TESTING THE ADEQUACY OF A SEMI-MARKOV PROCESS

DISSERTATION

Richard S. Seymour, Lt Col, USAF

AFIT-ENC-DS-15-S-003

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

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

AFIT-ENC-DS-15-S-003

TESTING THE ADEQUACY OF A SEMI-MARKOV PROCESS

DISSERTATION

Presented to the Faculty

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

in Partial Fulfillment of the Requirements for the

Degree of Doctoral of Philosophy in Applied Mathematics

Richard S. Seymour, B.S.O.R., M.S.C.S.

Lt Col, USAF

September 2015

AFIT-ENC-DS-15-S-003

TESTING THE ADEQUACY OF A SEMI-MARKOV PROCESS

Richard S. Seymour, B.S.O.R., M.S.C.S.
Lt Col, USAF

Committee Membership:

Dr. Christine M. Schubert Kabban
Chair

Dr. Gilbert L. Peterson
Member

Lt Col Richard L. Warr, PhD
Member

Lt Col Ryan D. Kappedal, PhD
Member

ADEDEJI B. BADIRU, Ph.D.
Dean, Graduate School of Engineering
and Management

AFIT-ENC-DS-15-S-003

Abstract

Due to the versatility of its structure, the semi-Markov process is a powerful modeling tool used to describe complex systems. Though similar in structure to continuous time Markov chains, semi-Markov processes allow for any transition time distribution which enables these processes to fit a wider range of problems than the continuous time Markov chain. While semi-Markov processes have been applied in fields as varied as biostatistics and finance, there does not exist a theoretically-based, systematic method to determine if a semi-Markov process accurately fits the underlying data used to create the model.

In fields such as regression and analysis of variance, the quality of the predictive model is judged in part by the goodness of fit of the model which relates the expected observation values with the actual observations. A similar methodology for semi-Markov processes would provide immediate insight in the efficacy of the fitted model and would allow competing models to be directly compared with one another.

This dissertation presents a methodology to measure the adequacy of a fitted semi-Markov process. To this end, a technique to assess the likelihood that a data sample could be generated by a specific semi-Markov process is developed, including a newly proposed goodness of fit metric. This technique relies on the covariance structure of the semi-Markov process; thus, a method to estimate the covariance structure is also proposed. The technique is applied to real and simulated data to demonstrate the goodness of fit metric's utility in model validation and its ability to identify potential covariate factors within the model.

Table of Contents

Page

List of Figures

List of Tables

TESTING THE ADEQUACY OF A SEMI-MARKOV PROCESS

I. Introduction

The area of stochastic processes known as multi-state modeling describes systems in a vast array of fields, from finance and health care to game theory and reliability. Typically, these models decompose a complex system into a series of connected states based on observed data. Although observed data provides the basis for the model, a modeler's judgment also plays a substantial role in model development. The variability inherent in observational data and the modeler's biases may result in a potentially inadequate system model. Other modeling arenas, such as regression and distribution fitting, utilize various quality metrics to confirm model adequacy; an analogous technique does not exist for general multi-state modeling.

Simple quality metrics determine the likelihood that an observed sample comes from a specific model, as illustrated by the analysis of variance F-statistic and Pearson's Chi-squared goodness of fit statistic. More advanced metrics, such as R^2, assess the degree of fit between the model and the observed data. All of these metrics are based in theory, simple to compute, and provide the modeler with immediate insight regarding the validity of a model.

Currently, multi-state modeling efforts omit any reference to model goodness of fit [1–3]. This is because the computational simplicity is believed to be lost due to the complexity of the model. In reality though, certain forms of quality metrics may be applied to multi-state models. However, past efforts have seen limited success [4] due to the subjective nature of many of the available metrics and nuances in data collection. Additionally, quality assessment tools often fail to provide definitive results in multi-state models due

to data collection techniques. Specifically, censoring within data collection methods means that actual state transition times are not observed and must be approximated, which adds additional uncertainty in any quality metric calculations. These drawbacks result in quality metrics losing their allure, especially when the modeler must test multiple competing models. To be of use for multi-state models, the quality metric must be able to differentiate between models in a quick and reliable manner.

The work of this dissertation utilizes a Chi-squared test statistic based on a model's covariance structure to create a goodness of fit metric for a semi-Markov process. The metric is developed and applied to a semi-Markov process since the semi-Markov process is a generalized class that contains any Markov process including discrete-time and continuous-time Markov chains. This metric is used to test the adequacy of the proposed model with respect to the observed data sample. The technique to calculate the metric and test model adequacy incorporates a variety of nuances contained within the structure of a semi-Markov process in order to elicit the required expectations for the metric. The testing technique is demonstrated through a series of well-defined models. Additionally, an application of this technique demonstrates its usefulness in model selection by deciding whether or not potential covariate factors significantly impact the model's performance.

This dissertation is structured into seven chapters. Following this first introductory chapter, background information is provided in Chapter 2 to describe the multi-state model including the semi-Markov process. Chapter 3 provides the theory for calculating the covariance of the Markov renewal process which is a surrogate for the covariance of a semi-Markov process. Chapter 4 presents the goodness of fit metric and its properties as well as provides a simulation study to demonstrate the robustness of the proposed metric. Chapter 5 demonstrates how the goodness of fit metric may be used to test the model structure for the presence of a possible covariate. Finally, Chapter 6 demonstrates the application of

the goodness of fit metric to examine trends in the tempo of Body Mass Index values in childhood and conclusions are provided in Chapter 7.

II. Background

2.1 Stochastic Processes

A stochastic process is a collection of random variables indexed on some set T. The evolution of these random variables can be used to represent a system changing over time. For example, let $X(t)$ be the price of a barrel of oil at time t, then $X = \{X(t), t \in T\}$ is a stochastic process that describes the price evolution of a barrel of oil over time T. T is also known as the index set. This index set can be discrete (the closing price each day) or continuous (the instanteous price throughout the day). An actual realization of X is known as a sample path.

Stochastic processes trace the evolution of a system as it assumes various states, s. A state, say s_1 or s_2, is an observable location of the stochastic process and the *state space* ($S = \{s_1, s_2, ...\}$) is the full set of possible states. Although S could be $\mathbb{R}$, in this work S will be limited to a discrete state space. Figure 2.1 shows a four state system with potential transitions between states indicated by arrows. Notice the system never returns after departing State 1 and never leaves after arriving at State 4.

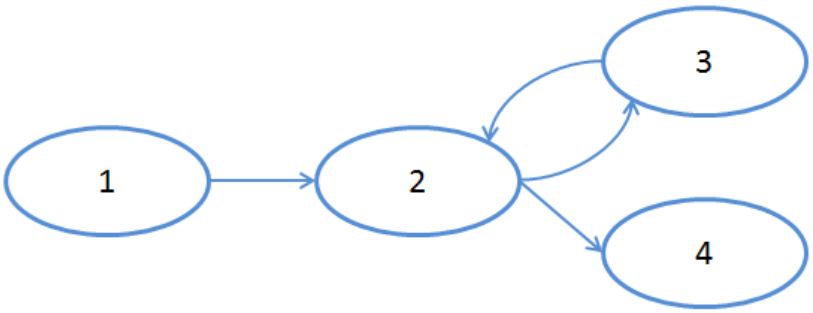

Figure 2.1: Basic state model

2.2 Markov Processes

A Markov process is a stochastic process for which the Markov property holds [5]. The Markov property is, given the entire history of the process, the conditional distribution of a future state is only dependent on the current state [6]. The Markov property reduces the required system memory to just the current state rather than an entire history of the system. Instead of tracking every state visited and the time spent in every state, a Markov process only needs to track the current system state. From a computational standpoint, this represents a significant reduction in physical memory requirements.

Markov processes may take on many forms depending on the structure of the system. A Markov process with a discrete state space is known as a Markov chain. A Markov chain is fully determined by its matrix of transition probabilities from State i directly to State j for each i and j in S. With this transition probability matrix, $\mathbf{P}$, and the current state, many features of the system can be determined, including the likelihood that the system will be in a particular state after the next transition or after the n-th transition, how many transitions will occur until a specific state is reached, and the amount of time until the system returns to the current state. Although the state space is discrete, the index set for a Markov chain can be either discrete or continuous based upon how time is measured and tracked.

This leads to the definition of a discrete time Markov chain. Let X_n denote the system state after the n-th transition.

Definition *Discrete Time Markov Chain [7]* A stochastic process $\{X_n, n \geq 0\}$ with countable state space S is a *discrete time Markov chain* if the following hold for all integer-valued $n \geq 0$, and for each i, j in S:

1. $X_n \in S$, and
2. $P(X_{n+1} = j | X_n = i, X_{n-1}, ..., X_0) = P(X_{n+1} = j | X_n = i)$.

Item 2 is the Markov property with respect to the discrete time Markov chain. Note that time is not explicitly tracked in the discrete time Markov chain. In this type of process,

each transition is considered a single time step. In effect, the n is the number of time steps and can be related back to the actual index set, T, for the process.

Definition *Continuous Time Markov Chain [7]* A stochastic process $\{X(t), t \geq 0\}$ with countable state space S is a *continuous time Markov chain* if the following are satisfied by the sequence $\{X_0, (X_n, Y_n), n \geq 1\}$ for all integer-valued $n \geq 0$, and for each i, j in S:

$$P(X_{n+1} = j, Y_{n+1} > y | X_n = i, Y_n, X_{n-1}, Y_{n-1}, ..., X_1, Y_1, X_0)$$
$$= P(X_1 = j, Y_1 > y | X_0 = i) = p_{ij} e^{-q_i y}.$$

In this formulation, X_n is the system state after the n-th transition, Y_n is the time the system requires to transition from X_{n-1} to X_n, also known as the *sojourn time*, p_{ij} is the probability of transitioning directly from State i to State j, and q_i is the exponential parameter for the sojourn time distribution for State i such that $q_i = \sum_{k \neq i} q_{ik}$ where k is any state in S.

In a continuous time Markov chain, every transition sojourn time follows an Exponential distribution which maintains the Markov property due to the memoryless nature of the distribution. Further, the transition probability matrix, $\mathbf{P}$, is equal to $[p_{ij}]$. The typical continuous time Markov chain is formulated so that $p_{ii} = 0$, for all i. This formulation eliminates immediate transitions back to the current state which simplifies moment calculations. If a system does allow self transitions, additional states can be added in order to represent the self transitions without altering the performance of the system. Often, the q_{ij}'s are placed in a matrix $\mathbf{Q} = [q_{ij}]$ which is known as the generator matrix, where $q_{ii} = -\sum_{k \neq i} q_{ik} = -q_i$. As a result, the row sums in $\mathbf{Q}$ are equal to 0. A continuous time Markov chain is fully described by $\mathbf{Q}$.

The continuous time Markov chain construct is restricted to modeling sojourn times using exponential distributions. To allow any sojourn time distribution, more complex models, such as a semi-Markov process, must be used. The semi-Markov process was originally defined by Lévy [8], Takács [9], and Smith [10] and further refined by Pyke [11].

Definition *Semi-Markov Process [7]* A stochastic process $\{X(t), t \geq 0\}$ with countable state space S is a *semi-Markov process* if the sequence $\{X_0, (X_n, Y_n), n \geq 1\}$ satisfies the following for all integer-valued $n \geq 0$, and for each i, j in S:

$$P(X_{n+1} = j, Y_{n+1} \leq y | X_n = i, Y_n, X_{n-1}, Y_{n-1}, ..., X_1, Y_1, X_0)$$

$$= P(X_1 = j, Y_1 \leq y | X_0 = i) = G_{ij}(y).$$

Here, X_n and Y_n have the same definition as a continuous time Markov chain, and

$$G_{ij} = p_{ij} F_{ij}(y),$$

which is the probability of transitioning from State i directly to State j (p_{ij}) times the cumulative distribution function of the sojourn time distribution between States i and j ($F_{ij}(y)$). The semi-Markov process kernel matrix is given by $\mathbf{G}(y) = [G_{ij}(y)]$ for i, j in S, $y \geq 0$.

As the formal definition implies, the semi-Markov process is similar to the continuous time Markov chain. The semi-Markov process consists of a transition probability matrix and distributions for the sojourn time between transitions. However, by relaxing the exponential sojourn distribution requirement, the Markov property no longer applies at every time t, but only at the actual transition times [5]. For example, suppose a system transitions into State 1 at time t and will transition into State 2 at some time $t + s$. If the system was a continuous time Markov chain, the transition time would follow an Exponential(λ) distribution and for any time $t + i, i < s$, the expected transition time would be λ. On the other hand, if the system was a semi-Markov process, the transition time would follow some general distribution that may not adhere to the memoryless property, and the expected transition time would not be the same for any time $t + i, i < s$. Therefore, the semi-Markov process must track the history between transitions by maintaining the time since the last transition.

The semi-Markov process construct is useful in providing added fidelity to real world systems because of the flexibility in choosing sojourn distributions. Although semi-Markov

processes are commonly associated with reliability and survival analyses [12], they have also been applied to a variety of problems in areas such as insurance [2], biology [13], finance [1], and computer science [14]. The basic approach to modeling with a semi-Markov process from observed data is the following:

1. Define the state space, S.

2. Estimate the transition probability distribution matrix, $\mathbf{P}$.

3. Estimate the cumulative distribution functions of the sojourn times for $\mathbf{F}$.

As a semi-Markov process evolves over time, the number of visits to each state forms a Markov renewal process as shown in Pyke [11].

Definition *Markov Renewal Process [7]* A process $\widetilde{N}$ is called a *Markov renewal process*, if the vector $\widetilde{N} = \{\widetilde{N}(t) = [N_j(t)], t \geq 0, \text{ for all } j \text{ in } S\}$,

where $N_j(t)$ records the number of visits the system has made to a State j at time t given an initial state at time 0. Figure 2.2 illustrates the relationship between semi-Markov processes and Markov renewal processes by showing the transitions for the semi-Markov process on the left and the counting process for State 3 on the right. States 1, 2, and 4 have similar Markov renewal process graphs. While the semi-Markov process tracks a single state occupancy value, i.e. the system is currently in State 4, the Markov renewal process tracks the number of visits to each of the S states. The link between the semi-Markov process and the Markov renewal process is exploited in the creation of the goodness of fit metric in Chapter 4 as the metric calculations are performed on the Markov renewal process instead of the actual semi-Markov process.

The Markov renewal process has a well-defined expectation. As illustrated in Pyke [15], the expected count in the Laplace domain at a fixed for each state is

$$\mathcal{L}(\mathrm{E}[\widetilde{N}(t)]) = (\mathbf{I} - \mathbf{q}(t))^{-1}\, \mathbf{q}(t), \tag{2.1}$$

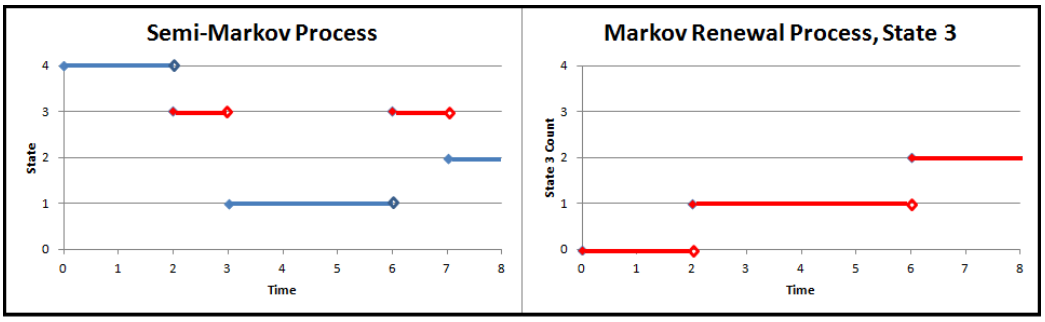

Figure 2.2: Semi-Markov process and related Markov renewal process for State 3

where $\mathbf{q}(t) = \mathbf{P} \circ \mathbf{f}(t)$, $\circ$ is the Hadamard product of two matrices, and $\mathbf{f}$ is a matrix of probability density functions that correspond to $\mathbf{F}$. For an observed semi-Markov process, the corresponding Markov renewal process, $\widetilde{N}$, has an expectation, $\widetilde{\Lambda}$ =E[$\widetilde{N}_j$], for all j in S. The vector value $\widetilde{M} = \widetilde{N} - \widetilde{\Lambda}$ represents the unexplained variation or error of the Markov renewal process which forms a Martingale.

A martingale, $M(t)$, is a stochastic process such that $\mathrm{E}[M(t)] < \infty$ for all t and $\mathrm{E}[M(t)|M(s)] = M(s)$ for $s < t$. Random walks and classical Brownian motion are two common examples of martingales. Submartingales and supermartingales are two extended classes of martingales. They are both stochastic processes with $\mathrm{E}[M(t)] < \infty$, $\forall\ t$, but a submartingale has $\mathrm{E}[M(t)|M(s)] \geq M(s)$, $s < t$, while a supermartingale has $\mathrm{E}[M(t)|M(s)] \leq M(s)$, $s < t$. A true martingale is, in fact, a special case of both a submartingale and a supermartingale. Martingales, submartingales, and supermartingales all belong to a larger class called semimartingales [16].

2.3 Semi-Markov Process Diagnostics

In most modeling realms, diagnostic tools are used to assess the model fit with respect to observed data. For multi-state models, including semi-Markov processes, model diagnostics are underdeveloped when compared with other areas such as linear modeling. While a variety of diagnostic tools have been developed, the majority only work with very

specific types of data and do not provide a clear metric for whether or not the model truly fits the data. Titman and Sharples [4] provide a comprehensive review of the various techniques for diagnosing model fit in multi-state models. Their review includes the shortcomings of most diagnostic tools with respect to panel-observed medical data.

