs quite poorly with regard to "statistical power". We are abusing terminology here, because we are not performing a statistical test. However, we use the term "statistical power" loosely as a description of how often the statistic can detect a false model. The A-D statistic performs better because it is actually using the data more than once. It uses the fact that if $F(X_i) \sim U(0,1)$ then $1 - F(X_i) \sim U(0,1)$. We use this property to get a modified statistic Q^{**},

$$Q^{**} = \frac{1}{2n} \sum_{i=1}^{n} \frac{\left(F(X_{(i)}) - i \ \ (n+1)\right)^2 + \left(1 - F(X_{(n-i+1)}) - i \ \ (n+1)\right)^2}{Var\left(F(X_{(i)})\right)}.$$

However, we find that $Q^{**} = Q^*$ for any sample. Therefore this reflection principle does not add any information because $F(X_i)$ and $1 - F(X_i)$ have a correlation of -1. For the reflection principle to help, we must transform the two samples. To improve our statistic we do the same as the A-D test, by using the negative logarithm transformation on $F(X_{(i)})$ and $1 - F(X_{(i)})$. Once again we need to find the expectation and variance of the transformed random variable $-\log(F(X_{(i)})$. Let $X \sim beta(\alpha, \beta)$ and $Y = -\log(X)$, therefore

$$f_Y(y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \left(e^{-y}\right)^{\alpha} \left(1 - e^{-y}\right)^{\beta-1}.$$

From here we can find the MGF $M_Y(s)$, take the first derivative with respect to s, and evaluate this at $s = 0$. This provides $E[Y] = \ _0(\alpha+\beta) - \ _0(\alpha) = \ _0(n+1) - \ _0(i)$,

where

$$_i(x) = \frac{d^{i+1}}{dx^{i+1}} \left[\log(\Gamma(x))\right].$$

Similarly, we find $Var[Y] = \ _1(\alpha) - \ _1(\alpha+\beta) = \ _1(i) - \ _1(n+1)$.

The mean and variance formulas may seem a little complicated, but another way to look at them is through the recurrence relation of the gamma function. We know $\alpha\Gamma(\alpha) = \Gamma(\alpha+1)$. Applying the logarithm to both sides and differentiating gives the recurrence relation, $\ _0(\alpha) + 1 \ \alpha = \ _0(\alpha+1)$, and differentiating one more time produces another recurrence relation, $\ _1(\alpha) - 1 \ \alpha^2 = \ _1(\alpha+1)$. Therefore the mean of $-\log(F(X_{(i)})$ is just $\sum_{j=i}^{n} 1 \ j$ and similarly the variance is $\sum_{j=i}^{n} 1 \ j^2$. These mean and variance formulas are very easy to calculate and are more intuitive then the $\ _i(x)$ function. Therefore, we call our statistic Q, and write it in its final form

$$Q = \sum_{i=1}^{n} \frac{\log\left(F(X_{(i)})\right) + \sum_{j=i}^{n} \frac{1}{j}\ ^2 + \ \log\left(1 - F(X_{(n-i+1)})\right) + \sum_{j=i}^{n} \frac{1}{j}\ ^2}{2n \sum_{j=i}^{n} \frac{1}{j^2}} \quad . \quad (6.1)$$

For the remainder of the paper we will refer to this statistic we have developed as Q.

Now we have a statistic that can distinguish a "false" model with similar statistical power as the A-D test, but has the property $E[Q] = 1$, regardless of the complexity of the model. We compare the A-D, C-vM, and the Q statistics in a few situations. All three statistics are expecting data from a $U(0,1)$ distribution since we are assuming F is the true model. We sample from a $U(0,1)$ 100,000 times to get an approximate 95% critical value. Then, we sample from three different beta distributions and see how often each of these statistics can detect the false models. We selected the beta distributions so that one has too heavy of tails, another has tails that are too light, and the final one has a heavy left tail with the right tail too light. The values of these simulations are given in Table 6.1. This verifies that in these generic situations the Q statistic competes favorably with some of the established goodness-of-fit statistics. From this table we can see that the C-vM test aggressively

Table 6.1: Simulated rejection percentages for 3 samples sizes and 3 different distributions for the Anderson-Darling, Q, and Cramér-von Mises Statistics.

Distribution	A-D	Q	C-vM
	$n = 10$		
$beta(0.8, 0.8)$	9.6%	10.4%	14.1%
$beta(2, 2)$	1.2%	1.5%	0.0%
$beta(0.9, 1.1)$	8.6%	8.5%	5.0%
	$n = 30$		
$beta(0.8, 0.8)$	11.6%	11.9%	24.6%
$beta(2, 2)$	11.4%	15.4%	0.0%
$beta(0.9, 1.1)$	15.6%	15.5%	5.2%
	$n = 100$		
$beta(0.8, 0.8)$	19.5%	19.4%	51.7%
$beta(2, 2)$	91.1%	93.8%	0.0%
$beta(0.9, 1.1)$	40.4%	40.2%	5.2%

identifies false models that do not adequately account for heavy tail behavior in the data. However, this test neglects to identify false models that suggest too heavy of tails that are not justified by the data. The A-D test and the Q statistic perform similarly.

So not only does Q have good statistical power but it is easy to interpret. An intuitive way to interpret Q is, if $Q < 1$, we can be fairly satisfied that we are modeling the process adequately. If $Q \geq 1$, we may want to consider an alternative model. This is simple and informal, but effective. Therefore, if $Q < 1$, then our proposed model has a smaller value of Q than the average value of Q for the true model.

If one chooses to use Q as a Fisherian test, then Table 6.2 provides estimated values to do so. Under the null hypothesis we assume our proposed model F is the true model. Therefore, $F(X_i) \sim U(0, 1)$ and we simulate n, $U(0, 1)$ random variables and calculate Q. We do this 1,000,000 times to get samples from Q which gives the approximated quantiles in Table 6.2. We emphasize that these values are only valid

Table 6.2: Approximated values of the statistic Q given the model (no estimation)

n	Mode	Mean	Median	75%	90%	95%	99%
5	0.32	1.00	0.63	1.21	2.19	3.06	5.54
7	0.35	1.00	0.67	1.24	2.14	2.92	5.18
10	0.39	1.00	0.69	1.24	2.08	2.81	4.81
13	0.41	0.99	0.71	1.24	2.04	2.72	4.61
16	0.43	1.00	0.73	1.24	2.03	2.71	4.49
20	0.44	1.00	0.73	1.25	2.02	2.66	4.37
25	0.45	1.00	0.74	1.25	2.00	2.64	4.32
30	0.46	1.00	0.75	1.25	1.99	2.61	4.21
40	0.46	1.00	0.75	1.24	1.97	2.57	4.10
55	0.47	1.00	0.76	1.25	1.96	2.55	4.06
75	0.48	1.00	0.76	1.25	1.96	2.55	4.04
100	0.49	1.00	0.76	1.24	1.95	2.52	3.99
150	0.49	1.00	0.77	1.25	1.95	2.52	3.98
200	0.49	1.00	0.77	1.25	1.94	2.53	3.95
1000	0.50	1.00	0.77	1.25	1.93	2.50	3.88

if the null hypothesis does not test if the data are from a parametric distribution with unknown parameters. Next, we generalize Q to apply in situations that are more complex.

6.2 Finding Q with censored data

Frequently, observational data have missing information. Often this comes in the form of censored data. For this model fit criterion to be useful in practice it must incorporate censored data.

In Chapter 3 we briefly introduced censored data. Throughout this dissertation, "censoring" refers to random censoring. As long as the censoring can be assumed to be random, we can use the techniques discussed in this section. Since right- and left-censoring are special cases of interval-censored data with out loss of generality,

we demonstrate the following methods on interval-censored data.

Let $\mathbf{D}$ represent the data when no censoring is present and let $\mathbf{D}^*$ be the data when censoring is present. We know that $Q|\mathbf{D}$ is a fixed constant, but when censoring occurs, $Q|\mathbf{D}^*$ is not a fixed constant but still random. Even though $Q|\mathbf{D}^*$ is random, we can use information about its expected value. We generalize the fitness criterion and use the expected value of Q as the statistic. Let $Q = E[Q|Data]$, whether the data are censored or not. If there is no censoring in the data $Q = E[Q|\mathbf{D}] = Q$. It is very difficult to calculate Q, so we need to find a way to estimate it. If F is the true model then the random part of the i^{th} censored observation is uniformly distributed on the interval $(F(a_i), F(b_i))$, where this observation is censored on the interval (a_i, b_i). This makes sampling from $Q|\mathbf{D}^*$ very easy. So if we simulate m samples from $Q|\mathbf{D}^*$ we can estimate Q using the mean of these samples. This estimate can be made as accurate as desired by increasing the sample size m. To approximate the accuracy of our estimated value of Q, if m is sufficiently large, we can apply the central limit theorem (CLT) and obtain a $100(1-\alpha)\%$ CI. Calculating a CI for Q takes only a few seconds on a desktop PC if $m = 10,000$.

