DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

# REVIEW OF THE METHODS USED FOR THE ESTIMATION OF SOJOURN TIME FOR SEMI-MARKOV PROCESS

by
Özgür DANIŞMAN

August, 2015
İZMİR

# REVIEW OF THE METHODS USED FOR THE ESTIMATION OF SOJOURN TIME FOR SEMI-MARKOV PROCESS

**A Thesis Submitted to the**
**Graduate School of Natural and Applied Sciences of Dokuz Eylül University**
**In Partial Fulfillment of the Requirements for the Master of Science in**
**Statistics, Statistics Program**

**by**
**Özgür DANIŞMAN**

**August, 2015**
**İZMİR**

**M.Sc THESIS EXAMINATION RESULT FORM**

We have read the thesis entitled **"REVIEW OF THE METHODS USED FOR THE ESTIMATION OF SOJOURN TIME FOR SEMI-MARKOV PROCESS"** completed by **ÖZGÜR DANIŞMAN** under supervision of **ASST. PROF. DR. UMAY UZUNOĞLU KOÇER** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Asst. Prof. Dr. Umay UZUNOĞLU KOÇER

Supervisor

Doç. Dr. Özlem EGE ORUÇ

(Jury Member)

Yrd. Doç. Dr. Halil TANIL

(Jury Member)

Prof. Dr. Ayşe OKUR

Director

Graduate School of Natural and Applied Sciences

# ACKNOWLEDGEMENTS

I would like to express my sincere gratitude and thanks to my advisor, Asst. Prof. Dr. Umay UZUNOĞLU KOÇER, for her guidance, continuous support and encouragement that she gave to me so I have accomplished my thesis study. Her suggestions played a very important role for my thesis and academic career.

I would like to thank to Assoc. Prof. Dr. Özlem EGE ORUÇ and Asst. Prof. Dr. Halil TANIL for their helpful and constructive comments as a member of my dissertation committee.

Special thanks to all of academic staff in Dokuz Eylül University Statistics Department especially to Okan İŞGÜDER, Ufuk BEYAZTAŞ and Beste SERTDEMİR for their advice, support and most importantly friendship.

Finally, I would like to express my deepest love, respect and gratitude to Hanife DANIŞMAN, my mother, to Rıfat DANIŞMAN, my father, to Rahşan DANIŞMAN GÜROCAK and Gülşan DANIŞMAN ÇİFTÇİ, my elder sisters, to Tanju TOKYAY, my aunt, to Ayhan TOKYAY, my uncle and to Çağlar TOKYAY ASIM, my cousin for their endless support, help and love. When I collapse or have a problem, they always support me in every way.

Özgür DANIŞMAN

**REVIEW OF THE METHODS USED FOR THE ESTIMATION OF SOJOURN TIME FOR SEMI-MARKOV PROCESS**

## ABSTRACT

Semi-Markov process is a stochastic process which can be applied in many areas such as survival analysis, queueing theory, inventory problems, seismological efforts or reliability analysis. The process makes transitions between states in accordance with the Markovian property while spending a random time in the states of the system. The random time that the system spends in any state is called sojourn time. However, according to the process, the sojourn times of the system can follow different distributions. When sojourn times of the process follow different distributions, the stochastic process is called semi-Markov process.

In this study, semi-Markov process which is a stochastic process have been investigated in detail. The methods that exists in the literature for estimation of distribution parameters of sojourn times for semi-Markov process are reviewed. Official match results and dates of a football team are characterized by semi-Markov process and EM-algorithm is used for the estimation of distribution parameters of sojourn times. After the estimation of parameters, stationary distributions and mean sojourn times for semi-Markov process are estimated by both empirical estimators and the way which is shown in the literature. Finally, a comparison is made between the results that obtained with given methods.