Of the diagnostic tools reviewed, prevalence counts and Chi-squared goodness of fit metrics show promise in assessing model fit but do not satisfy the requirements of a formal diagnostic test. Prevalence counts [17] compare the number of times a state is observed with the number of expected observations of the state for a series of fixed time. The comparison uses the metric

$$H_t = \frac{(O_t - E_t)^2}{E_t}, \tag{2.2}$$

where O is the observed count and E is the expected count. Large values of H indicate poor model fit, but prevalence counts cannot formally assess model fit because the distribution of H is unknown. Under a balanced observation scheme, meaning the observations occur at regular, set intervals,

$$H = \sum_t \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{(O_{tij} - E_{tij})^2}{E_{tij}} \tag{2.3}$$

follows a Chi-squared distribution where t is the set of observation times, i is the occupied state at t, and j is the occupied state at $t + 1$ [18]. Unfortunately, balanced observation schemes are rarely attained in data collection. While irregularly observed data can be tested with Equation 2.3 [19], the resulting statistic does not consistently follow the Chi-squared distribution in all cases.

The two metrics in Equations 2.2 and 2.3 are applications of the generalized Pearson Chi-squared test often used for assessing goodness of fit [20]. The test statistic is created by grouping the data into a series of n bins, then computing the following value:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \tag{2.4}$$

where O is the number of observed instances in bin i and E is the expected number of instances in bin i. The statistic follows a Chi-squared distribution with $n - p$ degrees

of freedom where p is the number of estimated parameters plus 1. Structural equation modeling uses the upper diagonal of a covariance matrix such that

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=i}^{n} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}, \tag{2.5}$$

where O_{ij} and E_{ij} are the observed and expected covariances between States i and j respectively [21]. This statistic follows a Chi-squared distribution with $\frac{1}{2}n(n+1)$ degrees of freedom.

The lack of a true formal test for multi-state models, including semi-Markov processes, motivates the development of the goodness of fit metric in following chapters. This metric builds off the previous attempts outlined in Equations 2.2-2.5 while incorporating the inherent structure of the semi-Markov process and its related Markov renewal process. The next chapter develops the covariance structure for a Markov renewal process as a surrogate for the semi-Markov process covariance.

III. The Covariance of a Markov Renewal Process

The Markov renewal process $\widetilde{N}(t)$, defined in Section 2.2, is represented by a vector containing a count of the number of transitions into each of the n states at time t. A transition probability matrix, **P**, and a sojourn time distribution matrix, **F**, uniquely define a specific Markov renewal process. While the moments of the Markov renewal process are well defined, see Appendix A, there is no explicit formulation for the covariance between the state counts. This covariance, Σ_{ij}, describes how the counts of States i and j within the Markov renewal process change together, or co-vary, and can be expressed as a function of the correlation between the states and the individual variances of the states. Traditionally, the covariance is used in goodness of fit testing and will be required for the metric developed in Chapter 4. This chapter describes how to compute the covariance for a Markov renewal process and illustrates the computation using a notional semi-Markov process.

3.1 Calculating the Markov Renewal Process Covariance

Let $\widetilde{N}(t)$ be a column vector of state counts, $\widetilde{\Lambda}(t)$ be a column vector of expected state counts, and $\widetilde{M}(t)$ be a column vector of state Martingales at time t as defined in Chapter 2. The covariance between state counts is then given by:

$$\begin{aligned}\Sigma[\widetilde{N}(t)] &= \mathrm{Cov}[\widetilde{N}(t), \widetilde{N}(t)] = \mathrm{E}[(\widetilde{N}(t) - \mathrm{E}[\widetilde{N}(t)])(\widetilde{N}(t) - \mathrm{E}[\widetilde{N}(t)])'] \\ &= \mathrm{E}[(\widetilde{N}(t) - \widetilde{\Lambda}(t))(\widetilde{N}(t) - \widetilde{\Lambda}(t))'] = \mathrm{E}[\widetilde{M}(t)\widetilde{M}(t)'] \end{aligned} \tag{3.1}$$

$$= \mathrm{E}\begin{bmatrix} M_1(t)^2 & M_1(t)M_2(t) & \dots & M_1(t)M_n(t) \\ M_1(t)M_2(t) & M_2(t)^2 & \dots & M_2(t)M_n(t) \\ \vdots & & \ddots & \vdots \\ M_1(t)M_n(t) & M_2(t)M_n(t) & \dots & M_n(t)^2 \end{bmatrix}$$

$$= \begin{bmatrix} \mathrm{E}[M_1(t)^2] & \mathrm{E}[M_1(t)M_2(t)] & \dots & \mathrm{E}[M_1(t)M_n(t)] \\ \mathrm{E}[M_1(t)M_2(t)] & \mathrm{E}[M_2(t)^2] & \dots & \mathrm{E}[M_2(t)M_n(t)] \\ \vdots & & \ddots & \vdots \\ \mathrm{E}[M_1(t)M_n(t)] & \mathrm{E}[M_2(t)M_n(t)] & \dots & \mathrm{E}[M_n(t)^2] \end{bmatrix}$$

$$= \begin{bmatrix} \mathrm{Var}[N_1(t)] & \mathrm{E}[M_1(t)M_2(t)] & \dots & \mathrm{E}[M_1(t)M_n(t)] \\ \mathrm{E}[M_1(t)M_2(t)] & \mathrm{Var}[N_2(t)] & \dots & \mathrm{E}[M_2(t)M_n(t)] \\ \vdots & & \ddots & \vdots \\ \mathrm{E}[M_1(t)M_n(t)] & \mathrm{E}[M_2(t)M_n(t)] & \dots & \mathrm{Var}[N_n(t)] \end{bmatrix}.$$

Thus the expectations of the product of two random variables, $\mathrm{E}[M_i(t)M_j(t)]$, comprise the off-diagonal portions of the Markov renewal process covariance matrix. While the variance of $\widetilde{N}(t)$ has a solution in the Laplace domain (Appendix A), $\mathrm{E}[M_i(t)M_j(t)]$ does not at this time.

Although $\Sigma(t)$ cannot currently be calculated directly, simulation will provide a reasonable approximation. This simulation involves collecting N iterations of a semi-Markov process out to a target time t and calculating the Markov renewal processes, $\widetilde{N}_i(t)$ where i goes from 1 to N. The simulated covariance for the Markov renewal process is then computed as:

$$\hat{\Sigma}(t) = \frac{1}{N-1}\sum_{l=1}^{N}(\widetilde{N}_l(t) - \mathrm{E}[\widetilde{N}(t)])(\widetilde{N}_l(t) - \mathrm{E}[\widetilde{N}(t)])^T \tag{3.2}$$

based on the covariance calculation for a sample [22]. For a given semi-Markov process that does not allow self transitions, $\mathrm{E}[\widetilde{N}(t)]$ of the related Markov renewal process [11] is a

constant based on t. In the Laplace domain,

$$\mathcal{L}(\mathrm{E}[\widetilde{N}(t)|X_0 = i]) = \left[(\mathbf{I} - q)^{-1} q\right]_{,i}, \tag{3.3}$$

where X_0 is the initial system state, $q = \mathbf{P} \circ \mathbf{f}(t)$ at each t, and $\mathbf{f}$ is a matrix of probability density functions that correspond to $\mathbf{F}$ [15]. An Euler inversion technique returns the time domain expected value,

$$\mathrm{E}[\widetilde{N}(t)|X_0 = i] = \left[\frac{2e^{at}}{\pi} \int_0^\infty \mathrm{Re}\left(\mathcal{L}(\mathrm{E}[a + iu])\right) \cos(ut)\, \mathrm{d}u\right]_{,i}, \tag{3.4}$$

where a is a point such that $\mathcal{L}(\mathrm{E}[a])$ is continuous to the right and $\mathrm{Re}(\mathcal{L}(\mathrm{E}[a + iu]))$ is the real portion of the complex $\mathcal{L}(\mathrm{E}[a + iu])$ [23]. This integral can be calculated numerically using a Fourier-series summation. Thus only the number of iterations, N, and the time, t, must be selected by the user to estimate the simulated covariance. The lower bounds on both of these parameters depend on the semi-Markov process being modeled.

First, consider a four-state semi-Markov process with a probability transition matrix

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix} \tag{3.5}$$

and a sojourn distribution matrix of

$$\mathbf{F} = \begin{bmatrix} F_{11} & F_{12} & F_{13} & F_{14} \\ F_{21} & F_{22} & F_{23} & F_{24} \\ F_{31} & F_{32} & F_{33} & F_{34} \\ F_{41} & F_{42} & F_{43} & F_{44} \end{bmatrix}. \tag{3.6}$$

Figure 3.1 illustrates the generic semi-Markov process defined by $\mathbf{P}$ and $\mathbf{F}$ from Equations 3.5 and 3.6. To compute the covariance matrix of this fully connected semi-

Markov process, simulate from the model as follows:

$$\widetilde{N} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{S\times 1} \quad \text{and } \Sigma[\widetilde{N}(t)] = \begin{bmatrix} 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 0 \end{bmatrix}_{S\times S};$$

$T = 0$ and $i = 0$;

Set t to the desired length of the simulation;

Set N to the desired number of simulation iterations;

Calculate $\mathrm{E}[\widetilde{N}(t)]$ using Equation 3.4;

while $i < N$ **do**

Select an initial state, S_c, based on the starting distribution of the system (i.e. the probability that the systems begins in a particular start);

while $T < t$ **do**

Draw a random variable, X_1 ~Uniform(0,1);

Use $\mathbf{P}$ to find the next state, S_n, based on X_1 and S_c;

Draw a random variable, X_2 ~Uniform(0,1);

Find the transition time, T_n, to S_n by making a random draw from $\mathbf{F}_{S_c S_n}$ based on X_2;

$T = T + T_n$;

if $T <= t$ **then**

$S_c = S_n$;

$\widetilde{N}_{S_n} = \widetilde{N}_{S_n} + 1$;

end

end

$\Sigma[\widetilde{N}(t)] = \Sigma[\widetilde{N}(t)] + \frac{1}{N-1}(\widetilde{N} - \mathrm{E}[\widetilde{N}(t)])(\widetilde{N} - \mathrm{E}[\widetilde{N}(t)])'$;

end

Algorithm 1: Covariance Simulation

At this point, $\widetilde{N}$ equals the total number of times each state was visited during the simulation prior to t.

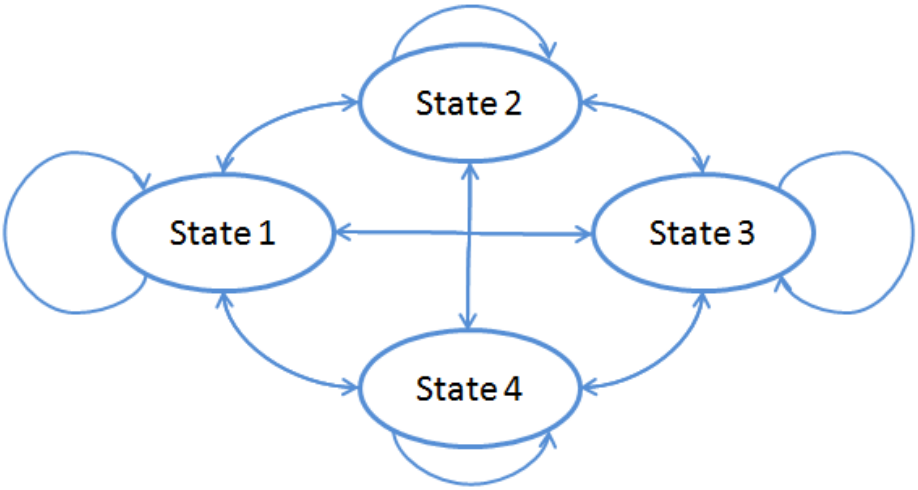

Figure 3.1: Fully connected four state model with self transitions

3.2 The Covariance of the Sample Problem

The Covariance Simulation is demonstrated for a sample problem using assumed values. This process does not allow instantaneous transitions back to the current state, State 1 cannot transition directly to State 3, and State 2 cannot transition directly to State 4. Let

$$\mathbf{P} = \begin{bmatrix} 0 & 0.75 & 0 & 0.25 \\ 0.5 & 0 & 0.5 & 0 \\ 0.25 & 0.45 & 0 & 0.3 \\ 0.15 & 0.35 & 0.5 & 0 \end{bmatrix} \tag{3.7}$$

and

$$\mathbf{F} = \begin{bmatrix} 0 & \Gamma(2,2) & 0 & \Gamma(2,2) \\ \Gamma(2,1) & 0 & \Gamma(0.5,4) & 0 \\ \Gamma(0.5,1) & \Gamma(0.5,4) & 0 & \Gamma(2,2) \\ \Gamma(0.5,4) & \Gamma(2,2) & \Gamma(2,1) & 0 \end{bmatrix}, \tag{3.8}$$

where

$$\Gamma(\alpha,\beta) = \int_0^x \frac{1}{\Gamma(\alpha)\beta^\alpha} u^{\alpha-1} e^{-\frac{u}{\beta}} du. \tag{3.9}$$

Running N =10,000 iterations of the Covariance Simulation at t =1,000 time units results in the following estimated covariance:

$$\hat{\Sigma}[\widetilde{N}(1000)] = \begin{bmatrix} 59.297 & 38.523 & -9.579 & -6.323 \\ 38.523 & 75.407 & 26.323 & -15.108 \\ -9.579 & 26.323 & 70.904 & 18.971 \\ -6.323 & -15.108 & 18.971 & 37.058 \end{bmatrix}. \tag{3.10}$$

The true variance of the Markov renewal process can be calculated using the equation for the $\text{Var}^*[\widetilde{N}(t)]$ found in Appendix A and

$$\text{Var}[\widetilde{N}(t)] = \frac{2e^{at}}{\pi} \int_0^\infty \text{Re}(\text{Var}^*[a+iu]) \cos(ut)\, \text{d}u. \tag{3.11}$$

Based on the semi-Markov process variances, the true $\Sigma[\widetilde{N}(1000)]$ diagonal values are 59.2, 75.5, 70.8, 37.1 respectively. The estimates produced by the Covariance Simulation are almost identical to the true values.

The choice of N will affect the estimate of $\Sigma(t)$. Increasing the number of iterations used to estimate the covariance matrix will result in a closer approximation, but at the cost of increased computational time. To illustrate the effect of the number of iterations, the covariance estimate of the sample problem is computed at t=1,000 time units using N =100, 500, 1,000, 5,000, 10,000, 50,000, and 100,000 iterations, repeated across R=1,000 samples. The average estimate of the covariance matrix and its corresponding 95% confidence interval around each $\hat{\Sigma}_{ij}$ value was computed. The results for the various iteration sizes are as follows:

$\hat{\Sigma}[\widetilde{N}(1,000)]$ and (95% confidence interval) based on N=100 iterations =

$$\begin{bmatrix} 59.8(47.2,74.1) & 38.7(26.8,50.7) & -9.9(-20.7,1.3) & -6.4(-14.5,1.2) \\ 38.7(26.8,50.7) & 76.0(60.1,94.3) & 26.6(14.3,39.7) & -15.1(-24.0,-6.5) \\ -9.9(-20.7,1.3) & 26.6(14.3,39.7) & 71.7(55.6,88.4) & 19.2(10.8,28.1) \\ -6.4(-14.5,1.2) & -15.1(-24.0,-6.5) & 19.2(10.8,28.1) & 37.3(29.1,46.0) \end{bmatrix} \tag{3.12}$$

$\hat{\Sigma}[\widetilde{N}(1,000)]$ and (95% confidence interval) based on N=500 iterations =

$$\begin{bmatrix} 59.1(52.7,65.5) & 38.6(33.1,44.5) & -9.6(-14.4,-4.4) & -6.4(-10.1,-3.2) \\ 38.6(33.1,44.5) & 75.6(67.8,83.9) & 26.2(20.5,31.8) & -15.3(-19.7,-11.1) \\ -9.6(-14.4,-4.4) & 26.2(20.5,31.8) & 70.9(64.0,78.3) & 19.0(15.0,23.1) \\ -6.4(-10.1,-3.2) & -15.3(-19.7,-11.1) & 19.0(15.0,23.1) & 37.2(33.3,41.3) \end{bmatrix}$$

$\hat{\Sigma}[\widetilde{N}(1,000)]$ and (95% confidence interval) based on N=1,000 iterations =

$$\begin{bmatrix} 59.2(54.9,63.8) & 38.6(34.6,42.5) & -9.7(-13.1,-6.4) & -6.4(-8.9,-4.0) \\ 38.6(34.6,42.5) & 75.6(70.2,81.1) & 26.2(22.4,29.9) & -15.3(-18.2,-12.4) \\ -9.7(-13.1,-6.4) & 26.2(22.4,29.9) & 70.9(65.8,76.0) & 19.0(16.1,21.7) \\ -6.4(-8.9,-4.0) & -15.3(-18.2,-12.4) & 19.0(16.1,21.7) & 37.2(34.5,40.1) \end{bmatrix}$$

$\hat{\Sigma}[\widetilde{N}(1,000)]$ and (95% confidence interval) based on N=5,000 iterations =

$$\begin{bmatrix} 59.1(57.3,61.0) & 38.5(36.8,40.2) & -9.7(-11.1,-8.2) & -6.4(-7.5,-5.3) \\ 38.5(36.8,40.2) & 75.5(73.2,78.0) & 26.2(24.4,27.9) & -15.3(-16.5,-14.0) \\ -9.7(-11.1,-8.2) & 26.2(24.4,27.9) & 70.9(68.5,73.2) & 19.0(17.8,20.2) \\ -6.4(-7.5,-5.3) & -15.3(-16.5,-14.0) & 19.0(17.8,20.2) & 37.2(35.9,38.4) \end{bmatrix}$$