By assuming F is the true model, we can use the same rule of thumb for Q as we did for Q. Clearly, by the rule of iterated expectations $E[Q] = E[E[Q|\mathbf{D}^*]] = E[Q] = 1$. If $Q < 1$, we should be satisfied our model is fairly reasonable.

Recall the construction engineering example from Chapter 3. There were 20 observations, however four companies were not measured on the $2 \to 3$ transition. These four observations are considered to be right-censored, if we are considering the model for the first passage time from state $0 \to 3$. The parameterization in this example was very arbitrary. We chose to model each transition with a Weibull distribution to demonstrate additional capabilities, but not necessarily to build a good model. How do we determine if the fitted model was adequate? Analyzing the frequentist model (with the MLEs) we simulate 1,000 samples from $Q|\mathbf{D}^*$. A 99%

Table 6.3: Values of Q for the construction engineering example

Distribution	$0 \to 1$	$1 \to 2$	$1 \to 3$	$2 \to 3$
gamma	1.847	1.297	0.285	1.223
Birnbaum-Saunders	1.856	1.074	0.362	1.212
lognormal	1.856	0.986	0.364	1.214
Weibull	1.955	1.152	0.261	1.156
inverse Gaussian	1.856	0.931	0.406	1.214
Fréchet	2.106	0.700	0.445	1.066
exponential	17.501	1.005	0.228	6.814

confidence interval for Q is $(1.530, 1.588)$. If this were a fixed model and we were conducting a statistical test Table 6.2 suggests we would not reject this at the 90% confidence level. However, using Q as a rule of thumb indicates we could probably find a better model for this process.

Can we find a better model using some of the popular lifetime distributions? We try using the Weibull, lognormal, gamma, inverse Gaussian, exponential, Fréchet, and the Birnbaum-Saunders distributions. From these seven candidate distributions (each evaluated at the MLEs), we choose the one that has the smallest value of Q. The values of Q can be found in Table 6.3. For the $0 \to 1$ transition the gamma seems to be the best fit. The Fréchet fits the $1 \to 2$ transition the best. The exponential models the $1 \to 3$ transition well, and we pick the Fréchet for the $2 \to 3$ transition. With this new parameterization, a 99% confidence interval for Q is $(1.089, 1.129)$. This is an improvement, but a more sophisticated method is needed for a more appropriate fit of this data, especially for the $0 \to 1$ transition.

This example shows how to find an estimate of Q when random censoring is present in the data. By finding the mean of the samples of Q we get an accurate estimate of Q. We believe similar techniques can be applied to Q when other types of missing data are present.

6.3 Finding Q in multistate aggregated data models

Some models do not work in the format we have developed thus far. An example is in reliability theory. Consider a system that has several components. Data might be available for a number of individual component tests, but there may not be many full system tests. Even though each component has plenty of data, we do not have a way to assess the fit for this full system model. So we would like to determine how well our overall system handles the data while using all of the component data. We label this type of data as *aggregated data*.

Definition An aggregated data model is a multistate model with data for many subjects observed on single transitions, which do not have corresponding data for other transitions.

Attempting to find Q for each component and then using these to develop an overall Q for the full system is problematic. Suppose we have a system with two components each with a value of Q; this information tells us little about the value Q for the overall system model. For example, if the true unconditional failure-time distribution of the two components is $Exp(\lambda_1)$ and $Exp(\lambda_2)$ respectively. We know the system fails if either component fails, which implies the true overall model for time to failure is an $Exp(\lambda_1 + \lambda_2)$, the minimum of an $Exp(\lambda_1)$ and $Exp(\lambda_2)$. We set $\lambda_1 = 3$ and $\lambda_2 = 1$ and use maximum likelihood estimation. Let Q_i be the value of Q for each component and Q_o be the Q for the overall system. In one simulation we have $Q_1 = 0.49$, $Q_2 = 0.14$ and $Q_o = 0.57$ and in the next we get $Q_1 = 0.81$, $Q_2 = 0.52$ and $Q_o = 0.44$. This simple situation demonstrates that the value of Q_o can fluctuate dramatically given the individual Q_is.

The solution we propose is similar to dealing with censored data. Again, using the fact that $E[Q] = 1$, we can randomly match component data together to have pseudo full system tests and then calculate Q. This is essentially bootstrapping (see Efron and Tibshirani (1993)) full system tests from the population of component tests. If we do this many times and average our values of Q we can again expect to have $E[Q] = 1$, if our model is correct.

Referring again to the same two component system from above, assume we have n_1 samples from component 1 and n_2 from component 2, with $n_1 < n_2$. Now take all the samples from component 1 and pair them randomly (with replacement) with observations from component 2. This gives us n_1 pseudo full system observations, and the remaining $n_2 - n_1$ observations from component 2 are treated as left-censored observations. With these observations Q can be calculated. If this is done many times, the average of all the sampled Q's is an estimate of Q and using the CLT we can again get a $100\%(1-\alpha)$ CI.

This method can present a significant increase in computation, especially if the CDF, F, for a given data value is difficult to calculate, then it will be hard to implement the method given above. However, for a fixed model we can calculate F at many values on the support and then interpolate needed values of F very quickly.

If there are censored data in an aggregated data model this becomes a little more complicated. Consider the case where (a_1, b_1) is an interval-censored observation from component 1 and is matched with another interval-censored observation (a_2, b_2) from component 2. If these intervals are not disjoint what value should we use for the pseudo full system test? Let F_1 be the distribution assigned as the failure time for component 1, and similarly F_2 for component 2. We then sample from the two intervals $(F_1(a_1), F_1(b_1))$ and $(F_2(a_2), F_2(b_2))$ and use the minimum value evaluated under F as an observation used to calculate Q. Using this method allows us to find a CI for Q even when we have a situation with aggregated and censored data.

SFGMs lend themselves naturally to aggregated data models. It is very important that Q be able to assess models with aggregated data.

6.4 Bayesian models

Since many flowgraphs are developed in a Bayesian setting we must consider Q for Bayesian models. The Q developed earlier hinges on the fact that the parameters are fixed for a given model. In Bayesian parametric models the parameters are *random*.

To handle Bayesian models we take an approach suggested in Box (1980). Consider a Bayesian model where $f(x|\theta)$ is the sampling distribution, $p(\theta)$ is the prior distribution, x is a single observation and θ is a vector of the parameters. Then we have

$$m(x) = \int f(x|\theta)p(\theta)\, d\theta,$$

which is a fixed model with no parameter estimation (if there are no random hyperparameters in the prior).

We use $m(x)$ as our proposed distribution with a given sample of size n. We find

$$F(x_i) = \int_{-}^{x_i} m(t)\, dt, \tag{6.2}$$

and use Q in the same manner as in the frequentist framework. However, in most cases we do not have an analytic solution for $F(x_i)$. If we obtain a large sample from the prior distribution of size n_p, we get the estimate,

$$m(x) \approx \frac{1}{n_p} \sum_{i=1}^{n_p} f(x|\theta_i)p(\theta_i).$$

With an approximate $m(x)$, we can replace the integral in (6.2) with a summation to find an approximate $F(x_i)$. We partition the support of $f(x|\theta)$ sufficiently such

that each observation x_i is a right endpoint of a partition. The number of partitions less that x_i is n_i where the width of each partition is w_j. Our approximation is

$$F(x_i) \approx \sum_{j=1}^{n_i} w_j \frac{1}{n_p} \sum_{k=1}^{n_p} f(x_j|\theta_k)p(\theta_k). \tag{6.3}$$

Box's method has drawn some skepticism from the Bayesian community because it is dependent on the prior distribution. It is true that $m(x)$ is very dependent on the prior; in fact, for diffuse priors the suggested fit is usually very poor. However, the mathematics clearly indicates that $m(x)$ is exactly the distribution one is assuming under the model. We show how influential a prior can be on $m(x)$ in an example to follow.

An algorithm to estimate Q in a Bayesian model is:

1. Define $f(x|\theta)$ and $p(\theta)$
2. Obtain a sample from the prior distribution, $p(\theta)$
3. Determine the points x_i, on the support of x, for sufficient resolution and accuracy
4. Calculate $m(x_i)$ and $F(x_i)$ for each x_i
5. Calculate Q using the frequentist methods

Consider a simple example where $f(x|\theta)$ has an $Exp(\theta)$ distribution with a conjugate prior $p(\theta) \sim gamma(a, b)$. Then

$$m(x) = \int_0^\infty (\theta \exp\ -\theta x\ ) \left(\frac{b^a}{\Gamma(a)} \theta^{a-1} \exp\ -b\theta\ \right)\ d\theta = \frac{ab^a}{(b+x)^{a+1}},$$

and

$$F(x) = \int_0^x m(t)\ dt = 1 - \left(\frac{b}{b+x} \right)^a.$$

Using this $F(x)$ we can find an estimate of our statistic Q.