**Keywords:** Semi-Markov process, sojourn time, EM-algorithm, parameter estimation

# YARI-MARKOV SÜRECİ İÇİN SÜRECİN HERHANGİ BİR DURUMDA KALMA SÜRESİNİ TAHMİN ETMEK AMACIYLA KULLANILAN YÖNTEMLERİN İNCELENMESİ

## ÖZ


Yarı-Markov süreci; sağkalım analizi, kuyruk analizi, envanter problemleri, sismolojik çalışmalar, güvenilirlik analizi gibi birçok alanda uygulamasının olanaklı olduğu bir stokastik süreçtir. Süreç Markov özelliğine göre durumlar arasında geçiş yapmakla birlikte, sistemin durumlarında rassal bir süre kalmaktadır. Herhangi bir durumda kaldığı bu rassal süreye oturma süresi denir. Bununla birlikte sistemin oturma süreleri, sürece bağlı olarak farklı dağılımlar gösterebilir. Bu durumda, stokastik süreç yarı-Markov süreci olarak adlandırılır.

Bu çalışmada bir stokastik süreç olan yarı-Markov süreci detaylı bir şekilde incelenmiştir. Yarı-Markov süreçlerinde oturma süresinin dağılım parametrelerini tahmin etmek için literatürde var olan yöntemler gözden geçirilmiştir. Bir futbol takımının maçlarda elde ettiği sonuçlar ve maç tarihleri yarı-Markov süreci olarak ifade edilmiş, oturma sürelerinin dağılım parametrelerini tahmin etmek için BM-algoritması kullanılmıştır. Parametre tahmininden sonra yarı-Markov süreci için durağan dağılım olasılıkları ve ortalama oturma süreleri hem deneysel tahminciler ile hem de literatürde gösterilen tahminciler ile tahmin edilmiştir. Son olarak, verilen yöntemler kullanılarak elde edilen sonuçlar arasında bir karşılaştırma yapılmıştır.



**Anahtar kelimeler:** Yarı-Markov süreci, oturma süresi, BM-algoritması, parametre tahmini

## CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

# CHAPTER ONE
# INTRODUCTION

The word stochastic means random and describes a situation that involves chance or probability. It is used as counterpart of the word deterministic. If a process or event is defined under the stochastic conditions, it means that it is not known exactly what will happen in the future but it can be predicted. Stochastic systems and processes are very important for statistics and mathematics specifically in probability theory.

Stochastic process is an interesting and developing area of probability and statistics. A stochastic process is a probabilistic process related to an index set $T$. It is defined as a collection of random variables. Generally, $t$ is interpreted as time and $X(t)$ represents the state of the process at time $t$. If the index set $T$ is a countable or finite set, process is a discrete-time stochastic process. If the index set $T$ is a continuous set, process is a continuous-time stochastic process. For instance, $X(t)$ may represent the number of people in a queue at time $t$ or the analysis of match results of a football club at time $t$ which is the main topic of this study.

A stochastic process satisfies the Markovian property if the conditional probability of future states of the stochastic process depends only on the present state and it is called Markov process. Markov process can be classified into four groups according to its state space and parameter space. If the state space and parameter space are discrete, we have a discrete-time Markov chain. If we have a continuous parameter space with a discrete state space, related process is a continuous-time Markov chain. If the state space is continuous and parameter space is discrete, process is called discrete-time Markov process. When the state space and parameter space is continuous, it is called continuous-time Markov process. Discrete-time Markov chain and continuous-time Markov chain are studied in detail in this thesis.

The random time that the system spends in any state is called sojourn time. According to the distribution of sojourn times in the states of the system, stochastic process can show different characteristics. A Markov process that makes transitions between states in accordance with a Markov chain and all the transition times are identically 1 is a discrete-time Markov chain. A Markov process that makes transitions between states in accordance with a Markov chain and sojourn time follow exponential distribution is called continuous-time Markov chain. Well, what if the sojourn times do not follow exponential distribution? Sojourn times can follow different distributions in a Markov process. In other words, it is said that distributions of the sojourn times are random. So, in a Markov process, if the process moves from states to states in accordance with a Markov chain and sojourn times are randomly distributed, this process is called semi-Markov process.