$\hat{\Sigma}[\widetilde{N}(1,000)]$ and (95% confidence interval) based on N=10,000 iterations =

$$\begin{bmatrix} 59.1(57.9,60.4) & 38.5(37.3,39.7) & -9.6(-10.7,-8.5) & -6.4(-7.2,-5.7) \\ 38.5(37.3,39.7) & 75.5(73.7,77.3) & 26.2(25.0,27.5) & -15.3(-16.2,-14.4) \\ -9.6(-10.7,-8.5) & 26.2(25.0,27.5) & 70.9(69.3,72.5) & 18.9(18.1,19.8) \\ -6.4(-7.2,-5.7) & -15.3(-16.2,-14.4) & 18.9(18.1,19.8) & 37.2(36.3,38.0) \end{bmatrix}$$

$\hat{\Sigma}[\widetilde{N}(1,000)]$ and (95% confidence interval) based on N=50,000 iterations =

$$\begin{bmatrix} 59.2(58.6,59.7) & 38.5(38.0,39.1) & -9.7(-10.1,-9.2) & -6.4(-6.8,-6.1) \\ 38.5(38.0,39.1) & 75.5(74.7,76.3) & 26.2(25.6,26.8) & -15.3(-15.7,-14.9) \\ -9.7(-10.1,-9.2) & 26.2(25.6,26.8) & 70.9(70.1,71.6) & 18.9(18.5,19.3) \\ -6.4(-6.8,-6.1) & -15.3(-15.7,-14.9) & 18.9(18.5,19.3) & 37.2(36.8,37.5) \end{bmatrix}$$

$\hat{\Sigma}[\widetilde{N}(1,000)]$ and (95% confidence interval) based on N=100,000 iterations =

$$\begin{bmatrix} 59.2(58.7,59.6) & 38.5(38.1,38.9) & -9.6(-10.0,-9.3) & -6.4(-6.6,-6.2) \\ 38.5(38.1,38.9) & 75.5(74.9,76.1) & 26.2(25.8,26.6) & -15.3(-15.5,-15.0) \\ -9.6(-10.0,-9.3) & 26.2(25.8,26.6) & 70.9(70.3,71.4) & 18.9(18.7,19.2) \\ -6.4(-6.6,-6.2) & -15.3(-15.5,-15.0) & 18.9(18.7,19.2) & 37.2(36.9,37.4) \end{bmatrix}$$

From the covariance estimates, the average diagonal values mirror the true variance values with 500 iterations, albeit with wide 95% confidence intervals of 10 units which indicates that a single set of 500 interations may not provide a good estimate of $\Sigma[\widetilde{N}(t)]$. Figure 3.2 shows the decreasing confidence interval sizes as the number of iterations increase. With 50,000 iterations, the confidence intervals drop to 1 unit. In this scenario, it takes the same amount of time to estimate the covariance with a single set of 50,000 iterations as it does to average 1,000 samples of 500 iterations each. However, the later method will provide a more accurate estimate of the true covariance matrix since it relies on averaging a series of estimates rather than using in single point estimate.

While t is dictated by the observed data sample, the minimum semi-Markov process runtime required for accurate covariance estimates depends largely on the particular model and how quickly it reaches a steady state on average. For instance, a model that averages 10 time units before the first transition cannot have an accurate covariance estimate for $\widetilde{N}(5)$ because the Markov renewal process counts will be highly dependent on the initial states. Meanwhile a small model that averages 0.1 time units between transitions can have an

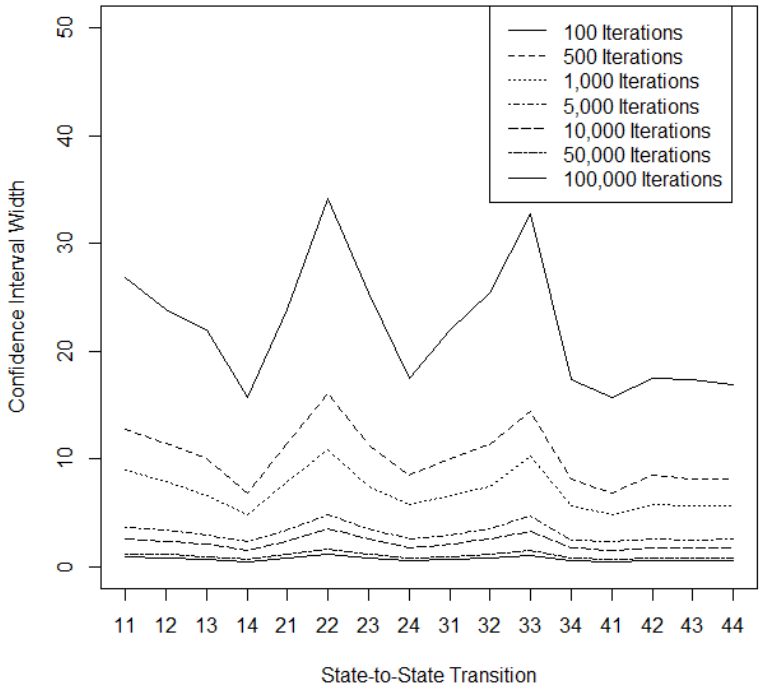

Figure 3.2: State-to-State 95% confidence interval widths for model covariance

an accurate covariance estimate for $\widetilde{N}(5)$ because the system will have enough transitions to overcome the effect of the initial state. To illustrate, the correlation matrix is used to normalize the covariance estimates across time. The correlation matrix is calculated by

$$\text{Corr}_{ij}[\widetilde{N}(t)] = \frac{\hat{\Sigma}_{ij}[\widetilde{N}(t)]}{\sqrt{\Sigma_{ii}[\widetilde{N}(t)]\Sigma_{jj}[\widetilde{N}(t)]}}. \tag{3.13}$$

Matrices of the average correlation with 95% confidence intervals around each Corr_{ij} value for $\widetilde{N}(1)$, $\widetilde{N}(20)$, $\widetilde{N}(50)$, and $\widetilde{N}(100)$ follow:

$\text{Corr}[\widetilde{N}(1)]$ and (95% confidence interval) =

$$\begin{bmatrix} 1(0.89,1.1) & -0.05(-0.11,0.01) & 0.02(-0.03,0.08) & 0(-0.05,0.06) \\ -0.05(-0.11,0.01) & 1(0.85,1.15) & 0.33(0.24,0.44) & -0.05(-0.09,-0.02) \\ 0.02(-0.03,0.08) & 0.33(0.24,0.44) & 1(0.86,1.14) & 0.03(-0.01,0.09) \\ 0(-0.05,0.06) & -0.05(-0.09,-0.02) & 0.03(-0.01,0.09) & 1(0.6,1.44) \end{bmatrix} \tag{3.14}$$

$\text{Corr}[\widetilde{N}(20)]$ and (95% confidence interval) =

$$\begin{bmatrix} 1(0.93,1.08) & 0.5(0.44,0.56) & -0.12(-0.17,-0.07) & -0.12(-0.17,-0.07) \\ 0.5(0.44,0.56) & 1(0.92,1.08) & 0.37(0.31,0.44) & -0.28(-0.33,-0.23) \\ -0.12(-0.17,-0.07) & 0.37(0.31,0.44) & 1(0.93,1.08) & 0.32(0.27,0.38) \\ -0.12(-0.17,-0.07) & -0.28(-0.33,-0.23) & 0.32(0.27,0.38) & 1(0.93,1.08) \end{bmatrix}$$

$\text{Corr}[\widetilde{N}(50)]$ and (95% confidence interval) =

$$\begin{bmatrix} 1(0.93,1.07) & 0.55(0.49,0.61) & -0.14(-0.19,-0.09) & -0.13(-0.18,-0.08) \\ 0.55(0.49,0.61) & 1(0.93,1.08) & 0.36(0.31,0.42) & -0.28(-0.34,-0.23) \\ -0.14(-0.19,-0.09) & 0.36(0.31,0.42) & 1(0.93,1.08) & 0.35(0.3,0.41) \\ -0.13(-0.18,-0.08) & -0.28(-0.34,-0.23) & 0.35(0.3,0.41) & 1(0.93,1.08) \end{bmatrix}$$

$\text{Corr}[\widetilde{N}(100)]$ and (95% confidence interval) =

$$\begin{bmatrix} 1(0.93,1.07) & 0.56(0.5,0.63) & -0.14(-0.2,-0.09) & -0.13(-0.19,-0.08) \\ 0.56(0.5,0.63) & 1(0.93,1.08) & 0.36(0.3,0.42) & -0.29(-0.34,-0.24) \\ -0.14(-0.2,-0.09) & 0.36(0.3,0.42) & 1(0.93,1.07) & 0.36(0.31,0.42) \\ -0.13(-0.19,-0.08) & -0.29(-0.34,-0.24) & 0.36(0.31,0.42) & 1(0.93,1.08) \end{bmatrix}.$$

By 50 time units, the average correlation matrix stabilizes and does not change as t increases; although, the confidence interval widths can improve slightly at higher t values, see Figure 3.3. The correlation matrix for $\widetilde{N}(20)$ approaches the correlation matrix for $\widetilde{N}(50)$. Meanwhile, the correlation matrix for $\widetilde{N}(1)$ provides a poor estimate with respect to the correlation matrices at larger t values. The underlying reason that the correlation matrix requires time to stabilize is due to the number of transitions that occur in a given

time. The particular model for the sample problem only has 0.37 expected transitions at $t = 1$, 7.44 expected transitions at $t = 20$, and 18.67 expected transitions at $t = 50$. Based on the testing in this example, t should be set so that the number of expected transitions is at least greater than twice the number of model states.

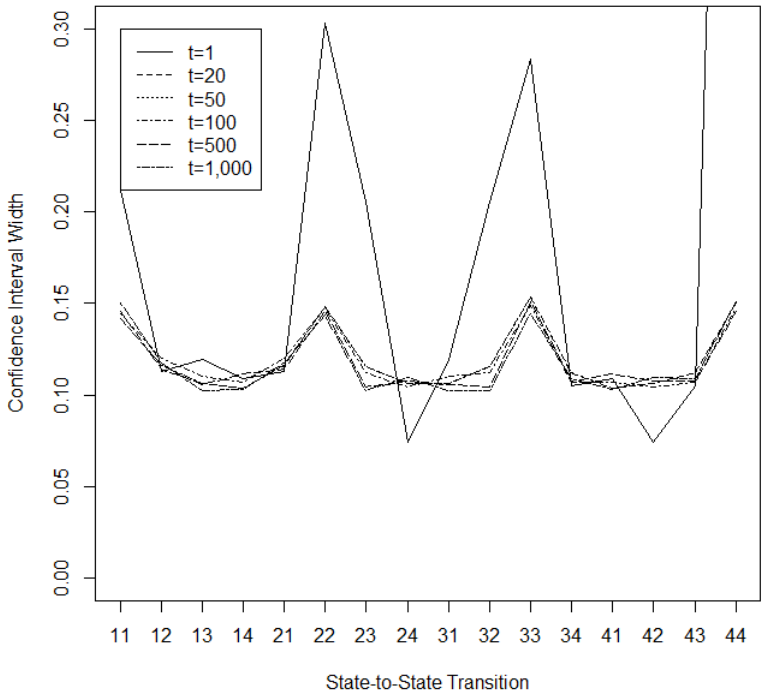

Figure 3.3: State-to-State 95% confidence interval widths for model correlation

Using Algorithm 1, the covariance of a Markov renewal process can be reasonably estimated. Assuming the data sample dictates the maximum value of t, the modeler is only required to specify the number of simulation iterations, N. Using the estimate of the model covariance, the goodness of fit of the theorized semi-Markov process with respect to the observed data sample can be determined.

IV. Goodness of Fit Metrics in Semi-Markov Processes

4.1 The Goodness of Fit Metric

The basic premise behind a goodness of fit metric is to compare observed data points with their expectations. When the two values are similar, the data is said to have a close fit with the model. In the case of a semi-Markov process, the observed process value at a time, t, is the state of the system at t, while the expected process value is the average state of the system at t. For a fixed time, t, a semi-Markov process only has one comparison available, namely the actual occupied state versus the expected occupied state. This lone comparison does not provide enough information to determine the level of similarity between the expected and observed values, especially when the expected occupied state is constant once the model reaches steady-state.

Consider a four state semi-Markov process that, on average, spends 40% of its time in State 1, 10% in State 2, 10% in State 3, and 40% in State 4. The long run expected value for the semi-Markov process would be 2.5 which is not an actual state and is the furthest possible value from the two most commonly occuring states in the system. In this scenario, 80% of the time the absolute value of the difference between expected and observed values is 1.5 and 20% of the time the absolute value is 0.5. Simply reordering the states by swapping State 1 with State 2 and State 3 with State 4 would maintain the expectation of 2.5, but the absolute value of the difference between expected and observed values would become 0.5 for 80% of the time and 1.5 for the remaining 20% of the time. This arbitrary change would result in different model fit estimates for two equivalent models.

On the other hand, as discussed in Section 2.2, the semi-Markov process possesses an underlying Markov renewal process that contains enough information to determine the level of similarity between the observed data and the model. At t, the observed Markov renewal process is a vector of the number of visits to each individual state. The corresponding

expected number of visits to each state can be calculated in the transform domain using the generating function, Ψ_z, defined in Pyke [15] as follows:

$$\Psi_z = 1 - (1 - z)(\mathbf{I} - \mathbf{q})^{-1}\mathbf{q}[z\mathbf{I} + (1 - z)\ {}_d\{(\mathbf{I} - \mathbf{q})^{-1}\}]^{-1}. \tag{4.1}$$

Appendix A shows the full derivation of the expected value and the variance for a Markov renewal process in the transform domain based on Ψ_z. The time domain expected value and variance are calculated via a Laplace transform inversion using Equations 3.4 and 3.11.

Once a proposed semi-Markov process is fit to observed data, a sample covariance matrix may be calculated from the observed and expected values of the corresponding Markov renewal process. The covariance matrix cannot be estimated using Equation 3.2 because there is only one $\widetilde{N}(t)$ vector value for any given t. Instead, the sample covariance calculation involves finding the expected Markov renewal process counts at every transition time during the observed data run. These expected counts are subtracted from the actual state counts at their respective times. Assuming the semi-Markov process contains n states and with N observed transitions, the result is a series of N vectors, each of length n. Then the covariance of the Markov renewal process at t_N is

$$\Sigma(t_N) = \frac{1}{N-1}\sum_{k=1}^{N}(O(t_k) - \mathrm{E}(t_k))(O(t_k) - \mathrm{E}(t_k))', \tag{4.2}$$

where $O_i(t_k)$ is the observed count and $\mathrm{E}_i(t_k)$ is the expected count of State i at the k-th transition time. Because the variances are time dependent, the covariance of the Markov renewal process is also time dependent. Traditional goodness of fit testing based on covariance matrices is not time dependent and results in a goodness of fit metric, GoF, as follows:

$$GoF = \sum_{i=1}^{n}\sum_{j=i}^{n}\frac{(\mathrm{O}(\Sigma_{ij}) - \mathrm{E}(\Sigma_{ij}))^2}{|\mathrm{E}(\Sigma_{ij})|}, \tag{4.3}$$

which follows a Chi-squared distribution with $\nu = \frac{1}{2}n(n+1)$ degrees of freedom. However, the time dependency of the Markov renewal process alters the GoF equation slightly so

that

$$GoF(t) = \sum_{i=1}^{n} \sum_{j=i}^{n} \frac{(\mathrm{O}(\Sigma_{ij}(t)) - \mathrm{E}(\Sigma_{ij}(t)))^2}{|\mathrm{E}(\Sigma_{ij}(t))|}. \tag{4.4}$$

While this seems like a very minor change, as Section 4.3 will show, the inclusion of time has major ramifications for the $GoF(t)$ distribution.