We use this example to demonstrate how our prior influences $m(x)$. Consider trying to model the lifetime of an electronic component and it appears the only information we have about the component is our test data. We choose to use a diffuse prior, $p(\theta) \sim gamma(1, 1\ \ 50)$. This prior on θ implies we basically have no idea about the value of the mean time to failure (MTTF). Let the true component failure time distribution be a $Weibull(0.9, 15)$ with a mean of about 15.78 years. Therefore with this vague prior, $m(x)$ looks very flat, almost identical to the prior. Choosing this vague prior means we are deliberately choosing a very poor distribution to model the data. In fact this prior is saying that we are 98% certain that the MTTF is less than one year! We should be able to use some information about other electronic components similar to this new one that we are testing. We know that the old component which was used for the same purpose had a MTTF of 12 years. We believe the new component should have a higher MTTF because the manufacturing process has improved. Using this information we develop a more informative prior for θ and say we are quite certain the MTTF is between 10 and 25 years. We can choose $a = 35$ and $b = 513$ to roughly satisfy these conditions. In Figure 6.1 we see that $m(x)$ with informative prior information is more appealing. Often it is difficult to get informative prior distributions in complicated models. Bedrick et al. (1996) provides a systematic way to obtain prior distributions for regression parameters, similar ideas can be applied to other situations.

If one insists on using vague priors, Q will not be of much use. Although we do not explore them, there are other alternatives to find $F(X_i)$ in Bayesian models. Other ideas about goodness-of-fit testing in Bayesian reliability models can be found in Hamada et al. (2008).

One item worth noting in a Bayesian setting is that in some cases modelers may want to relax the rule of thumb that $Q < 1$. The primary reason for this is if the prior

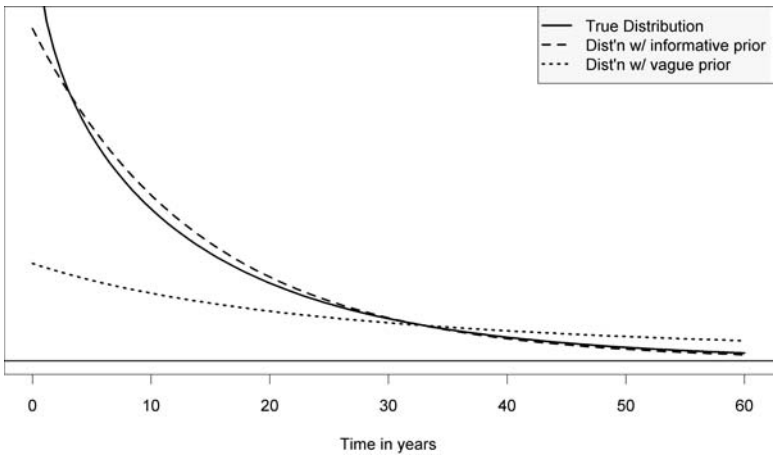


Figure 6.1: An example of a sampling distribution with an informative prior and a vague prior.

information is somewhat contrary to the data, the modeler may be satisfied even if the model does not fit the data extremely well, because the prior has influenced the model away from the data. This could be the case with strong prior information and where only a small sample is available. However, frequentists should not interpret Q as leniently because the data is the only information they are willing to formally take into account.

6.5 Penalizing Q for estimated parameters

If we have more than one proposed model we also need to adjust Q for the number of parameters that are being estimated. If we use many parameters, in theory we could force our model to have an arbitrarily small value of Q. Obviously, if we did this we would be over-fitting the data. To ensure we do not favor over-fitted models we penalize Q for models that estimate parameters. Assume we have p parameters in our proposed model, then we adjust Q by adding the term $p\ \ (p+1)$. We use the

notation

$$Q_p = Q + \frac{p}{p+1}$$

to indicate that we have adjusted for the estimation of parameters. Then we use this particular penalization term to prevent blatant over-fitting of the data.

Now that we have penalized this criterion we can use Q_p simultaneously for both model selection and model adequacy. If $Q_p < 1$ we assume the model is reasonable and proceed with prediction. If more than one model has a value of $Q_p < 1$ we select the model with the smallest Q_p value.

So how does the criterion $Q_p < 1$ perform? If we assume a true model, simulate a sample from it, and use Q_p to select a model, how good will our predictions be? In practice we have data from a true model and we propose a false model. If this false model is close enough to the ECDF we declare that it is adequate and use this false model for prediction. We give a simulated example to demonstrate the benefits of using Q_p.

Consider the process with distribution $Weibull(2, 2)$. If we obtain a sample of size $n = 10$, fit a model, determine if the model is adequate with Q_p, and if it is, predict one observation from this fitted model. In this situation how good will our prediction be? Table 6.5 shows how this procedure performs in 50,000 trials and Figure 6.2 shows a histogram of the predictions of the two models compared to the true model. Clearly the gamma parameterization appears to be better than the lognormal, but for both if we use Q_p, we do better than if we use no model validation criterion. Using Q_p and the lognormal distribution we only accept 23.3% of the models whereas using the gamma we accept 30.4%. We also compare these parameterizations using Q_p with a crude non-parametric method of sampling from the ECDF (with an exponential tail). Even though we propose incorrect models (the gamma and lognormal) we are able increase our prediction accuracy.

Table 6.4: A simulation study using Q_p as a model validation tool.

	Median	Mean	Variance	Percent
Weibull(2,2)	1.665	1.772	0.858	
Gamma with Q_p	1.612	1.773	0.967	30.4
Gamma w/o Q_p	1.611	1.774	0.988	
Lognormal with Q_p	1.551	1.781	1.132	23.3
Lognormal w/o Q_p	1.541	1.822	1.782	
ECDF	1.653	1.841	1.470	

6.6 Simulated recurring illness process example (continued)

We return to the recurring illness process of Figure 2.4. In the Chapter 3 we tried a few additional parameterizations, which looked appropriate on the individual transitions but did not look as good for the overall model of the first passage from state 0 to state 2. These assessments were very subjective which involved just comparing the candidate distributions with a histogram of the data.

So now we would like to see if the criterion we have developed can assist in this example. We can start by looking at the two proposed models in (2.8) and (3.7). Both have similar parameterizations so it is useful to see how the model in (2.8) compares with the model in (3.7). The model in (2.8) has $Q_p = 0.868$, and the model in (2.8) has $Q_p = 0.933$. This confirms our conjecture in Chapter 3 that the model in (2.8) seemed more appropriate.

Can we improve our model by using Q_p to find the "best" distribution for each branch? Looking at each branch we find the lowest Q_p for each transition. We chose the gamma, Fréchet, and inverse Gaussian distributions for the $0 \to 1$, $1 \to 0$, and $0 \to 2$ transitions respectively. However, this new model does not get a lower overall value of Q_p, with $Q_p = 0.870$. We compare the three proposed models in Figure 6.3.

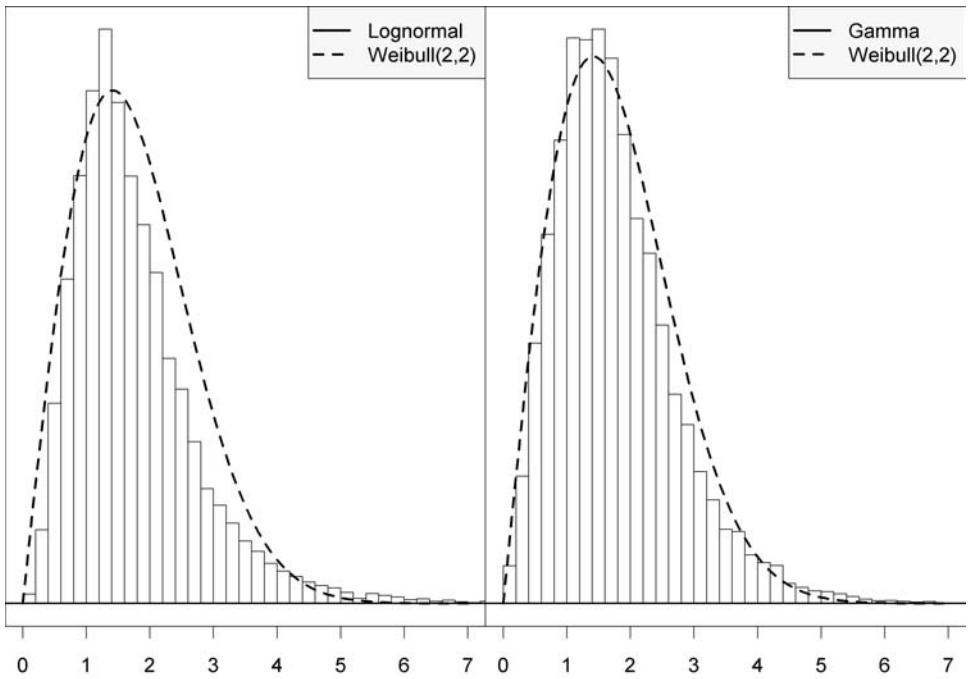


Figure 6.2: Histograms of two predictive models when Q_p is used, compared with the true model.