Semi-Markov process was first introduced by Lévy (1954) and Smith (1954) independently at the International Congress of Mathematicians held at Amsterdam. Takács (1954) also presented an introduction about semi-Markov process and in addition to this, he studied on some problems in Counter theory. Pyke (1961) gave some definitions and basic properties about Markov renewal process and described the relationship between Markov renewal process and semi-Markov process. These studies have been followed about the introduction of semi-Markov process by Ezhov and Korolyuk (1967). Athreya, McDonald and Ney (1978) studied the limit theorems for semi-Markov process. Afterwards, it has been started to introduce some of applications and parameter estimation for semi-Markov process. There are some studies by Voelkey and Crowley (1984) and Ouhbi and Limnios (1999) which is about nonparametric inference based on semi-Markov process. Semi-Markov process has been used to characterize earthquake data and used for earthquake forecasting (Altinok, & Kolcak, 1999; Alvarez, 2004). A semi-Markov model for viewing a manpower system has been constructed by Yadavalli and Natarajan (2001). An application with Weibull sojourn times to power-plant reliability analysis has been presented (Perman, Senagacnik, & Tuma, 1997). An empirical estimator is proposed by Limnios, Ouhbi and Sadek (2005). Pons (2007) and Ugwuowo and Udoumoh (2011) constructed several semi-Markov models and made parameter estimation of the distribution of sojourn times with censored variables. A non-homogeneous parametric semi-Markov model to estimate and identify the precipitation occurences has been adopted (Masala, 2014).

EM-algorithm is explained and given its name with a paper by Dempster, Laird and Rubin (1977) but they had emphasized that it had been proposed many times in different and special conditions by earlier authors. The convergence studies of the Dempster, Laird and Rubin (1977) is flawed and correct convergence properties are intruduced by Wu (1983). It is a general iterative algorithm for parameter estimation that uses maximum likelihood approach when some of the variables are not observed or inappropriate such as missing or censored.

The purposes of this thesis are to characterize a data set with semi-Markov process, estimating the distribution parameters of the sojourn times with EM-algorithm, finding the mean sojourn times and the stationary distributions of the semi-Markov process with two different methods and then comparing the results and viewing the goodness of estimators.

The second chapter of the study involves a detailed introduction of stochastic process. Main properties of stochastic process are given and Markov process which is a special type of stochastic process is examined. Semi-Markov process and its properties are studied inclusively and the relationship with other processes is emphasized. Since semi-Markov process is the main idea of this study and it is necessary to adopt these properties. The purpose of the third chapter is to investigate the methods for the estimation of sojourn time distribution. The fourth chapter includes two different parts: parameter estimation with EM-algorithm and the empirical estimator. After the characterization of a real data set with semi-Markov process, transition probability matrix and limiting distributions for Markov chain is constructed. Then, using the estimators, estimations for distribution paramaters and mean sojourn times are procured. After calculating the estimations found by EM-algorithm and empirical estimator, the results are compared and their goodness are examined. Finally, the conclusion remarks are presented in the last section.

# CHAPTER TWO
# STOCHASTIC PROCESSES

Stochastic process is a very important topic that has a large variety of usage area such as finance, physics, traffic events, precipitation analysis or queueing systems. In this chapter, basic definitions and characterization of stochastic processes are presented. Also, the classification of stochastic processes is introduced. The main focus of the chapter is semi-Markov process.

## 2.1 Basic Preliminaries

Stochastic process $\{X(t), t \in T\}$ is a process which is a family of random variables. Here, $X(t)$ is a random variable for each $t$ in the index set $T$ because, the process takes random values corresponding to various times or time points. The variable $t$ generally represents the time and so, $X(t)$ is the state of the stochastic process at time $t$.

$T$ is the *index set* of the stochastic process. It is called discrete-time stochastic process if $T$ is a *countable index set* and it is called continuous-time stochastic process if the index set $T$ is a *continuum*.