4.2 Baseline Model

A baseline semi-Markov process from the sample problem of Section 3.2 will be used to illustrate the behavior of the $GoF(t)$ in Equation 4.4. Recall that the baseline semi-Markov process consisted of four states that are nearly fully connected. A visualization of the baseline semi-Markov process appears in Figure 4.1. The baseline semi-Markov process was defined by the transition probability matrix

$$\mathbf{P} = \begin{bmatrix} 0 & 0.75 & 0 & 0.25 \\ 0.5 & 0 & 0.5 & 0 \\ 0.25 & 0.45 & 0 & 0.3 \\ 0.15 & 0.35 & 0.5 & 0 \end{bmatrix} \tag{4.5}$$

and a sojourn distribution matrix of

$$\mathbf{F} = \begin{bmatrix} 0 & \Gamma(2,2) & 0 & \Gamma(2,2) \\ \Gamma(2,1) & 0 & \Gamma(0.5,4) & 0 \\ \Gamma(0.5,1) & \Gamma(0.5,4) & 0 & \Gamma(2,2) \\ \Gamma(0.5,4) & \Gamma(2,2) & \Gamma(2,1) & 0 \end{bmatrix}, \tag{4.6}$$

where

$$\Gamma(\alpha,\beta) = \int_0^x \frac{1}{\Gamma(\alpha)\beta^{\alpha}} u^{\alpha-1} e^{-\frac{u}{\beta}} du. \tag{4.7}$$

The result is a baseline kernel matrix of

$$\mathbf{G(t)} = \begin{bmatrix} 0 & 0.1875te^{-\frac{t}{2}} & 0 & 0.0625te^{-\frac{t}{2}} \\ 0.5te^{-t} & 0 & 1.7725t^{-0.5}e^{-\frac{t}{4}} & 0 \\ 0.8862t^{-0.5}e^{-t} & 1.5952t^{-0.5}e^{-\frac{t}{4}} & 0 & 0.075te^{-\frac{t}{2}} \\ 0.5317t^{-0.5}e^{-\frac{t}{4}} & 0.0875te^{-\frac{t}{2}} & 0.5te^{-t} & 0 \end{bmatrix}. \tag{4.8}$$

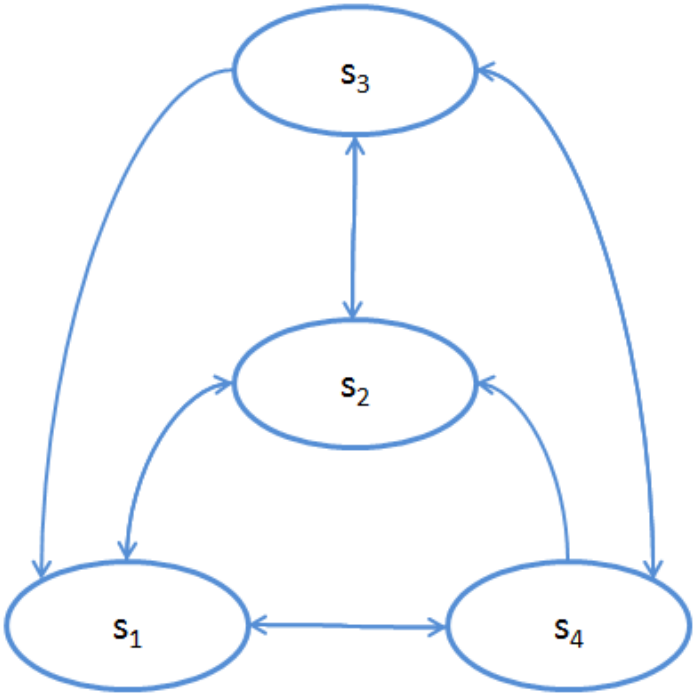

Figure 4.1: The baseline semi-Markov process connected graph

Using the baseline semi-Markov process, the covariance matrix of the related Markov renewal process at time t is calculated as

$$\Sigma[\widetilde{N}(t)] = \begin{bmatrix} \mathrm{Var}(\widetilde{N}_1(t)) & \mathrm{Cov}(\widetilde{N}_1(t)\widetilde{N}_2(t)) & \mathrm{Cov}(\widetilde{N}_1(t)\widetilde{N}_3(t)) & \mathrm{Cov}(\widetilde{N}_1(t)\widetilde{N}_4(t)) \\ \mathrm{Cov}(\widetilde{N}_2(t)\widetilde{N}_1(t)) & \mathrm{Var}(\widetilde{N}_2(t)) & \mathrm{Cov}(\widetilde{N}_2(t)\widetilde{N}_3(t)) & \mathrm{Cov}(\widetilde{N}_2(t)\widetilde{N}_4(t)) \\ \mathrm{Cov}(\widetilde{N}_3(t)\widetilde{N}_1(t)) & \mathrm{Cov}(\widetilde{N}_3(t)\widetilde{N}_2(t)) & \mathrm{Var}(\widetilde{N}_3(t)) & \mathrm{Cov}(\widetilde{N}_3(t)\widetilde{N}_4(t)) \\ \mathrm{Cov}(\widetilde{N}_4(t)\widetilde{N}_1(t)) & \mathrm{Cov}(\widetilde{N}_4(t)\widetilde{N}_2(t)) & \mathrm{Cov}(\widetilde{N}_4(t)\widetilde{N}_3(t)) & \mathrm{Var}(\widetilde{N}_4(t)) \end{bmatrix}. \tag{4.9}$$

Based on Equation 3.1, $\Sigma[\widetilde{N}(t)] = [\mathrm{E}[M_i(t)M_j(t)]] = [\mathrm{Cov}[N_i(t)N_j(t)]] = [\mathrm{E}[(N_i(t) - \mathrm{E}[N_i(t)])(N_j(t) - \mathrm{E}[N_j(t)])]]$ which can be simulated. Thus, the resulting $GoF(t)$ metric for a sample run of the baseline model is

$$GoF(t) = \sum_{i=1}^{4}\sum_{j=i}^{4} \frac{(\mathrm{O}(\Sigma_{ij}(t)) - \mathrm{E}(\Sigma_{ij}(t)))^2}{|\mathrm{E}(\Sigma_{ij}(t))|}. \tag{4.10}$$

Since there are ten elements along the upper diagonal of each matrix used in the $GoF(t)$ calculation, the initial expectation is for the $GoF(t)$ to follow a Chi-squared distribution

with nine degrees of freedom. However, the inclusion of the time aspect actually results in the distribution of $GoF(t)$ more closely following a transformed Chi-squared distribution.

4.3 Goodness of Fit Metric Distribution

The distribution of $GoF(t)$ is easily observed through simulation. The simulation involves collecting a sample run of the baseline model out to time t. For each transition, k, from time 0 to t, the Markov renewal process count $\widetilde{N}(t_k)$ and the $\mathrm{E}[\widetilde{N}(t_k)]$ are calculated, where t_k is the time of the k-th transition. The observed covariance between States i and j for the sample run becomes

$$\mathrm{O}(\Sigma_{ij}(t))) = \frac{1}{k-1}\sum_{l=1}^{k}(N_i(t_l) - \mathrm{E}(N_i(t_l))(N_j(t_l) - \mathrm{E}(N_j(t_l)). \tag{4.11}$$

The expected covariance matrix is calculated through simulation as described in Section 3.1. The $GoF(t)$ metric is found via Equation 4.10.

Figure 4.2 illustrates the distribution of $GoF(t)$ for the baseline model at t=50 after collecting 10,000 samples. As a point of reference, the Chi-squared distribution with 9 degrees of freedom also appears on the graph. While the metric distribution and theoretical distribution are similarly shaped, the metric definitely does not follow a Chi-squared distribution with 9 degrees of freedom due to the shift along the x-axis. In addition, the distribution of the metric changes over time as illustrated in Figure 4.3 which has completely different x-axes in each portion of the figure, attributable to the changing distribution.

Although the $GoF(t)$ metric distribution differs for the various runtimes by shifting along the x-axis, the shapes of the distributions follow a general pattern. Figure 4.4 illustrates the upper and lower 95% confidence interval cutoffs, means, and medians for the metric distributions at various runtimes. All four measures appear linear with respect to the model runtime, clearly demonstrating the distribution shift with respect to time. Since the

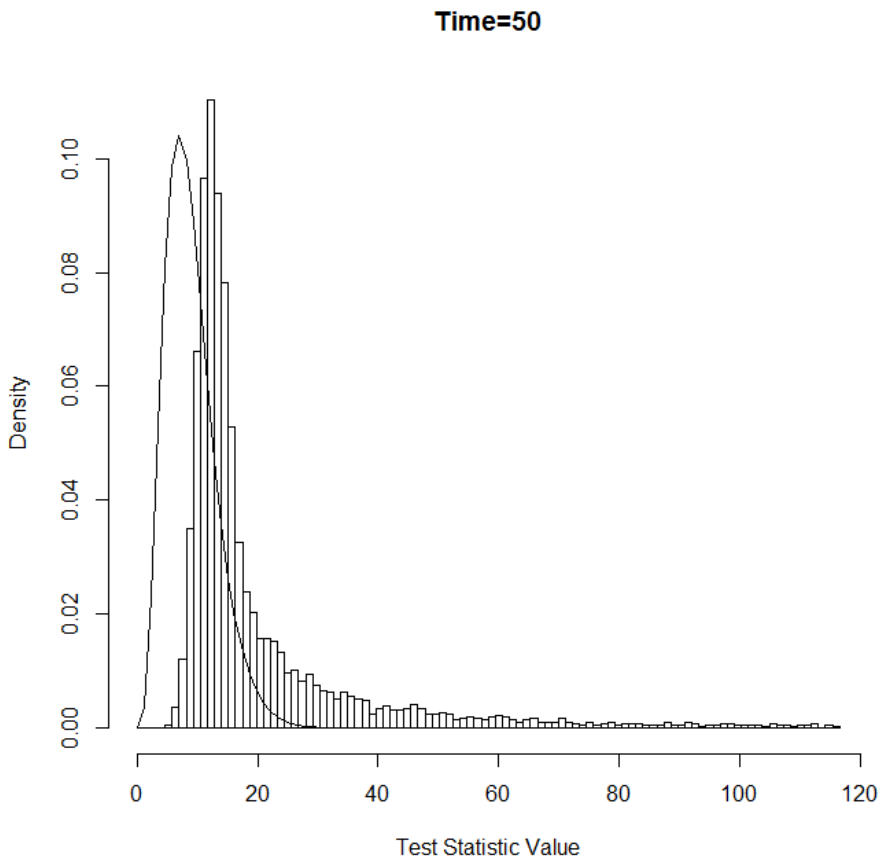

Figure 4.2: $GoF(t)$ distribution for the baseline model at time 50 after 10,000 iterations

$GoF(t)$ distribution does not follow a true Chi-squared distribution, the testing in Chapters 4-6 utilize simulation in place of a theoretical distribution to evaluate models.

4.4 $GoF(t)$ Metric Sensitivity

The ability of the $GoF(t)$ metric to distinguish between changes in the model is also demonstrated through simulation. This sensitivity testing involves collecting data from a similar, known model and treating it as if it came from the target model. To accomplish the sensitivity testing, a simulation from an altered version of the baseline model, B, is run, where at least one value in $\mathbf{P}$ or at least one sojourn distribution has been changed. The

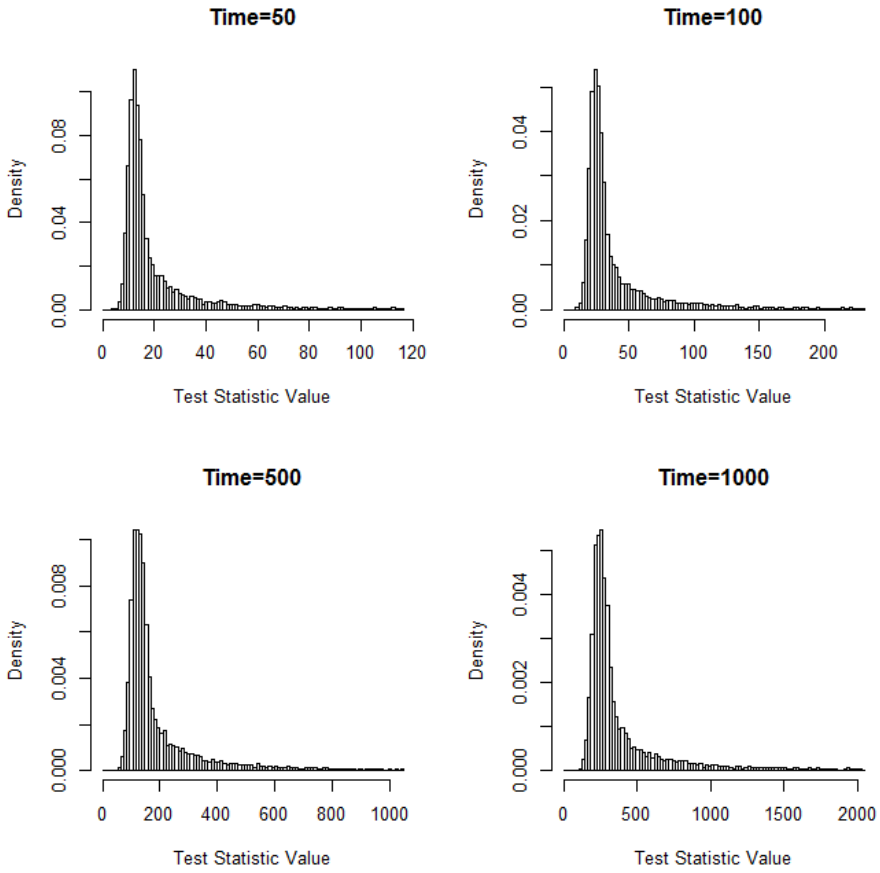

Figure 4.3: $GoF(t)$ distribution at various times after 10,000 iterations

altered model is called the test model, T. Recall that a semi-Markov process is completely defined by $\mathbf{P}$ and the sojourn distributions in $\mathbf{F}$. After running T out to time t, the Markov renewal process count $\widetilde{N_{\mathrm{T}}}(t_k)$ and the $\mathrm{E}[\widetilde{N_{\mathrm{B}}}(t_k)]$ are calculated for the T and B models at the time of the k-th transition (t_k). For testing purposes, the transitions in the test model sample run will determine the individual t_k values. The observed covariance between States i and j for the sample run becomes

$$\mathrm{O}(\Sigma_{ij}(t))) = \frac{1}{k-1}\sum_{l=1}^{k}(N_{T,i}(t_l) - \mathrm{E}(N_{B,i}(t_l))(N_{T,j}(t_l) - \mathrm{E}(N_{B,j}(t_l)), \tag{4.12}$$

where $N_{T,i}(t_k)$ is the Markov renewal process count for the i-th state of the test model at time t_k and $N_{B,i}(t_k)$ is the Markov renewal process count for the i-th state of the baseline

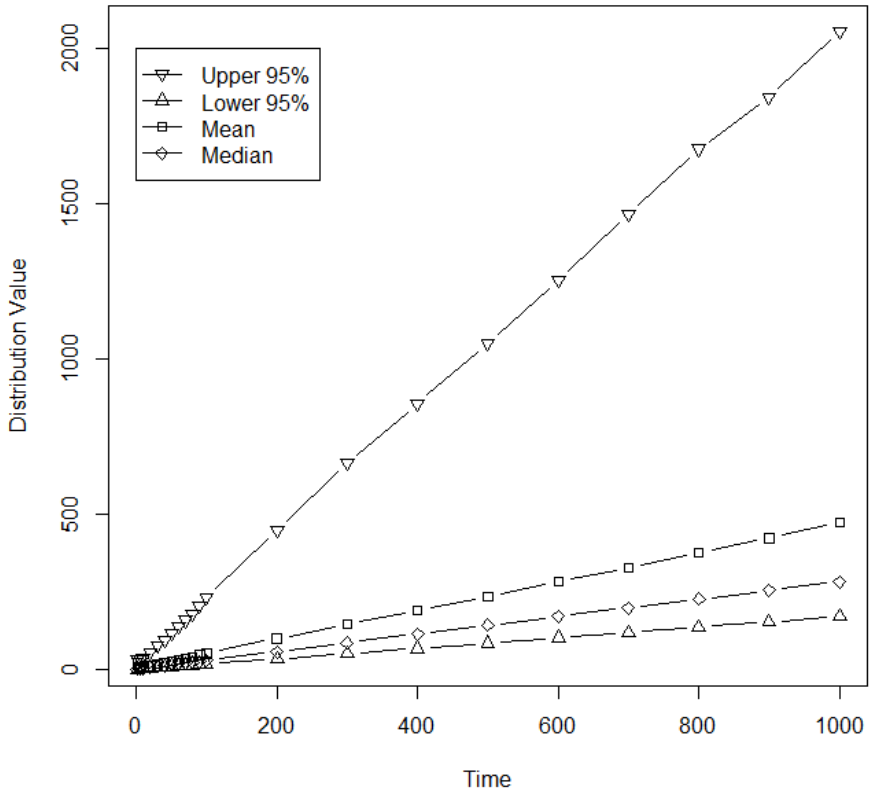

Figure 4.4: $GoF(t)$ distribution measures of central tendency and dispersion

model at time t_k. The expected covariance matrix is still based solely on the baseline model from Equation 3.2 and $GoF(t)$ is found via Equation 4.10.

4.4.1 Simulation.

In this simulation, the baseline model consists of **P** defined in Equation 4.5 and the sojourn distributions defined in Equation 4.6. There are four unique sojourn distributions ($\Gamma(2,2)$, $\Gamma(2,1)$, $\Gamma(0.5,4)$, and $\Gamma(0.5,1)$) which can be defined as a vector of α values (2,2,0.5,0.5) and a vector of β values (2,1,4,1) respectively. For testing purposes, 6 additional α and β vectors were created which can be found in Table 4.1. The $\alpha 2$, $\alpha 3$, $\beta 2$, and $\beta 3$ vectors contain large changes to the sojourn distributions that the $GoF(t)$ metric

should readily identify because the altered sojourn times will result in much higher or much lower state visit counts than in the baseline model. The remaining vectors have subtle sojourn distribution changes that are more difficult for the metric to identify because the state visit counts will not be affected as much. Figure 4.5 illustrates how the sojourn distributions change with the various parameterizations.

Table 4.1: Sojourn Distribution Parameters

α Vectors		β Vectors	
$\alpha 1$ (Baseline)	(2,2,0.5,0.5)	$\beta 1$ (Baseline)	(2,1,4,1)
$\alpha 2$	(4,4,1,1)	$\beta 2$	(4,2,8,2)
$\alpha 3$	(1,1,0.25,0.25)	$\beta 3$	(1,0.5,2,0.5)
$\alpha 4$	(4,2,0.5,0.5)	$\beta 4$	(4,1,4,1)
$\alpha 5$	(2,8,0.5,0.5)	$\beta 5$	(2,4,4,1)
$\alpha 6$	(2,2,5,0.5)	$\beta 6$	(2,1,40,1)
$\alpha 7$	(2,2,0.5,50)	$\beta 7$	(2,1,4,100)

Each combination of α- and β-vectors is examined by creating a model with those specific parameters. Each test run simulates the model to t =1,000 time units. The test run covariance matrix is created using the Markov renewal process counts of the test model, the expected Markov renewal counts of the baseline model, and Equation 4.12 for each transition in the test run. An expected covariance matrix is calculated using 500,000 iterations of the baseline model and the Covariance Simulation Algorithm described in Section 3.1. The $GoF(t)$ metric for the sample run is computed as in Equation 4.10. This process is repeated for 5,000 iterations to build a distribution for the $GoF(t)$ metric of the test model versus the baseline model.