So what do we gain by using this fitness criterion Q_p? If we look at the plot and determine which is better, does Q_p help? We argue that it does, by providing a measure of fitness, which confirms what we see and may reveal information we do not see in a plot.

6.7 Summary

In this chapter we have introduced a general model fitness criterion that can be applied to a variety of models, including SFGMs. Model assessment is a critical aspect of any model. Using this criterion helps ensure that the model assumptions of a

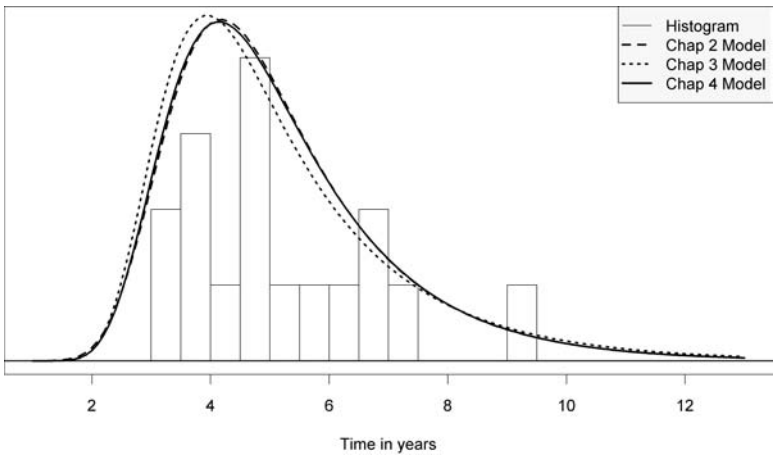


Figure 6.3: A histogram of the simulated recurring illness process data with plots of the models from chapters 2, 3 and 6.

SFGM are reasonable before implementation. We have proposed assessing goodness-of-fit using a criterion versus a statistical test. We do this with the justification that we know our proposed model is wrong, but want to know if it is adequate with the information we have. The criterion Q_p gives us some measure of model goodness-of-fit. We generalized the criterion for censored data and models with aggregate data.

The development of this criterion is still in its infancy. Other modifications and simulations must be run before we could clearly determine if this philosophy is reasonable and if this particular goodness-of-fit criterion is of practical use.

Chapter 7

Conclusions

In this dissertation we discussed several research topics involving SFGMs. This chapter summarizes the accomplishments of this work, then presents avenues for future research. We have

- Enabled SFGMs to be parameterized using any smooth distribution
- Provided a more general way to handle incomplete data in SFGMs
- Improved computation time in Bayesian SFGMs
- Demonstrated a Bayesian non-parametric method (MPTs) in SFGMs
- Incorporated covariates into the flowgraph framework using accelerated failure time models
- Presented the methodology to include time-varying covariates in semi-Markov models
- Suggested how to assess model goodness-of-fit using the Q_p criterion

7.1 Summary of results and contributions

In Chapter 3 we introduced the techniques to model any smooth distribution in SFGMs. This is a fundamental achievement that opens many additional modeling possibilities in SFGMs. Two primary distributions that had not been used in SFGMs were the lognormal and the Weibull. Now they can be fully incorporated in a SFGM, this will encourage the use of SFGMs in areas where these distributions are popular. With these techniques we can also find the CDF of a first passage time, which is critical for timely computation needed in chapters 6 and 5.

In Chapter 4 we looked at estimating the Bayesian posterior predictive density for flowgraphs in a new way. This provides a faster way to compute SFGMs in the Bayesian framework and increases the modeling flexibility. With this method we can use any time-to-event distribution without smoothness constraints. We demonstrate this capability in a Bayesian non-parametric model. This is the first application of Bayesian non-parametrics in SFGMs. Using this method of sampling from the PPD we also introduce a medical example with a time-varying covariate. This is a fairly complex model that may not be able to be implemented in the traditional Bayesian SFGM paradigm. This is a significant accomplishment as a general way to include time-varying covariates in semi-Markov models. Using this method to find the PPD enables significant new applications in Bayesian SFGMs. In most cases this method also bypasses the requirement to use MGFs or complex LTs in SFGMs. However, for the cases with incomplete data, we use the SFGM framework to produce the likelihood function. To date, SFGMs have dealt exclusively with continuous time processes because of the smoothness constraint. When inversion of MGFs is not required we have a way to deal with discrete time SFGMs. Removing the need for working in the MGF domain greatly expands the applications where SFGMs can be applied. We also demonstrate how covariates can be incorporated into SFGMs using the AFT model framework.

Chapter 5 extended previous work to model incomplete data. We provided a more general definition of incomplete data. We developed new techniques to model the diabetic retinopathy data from Chapter 4. This is an important contribution for medical and other types of research. Often a patient or system is only observed at a few discrete times. In the past several assumptions were necessary to model this type of data. Using the techniques from this chapter we can drop some of these assumptions and improve our inference and prediction from the data.

In Chapter 6 we proposed a new goodness-of fit criterion. This fit criterion is general enough to be applied to most univariate continuous models. The straightforward interpretation of the proposed fit criterion statistic Q_p lends itself to assess the fit of complicated multistate models such as SFGMs. This criterion can be applied in many situations and is simple when the CDF of the proposed model can be found at the data points. Q_p is applicable regardless of the sample size, because no asymptotic assumptions are made. This is a new and intuitive way to look at model goodness-of-fit. With this research we can now assess models and determine if the proposed model is reasonable. This addresses one of the most fundamental problems in statistics and is crucial in applied statistical models such as flowgraphs. There is a possibility that this statistic could be adapted to the multivariate case. The direct probability integral transform methods are not able to handle multivariate data, because it is not generally true that the distribution function F(X,Y) is not $U(0,1)$, even when F is continuous, see Genest and Rivest (2001). However, it may be possible to conduct a multivariate version of this fit criterion using the Rosenblatt transformation, proposed in Rosenblatt (1952).

This work has made significant contributions to state-of-the-art methods in statistical flowgraph models. Flowgraphs are inherently computational. As advances in processor speed and efficiency of algorithms advance these will naturally enhance the computational methods of SFGMs. Next, we discuss some open problems that

are of interest.

7.2 Future work

Statistical flowgraphs were introduced in literature about fifteen years ago. Many areas in flowgraphs are still not explored. This section focuses on open application areas and research problems in SFGMs. We briefly list several problems and explain why they are important.

One of the most critical theoretical areas is showing that SFGMs are a counting process. Of one the reasons for the success of survival models was the work of Aalen (1975) and Aalen (1978). Aalen showed that survival models are counting processes, thus providing the mathematical and stochastic foundation for survival models. If this could be done for SFGMs, the framework would receive wider acceptance and use in the academic community. Even though the flowgraph framework is essentially an applied method, providing the theoretical background is essential. This is a primary research area for SFGMs that would have many benefits once accomplished.

In Chapter 3 we showed how any smooth distribution with positive support can be used in a SFGM. To accomplish this we used the EULER method to invert the first passage complex LT to a density. This method is effective, but can be improved. Abate and Whitt (1995) recommends that when inverting the complex LT at many points, to use the fast Fourier transform (FFT) instead of the EULER method. This is exactly what we are doing when we use the EULER method. We discretize the support and invert the complex LT at each discretized point. Using the FFT instead of the EULER method would be computationally advantageous. Not only can the FFT be used for inversion, but it can also be used to find the complex LT. This would eliminate the need for numerical integration to compute the complex LT in cases such as the lognormal and the Weibull. From initial attempts of using the FFT

it appears that a standard FFT implementation many not work with SFGMs that contain loops and parallel paths. It may be possible to use the FFT for any smooth SFGM if the FFT was implemented with this in mind.

Now that covariates have been incorporated into SFGMs in Huzurbazar and Williams (2010), it is natural to extend these ideas. This could be done by including random effects into SFGMs. When there is a factor with many levels, such as, at what hospital an individual was treated, it is often difficult to incorporate and estimate too many of these coefficients into the model. One way to include this effect in the model is by treating it as a random effect that has a probability distribution. If we do this the only additional parameters that need to be estimated for that level are the parameters for the probability distribution we assigned to the random effect. Survival models such as this have been termed frailty models. Using this concept in SFGMs could enhance their modeling capabilities for survival times and in reliability.

One of the main assumptions in SFGM is the semi-Markov property. This is very restrictive in some situations. It may not be realistic to model two events a person has as being independent. If we do not have independence then the current SFGM framework is unable to properly model that process. Huzurbazar and Williams (2010) suggests a technique to model some dependencies between neighboring transitions. Computationally this may be implemented in simple situations but would be difficult if the correlation structure of the data was not trivial. A large area of work must be done in modeling covariance structures in multistate time-to-event data. Assuming a covariance structure of a multistate process would open up SFGMs to many more application areas where the semi-Markov property is not valid.