The *state space* of the stochastic process is a set that contains the possible values which $X(t), t \in T$ can be assumed.

Note that for a continuous-time stochastic process $\{X(t), t \in T\}$, if for all of $t_0 < t_1 < t_2 < \ldots < t_n$,

$$X(t_1) - X(t_0), X(t_2) - X(t_1), \ldots, X(t_n) - X(t_{n-1}), \ldots$$

are mutually independent, the stochastic process has *independent increments*. In addition to this, if the distributions of $X(t_2) - X(t_1)$ and $X(t_2 + s) - X(t_1 + s)$ are identical for all $t_1, t_2 \in T$ and $s > 0$, the stochastic process has *stationary independent increments*.

Now, let us introduce the stochastic process $\{N(t), t \geq 0\}$. $N(t)$ represents the total number of events occurred by time $t$. This process is said to be a *counting process* if $N(t)$ is integer-valued and non-decreasing with time.

Stochastic process has two major categories: arrival type processes and Markov processes. In this study, because of the importance for semi-Markov process and to understanding semi-Markov process better, definition and properties of Markov process is presented.

**2.2 Markov Processes**

In the simplest term, Markov process is a stochastic process where the next state is dependent only the previous state. Markov process can be applied in many areas such as communication systems, biological systems, economic mobility or population studies. For instance, a stock market model that price of stocks depends on changes in price on the previous day. After introducing the basic logic of the Markov process, it can be given a statistical definition for Markov process.

If conditional probability distribution of a stochastic process' depends only the most recent value of the process, this process is called *Markov process*. That is, when the value of the $X(t)$ is given, the probability of future state $X(k+t)$, $k > 0$, is independent of values of the past $X(h)$, $h > t$. Then, the stochastic process is called Markov process if,

$$P\{X(t) \leq x \mid X(t_1) = x_1, X(t_2) = x_2, X(t_3) = x_3, \ldots, X(t_n) = x_n\} \tag{2.1}$$
$$= P\{X(t) \leq x \mid X(t_n) = x_n\}$$

where $t_1 < t_2 < t_3 < \ldots < t_n < t$.

Markov process has four major categories according to its parameter space and state space. The following Table 2.1 shows the classification of Markov process.

Table 2.1 Classification of the Markov process

| | State Space | |
|---|---|---|
| Parameter Space | Discrete | Continuous |
| Discrete | Discrete-time Markov chain | Discrete-time Markov process |
| Continuous | Continuous-time Markov chain | Continuous-time Markov process |

Discrete-time Markov chains and continuous-time Markov chains are explained in this study.

### *2.2.1 Discrete-Time Markov Chains (DTMC)*

A Markov process with *discrete* parameter and state space is called DTMC. The state space of the process takes on a finite or countable infinite number of possible values. The process makes transitions at time epochs $n = 0,1,2,3,\ldots$. This stochastic process is called discrete-time Markov chain if it satisfies the following Markov property,

$$P\{X_n = i_n \mid X_0 = i_0, X_1 = i_1, X_2 = i_2, \ldots, X_{n-1} = i_{n-1}\} \qquad (2.2)$$
$$= P\{X_n = i_n \mid X_{n-1} = i_{n-1}\}$$

$i_l$ is the state of the process for $l = 0,1,2,\ldots n$. A family of random variables of a discrete time process are commonly shown by $\{X_n, n = 0,1,2,\ldots\}$ with state space $S = \{0,1,2,\ldots\}$.

A discrete-time stochastic process is *time-homogeneous* if the probabilities of moving from a state to an another state does not change through time. In other words, it is independent of $n$ and the DTMC is said to have stationary transition probabilities.