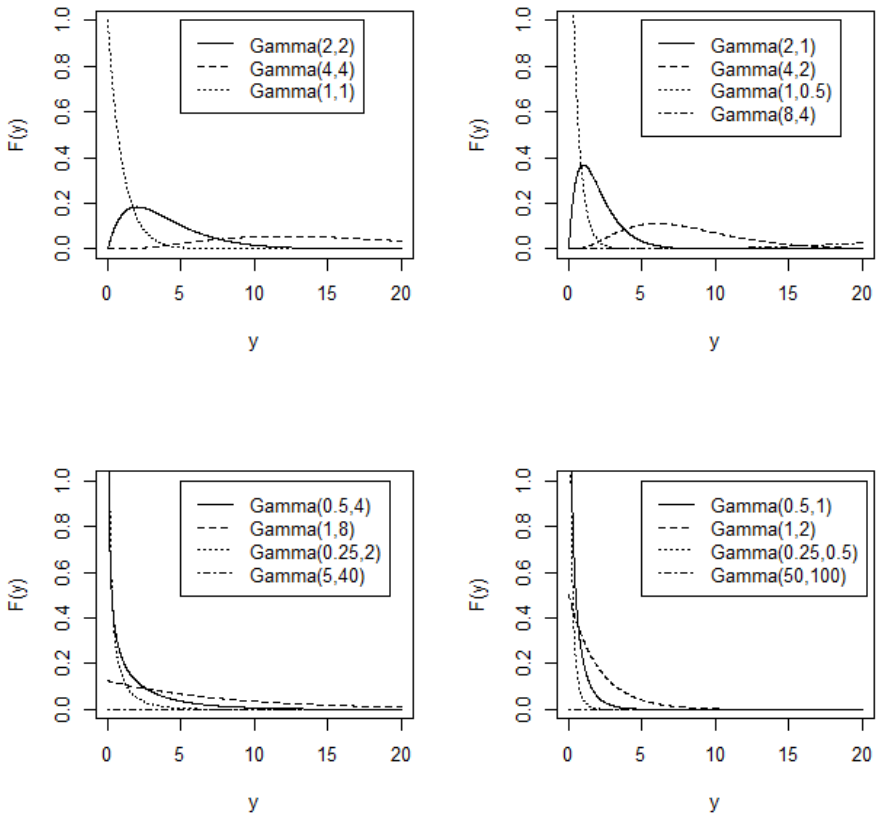

Figure 4.5: Sojourn probability density function relationships

To establish rejection criteria, 100,000 iterations of the baseline model versus itself were conducted with the upper 5% of the test statistic values considered as the rejection region for the test. This rejection region establishes a type I error rate (incorrectly rejecting a true model) of 0.05 for the test. Table 4.2 illustrates the type II error rates (incorrectly failing to reject a false model) for the test model simulations. In this case, the $\alpha 1$-$\beta 1$ entry is the baseline model which will not have a type II error, but does have a type I error set at 0.05. Of the other 48 test scenarios, 42 scenarios have type II error rates less than 0.05 and only 4 scenarios have type II error rates greater than 0.25. These 4 scenarios are the $\alpha 2 - \beta 3$ model, the $\alpha 3 - \beta 2$ model, the $\alpha 7 - \beta 3$ model, and the $\alpha 7 - \beta 7$ model.

Table 4.2: Model Type II Error Rates

	α Vectors						
β vectors	$\alpha 1$	$\alpha 2$	$\alpha 3$	$\alpha 4$	$\alpha 5$	$\alpha 6$	$\alpha 7$
$\beta 1$	0	0	0	0	0	0	0
$\beta 2$	0	0	0.879	0	0	0	0
$\beta 3$	0	0.973	0	0.002	0.0006	0.0004	0.666
$\beta 4$	0.0002	0	0.041	0	0	0	0
$\beta 5$	0	0	0.023	0	0	0	0
$\beta 6$	0	0	0.068	0	0	0	0
$\beta 7$	0.0026	0	0.252	0	0	0	0.062

Analyzing the effects of the α- and β-vectors explains the higher type II error rates in the 4 non-baseline models. For the $\alpha 3 - \beta 2$ model, the α-vector is exactly half the baseline α-vector while the β-vector is exactly double the baseline β-vector. For these Γ-distribution parameterizations, the means of the distributions equal $\alpha\beta$ and the variances equal $\alpha\beta^2$. These parameterizations imply that the $\alpha 1 - \beta 1$ and $\alpha 3 - \beta 2$ models possess identical sojourn distribution means but the $\alpha 3 - \beta 2$ model has twice the sojourn distribution variances. The additional variance results in the lower type II error rate of 0.879 when compared with the $\alpha 2 - \beta 3$ model's type II error rate of 0.973. The $\alpha 2 - \beta 3$ model has half the sojourn distribution variances compared with the baseline model while maintaining the same means. In this case, the test model performs "better" than the true model in that fewer model runs will be rejected even though the model is wrong.

The $\alpha 7 - \beta 3$ and $\alpha 3 - \beta 7$ models have type II error rates greater than 0.1 due to the balance between the new parameterizations. The difference between these two models and the baseline model is the $\Gamma(0.5,1)$ distribution which becomes a $\Gamma(50,0.5)$ distribution with all other β values cut in half in the $\alpha 7 - \beta 3$ model and a $\Gamma(0.25,100)$ with all other α values cut in half in the $\alpha 3 - \beta 7$ model. This results in a larger mean and variance for

one distribution (the Γ(50,0.5) in the $\alpha 7 - \beta 3$ scenario and the Γ(0.25,100) in the $\alpha 3 - \beta 7$ scenario) while the other three distributions from **F** in each respective scenario are reduced. The typical run with either of these models has relatively few transitions due to the large mean of the associated Γ-distribution. However, the reduced means and variances in the other distributions counterbalance the more misspecified distribution.

Figure 4.6 shows the rejection probabilities for each test model versus the baseline model over the time length of the test run. Most of the scenarios stabilize with rejection rates above 0.9 by 200 to 500 time units. This equals a type II error of less than 0.1. The baseline model plus the five test models discussed above correspond to the five lines that remain below 0.9 all the way out to 1,000 time units. Naturally, a wide range of models would perform similarly to the baseline model as illustrated by these test models.

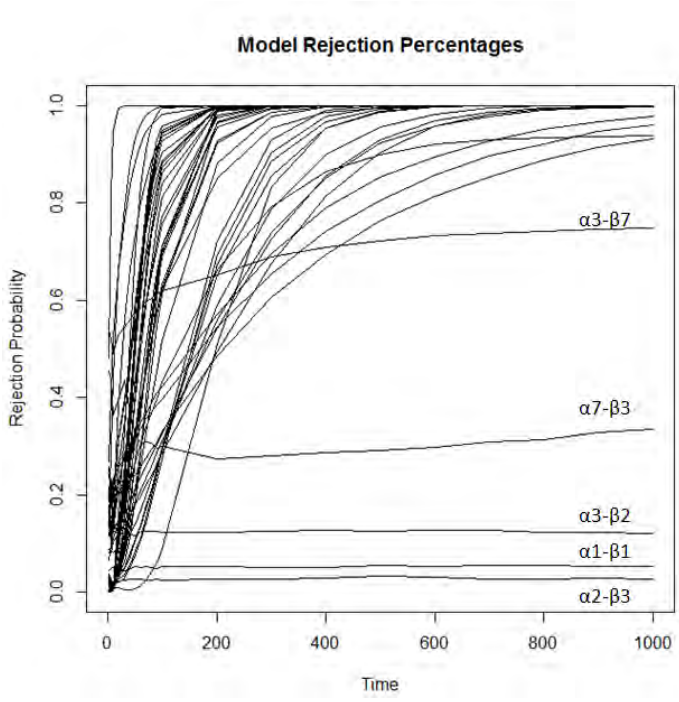

Figure 4.6: Rejection percentages for various sojourn distributions over time

In addition to the various sojourn distribution parameterizations, testing included a variety of probability transition matrices. Appendix B contains the rejection percentage matrices for seven additional **P** matrices. The end result is that this test is sensitive against even slight changes to the probability transition matrix. Even the $\alpha_1 - \beta_1$ parameterizations have a much higher rejection rate when the various **P** matrices are used. In these cases, the fact that the sojourn time distributions are identical is offset by the differences in the transition probabilities. These differences result in highly altered sample paths that are unlikely to be produced by the baseline model. In most of the cases, the $\alpha_7 - \beta_7$ parameterization has a slightly lower rejection rate in the 80-90% range because the misspecified $\Gamma(50, 100)$ sojourn distribution reduces the Markov renewal process counts to the point that the altered **P** is masked.

The reason that perturbations of **P** have such a profound effect, in most cases, on the ability to distinguish between models is due to the portions of the covariance matrix that are influenced by **P**. **F** effects the time between transitions. Suppose the distributions in **F** were shifted along the x-axis so that their means were doubled. The impact of this change would be to halve the Markov renewal process counts and expectations for a given time. Ultimately, the model would behave similarly to the original model, it would just take longer to transition. **P** effects the actual behavior of the model. For the baseline model, doubling the probability of transitioning from State 1 to State 4 (from 0.25 to 0.5) would also reduce the probability of transitioning from State 1 to State 2 (from 0.75 to 0.5). Not only does this change alter the Markov renewal process counts and expectations, but it also impacts the correlation matrix for the model by reducing the correlation among States 1 and 2 and increasing the correlation between States 1 and 4. For a given t, the new correlation matrix results in a new covariance matrix as well since left hand side of Equation 3.13 has changed.

The $GoF(t)$ metric developed in this chapter can detect differences between the observed data and expected data for an underlying, hypothesized semi-Markov process. The $GoF(t)$ metric detected model differences in 92% of the models tested in this chapter. A natural extension of this goodness of fit testing is to determine whether or not a potential covariate has an impact on the hypothesized model. The next chapter examines covariate detection which impacts data collection requirements and overall model definition.

V. Testing a Semi-Markov Process for the Presence of Covariates

The $GoF(t)$ metric developed in Chapter 4 can be used to determine whether or not a significant covariate is present within a model. The process involves calculating two $GoF(t)$ metric values for a single observed data sample. The first $GoF(t)$ value comes from comparing the data to a model that contains the covariate and determining where this $GoF(t)$ value falls in the distribution of $GoF(t)$ values for this covariate model. The second $GoF(t)$ value is created by comparing the data to a model that does not contain the covariate and determining where this second $GoF(t)$ value falls in the $GoF(t)$ distribution of the non-covariate model. The major difference between the models will be the state space representations which will alter the expectations and covariance matrices used to create the $GoF(t)$ metric. The changes in the state space representation require the covariates to be categorical rather than continuous, as a continuous covariate would result in an uncountably infinite state space.

Before demonstrating the test for covariates, consider the baseline model to add context. Suppose the four state model in Figure 4.1 tracks an individual as he or she progresses through the various stages of a disease where State 1 = Healthy, State 2 = Diseased, State 3 = Recovering, State 4 = Disease No Longer Present. In this scenario, the Healthy State only occurs prior to an individual being diagnosed with the disease for the first time. Any transitions back to the Healthy State from the other three states represent the death of the current subject and the admission of a new subject into the system.

Within this scenario, the subject's sex could be a covariate in the model as disease progression may substantially vary between men and women. The eight state model in Figure 5.1 represents a state space expansion to account for a subject's sex. In the covariate model, let State 1 = Healthy Female, State 2 = Diseased Female, State 3 = Recovering Female, State 4 = Disease No Longer Present Female, State 5 = Healthy Male, State 6 =

Diseased Male, State 7 = Recovering Male, State 8 = Disease No Longer Present Male. The dotted arrows represent transitions when a subject dies and the next subject is of the opposite sex. Effectively, there are two separate four-state models that are connected at specific transition points.

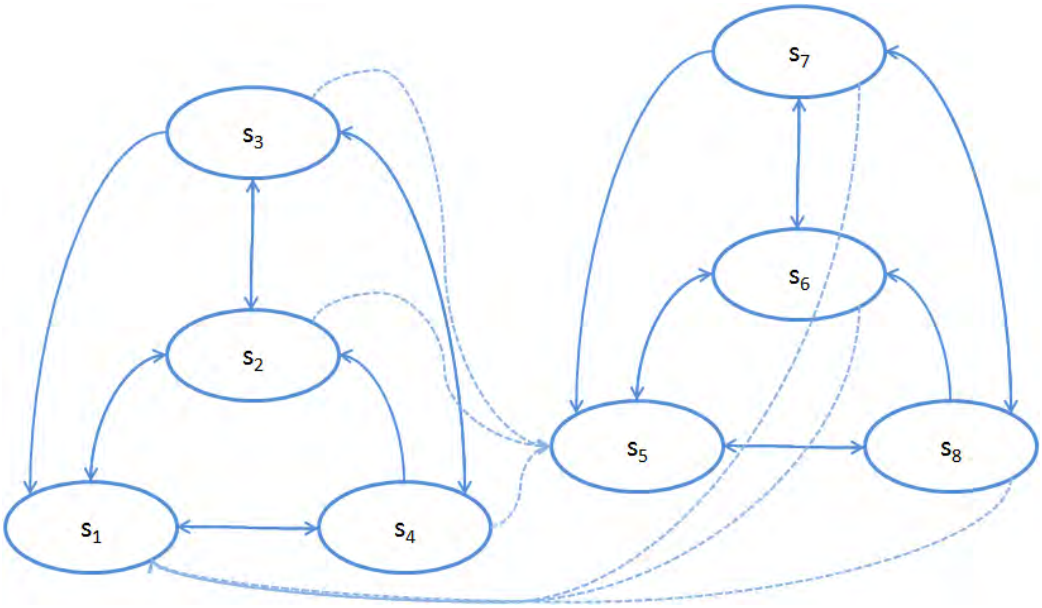

Figure 5.1: The covariate baseline semi-Markov process connected graph

Ultimately, the state space changes required for the covariate model result in changes to the transition probability matrix, the sojourn distribution matrix, and the kernel matrix. Let π represent the probability that a given subject is covariate level 1 and 1-π represent the probability that a subject is covariate level 2. The starting covariate model transition

probability matrix for each state, s_i, in the covariate baseline, CB, model is

$$\mathbf{P}_{\text{CB}} = \begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \\ s_7 \\ s_8 \end{array} \begin{array}{c} \begin{array}{cccccccc} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 & s_8 \end{array} \\ \begin{bmatrix} 0 & 0.75 & 0 & 0.25 & 0 & 0 & 0 & 0 \\ 0.5\pi & 0 & 0.5 & 0 & 0.5(1-\pi) & 0 & 0 & 0 \\ 0.25\pi & 0.45 & 0 & 0.3 & 0.25(1-\pi) & 0 & 0 & 0 \\ 0.15\pi & 0.35 & 0.5 & 0 & 0.15(1-\pi) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.75 & 0 & 0.25 \\ 0.5\pi & 0 & 0 & 0 & 0.5(1-\pi) & 0 & 0.5 & 0 \\ 0.25\pi & 0 & 0 & 0 & 0.25(1-\pi) & 0.45 & 0 & 0.3 \\ 0.15\pi & 0 & 0 & 0 & 0.15(1-\pi) & 0.35 & 0.5 & 0 \end{bmatrix} \end{array} \tag{5.1}$$

while the starting covariate model sojourn distribution matrix from the example for the CB model is

$$\mathbf{F}_{\text{CB}} = \begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \\ s_7 \\ s_8 \end{array} \begin{array}{c} \begin{array}{cccccccc} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 & s_8 \end{array} \\ \begin{bmatrix} 0 & F_1 & 0 & F_1 & 0 & 0 & 0 & 0 \\ F_2 & 0 & F_3 & 0 & F_2 & 0 & 0 & 0 \\ F_4 & F_3 & 0 & F_1 & F_4 & 0 & 0 & 0 \\ F_3 & F_1 & F_2 & 0 & F_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & F_1 & 0 & F_1 \\ F_2 & 0 & 0 & 0 & F_2 & 0 & F_3 & 0 \\ F_4 & 0 & 0 & 0 & F_4 & F_3 & 0 & F_1 \\ F_3 & 0 & 0 & 0 & F_3 & F_1 & F_2 & 0 \end{bmatrix} \end{array}, \tag{5.2}$$

where F_1 is a $\Gamma(2,2)$, F_2 is a $\Gamma(2,1)$, F_3 is a $\Gamma(0.5,4)$, and F_4 is a $\Gamma(0.5,1)$. The CB subscript for matrices is used to denote the covariate baseline model matrices as opposed to the original matrices used in Chapter 4 and the s_i notation tracks the states of the covariate model. The covariate baseline kernel matrix becomes

$$\mathbf{G}_{\text{CB}} = \begin{bmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{bmatrix} \tag{5.3}$$

where

$$\mathbf{A} = \begin{bmatrix} 0 & 0.1875xe^{-\frac{x}{2}} & 0 & 0.0625xe^{-\frac{x}{2}} \\ 0.5\pi xe^{-x} & 0 & 1.7725x^{-0.5}e^{-\frac{x}{4}} & 0 \\ 0.8862\pi x^{-0.5}e^{-x} & 1.5952x^{-0.5}e^{-\frac{x}{4}} & 0 & 0.075xe^{-\frac{x}{2}} \\ 0.5317\pi x^{-0.5}e^{-\frac{x}{4}} & 0.0875xe^{-\frac{x}{2}} & 0.5xe^{-x} & 0 \end{bmatrix},$$

$$\mathbf{B} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0.5(1-\pi)xe^{-x} & 0 & 0 & 0 \\ 0.8862(1-\pi)x^{-0.5}e^{-x} & 0 & 0 & 0 \\ 0.5317(1-\pi)x^{-0.5}e^{-\frac{x}{4}} & 0 & 0 & 0 \end{bmatrix},$$

$$\mathbf{C} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0.5\pi xe^{-x} & 0 & 0 & 0 \\ 0.8862\pi x^{-0.5}e^{-x} & 0 & 0 & 0 \\ 0.5317\pi x^{-0.5}e^{-\frac{x}{4}} & 0 & 0 & 0 \end{bmatrix},$$

$$\mathbf{D} = \begin{bmatrix} 0 & 0.1875xe^{-\frac{x}{2}} & 0 & 0.0625xe^{-\frac{x}{2}} \\ 0.5(1-\pi)xe^{-x} & 0 & 1.7725x^{-0.5}e^{-\frac{x}{4}} & 0 \\ 0.8862(1-\pi)x^{-0.5}e^{-x} & 1.5952x^{-0.5}e^{-\frac{x}{4}} & 0 & 0.075xe^{-\frac{x}{2}} \\ 0.5317(1-\pi)x^{-0.5}e^{-\frac{x}{4}} & 0.0875xe^{-\frac{x}{2}} & 0.5xe^{-x} & 0 \end{bmatrix}.$$

A single observed data set can be tested against either a four-state model or an eight-state model depending on whether or not the covariate is taken into account. Consider a test run where four subjects are female, male, male, and female respectively. A potential sample path given these four subjects appears in Table 5.1. If the covariate was insignificant, as is the case with the covariate baseline model depicted above, data collected from the eight-state model will pass the goodness of fit test for the four-state model once the data is collapsed to four states. This will be demonstrated in Section 5.1.