The techniques used in SFGMs can be used in other areas. The work done in Chapter 4 allows the use of any distribution, whether smooth or not, in Bayesian SFGMs. This could applied in areas such as dynamic linear models (DLMs), see West and Harrison (1997). DLMs usually assume normal error terms, which allows

for easy analytic and computational calculations, however it is possible to use other error terms. The SFGM methodology provides a way to compute the convolution of these random variables.

Once a SFGM is applied and prediction is made, is it possible to use one more observation obtained after the fact, without redoing the complete analysis? This may be possible using the Bayesian framework and sequential Monte Carlo (SMC) techniques (see Cappé et al. (2005)). This would also provide the methodology to implement on-line (real-time) SFGMs. These could be used in many applications such as real-time system monitoring in reliability, financial models, and environmental models.

Another area that could be explored is finding the extreme values of a flowgraph. One that may be of particular interest is what is the distribution of the shortest time to an event for an given number trials? This could be of some interest when attempting to model extreme failure events such as worst case scenarios in reliability and epidemiology. Other extreme events, such as in hydrology and climate, could be modeled with SFGMs if an extreme first passage distribution could be found.

Statistical flowgraphs are a natural fit for generalizing the project evaluation and review technique (PERT), see Whitehouse (1973). PERT is used primarily in project management to determine the time required to complete a project. It uses generalized beta distributions that are specified using prior information about each task. This information is combined to identify the mean time until completion. SFGMs can add a whole new level of flexibility to the PERT by using other distributions and providing probability intervals as opposed to just point estimates. PERT identifies the critical path that is the shortest time path in the graph, this would also be of interest if SFGMs were used to generalize PERT.

Flowgraph methodology could also be applied in Hidden Markov models (HMMs),

see Zucchini and MacDonald (2009) for an assessable introduction. HMMs are flexible models that enable us to model things we cannot observe using related events that we can observe. However, the Markov property can be restrictive so SFGMs could naturally extend this assumption in the field of hidden semi-Markov models (HSMMs). HSMMs were introduced by Ferguson (1980), are even more general than HMMs but more difficult to compute. SFGM may provide a way to compute HSMMs in an easier fashion, which could be an interesting research opportunity.

Another area of significant interest is identifying the model uncertainty of SFGMs. Currently, SFGMs do not have a way to identify the uncertainty of the proposed model. It would be beneficial to apply a method to determine some confidence bounds on the estimated CDF or survival curve of a SFGM. Having this confidence band on the survival function could tell the researcher how much credence we could place on our selected model.

There are a multitude of opportunities for exciting applications and research in SFGMs. As SFGM literature grows, application areas will also increase. The SFGM theory developed thus far elegantly provides a way to model a finite state problem where the semi-Markov assumption is reasonable.

Appendix A

R code for recurring illness process

The R code provided in this appendix is intended to be used as a reference guide for others wanting to implement SFGMs. The code is fairly rough and compact with limited comments. We attempt to annotate the input and output of each function or module but do not comment line by line. We make no assertions that the code is optimized, in fact the primary focus was on ease of programming and not speed. We also warn that there may be errors that are unknown to the author.

A.1 Selected code from Chapter 2

```
############################################################
###Generating the data
  set.seed(1)
  #parameters for the 3 lognormal transitions and the probability p
  #1 is for the 0-1, 2 is for the 1-0, and 3 is for the 1-2 transition
  mu1 = 0; sg1 = 1
  mu2 = -4; sg2 = 0.5
  mu3 = 1; sg3 = 0.3
  p = 0.4
```

```
  obs = as.list(rlnorm(20,mu1,sg1))

  for (i in 1:20) {
    x = rbinom(1,1,p)
    while(x==0){
      obs[[i]] = c(obs[[i]],rlnorm(1,mu2,sg2),rlnorm(1,mu1,sg1))
      x = rbinom(1,1,p)
    }
    obs[[i]] = c(obs[[i]],rlnorm(1,mu3,sg3))
  }

  #determining which observations to use as incomplete data
  incomp <- sample(1:20, 3)
  samps <- 1:20

  #putting the transition data into vectors transXX
  trans01 <- NULL; trans10 <- NULL; trans12 <- NULL
  for (i in samps[-incomp]) {
    for (j in 1:(length(obs[[i]])-1)) {
      if ( (j %% 2) == 1 ) {trans01=c(trans01,obs[[i]][j])
      } else {trans10=c(trans10,obs[[i]][j])}
    }
    trans12 <- c(trans12,obs[[i]][length(obs[[i]])])
  }

############################################################
###Defining the functions to find the MLEs, PDFs and CDFs
  library(statmod)

  #Inverse Gaussian Distribution MLEs
  IGmle <- function(x) {
    n <- length(x)
    m <- mean(x)
    lam <- n*m^2/sum((x-m)^2/x)
    return(c(m,lam))
  }

  #function to find the PDF of the Inverse Gaussian Distribution
  dinvgauss <- function(x,mu,la) {
    n <- length(x)
    out <- rep(NA,n)
```

```
  for (i in 1:n) {
    out[i] <- sqrt(la/(2*pi*x[i]^3))*exp(-1/2*la*(x[i]-mu)^2/(mu^2*x[i]))
  }
  return(out)
}

#function to find the CDF of the Inverse Gaussian Distribution
pinvgauss <- function(x,mu,la) {
  pnorm(sqrt(la/x)*(x/mu-1))+exp(2*la/mu)*pnorm(-sqrt(la/x)*(x/mu+1))
}

#Weibull Distribution MLEs
weimle <- function(x) {
  n <- length(x)
  g <- function(v) {1/v+sum(log(x))/n-sum(x^v*log(x))/sum(x^v)}
  v <- uniroot(g,c(0.01,20))$root
  w <- (n/sum(x^v))^(1/v)
  return(c(v,1/w))
}

#Lognormal Distribution MLEs
lnormle <- function(x) {
  n <- length(x)
  mu <- mean(log(x))
  sg <- sqrt(sum((log(x)-mu)^2)/n)
  return(c(mu,sg))
}

#Gamma Distribution MLEs
gammamle <- function(x) {
  n <- length(x)

  da <- function(a) { -n*digamma(a)+n*log(a/mean(x))+sum(log(x)) }
  ahat <- uniroot(da,c(0.00001,10000))$root
  bhat <- ahat/mean(x)
  return(c(ahat,bhat))
}

#Exponential Distribution MLEs
expmle <- function(x) { 1/mean(x) }
```

```
#function to find the PDF of the Frechet distribution
dfrechet <- function(x,sg,xi){ n<-length(x); output<-rep(NA,n)
  for (i in 1:n) { if (x[i] < 0) output[i] <- 0
    else output[i] <- xi/sg*x[i]^(-xi-1)*exp(-x[i]^(-xi)/sg) }
  return(output)
}

#function to find the CDF of the Frechet distribution
pfrechet <- function(x,sg,xi){
  n <- length(x); output <- rep(NA,n)
  for (i in 1:n) { if (x[i] < 0) output[i] <- 0
     else output[i] <- exp(-x[i]^(-xi)/sg)  }
  return(output)
}

#Frechet Distribution MLEs
frechmle <- function(x) {
  n <- length(x)
  dz <- function(z) { n/z-sum(log(x))+n*sum(x^(-z)*log(x))/sum(x^(-z))}
  zhat <- uniroot(dz,c(0.2,20))$root
  sighat <- mean(x^(-zhat))
  return(c(sighat,zhat))
}

#function to find the PDF of the Birnbaum-Saunders distribution
dbs <- function(x,alf,lam){ n<-length(x); output<-rep(NA,n)
  for (i in 1:n) { if (x[i] < 0) output[i] <- 0
    else output[i] <- (sqrt(lam*x[i])+1/sqrt(lam*x[i]))/
                       (2*alf*x[i]*sqrt(2*pi))*
                     exp(-1/(2*alf^2)*((sqrt(lam*x[i])-
                       1/sqrt(lam*x[i])))^2) }
  return(output)
}

#function to find the CDF of the Frechet distribution
pbs <- function(x,alf,lam){ pnorm((sqrt(lam*x)-1/sqrt(lam*x))/alf) }

#BirnbaumSaunders Distribution MLEs
BSmle <- function(x) {
  n <- length(x)
  dl <- function(l) { 1/n*sum(x/(l*x+1))-
```

```
                        (l*sum(x)-n)/(l^2*sum(x)-2*n*l+sum(1/x)) }
    lhat <- uniroot(dl,c(0.01,200))$root
    ahat <- sqrt(lhat*mean(x)-2+sum(1/x)/(n*lhat))
    return(c(ahat,lhat))
  }

#############################################################
###finding the MLEs for each transition and parameterization
  w01 <- weimle(trans01)
  w10 <- weimle(trans10)
  w12 <- weimle(trans12)
  i01 <- IGmle(trans01)
  i10 <- IGmle(trans10)
  i12 <- IGmle(trans12)
  g01 <- gammamle(trans01)
  g10 <- gammamle(trans10)
  g12 <- gammamle(trans12)
  e01 <- expmle(trans01)
  e10 <- expmle(trans10)
  e12 <- expmle(trans12)
  l01 <- lnormle(trans01)
  l10 <- lnormle(trans10)
  l12 <- lnormle(trans12)
  b01 <- BSmle(trans01)
  b10 <- BSmle(trans10)
  b12 <- BSmle(trans12)
  f01 <- frechmle(trans01)
  f10 <- frechmle(trans10)
  f12 <- frechmle(trans12)

  #MLE for p
  phat <- length(trans10)/(length(trans10)+length(trans12))

#############################################################
###Plotting a histogram with the some of the MLE fits for the
  #  0-1 transition
  # Similar plots can be found using similar code
  # to write the plot to a .png file, uncomment
  # the next and last line of the paragraph
  #png(filename="figure1-6.png",width=3600,height=2100,res=300)
  par(oma=c(0,0,0,0),mar=c(4.2,0,0,0))  #c(bottom, left, top, right)
```

```
  hist(trans01,br=20,xlim=c(0,5),ylim=c(0,6),main="",
                xlab="Time in years",ylab="",axes=F)
  par(oma=c(0,0,0,0),mar=c(4.2,0,0,0),new=T)
  curve(dinvgauss(x,i01[1],i01[2]),xlim=c(0,5),ylim=c(0,0.85),
                lty=1,axes=F,lwd=3,xlab="",ylab="")
  par(oma=c(0,0,0,0),mar=c(4.2,0,0,0),new=T)
  curve(dweibull(x,w01[1],w01[2]),xlim=c(0,5),ylim=c(0,0.85),
                lty=4,axes=F,lwd=3,xlab="",ylab="")
  par(oma=c(0,0,0,0),mar=c(4.2,0,0,0),new=T)
  curve(dgamma(x,g01[1],g01[2]),xlim=c(0,5),ylim=c(0,0.85),
                lty=2,axes=F,lwd=3,xlab="",ylab="")
  par(oma=c(0,0,0,0),mar=c(4.2,0,0,0),new=T)
  curve(dexp(x,e01[1]),xlim=c(0,5),ylim=c(0,0.85),lty=3,axes=F,
                lwd=3,xlab="",ylab="")
  axis(1, 0:5)
  legend("topright", c("inverse Gaussian", "Weibull",
                       "gamma","exponential"),
       lty = c(1,4,2,3), pch = c(-1,-1,-1,-1), bg = 'gray97',
                lwd=c(3,3,3,3))
  abline(h=0,lwd=2); box()
  #dev.off()
############################################################
```