Table 5.1: Example Markov Renewal Process Sample Paths

Subject	Four-state Markov Renewal Process Sample Path Without Covariate	Eight-State Markov Renewal Process Sample Path With Covariate
1	1 2 3	1 2 3
2	1 4 2	5 8 6
3	1 2 3	5 6 7
4	1 4 3	1 4 3

A significant covariate would necessitate a change to at least one of the values or distributions in the $\mathbf{P}_{CB}$ or $\mathbf{F}_{CB}$ matrices which also alters the $\mathbf{G}_{CB}$ matrix. For testing purposes, assume that the four-state baseline model does not change. This assumption forces the covariate model changes to be balanced within the model. For instance, starting with $\mathbf{P}_{CB}$, if the probability of going from State 5 to State 8 is changed to 0.5, then at least one of the other transition probabilities in the State 5 row must change since rows must sum to 1. Assume that the State 5 to State 6 transition probability also becomes 0.5 to satisfy this constraint. In addition, the State 1 to State 2 and the State 1 to State 4 transitions would change based on π so that their average transition probabilities when the covariate is not

considered are 0.75 and 0.25 respectively. The new State 1 to State 2 and State 1 to State 4 transition probabilities for the eight-state model can be calculated by solving the following series of equations:

$$0.75 = p_{12}\pi + p_{56}(1-\pi) = p_{12} * \pi + 0.5 * (1-\pi)$$

$$0.25 = p_{14}\pi + p_{58}(1-\pi) = p_{14} * \pi + 0.5 * (1-\pi)$$

$$1 = p_{12} + p_{14}$$

$$0 \leq p_{12}, p_{14}, p_{56}, p_{58} \leq 1$$

Generally, a covariate probability value π will either be known or assumed, as such, there remain only three equations and two unknowns. This over-specified system of equations limits the amount by which any of the individual probabilities in the $\mathbf{P}_{\text{CB}}$ matrix may change. In this case, assuming $\pi = 0.5$, then $p_{12} = 1$ and $p_{14} = 0$. If one tries to make the State 5 to State 8 transition probability 0.6 then

$$0.75 = p_{12} * 0.5 + 0.4 * 0.5$$

$$0.25 = p_{14} * 0.5 + 0.6 * 0.5$$

$$1 = p_{12} + p_{14}$$

implies $p_{12} = 1.1$ and $p_{14} = -0.1$. The probabilities greater than 1.0 and less than 0.0 violate the laws of probability which means this particular change to the State 5-State 8 transition cannot occur in this model.

Similarly, a change to one of the sojourn distributions forces at least one other sojourn distribution to change if the covariate model is going to maintain balance with the smaller baseline model. In this case, a balance exists between States 1 and 5 in the covariate model versus State 1 in the baseline model. Additionally, States 2 and 6, States 3 and 7, and States 4 and 8 in the covariate model are balanced with States 2, 3, and 4, respectively, in the baseline model. This balance is based on the assumption that the original four-state model provides an adequate fit of the data and that the larger covariate model is now

under test. If the covariate model distributions were not balanced with their baseline model counterparts, then the covariate model would reduce to a four state model that is not the baseline model. This concept of balance is a function of artificial semi-Markov processes used in this testing. In normal implementation, the data would dictate the model parameters for both the covariate model and the reduced model.

To demonstrate the sojourn distribution balance, if the sojourn time for transitioning from State 5 to State 8 followed a $\Gamma(3,2)$ distribution, then the transition distribution from State 1 to State 4 in the covariate model would also have to change to ensure the average sojourn distribution between the two transitions remains a $\Gamma(2,2)$ distribution which is the sojourn distribution between States 1 and 4 in the baseline model. While there are a variety of ways to balance the distributions, for simplicity, the distributional mean is being kept consistent. In the case of a $\Gamma(2,2)$ distribution changing to a $\Gamma(3,2)$ distribution, the State 1 to State 4 sojourn distribution would become a $\Gamma(1,2)$ distribution assuming $\pi = 0.5$. Using this pair of distributions, the average of all State 1 to State 4 and all State 5 to State 8 transitions is 4 time units, matching the original mean of the $\Gamma(2,2)$ distribution. This method does not maintain the distributional variance as the combination of the $\Gamma(1,2)$ distribution and the $\Gamma(3,2)$ distribution has a variance of 12 while the original variance of the $\Gamma(2,2)$ distribution is 8. In reality, the sample data itself would be used to determine the **P** and **F** matrices for the model with covariates and the model without covariates.

5.1 Covariate Testing

Once the two models (baseline and covariate) are selected, testing proceeds by calculating the $GoF(t)$ metric of the sample data for both models using the methods outlined in Chapter 4. The $GoF(t)$ values indicate whether or not either of the models can be rejected as poor fits for the data. In the case where both models fail to reject, a preference must be made with respect to either the larger covariate model or the model without covariates. This decision is ultimately up to the modeler who must balance data

collection requirements and computational complexity with the added model fidelity that the covariates provide.

In this testing, the baseline four-state model from Chapter 4 is the model without covariates and a suite of models with covariates was created to illustrate the ability of the $GoF(t)$ metric test to differentiate between the models. The first round of testing demonstrates the equivalence between the four-state baseline model from Chapter 4 and eight-state covariate baseline model defined by $\mathbf{P}_{\text{CB}}$ and $\mathbf{F}_{\text{CB}}$. The testing involves drawing N=5,000 iterations from the covariate baseline model using $\pi = 0.5$. The eight-state sample paths are reduced down to four states by making all State 5 instances in the sample path State 1's, all State 6's instances State 2's, all State 7's instances State 3's, and all State 8's instances State 4's. This transition effectively ignores the potential presence of the covariate. Ignoring the covariate, 4.58% of the runs are rejected based on their $GoF(t)$ metric values which is in line with the expected 5% that should be incorrectly rejected. For all of the covariate testing, a type I error rate of 5% is used.

Ideally, any covariate model that is properly balanced with respect to a four-state model will have a rejection rate around 5% when it is reduced to the four-state model without covariates. The difference in models can be detected by comparing the larger covariate models to one another. The first step is a demonstration that the baseline model and covariate baseline model are equivalent to one another in the larger eight-state domain. This round of testing draws N=5,000 iterations from the four-state baseline model with one slight difference. Each time the simulation returns to State 1, signifying a new subject enters the system, the simulation determines whether or not the new subject is male or female according to covariate level probability, π. After the simulation concludes, the Markov renewal process is expanded to eight states by making States 1, 2, 3, and 4 into States 5, 6, 7, and 8 if a particular subject is male. Note that for an assumed $\pi = 0.5$, $p_{ij} = p_{(i+4)(j+4)}$ and $F_{ij} = F_{(i+4)(j+4)}$ for $1 \leq i, j \leq 4$. $GoF(t)$ metric values are then calculated for the runs by

comparing them directly with the eight-state covariate baseline model. In this case, 5.04% of the samples are rejected which is in line with the expected 5% that should be incorrectly rejected.

From this point forward, any potential covariate model is tested directly against the covariate baseline model. The underlying analysis is that if the data is not significantly different from the covariate baseline model, the proposed data model can be reduced down to the four-state baseline model without losing any fidelity at the modeler's discretion. Practically, this means it is not necessary to track the presence of the covariate because it does not actually add any additional information about the process.

Testing focused on two separate probability transition matrices, $\mathbf{P}_{C1}$ and $\mathbf{P}_{C2}$, and four different sets of Gamma-distribution parameters, $\alpha_2 - \beta_2,\ \alpha_3 - \beta_3,\ \alpha_4 - \beta_4$, and $\alpha_5 - \beta_5$. The first probability transition matrix,

$$\mathbf{P}_{\mathrm{C1}} = \begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \\ s_7 \\ s_8 \end{array} \begin{array}{c} \begin{array}{cccccccc} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 & s_8 \end{array} \\ \left[\begin{array}{cccccccc} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.5\pi & 0 & 0.5 & 0 & 0.5(1-\pi) & 0 & 0 & 0 \\ 0.25\pi & 0.45 & 0 & 0.3 & 0.25(1-\pi) & 0 & 0 & 0 \\ 0.15\pi & 0.35 & 0.5 & 0 & 0.15(1-\pi) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.5 & 0 & 0.5 \\ 0.5\pi & 0 & 0 & 0 & 0.5(1-\pi) & 0 & 0.5 & 0 \\ 0.25\pi & 0 & 0 & 0 & 0.25(1-\pi) & 0.45 & 0 & 03 \\ 0.15\pi & 0 & 0 & 0 & 0.15(1-\pi) & 0.35 & 05 & 0 \end{array}\right] \end{array}, \tag{5.4}$$

represents a drastic change in comparison with the covariate baseline model for the probability transitions of States 1 and 5. The second probability transition matrix,

$$\mathbf{P}_{C2} = \begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \\ s_5 \\ s_6 \\ s_7 \\ s_8 \end{array} \begin{array}{c} \begin{array}{cccccccc} s_1 & s_2 & s_3 & s_4 & s_5 & s_6 & s_7 & s_8 \end{array} \\ \left[\begin{array}{cccccccc} 0 & 0.875 & 0 & 0.125 & 0 & 0 & 0 & 0 \\ 0.5\pi & 0 & 0.5 & 0 & 0.5(1-\pi) & 0 & 0 & 0 \\ 0.25\pi & 0.45 & 0 & 0.3 & 0.25(1-\pi) & 0 & 0 & 0 \\ 0.15\pi & 0.35 & 0.5 & 0 & 0.15(1-\pi) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.625 & 0 & 0.375 \\ 0.5\pi & 0 & 0 & 0 & 0.5(1-\pi) & 0 & 0.5 & 0 \\ 0.25\pi & 0 & 0 & 0 & 0.25(1-\pi) & 0.45 & 0 & 03 \\ 0.15\pi & 0 & 0 & 0 & 0.15(1-\pi) & 0.35 & 05 & 0 \end{array}\right] \end{array}, \tag{5.5}$$

represents a smaller change for the probability transitions of States 1 and 5. For the Γ-distribution parameters, let the upper and lower halves of the $\mathbf{F}_{CB}$ matrix each have four separate distributions (Γ(2,2), Γ(2,1), Γ(0.5,4), Γ(0.5,1)). Following the notation used in Section 4.4.1, define a vector of α and β values for all eight states as (2,2,0.5,0.5,2,2,0.5,0.5) and (2,1,4,1,2,1,4,1) respectively. Table 5.2 shows the original baseline α and β vectors and the four additional sets of vectors tested. Of the four test vectors, the first three only change a single pair of α or β values, while the fourth set represents a more drastic change across all distributions. The distributions were selected to balance the baseline distributions with respect to the combined average of each pair of sojourn distributions when compared with the same pair in the covariate baseline model. Basically, the average among all State 1 to State 2 transitions and all State 5 to State 6 transitions is the same whether the covariate baseline vectors or any of the four additional vector pairs are used. The need to maintain this balance eliminates the need to run the combinatorial analysis with every $\mathbf{P} - \alpha - \beta$ permutation as seen in Chapter 4.

Table 5.2: Covariate Sojourn Distribution Parameters

α Vectors		β Vectors	
$\alpha 1$ (Baseline)	(2,2,0.5,0.5,2,2,0.5,0.5)	$\beta 1$ (Baseline)	(2,1,4,1,2,1,4,1)
$\alpha 2$	(3,2,0.5,0.5,1,2,0.5,0.5)	$\beta 2$	(2,1,4,1,2,1,4,1)
$\alpha 3$	(2,2,0.5,0.5,2,2,0.5,0.5)	$\beta 3$	(2,1,1,1,2,1,7,1)
$\alpha 4$	(60,2,0.5,0.5,0.1,2,0.5,0.5)	$\beta 4$	(0.1,1,4,1,20,1,4,1)
$\alpha 5$	(3,1,1,0.25,3,3.5,0.5,0.75)	$\beta 5$	(1,0.5,3,0.25,1.67,1,2,1.25)

Comparing the $\mathbf{P}_{C1}\alpha_1\beta_1$ model to the baseline covariate model ($\mathbf{P}_{CB}\alpha_1\beta_1$), the type II error rate is 0.499 while the $\mathbf{P}_{C2}\alpha_1\beta_1$ model has a type II error rate of 0.891 when compared with the baseline covariate model. The disparity in type II error rates is expected because the $\mathbf{P}_{C1}\alpha_1\beta_1$ model represents a larger change to $\mathbf{P}$ than the $\mathbf{P}_{C2}\alpha_1\beta_1$ model. For the $\mathbf{P}_{CB}\alpha_2\beta_2$, $\mathbf{P}_{CB}\alpha_3\beta_3$, $\mathbf{P}_{CB}\alpha_4\beta_4$, and $\mathbf{P}_{CB}\alpha_5\beta_5$ models, the type II error rates when compared with the baseline covariate model are 0.934, 0.933, 0.392, and 0.948, respectively. Figure 5.2 shows the various distributions used in the covariate baseline model and the $\mathbf{P}_{CB}\alpha_2\beta_2$ and $\mathbf{P}_{CB}\alpha_3\beta_3$ models. In both graphs, the solid line represents the distribution used in the covariate baseline model while the two dashed lines represent the balanced distributions from the test models. The large degree of overlap among the various distributions results makes it impossible for the $GoF(t)$ metric to discern between these models. On the other hand, the distributions used in the $\mathbf{P}_{CB}\alpha_4\beta_4$ model differ significantly from the covariate baseline model as illustrated in Figure 5.3. These differences are picked up in over 60% of the models tested.

While it may appear that covariate testing lacks power to differentiate between models based on some of the scenarios, the reality is that it can detect significant differences such those shown in Figure 5.3. The scenarios used in this chapter were specifically crafted to ensure the average across the covariate would collapse down to the baseline model. The

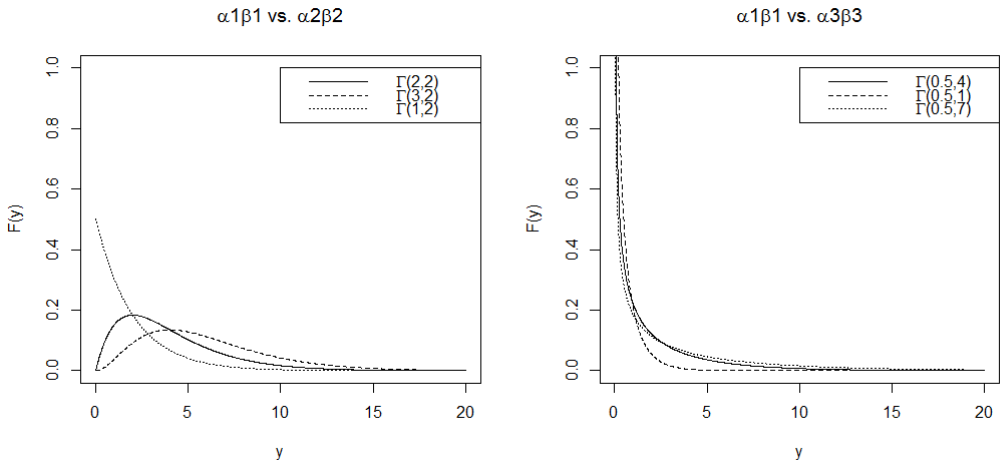

Figure 5.2: Distributions from the baseline covariate, $\alpha_2\beta_2$, and $\alpha_3\beta_3$ models

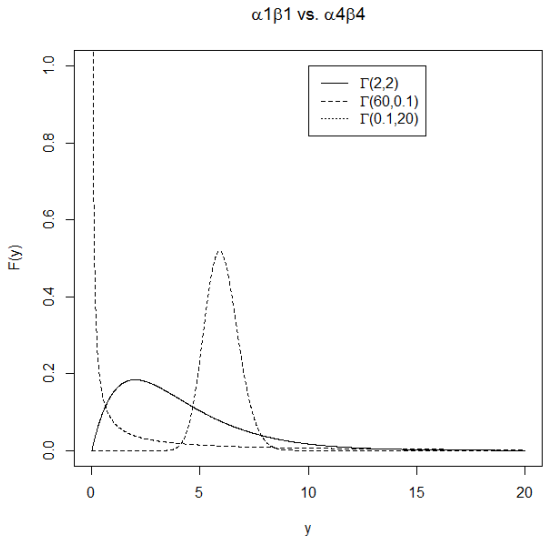

Figure 5.3: Distributions from the baseline covariate and $\alpha_4\beta_4$ models

end result in several scenarios was a small enough difference that the covariate proved insignificant. For an actual data sample, the baseline and covariate models would be dictated by the sample and the underlying theory which results in more realistic differences

between models. To this end, the next chapter examines a real world application of the $GoF(t)$ metric with an emphasis on covariate detection.

VI. BMI Model Analysis

According to the Centers of Disease Control and Prevention (CDC) website [24], childhood obesity affects 17% of US children between the ages of 2 and 19, with 20.5% of US children between the ages of the 12 and 19 considered obese as of 2011 based on Body Mass Index (BMI). Effectively, one out of every five American teenagers is above the 95th percentile of their BMI-for-age growth chart. A recent study using regression techniques [25] concluded that obesity trends began rising in the 1980s while a study using prevalence rates [26] identified an increasing trend over a span of merely four years. Previous studies have drawn correlations between childhood obesity and various childhood and adult afflictions including heart disease [27] and diabetes [28]. Due to these life-threatening potential consequences, many social initiatives have started in recent years targeting poor dieting and lack of exercise, two of the primary causal factors in childhood obesity according to the CDC.

Whether it is the National Football League's Play 60 exercise campaign or the Healthy, Hunger-Free Kids Act (the school lunch reform initiative championed by First Lady Michelle Obama), analysts and policy makers require robust tools to determine if the various initiatives are impacting the childhood obesity rates. Quickly identifying improvements in obesity rates can lead to increased focus and funding for current initiatives, while identifying a lack of improvement can result in renewed efforts to create new initiatives to combat obesity. However, the community requires a tool that can differentiate between improving obesity rates, stagnating obesity rates, escalating obesity rates and spurious fluctuations in the data or noise that can lead to false conclusions. To this end, the $GoF(t)$ metric outlined in Equation 4.3 can help determine the tempo of current childhood BMI trends. While the aforementioned studies focus on whether or not the obesity trends appear to be changing over time, the $GoF(t)$ metric can help determine

whether the changes are possibly random noise, merely a consistency in association, versus a true population shift.

Data from the Fels Longitudinal Study, which contains the age and BMI values for children from the 1930s to the present, was examined to explore the population trends and transitions to obesity in children over time. This data consisted of repeated measurements of 1146 children (583 boys and 563 girls) at 6 month intervals from age 2 through age 19 for a total of 10,771 boy observations and 10,434 girl observations. On average, the boys spent 11.27 years in the study while the girls spent 11.38 years in the study. Details regarding the history and methods of the Fels Longitudinal Study can be found in Roche [29]. Based on the BMI-for-age growth charts and the traditional BMI-based categories of Underweight (less than the 5th percentile), Normal (5th to 85th percentile), Overweight (85th to 95th percentile), and Obese (greater than the 95th percentile), each child is tracked as he or she transitions from one category to the next. Figure 6.1 shows a simple state model of the four BMI-based categories. The solid arrows represent moving among adjoining states while the dashed arrows indicate a jump of two or more states.

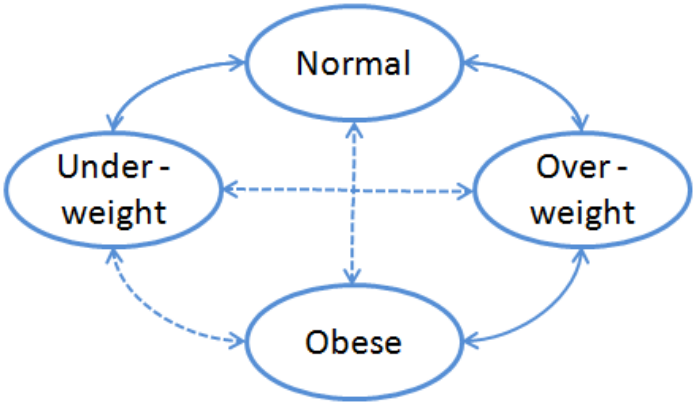

Figure 6.1: Four state weight classification model

While skipping a state generally does not occur in real life without extenuating factors like weight loss surgery or growth hormones, the modeling procedure and the six month

(discrete time) intervals in data collection allow this phenomenon to occur in this model. For this model, a child enters the system at the category of their first data measurement. The model changes states when the child's category changes, as calculated from subsequent measurements. After the child's final visit, a new child enters the system at his or her current category state and the system proceeds. If the new child's starting state is the same as the previous child's ending state, the system will continue accruing time in that particular state without transitioning. In the data set, this occurs 227 times among boys and 197 times among girls. On the other hand, if the new child starts as Underweight and the previous child ended up as Obese, the system would undergo a transition from Obese to Underweight, skipping Overweight and Normal weight in the process. If a child only had one measurement, that child was removed from further analysis.

Based on the observation time frames of the children, two separate models were created from the dataset. The 1930-1960 Model includes all of the measurements prior to 1960 and the 1980-2007 Model contains every measurement after 1979. In the 1930-1960 Model,

$$\mathbf{P}_{1930-1960Model} = \begin{array}{c} \\ \text{UW} \\ \text{N} \\ \text{OW} \\ \text{O} \end{array} \begin{array}{c} \begin{array}{cccc} \text{UW} & \text{N} & \text{OW} & \text{O} \end{array} \\ \begin{bmatrix} 0 & 0.992 & 0.008 & 0 \\ 0.338 & 0 & 0.622 & 0.040 \\ 0.007 & 0.850 & 0 & 0.143 \\ 0.019 & 0.264 & 0.717 & 0 \end{bmatrix} \end{array} \tag{6.1}$$

and

$$\mathbf{F}_{1930-1960Model} =$$

$$\begin{array}{c} \\ \text{UW} \\ \text{N} \\ \text{OW} \\ \text{O} \end{array} \begin{array}{c} \begin{array}{cccc} \text{UW} & \text{N} & \text{OW} & \text{O} \end{array} \\ \begin{bmatrix} 0 & \Gamma(0.36, 4.82) & \Gamma(0.20, 4.00) & 0 \\ \Gamma(0.41, 15.77) & 0 & \Gamma(0.43, 19.38) & \Gamma(0.69, 7.66) \\ \Gamma(0.20, 4.00) & \Gamma(0.60, 1.91) & 0 & \Gamma(1.75, 1.02) \\ \Gamma(0.20, 4.00) & \Gamma(0.40, 4.98) & \Gamma(0.40, 2.92) & 0 \end{bmatrix} \end{array}, \tag{6.2}$$

where UW is Underweight, N is Normal, OW is Overweight, and O is Obese. For the 1980-2007 Model,

$$\mathbf{P}_{1980-2007Model} = \begin{array}{c} \\ \text{UW} \\ \text{N} \\ \text{OW} \\ \text{O} \end{array} \begin{array}{c} \begin{array}{cccc} \text{UW} & \text{N} & \text{OW} & \text{O} \end{array} \\ \begin{bmatrix} 0 & 0.969 & 0.031 & 0 \\ 0.364 & 0 & 0.547 & 0.089 \\ 0.016 & 0.607 & 0 & 0.377 \\ 0.032 & 0.411 & 0.557 & 0 \end{bmatrix} \end{array} \tag{6.3}$$

and

$$\mathbf{F}_{1980-2007Model} =$$

$$\begin{array}{c} \\ \text{UW} \\ \text{N} \\ \text{OW} \\ \text{O} \end{array} \begin{array}{c} \begin{array}{cccc} \text{UW} & \text{N} & \text{OW} & \text{O} \end{array} \\ \begin{bmatrix} 0 & \Gamma(0.76, 2.80) & \Gamma(0.25, 5.00) & 0 \\ \Gamma(0.45, 19.90) & 0 & \Gamma(0.66, 17.82) & \Gamma(1.51, 8.34) \\ \Gamma(0.20, 4.00) & \Gamma(1.08, 1.62) & 0 & \Gamma(1.42, 1.44) \\ \Gamma(0.64, 8.58) & \Gamma(1.09, 3.40) & \Gamma(0.87, 2.34) & 0 \end{bmatrix} \end{array}. \tag{6.4}$$

From a modeling standpoint, the impact of increasing obesity rates would be most noticeable in two main places, the last column of **P** and the last row of **F**. The final **P** column contains the probabilities of transitioning into the Obese category from one of the other categories. If obesity rates increase, one or more of these transition probabilities

could increase which would indicate that a higher portion of the population reaches the Obese state. Meanwhile, the bottom row of **F** contains the distributions of the sojourn times for the system to leave the Obese state once it arrives there. Increases in sojourn times for the Obese state would indicate that the average child that enters this state spends longer there and that there is a greater chance for the next child to enter the system in the Obese category which would not require a state change.

Comparing the two models above, the probability of reaching the Obese state is lowest for the 1930-1960 Model and highest for the 1980-2007 Model. From Table 6.1, the 1980-2007 Model has the largest average sojourn times for every state. Figure 6.2 shows the difference in the sojourn time distribution going from State 2 to State 4 in the two models. The 1930-1960 Model has typically shorter sojourn times for this transition when compared with the 1980-2007 Model. Taken together, the 1980-2007 Model implies that a larger portion of the child population reaches the Obese category and on average stays there for a longer period of time. The larger question is whether the differences in the models are significant. The 1930-1960 and 1980-2007 Models were tested against each other after t=1000 person years, approximately 5 actual years of data collection. For this study, a person year is the collective amount of time the study participants age during observation which is indicative of the number of measurements recorded. For instance, tracking 20 individuals for 5 years would contribute 100 person years to the study which corresponds to approximately 200 individual measurements. The data set contained 1146 children and 12,980 person years. The 1930-1960 Model has a type II error of 0.441 against the 1980-2007 Model. This type II error rate deceases to 0.01 after 2000 person years (roughly 10 years of data collection). The actual 1930-1960 data sample is also rejected against the 1980-2007 Model. The 1980-2007 Model has a type II error rate of 0.076 against the 1930-1960 Model after 1000 person years and the actual 1980-2007 data sample is rejected when compared with the 1930-1960 Model.

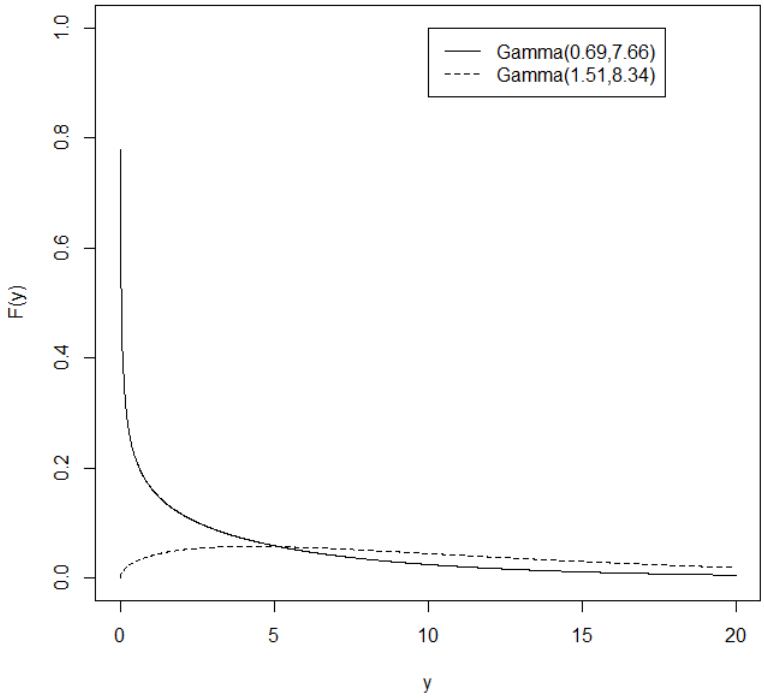

Figure 6.2: 1930-1960 $\mathbf{F}_{24}$ distribution versus the 1980-2007 $\mathbf{F}_{24}$ distribution

Table 6.1: Average State Sojourn Time (Years)

	1930-1960 Model	1980-2007 Model
Underweight	1.72	2.10
Normal	7.58	10.81
Overweight	1.23	1.85
Obese	1.38	2.83

The disparity in the type II error rates is due to the behavior of the $GoF(t)$ metric distribution for the two model comparisons. Specifically, over 40% of the 1930-1960 Model $GoF(t)$ metric values still appear to be from the 1980-2007 Model. In these cases, there are enough visits to the Overweight and Obese States to mimic the 1980-2007 Model behavior. As t increases, the 1930-1960 Model $GoF(t)$ metric distribution separates itself

from the 1980-2007 Model $GoF(t)$ metric distribution and the type II error rate decreases to 0.01. The bottom line is there exists strong evidence that the underlying BMI model has changed significantly from the mid-1900s to the late-1900s.

It is worth noting that the changing **P** and **F** matrices between the 1930-1960 Model and the 1980-2007 Model offer a stark contrast with the matrices from the scenarios in Chapter 5. While the individual changes to **P** and **F** are relatively small between the two BMI models, the cumulative effect is two models that are relatively easy to distinguish between. In the Chapter 5 scenarios, each test model was carefully created to ensure a balance with respect to the baseline model, with the added result that competing models were often difficult to differentiate between. This is a case where the real data example proved to be an easier demonstration of the $GoF(t)$ metric than the simulated scenario, but simplifying assumptions regarding the modeling of transitions between states from child to child affect the results. Further model refinement would be necessary to remove these effects entirely for less biased estimates of the transition probabilities and distributions between these two cohorts of children.

For analysts and policy-makers the utility of this goodness of fit metric comes from the ability to determine if a data sample could be the product of a given model. In the case of childhood obesity, the metric can test whether current obesity trends differ from previous generations. Suppose analysts are interested in the effectiveness of Play 60, the Healthy, Hunger-Free Kids Act, and all of the other recent programs targeting obesity. After collecting five years worth of pediatric visits for a group of patients, the summary statistics can tell whether the obesity rates are decreasing. However, comparing this data to the 1980-2007 Model can tell whether there is an actual change in population behavior that warrants a new model, or if the current trends are merely the product of the stochastic noise that is expected based on the 1980-2007 Model. In the former case, policy-makers may elect to continue with the current activities, while large-scale program changes may be warranted

in the later case. In either scenario, policy-makers have an improved grasp of what the current trends mean with respect to the previous childhood obesity model behavior.

VII. Conclusion

Assessing the fit of a multi-state model with respect to the observed data used to create the model is an important step in the diagnostics phase of modeling. Previous diagnostics tools using prevalence counts and Chi-squared goodness of fit metrics tended to possess limited utility when compared with the wide range of models and data collection techniques used. The Goodness of Fit metric outlined in this dissertation has been shown to overcome the limitations of previously available diagnostic tools by providing a straightforward method of relating the observed data directly to a hypothesized multi-state model.

While validating the goodness of fit for a model is an important diagnostic step, deciding among competing models is the ultimate goal of model building with respect to providing a decision maker with the best possible information. To this end, this work illustrates how models that contain covariate factors can be tested against one another to determine which model best represents the true observed data. Proper model selection can impact model size, model complexity, and data collection efforts which in turn can have a profound impact on the time required to analyze a particular model.

Future work may explore the impact of sojourn distributions that lack a linear Laplace transformation function. In these cases, the Laplace transform inversion technique used to calculate expected values becomes a numerical integration of a numerical integration of a function rather than the single numerical integration of a function demonstrated in this work. Additionally, more work is required in fitting the tails of the $GoF(T)$ metric's distribution with respect to the transformed Chi-squared distribution. The ability to test an observed data sample against a theoretical distribution rather than a simulated distribution would greatly reduce the computational workload in determining whether or not a model is likely to produce an observation.

Appendix A: Markov Renewal Process Moments

The covariance calculations in Chapter 3 require the expected value and variance of a Markov renewal process at a specific time, t. The following is a derivation of both values in the transform domain based on the generating function

$$\Psi_z = \mathbf{1} - (1-z)(\mathbf{I}-q)^{-1}q[z\mathbf{I} + (1-z)\ {}_d\{(\mathbf{I}-q)^{-1}\}]^{-1}, \tag{A.1}$$

defined by Pyke in Corollary 5.1 [15]. In this case, z is a random value between 0 and 1 and $\mathbf{q} = \mathbf{q}(t) = \mathbf{P} \circ \frac{d}{dt}\mathbf{F}(t)$. Pyke uses Ψ_z to derive $\mathcal{L}(\mathrm{E}[\widetilde{N}(t)])$ in Theorem 5.2 [15]. Below is a derivation of $\mathcal{L}(\mathrm{E}[\widetilde{N}(t)])$, $\mathcal{L}(\mathrm{E}[\widetilde{N}(t)^2])$, and $\mathrm{Var}[\widetilde{N}(t)]$.

$$\mathcal{L}(\mathrm{E}[\widetilde{N}(t)]) = \frac{\partial}{\partial z}\Psi_z\Big|_{z=1}.$$

Consider

$$\frac{\partial}{\partial z}\Psi_z = \frac{\partial}{\partial z}\mathbf{1} - (1-z)(\mathbf{I}-q)^{-1}q[z\mathbf{I} + (1-z)\ {}_d\{(\mathbf{I}-q)^{-1}\}]^{-1}.$$