We omit the code to find first passage distribution using the saddlepoint method. The details for this can be found in Huzurbazar (2005c). The next section shows how to find the first passage using the EULER method. We also omit most of the code to plot the results, which can be accomplished without too much trouble.

A.2 Selected code from Chapter 3

The euler function is an implementation of the EULER algorithm from Abate and Whitt (1995) (adapted for R). The *euler* function requires two inputs, the first is another function *fnRf* that returns the real portion of the complex LT, $L(z)$, you are trying to invert (i.e. $fnRf(X,Y) = Re[L(z)]$, where $X = Re(z)$ and $Y = Im(z)$).

The second input is the value t, the time point at which the PDF of the complex LT, $f(t)$, is evaluated. The output is the value of $f(t)$. It is worth noting that this algorithm cannot be evaluated at $f(0)$.

```
  euler <- function(fnRf,T,A = 18.4,Ntr = 15,num=11) {
    w = c(1/2,rep(1,Ntr-1),
        rev(cumsum(choose((num),0:(num))))/(2^(num)))
    SU <- rep(NA,Ntr+num+1);
    for (i in 0:(Ntr+num)) { SU[i+1] <- fnRf(A/(2*T), i*pi/T)}
    return(exp(A/2)/T*sum(w*(-1)^(0:(Ntr+num))*SU ))
  }

############################################################
###Functions needed for numerical intergration
  reLT <- function(t,v,w,x,y,pdf) {pdf(t,v,w)*exp(-x*t)*cos(y*t)}
  imLT <- function(t,v,w,x,y,pdf) {pdf(t,v,w)*exp(-x*t)*sin(y*t)}
  int <- function(funct,v,w,x,y,pdf) {
            integrate(funct,0,Inf,v=v,w=w,x=x,y=y,pdf=pdf,
            subdivisions=10000,rel.tol=1e-10,stop.on.error=FALSE)$value
  }

############################################################
###The matrix and vector defining the flowgraph parameterization
  param_input <- matrix(c(b01, f10, l12), byrow=T,ncol=2)
  pdf_input <- c(dbs,dfrechet,dlnorm)

############################################################
###The function to find the real portion of the complex LT
  fnRf <- function(X,Y) {
    val <- rep(NA,3)
    for (i in 1:length(pdf_input)) {
      val[i] <- int(reLT,param_input[i,1],param_input[i,2],
                    X,Y,pdf_input[[i]])-
          1i*int(imLT,param_input[i,1],param_input[i,2],X,Y,pdf_input[[i]])
    }
    return(Re((1-phat)*val[1]*val[3]/(1-phat*val[1]*val[2])))
  }

############################################################
###Vectors defining the support and values of the first passage PDF
```

```
  supp1 <- (10:130)/10; out1 <- rep(NA,121)

############################################################
###Code to call the euler function and invert the complex LT to a PDF
  for (i in 1:121) { out1[i] <- euler(fnRf,supp1[i])[1] }
```

A.3 Selected code from Chapter 4

```
############################################################
###Defining the matrix for the posterior samples
  n <- 100000
  post <- matrix(NA,nrow=n,ncol=7)

############################################################
###Function to simulate from an inverse gamma distribution
  rinvgamma <- function(n,a,b) {1/rgamma(n,a,rate=b)}

############################################################
###Definition of the hyperparamters
  # n is for normal a is mean b is variance
  # ig is for inverse gamma a and b are the parameters
  # b is for beta
  an01 <- 0
  bn01 <- 100
  aig01 <- 1/20
  big01 <- 1
  an10 <- 0
  bn10 <- 100
  aig10 <- 1/20
  big10 <- 1
  an12 <- 0
  bn12 <- 100
  aig12 <- 1/20
  big12 <- 1
  ab <- 1
  bb <- 1

############################################################
```

```
###Definition of the values needed in the conditionals
  n01 <- length(trans01)
  x01 <- sum(log(trans01))
  x012 <- sum((log(trans01))^2)
  n10 <- length(trans10)
  x10 <- sum(log(trans10))
  x102 <- sum((log(trans10))^2)
  n12 <- length(trans12)
  x12 <- sum(log(trans12))
  x122 <- sum((log(trans12))^2)

############################################################
###Starting values for the MCMC
  post[1,] <- c(0,1,0,1,0,1,1/2)
  set.seed(22)

############################################################
###Conducting the Gibbs sampler
  for (i in 2:n) {
    post[i,1] <- rnorm(1,(x01/post[i-1,2])/(n01/post[i-1,2]+1/bn01),
                       1/(n01/post[i-1,2]+1/bn01))
    post[i,2] <- rinvgamma(1,n01/2+aig01,(x012-2*post[i,1]*x01+n01*
                       (post[i,1])^2)/2+big01)
    post[i,3] <- rnorm(1,(x10/post[i-1,4])/(n10/post[i-1,4]+1/bn10),
                       1/(n10/post[i-1,4]+1/bn10))
    post[i,4] <- rinvgamma(1,n10/2+aig10,(x102-2*post[i,3]*x10+n10*
                       (post[i,3])^2)/2+big10)
    post[i,5] <- rnorm(1,(x12/post[i-1,6])/(n12/post[i-1,6]+1/bn12),
                       1/(n12/post[i-1,6]+1/bn12))
    post[i,6] <- rinvgamma(1,n12/2+aig12,(x122-2*post[i,5]*x12+n12*
                       (post[i,5])^2)/2+big12)
  }
  post[,7] <- rbeta(n,length(trans10)+1,length(trans12)+1)

############################################################
###Sampling from the PPD
  ppd <- rep(0,n)
  for (i in 1:n) {
    cycs <- rnbinom(1,1,1-post[i,7])
    ppd[i] <- sum(c(rlnorm(cycs+1,post[i,1],sqrt(post[i,2])),
                    rlnorm(cycs,post[i,3],sqrt(post[i,4])),
```

```
                    rlnorm(cycs+1,post[i,5],sqrt(post[i,6])) ))
  }
```