Let $c = {}_d\{(\mathbf{I}-q)^{-1}\}$, then

$$\begin{aligned}\frac{\partial}{\partial z}\Psi_z &= (\mathbf{I}-q)^{-1}q[z\mathbf{I} + (1-z)c]^{-1} - (1-z)(\mathbf{I}-q)^{-1}q(-1)* \\ &\qquad [z\mathbf{I} + (1-z)c]^{-1}[\mathbf{I}-c][z\mathbf{I} + (1-z)c]^{-1}. \\ &= (\mathbf{I}-q)^{-1}q[z\mathbf{I} + (1-z)c]^{-1} + (1-z)(\mathbf{I}-q)^{-1}q[z\mathbf{I} + (1-z)c]^{-1}* \\ &\qquad [\mathbf{I}-c][z\mathbf{I} + (1-z)c]^{-1} \text{ and}\end{aligned} \tag{A.2}$$

$$\begin{aligned}\mathcal{L}(\mathrm{E}[\widetilde{N}(t)]) &= (\mathbf{I}-q)^{-1}q[z\mathbf{I} + (1-z)c]^{-1} + (1-z)(\mathbf{I}-q)^{-1}q[z\mathbf{I} + (1-z)c]^{-1}* \\ &\qquad [\mathbf{I}-c][z\mathbf{I} + (1-z)c]^{-1}\Big|_{z=1}. \\ &= (\mathbf{I}-q)^{-1}q.\end{aligned}$$

$$\mathcal{L}(\mathrm{E}[\widetilde{N}(t)^2]) = \frac{\partial^2}{\partial z^2}\Psi_z\Big|_{z=1} + \frac{\partial}{\partial z}\Psi_z\Big|_{z=1}.$$

$$\frac{\partial^2}{\partial z^2}\Psi_z = \frac{\partial}{\partial z}(\mathbf{I}-q)^{-1}q[z\mathbf{I}+(1-z)c]^{-1} + (1-z)(\mathbf{I}-q)^{-1}q*$$
$$[z\mathbf{I}+(1-z)c]^{-1}[\mathbf{I}-c][z\mathbf{I}+(1-z)c]^{-1}.$$
$$= (-1)(\mathbf{I}-q)^{-1}q[z\mathbf{I}+(1-z)c]^{-1}(\mathbf{I}-c)[z\mathbf{I}+(1-z)c]^{-1}-$$
$$(\mathbf{I}-q)^{-1}q[z\mathbf{I}+(1-z)c]^{-1}[\mathbf{I}-c][z\mathbf{I}+(1-z)c]^{-1}+$$
$$(1-z)(\mathbf{I}-q)^{-1}q(-1)[z\mathbf{I}+(1-z)c]^{-1}[\mathbf{I}-c][z\mathbf{I}+(1-z)c]^{-1}*$$
$$[\mathbf{I}-c][z\mathbf{I}+(1-z)c]^{-1} + (1-z)(\mathbf{I}-q)^{-1}q[z\mathbf{I}+(1-z)c]^{-1}*$$
$$[\mathbf{I}-c](-1)[z\mathbf{I}+(1-z)c]^{-1}[\mathbf{I}-c][z\mathbf{I}+(1-z)c]^{-1}.$$
$$= -2(\mathbf{I}-q)^{-1}q[z\mathbf{I}+(1-z)c]^{-1}(\mathbf{I}-c)[z\mathbf{I}+(1-z)c]^{-1}-$$
$$2(1-z)(\mathbf{I}-q)^{-1}q[z\mathbf{I}+(1-z)c]^{-1}[\mathbf{I}-c][z\mathbf{I}+(1-z)c]^{-1}*$$
$$[\mathbf{I}-c][z\mathbf{I}+(1-z)c]^{-1}, \text{ then}$$

$$\frac{\partial^2}{\partial z^2}\Psi_z\Big|_{z=1} = -2(\mathbf{I}-q)^{-1}q[z\mathbf{I}+(1-z)c]^{-1}(\mathbf{I}-c)[z\mathbf{I}+(1-z)c]^{-1}-$$
$$2(1-z)(\mathbf{I}-q)^{-1}q[z\mathbf{I}+(1-z)c]^{-1}[\mathbf{I}-c][z\mathbf{I}+(1-z)c]^{-1}*$$
$$[\mathbf{I}-c][z\mathbf{I}+(1-z)c]^{-1}\Big|_{z=1}.$$
$$= -2(\mathbf{I}-q)^{-1}q(I-c).$$

$$\mathcal{L}(\mathrm{E}[\widetilde{N}(t)^2]) = -2(\mathbf{I}-q)^{-1}q(I-c) + (\mathbf{I}-q)^{-1}q.$$

$$\mathrm{Var}[\widetilde{N}(t)] = \mathcal{L}^{-1}(\mathrm{E}[\widetilde{N}(t)^2]) - \left(\mathcal{L}^{-1}(\mathrm{E}[\widetilde{N}(t)])\right)^2.$$

Appendix B: Probability Transition Matrix Test Results

Using the test construct from Chapter 4, $GoF(t)$ metric tests utilizing differing **P**-matrices were also conducted. Each table below corresponds to one of seven **P**-matrices.

$$\mathbf{P}_1 = \begin{bmatrix} 0 & 0.75 & 0 & 0.25 \\ 0.5 & 0 & 0.5 & 0 \\ 0.25 & 0.45 & 0 & 0.3 \\ 0.15 & 0.35 & 0.5 & 0 \end{bmatrix} \tag{B.1}$$

$$\mathbf{P}_2 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix} \tag{B.2}$$

$$\mathbf{P}_3 = \begin{bmatrix} 0 & 0.65 & 0.10 & 0.25 \\ 0.4 & 0 & 0.4 & 0.2 \\ 0.25 & 0.45 & 0 & 0.3 \\ 0.15 & 0.35 & 0.5 & 0 \end{bmatrix} \tag{B.3}$$

$$\mathbf{P}_4 = \begin{bmatrix} 0 & 0.25 & 0 & 0.75 \\ 0.5 & 0 & 0.5 & 0 \\ 0.45 & 0.25 & 0 & 0.3 \\ 0.5 & 0.35 & 0.15 & 0 \end{bmatrix} \tag{B.4}$$

$$\mathbf{P}_5 = \begin{bmatrix} 0 & 0.15 & 0 & 0.85 \\ 0.5 & 0 & 0.5 & 0 \\ 0.25 & 0.45 & 0 & 0.3 \\ 0.15 & 0.35 & 0.5 & 0 \end{bmatrix} \tag{B.5}$$

$$\mathbf{P}_6 = \begin{bmatrix} 0 & 0.75 & 0 & 0.25 \\ 0.05 & 0 & 0.95 & 0 \\ 0.25 & 0.45 & 0 & 0.3 \\ 0.15 & 0.35 & 0.5 & 0 \end{bmatrix} \tag{B.6}$$

$$\mathbf{P}_7 = \begin{bmatrix} 0 & 0.75 & 0 & 0.25 \\ 0.5 & 0 & 0.5 & 0 \\ 0.75 & 0.15 & 0 & 0.1 \\ 0.15 & 0.35 & 0.5 & 0 \end{bmatrix} \tag{B.7}$$

$$\mathbf{P}_8 = \begin{bmatrix} 0 & 0.75 & 0 & 0.25 \\ 0.5 & 0 & 0.5 & 0 \\ 0.25 & 0.45 & 0 & 0.3 \\ 0.145 & 0.85 & 0.005 & 0 \end{bmatrix} \tag{B.8}$$

Table B.1: Type II Error For $\mathbf{P}_2$

	α Vectors						
β vectors	$\alpha 1$	$\alpha 2$	$\alpha 3$	$\alpha 4$	$\alpha 5$	$\alpha 6$	$\alpha 7$
$\beta 1$	0	0	0	0	0	0	0
$\beta 2$	0	0	0	0	0	0	0
$\beta 3$	0	0	0	0	0	0	0
$\beta 4$	0	0	0	0	0	0	0
$\beta 5$	0	0	0	0	0	0	0
$\beta 6$	0	0	0	0	0	0	0
$\beta 7$	0	0	0	0	0	0	0

Table B.2: Type II Error For $\mathbf{P}_3$

	α Vectors						
β vectors	$\alpha 1$	$\alpha 2$	$\alpha 3$	$\alpha 4$	$\alpha 5$	$\alpha 6$	$\alpha 7$
$\beta 1$	0.154	0	0	0	0	0	0
$\beta 2$	0	0	0.128	0	0	0	0
$\beta 3$	0	0.148	0	0	0	0.0006	0
$\beta 4$	0.0002	0	0	0	0	0	0
$\beta 5$	0.0002	0	0	0	0	0	0
$\beta 6$	0	0	0.046	0	0	0	0
$\beta 7$	0	0	0.017	0	0	0	0.137

Table B.3: Type II Error For $\mathbf{P}_4$

	α Vectors						
β vectors	$\alpha 1$	$\alpha 2$	$\alpha 3$	$\alpha 4$	$\alpha 5$	$\alpha 6$	$\alpha 7$
$\beta 1$	0	0	0	0	0	0	0
$\beta 2$	0	0	0	0	0	0	0
$\beta 3$	0	0	0	0	0	0	0
$\beta 4$	0	0	0	0	0	0	0
$\beta 5$	0	0	0	0	0	0	0
$\beta 6$	0	0	0	0	0	0	0
$\beta 7$	0	0	0	0	0	0	0.113

Table B.4: Type II Error For $\mathbf{P}_5$

	α Vectors						
β vectors	$\alpha 1$	$\alpha 2$	$\alpha 3$	$\alpha 4$	$\alpha 5$	$\alpha 6$	$\alpha 7$
$\beta 1$	0	0	0	0	0	0	0
$\beta 2$	0	0	0	0	0	0	0
$\beta 3$	0	0	0	0	0	0	0
$\beta 4$	0	0	0	0	0	0	0
$\beta 5$	0	0	0	0	0	0	0
$\beta 6$	0	0	0	0	0	0	0
$\beta 7$	0	0	0	0	0	0	0.070

Table B.5: Type II Error For $\mathbf{P}_6$

	α Vectors						
β vectors	$\alpha 1$	$\alpha 2$	$\alpha 3$	$\alpha 4$	$\alpha 5$	$\alpha 6$	$\alpha 7$
$\beta 1$	0	0	0	0	0	0	0
$\beta 2$	0	0	0	0	0	0	0
$\beta 3$	0	0	0	0	0	0	0
$\beta 4$	0	0	0	0	0	0	0
$\beta 5$	0	0	0	0	0	0	0
$\beta 6$	0	0	0	0	0	0	0
$\beta 7$	0	0	0	0	0	0	0.063

Table B.6: Type II Error For $\mathbf{P}_7$

	α Vectors						
β vectors	$\alpha 1$	$\alpha 2$	$\alpha 3$	$\alpha 4$	$\alpha 5$	$\alpha 6$	$\alpha 7$
$\beta 1$	0.006	0	0	0	0	0	0
$\beta 2$	0	0	0.005	0	0	0	0
$\beta 3$	0	0.007	0	0	0	0	0
$\beta 4$	0	0	0	0	0	0	0
$\beta 5$	0	0	0	0	0	0	0
$\beta 6$	0	0	0.007	0	0	0	0
$\beta 7$	0	0	0.003	0	0	0	0.190

Table B.7: Type II Error For $\mathbf{P}_8$

	α Vectors						
β vectors	$\alpha 1$	$\alpha 2$	$\alpha 3$	$\alpha 4$	$\alpha 5$	$\alpha 6$	$\alpha 7$
$\beta 1$	0.532	0	0	0	0	0	0
$\beta 2$	0	0	0.453	0	0	0	0
$\beta 3$	0	0.589	0	0.004	0	0	0.346
$\beta 4$	0	0	0.042	0	0	0	0
$\beta 5$	0	0	0	0	0	0	0
$\beta 6$	0	0	0.020	0	0	0	0
$\beta 7$	0.0004	0	0.100	0	0	0	0.070

Bibliography

[1] Guglielmo, D., Manca, R., and Salvi, G., "Bivariate Semi-Markov Process for Counterparty Credit Risk," Tech. rep., arXiv. org, 2012.

[2] Buchardt, K., Møller, T., and Schmidt, K. B., "Cash flows and policyholder behaviour in the semi-Markov life insurance setup," *Department of Mathematical Sciences, University of Copenhagen and PFA Pension*, 2013.

[3] Xia, J. C., Zeephongsekul, P., and Packer, D., "Spatial and temporal modeling of tourist movements using Semi-Markov processes," *Tourism Management*, Vol. 32, No. 4, 2011, pp. 844–851.

[4] Titman, A. C. and Sharples, L. D., "Model diagnostics for multi-state models," *Statistical methods in medical research*, 2009.

[5] Kao, E. P., *An Introduction to Stochastic Processes*, Vol. 4, Duxbury Press NY, 1997.

[6] Ross, S. M., *Stochastic Processes, 2nd*, New York: Wiley, 1996.

[7] Kulkarni, V. G., *Modeling and Analysis of Stochastic Systems*, CRC Press, 1996.

[8] Lévy, P., "Systémes Semi-Markoviens," *Proc. Int. Congr. Math. (Amsterdam)*, Vol. 3, 1954, pp. 416–426.

[9] Takács, L., "Some investigations concerning recurrent stochastic processes of a certain type," *Magyar Tud. Akad. Mat. Kutato Int. Kzl*, Vol. 3, 1954, pp. 115–128.

[10] Smith, W. L., "Regenerative stochastic processes," *Proc. Roy. Soc. (London)*, Vol. 232, 1955, pp. 6–31.

[11] Pyke, R., "Markov renewal processes: definitions and preliminary properties," *The Annals of Mathematical Statistics*, 1961, pp. 1231–1242.

[12] Barlow, R. E. and Proschan, F., *Mathematical Theory of Reliability*, No. 17, Siam, 1996.

[13] Gorissen, M. and Vanderzande, C., "Semi-Markov models of mRNA-translation," *arXiv preprint arXiv:1104.0131*, 2011.

[14] Xie, Y. and Yu, S.-Z., "A large-scale hidden semi-Markov model for anomaly detection on user browsing behaviors," *Networking, IEEE/ACM Transactions on*, Vol. 17, No. 1, 2009, pp. 54–65.

[15] Pyke, R., "Markov renewal processes with finitely many states," *The Annals of Mathematical Statistics*, 1961, pp. 1243–1259.

[16] He, S.-w., Wang, J.-g., and Yan, J.-a., *Semimartingale Theory and Stochastic Calculus*, Taylor & Francis, 1992.

[17] Gentleman, R., Lawless, J., Lindsey, J., and Yan, P., "Multi-state Markov models for analysing incomplete disease history data with illustrations for hiv disease," *Statistics in medicine*, Vol. 13, No. 8, 1994, pp. 805–821.

[18] De Stavola, B. L., "Testing departures from time homogeneity in multistate Markov processes," *Applied Statistics*, 1988, pp. 242–250.

[19] Aguirre-Hernández, R. and Farewell, V., "A Pearson-type goodness-of-fit test for stationary and time-continuous Markov regression models," *Statistics in medicine*, Vol. 21, No. 13, 2002, pp. 1899–1911.

[20] Pearson, K., "X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, Vol. 50, No. 302, 1900, pp. 157–175.

[21] Bentler, P. M. and Bonett, D. G., "Significance tests and goodness of fit in the analysis of covariance structures." *Psychological bulletin*, Vol. 88, No. 3, 1980, pp. 588.

[22] Mendenhall, W., Beaver, R., and Beaver, B., *Introduction to probability and statistics*, Cengage Learning, 2012.

[23] Abate, J. and Whitt, W., "Numerical inversion of Laplace transforms of probability distributions," *ORSA Journal on computing*, Vol. 7, No. 1, 1995, pp. 36–43.

[24] "Centers for Disease Control and Prevention: Childhood Overweight and Obesity," World Wide Web Page, Available at *http://www.cdc.gov/obesity/childhood/index.html*.

[25] von Hippel, P. T. and Nahhas, R. W., "Extending the history of child obesity in the United States: The fels longitudinal study, birth years 1930-1993," *Obesity*, Vol. 21, No. 10, 2013, pp. 2153–2156.

[26] Lobstein, T. and Jackson-Leach, R., "Child overweight and obesity in the USA: prevalence rates according to IOTF definitions," *International Journal of Pediatric Obesity*, Vol. 2, No. 1, 2007, pp. 62–64.

[27] Freedman, D. S., Mei, Z., Srinivasan, S. R., Berenson, G. S., and Dietz, W. H., "Cardiovascular risk factors and excess adiposity among overweight children and adolescents: the Bogalusa Heart Study," *The Journal of pediatrics*, Vol. 150, No. 1, 2007, pp. 12–17.

[28] Whitlock, E. P., Williams, S. B., Gold, R., Smith, P. R., and Shipman, S. A., "Screening and interventions for childhood overweight: a summary of evidence for

the US Preventive Services Task Force," *Pediatrics*, Vol. 116, No. 1, 2005, pp. e125–e144.

[29] Roche, A. F., *Growth, maturation, and body composition: the Fels Longitudinal Study 1929-1991*, No. 9, Cambridge University Press, 1992.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704–0188

1. REPORT DATE *(DD–MM–YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From — To)*
17–09–2015	Dissertation	October 2012 - September 2015

4. TITLE AND SUBTITLE

Testing the Adequacy of a Semi-Markov Process

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Seymour, Richard S., Lt Col, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way
Wright-Patterson AFB, OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENC-DS-15-S-003

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Withheld

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Distribution Statement A:
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Due to the versatility of its structure, the semi-Markov process is a powerful modeling tool used to describe complex systems. Though similar in structure to continuous time Markov chains, semi-Markov processes allow for any transition time distribution which enables these processes to fit a wider range of problems than the continuous time Markov chain. While semi-Markov processes have been applied in fields as varied as biostatistics and finance, there does not exist a theoretically-based, systematic method to determine if a semi-Markov process accurately fits the underlying data used to create the model.

In fields such as regression and analysis of variance, the quality of the predictive model is judged in part by the goodness of fit of the model which relates the expected observation values with the actual observations. A similar methodology for semi-Markov processes would provide immediate insight in the efficacy of the fitted model and would allow competing models to be directly compared with one another.

This dissertation presents a methodology to measure the adequacy of a fitted semi-Markov process. To this end, a technique to assess the likelihood that a data sample could be generated by a specific semi-Markov process is developed, including a newly proposed goodness of fit metric. This technique relies on the covariance structure of the semi-Markov process; thus, a method to estimate the covariance structure is also proposed. The technique is applied to real and simulated data to demonstrate the goodness of fit metric's utility in model validation and its ability to identify potential covariate factors within the model.

15. SUBJECT TERMS

Semi-Markov Process, Goodness-of-Fit

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Christine M. Schubert Kabban, AFIT/ENC
U	U	U	UU	78	19b. TELEPHONE NUMBER *(include area code)* (937) 255-3636x4549; christine.schubertkabban@afit.edu