A.4 Selected code from Chapter 5

```
############################################################
###The matrix and vector defining the flowgraph parameterization
  param_input <- matrix(c(g01,i10,g12), byrow=T,ncol=2)
  pdf_input <- c(dgamma,dinvgauss,dgamma)

############################################################
###This function returns the real value of the first passage CDF
  FnRf <- function(X,Y) {
    val1 <- (1+(X+Y*1i)/param_input[1,2])^(-param_input[1,1])
    val2 <- exp(param_input[2,2]/param_input[2,1])*
            (1-sqrt(1+2*param_input[2,1]^2*(X+Y*1i)/param_input[2,2]))
    val3 <- (1+(X+Y*1i)/param_input[3,2])^(-param_input[3,1])
  output <- ((1-p1)*val1*val3)/(1-p1*val1*val2)
  return(Re(output/(X+1i*Y)))
  }

############################################################
###This function finds the CDF of an incomplete observation
  cdfob <- function(p,m1,s1,m2,s2,m3,s3,t) {
    param_input <<- matrix(c(m1,s1,m2,s2,m3,s3), byrow=T,ncol=2)
    p1 <<- p
    return(euler(FnRf,t)[1])
  }

############################################################
###The incomplete observations are 1,14,15
# let's say for obs1 we know it occured before time T=5
#    (left censored) -- F_{02*}(5)
# let's say for obs14 we saw it in state 1 at time T=4
#    (right censored) -- (1-F_{02*}(4))
# let's say for obs15 we saw it in state 0 at time T=3
#     and at state 2 at time T=14 (interval censored) --
#     (F_{02*}(14)-F_{02*}(3))
```

```
############################################################
###The loglikelihood function
  loglike <- function(p,m1,s1,m2,s2,m3,s3) {
   sum(dgamma(trans01,m1^2/s1,m1/s1,log=T)) +
   sum(log(dinvgauss(trans10,m2,s2))) +
   sum(dgamma(trans12,m3^2/s3,m3/s3,log=T)) +
   log(max(cdfob(p,m1^2/s1,m1/s1,m2,s2,m3^2/s3,m3/s3,5),0)) +
   log(max(1-cdfob(p,m1^2/s1,m1/s1,m2,s2,m3^2/s3,m3/s3,4),0)) +
   log(max(cdfob(p,m1^2/s1,m1/s1,m2,s2,m3^2/s3,m3/s3,14) -
       cdfob(p,m1^2/s1,m1/s1,m2,s2,m3^2/s3,m3/s3,3),0)) +
   9*log(p) + 17*log(1-p)
  }
```

A.5 Selected code from Chapter 6

The function *calcQ* takes the sample $F(x_i)$ and returns the value of Q. The function *censcalcQ* expects the input to be a matrix with each row the values $F(a_i), F(b_i)$, where the interval-censored observation $X_i \quad (a_i, b_i)$. This function returns a confidence interval for Q. The CI can be shrunk by increasing the size of m and the confidence level can be altered using *alpha*.

```
############################################################
###Function to calculate Q
  calcQ <- function(x) {
    n <- length(x); x <- sort(x)
    samp1 <- -log(x); samp2 <- -log(1-x[n:1])
    mean1 <- rev(cumsum(1/(n:1))); var1 <- rev(cumsum(1/(n:1)^2))
    return(1/(2*n)*sum(((samp1-mean1)^2+(samp2-mean1)^2)/var1))
  }

############################################################
###Function to calculate Q hat
  censcalcQ <- function(dat,m=1000,alpha=0.01) {
  # dat is a nx2 matrix with the left and right endpoints
  #   of the censored data (in the interval [0,1])
```

```
  # m is the number of samples taken to estimate the mean of Q
  # alpha provides input for the (1-alpha)*100\% CI for the mean of Q
    n <- length(dat[,1])
    out <- rep(NA,m)
    for (i in 1:m) {
      samp <- runif(n,dat[,1],dat[,2])
      out[i] <- calcQ(samp)
    }
    s1 <- qnorm(1-alpha)*sd(out)/sqrt(m)
    m1 <- mean(out)
    return(c(m1-s1,m1+s1))
  }

############################################################
###Function to calculate the CDF
  FnRf <- function(X,Y) {
    val <- rep(NA,3)
    for (i in 1:length(pdf_input)) {
      val[i] <- int(reLT,param_input[i,1],param_input[i,2],X,Y,
                    pdf_input[[i]])-
        1i*int(imLT,param_input[i,1],param_input[i,2],X,Y,pdf_input[[i]])
    }
    return(Re((1-phat)*val[1]*val[3]/((1-phat*val[1]*val[2])*(X+1i*Y))))
  }

############################################################
###Putting the overall data into a vector
  new1 <- lapply(obs,sum)[-incomp]
  out <- rep(NA, 17); for (i in 1:17) {out[i] <- new1[[i]]}

############################################################
###Value of F(X) for model in Chapter 3
  param_input <- matrix(c(b01, f10, l12), byrow=T,ncol=2)
  pdf_input <- c(dbs,dfrechet,dlnorm)
  Fchap3 <- rep(NA,17)
  for (i in 1:17) {Fchap3[i] <- euler(FnRf,out[i])[1] }

############################################################
###Value of Q for model in Chapter 3
  calcQ(Fchap3)
```

Glossary

Acronyms

A-D	Anderson-Darling goodness-of-fit statistic
AIC	Akaike information criterion
C-vM	Cramér-von Mises goodness-of-fit statistic
CDF	Cumulative distribution function
CDH	Censored data histogram
CF	Characteristic function
CI	Confidence interval
CLT	Central Limit Theorem
DIC	Deviance information criterion
ECDF	Empirical cumulative distribution function
FPT	Finite Polya tree
K-S	Kolmogorov-Smirnov goodness-of-fit statistic
LT	Laplace transform

MCMC	Markov chain Monte Carlo
MGF	Moment generating function
MLE	Maximum likelihood estimator
MOM	Method of moments
MPT	Mixture of finite Polya trees
SFGM	Statistical flowgraph model
PDF	Probability density function
PPD	Posterior predictive distribution

Notation

$f_{ij}(t)$	The PDF of the direct passage from state i to state j
$f_{ij*}(t)$	The PDF of the first passage from state i to state j
$F_{ij}(t)$	The CDF of the direct passage from state i to state j
$F_{ij*}(t)$	The CDF of the first passage from state i to state j
I	The identity matrix (assuming the appropriate dimension)
Q	The goodness-of-fit criterion developed in chapter 6
Q	An estimate of the goodness-of-fit criterion developed in chapter 6
Q_p	A penalized estimate of the goodness-of-fit criterion developed in chapter 6
$X(t)$	The state of a stochastic process at time t

References

Aalen, O. O. (1975). *Statistical inference for a family of counting processes*, PhD thesis, University of California, Berkeley, CA.

Aalen, O. O. (1978). Nonparametric inference for a family of counting processes, *The Annals of Statistics* **6**(4): 701 726.

Abate, J. and Whitt, W. (1992). The Fourier-series method for inverting transforms of probability distributions, *Queuing Systems* **10**: 5 88.

Abate, J. and Whitt, W. (1995). Numerical inversion of Laplace transforms of probability distributions, *INFORMS Journal on Computing* **7**: 36 43.

Akaike, H. (1974). A new look at the statistical model identification, *IEEE Transactions on Automatic Control* **19**(6): 716 723.

Albert, J. (2007). *Bayesian Compuation with R*, Springer, New York.

Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A. and Sorensen, D. (1999). *LAPACK Users' Guide*, third edn, Society for Industrial and Applied Mathematics, Philadelphia, PA.

Barbu, V. S. and Liminios, N. (2008). *Semi-Markov Chains and Hidden Semi-Markov Models Toward Applications: Their Use in Reliability and DNA Analysis*, Springer, New York.

Bedrick, E. J., Christensen, R. and Johnson, W. (1996). A new perspective on priors for generalized linear models, *Journal of the American Statistical Association* **91**: 1450 1460.

Billingsley, P. (1995). *Probability and Measure*, 3rd edn, Wiley, New York.

Birnbaum, Z. W. and Saunders, S. C. (1969). A new family of life distributions, *Journal of Applied Probability* **6**: 319 327.

Box, G. E. P. (1980). Sampling and bayes' inference in scientific modelling and robustness, *Journal of the Royal Statistical Society, Series A* **143**: 383 404.

Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces*, Wiley.

Butler, R. W. and Huzurbazar, A. V. (1997). Stochastic network models for survival analysis, *J. Amer. Statist. Assoc.* **92**: 246 257.

Butler, R. W. and Huzurbazar, A. V. (2000). Bayesian prediction of waiting times in stochastic models, *Canad. J. Statist.* **28**(2): 311 325.

Cappé, O., Moulines, E. and Rydén, T. (2005). *Inference in Hidden Markov Models (Springer Series in Statistics)*, Springer, New York.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*, second edn, Duxbury, Pacific Grove, CA.

Çinlar, E. (1969). On semi-markov processes on arbitrary spaces, *Mathematical Proceedings of the Cambridge Philosophical Society* **66**: 381 392.

Christensen, R. (2002). *Plane Answers to Complex Questions*, third edn, Springer, New York.

Christensen, R., Hanson, T. and Jara, A. (2008). Parametric nonparametric statistics: An introduction to mixtures of finite polya trees, *The American Statistician* **62**(4): 296 306.

Christensen, R., Johnson, W., Branscum, A. and Hanson, T. E. (2010). *Bayesian Ideas and Data Analysis: An Introduction for Scientists and Statisticians*, CRC Press, Boca Raton, FL.

Collett, D. (2003). *Modelling Survival Data in Medical Research*, second edn, CRC Press, Boca Raton, FL.

Collins, D. H. (2009). *Nonparametric Estimation of First Passage Time Distributions in Flowgraph Models*, PhD thesis, University of New Mexico, Albuquerque, NM.

Collins, D. H. and Huzurbazar, A. V. (2008). System reliability and safety assessment using nonparametric flowgraph models, *Journal of Risk and Reliability* **222**: 667 674.

Daniels, H. (1954). Saddlepoint approximations in statistics, *Annals of Mathematical Statistics* **25**: 631 650.

Edwards, D. (2000). *Introduction to Graphical Modelling*, second edn, Springer, New York.

Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*, Chapman & Hall, New York.

Ferguson, J. D. (1980). Variable duration models for speech, *In Proceedings of the Symposium on the Application of Hidden Markov Models to Text and Speech (Princeton, NJ)* pp. 143 179.

Fix, E. and Neyman, J. (1951). A simple stochastic model of recovery, relapse and loss of patients, *Human Biology* **23**: 205 241.

Genest, C. and Rivest, L.-P. (2001). On the multivariate probability integral transformation, *Statistics and Probability Letters* **53**(4): 391 399.

Givens, G. H. and Hoeting, J. A. (2005). *Computational Statistics*, Wiley, New Jersy.

Grimmett, G. R. and Stirzaker, D. R. (2001). *Probability and Random Processes*, Oxford University Press.

Gross, A. J., Clark, V. A. and Liu, V. (1971). Estimation of survival parameters where one of two organs must function for survival, *Biometrics* **27**: 369 377.

Hamada, M. S., Wilson, A. G., Reese, C. S. and Martz, H. F. (2008). *Bayesian Reliability*, Springer, New York.

Hornik, K. (2009). The R FAQ. ISBN 3-900051-08-9.
URL: *http://CRAN.R-project.org/doc/FAQ/R-FAQ.html*

Hougaard, P. (1999). Multi-state models: A review, *Lifetime Data Analysis* pp. 239 264.

Huzurbazar, A. V. (1999a). Flowgraph models for generalized phase type distributions with non-exponential waiting times, *Scandinavian Journal of Statistics* **26**: 145 157.

Huzurbazar, A. V. (2000). Modeling and analysis of engineering systems data using flowgraph models, *Technometrics* **42**: 300 306.

Huzurbazar, A. V. (2002). Modeling time-to-event data using flowgraph models, *Advances on methodological and applied aspects of probability and statistics (Hamilton, ON, 1998)*, Taylor & Francis, London, pp. 561 571.

Huzurbazar, A. V. (2004a). Modelling survival data using flowgraph models, *Advances in survival analysis*, Vol. 23 of *Handbook of Statist.*, Elsevier, Amsterdam, pp. 729 746.

Huzurbazar, A. V. (2004b). Multistate models, flowgraph models, and semi-Markov processes, *Comm. Statist. Theory Methods* **33**(3): 457 474.

Huzurbazar, A. V. (2005a). A censored data histogram, *Comm. Statist. Simulation Comput.* **34**(1): 113 120.

Huzurbazar, A. V. (2005b). Flowgraph models: a Bayesian case study in construction engineering, *J. Statist. Plann. Inference* **129**(1-2): 181 193.

Huzurbazar, A. V. (2005c). *Flowgraph Models for Multistate Time-to-Event Data*, Wiley, New York.

Huzurbazar, A. V. and Williams, B. J. (2005). Flowgraph models for complex multistate system reliability, *Modern statistical and mathematical methods in reliability*, Vol. 10 of *Ser. Qual. Reliab. Eng. Stat.*, World Sci. Publ., Singapore, pp. 247 262.

Huzurbazar, A. V. and Williams, B. J. (2010). Incorporating covariates in flowgraph models: Applications to recurrent event data, *Technometrics* . To appear.

Huzurbazar, S. (1999b). Practical saddlepoint approximations, *The American Statistician* **53**: 225 232.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*, second edn, Springer.

Kutner, M. H., Nachtsheim, C. J., Neter, J. and Li, W. (2005). *Applied Linear Statistical Models*, fifth edn, McGraw-Hill, New York.

Lagakos, S. W. (1976). A stochastic model for censored-survival data in the presence of an auxiliary variable, *Biometrics* **32**(3): 551 559.

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*, third edn, Springer, New York.

Lévy, P. (1954). Processus semi-Markoviens, *In Proc. of International Congress of Mathematics (Amsterdam)* .

Limnios, N. and Oprişan, G. (2001). *Semi-Markov Processes and Reliability*, Birkhauser, Boston.

Longini, I. M., Clark, W., Byers, R., Ward, J., Darrow, W., Lemp, G. and Hethcote, H. (1989). Statistical analysis of the stages of HIV infection using a markov model, *Statisitcs in Medicine* **8**: 831 843.

Marin, J.-M. and Robert, C. P. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*, Springer, New York.

Marshall, G., Guo, W. and Jones, R. H. (1995). Markov: A computer program for multi-state markov models with covariables, *Computer Methods and Programs in Biomedicine* **47**(2): 147 156.

Marshall, G. and Jones, R. H. (1995). Multi-state models and diabetic retinopathy, *Statistics in Medicine* **14**(18): 1975 1983.

Mason, S. J. (1953). Feedback theory: some properties of signal flow graphs, *Proceedings of the Institute of Radio Engineers* **41**: 1144 1156.

Miner, M. A. (1945). Cumulative damage in fatigue, *Journal of Applied Mechanics* **12**: A159 A164.

Petersen, T. (1986). Fitting parametric survival models with time-dependent covariates, *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **35**(3): 281 288.

Pyke, R. (1961a). Markov renewal processes: Definitions and preliminary properties, *The Annals of Mathematical Statistics* **32**(4): 1231 1242.

Pyke, R. (1961b). Markov renewal processes with finitely many states, *The Annals of Mathematical Statistics* **32**(4): 1243 1259.

Pyke, R. and Schaufele, R. (1964). Limit theorems for markov renewal processes, *The Annals of Mathematical Statistics* **35**(4): 1746 1764.

Pyke, R. and Schaufele, R. (1966). The existence and uniqueness of stationary measures for markov renewal processes, *The Annals of Mathematical Statistics* **37**(6): 1439 1462.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.
URL: *http://www.R-project.org*

Reid, N. (1988). Saddlepoint methods and statistical inference, *Statistical Science* **3**(2): 213 227.

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*, second edn, Springer, New York.

Rosenblatt, M. (1952). Remarks on a multivariate transformation, *The Annals of Mathematical Statistics* **23**(3): 470 472.

Ross, S. (1996). *Stochastic Processes*, second edn, Wiley, New York.

Rudin, W. (1976). *Principles of Mathematical Analysis*, third edn, McGraw-Hill.

Smith, B. (2005). Bayesian output analysis program (boa) for MCMC. R package version 1.1.5.
URL: *http://www.public-health.uiowa.edu/boa*

Smith, W. L. (1955). Regenerative stochastic processes, *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* **232**(1188): 6 31.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002). Bayesian measures of model complexity and fit, *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64**(4): 583 639.

Stroock, D. W. (2005). *An Introduction to Markov Processes*, Springer, New York.

Takacs, L. (1954). Some investigations concerning recurrent stochastic processes of a certain type, *Magyar Tud. Akad. Mat. Kutato Int. Kzl.* **3**: 115 128.

Takacs, L. (1959). On a sojourn time problem in the theory of stochastic processes, *Transactions of the American Mathematical Society* **93**(3): 531 540.

Taylor, H. M. and Karlin, S. (1998). *An Introduction to Stochastic Modeling, Third Edition*, 3 edn, Academic Press, London.

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*, Springer, New York.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*, second edn, Springer, New York.

Whitehouse, G. (1973). *Systems Analysis and Design Using Network Techniques*, Prentice Hall, NJ.

Williams, B. J. and Huzurbazar, A. V. (2006). Posterior sampling with constructed likelihood functions: an application to flowgraph models, *Appl. Stoch. Models Bus. Ind.* **22**(2): 127 137.

Wolstenholme, L. C. (1999). *Reliability Modelling: A Statistical Approach*, Chapman & Hall/CRC, Boca Raton, FL.

Yau, C. L. and Huzurbazar, A. V. (2002). Analysis of censored and incomplete survival data using flowgraph models, *Statistics in Medicine* **21**: 3727 3743.

Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov Models for Time Series: an Introduction Using R*, CRC Press, Boca Raton, FL.