Let $X_0 = i_0$ be the initial state of the process for $t = t_0$. In a discrete-time Markov chain, the process makes transitions from one state $i$ to another state $j$ with a transition probability $P_{ij}$ at times $t_n$, $n = 0,1,2,\ldots$ where,

$$P_{ij} \geq 0, \qquad i, j \in S;$$
$$\sum_{j=0}^{\infty} P_{ij} = 1, \qquad i \in S.$$

and so, the probability of making a transition from state $i$ to state $j$ in one-step is denoted by $P_{ij}^{(1)}$ and,

$$P_{ij}^{(1)} = P\{X_{n+1} = j \mid \mathrm{X}_{\mathrm{n}} = i\} \tag{2.3}$$

where $n, i, j = 0,1,2,3,\ldots$ .

Let $\boldsymbol{P}$ show the transition probability matrix of the probabilities $P_{ij}$, then

$$\boldsymbol{P} = \begin{bmatrix} P_{00} & P_{01} & P_{02} & \cdots \\ P_{10} & P_{11} & P_{12} & \cdots \\ \vdots & & & \\ P_{i0} & P_{i1} & P_{i2} & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}.$$

Let $P_{ij}^{(n)}$ be the probability that the process makes transition from state $i$ to state $j$ in n steps, so

$$P_{ij}^{(n)} = P\{X_n = j \mid \mathrm{X}_0 = i\} \tag{2.4}$$

and as we know, the discrete-time Markov chain is time-homogeneous so,

$$P_{ij}^{(n)} = P\{X_{m+n} = j \mid \mathrm{X}_{\mathrm{m}} = i\} \tag{2.5}$$

where $m > 0$.

*2.2.1.1 Chapman-Kolmogorov Equations*

The Chapman-Kolmogorov equations provide a method for computing these $n$-step transition probabilities. These equations are

$$P_{ij}^{n+m} = \sum_{k=0}^{\infty} P_{ik}^{n} P_{kj}^{m} \qquad \text{for all } n, m \geq 0, \text{ all } i, j$$

and are most easily understood by noting that $P_{ik}^n P_{kj}^m$ represents the probability that starting in $i$ the process will go to state $j$ in $n+m$ transitions through a path which takes it into state $k$ at the $n$th transition. Hence, summing over all intermediate states $k$ yields the probability that the process will be in state $j$ after $n+m$ transitions. Formally, we have

$$\begin{aligned} P_{ij}^{n+m} &= P\{X_{n+m} = j \mid X_0 = i\} \\ &= \sum_{k=0}^{\infty} P\{X_{n+m} = j, X_n = k \mid X_0 = i\} \\ &= \sum_{k=0}^{\infty} P\{X_{n+m} = j \mid X_n = k, X_0 = i\} P\{X_n = k \mid X_0 = i\} \\ &= \sum_{k=0}^{\infty} P_{kj}^m P_{ik}^n \ . \end{aligned}$$

(Ross, 2007, p.189)

Let $\boldsymbol{P}^{(n)}$ be the $n$-step transition probability matrix and $\boldsymbol{P}^{(m)}$ be the $m$-step transition probability matrix, then according to the Chapman-Kolmogorov equations,

$$\boldsymbol{P}^{(n+m)} = \boldsymbol{P}^{(n)} . \boldsymbol{P}^{(m)} \tag{2.6}$$

and where, the dot represents the matrix multiplication.

*2.2.1.2 Classification of States*

The classification of states according to their transition properties is important in a Markov chain since it gives the ergodicity conditions. State $j$ is said to be *accessible* from state $i$ if when the Markov chain is in state $i$, it is possible to reach state $j$ in some number of steps. It means that $P_{ij}^{(n)} > 0$ for some $n \geq 0$. If two states $i$ and $j$ are accessible to each other, it is said that they are to *communicate*. This communication is indicated with $i \leftrightarrow j$. It should be known that any state communicates with itself because, by definition,

$$P_{ii}^0 = P\{X_0 = i \mid X_0 = i\} = 1$$

For two state that communicates with each other, the equivalence relation of communication satisfies some properties:

- Any state $i$ communicates with itself, $i \leftrightarrow i$ for each state $i$. (*reflexivity* property)
- If $i$ communicates with $j\,(i \leftrightarrow j)$, $j$ communicates with $i$ too $(j \leftrightarrow i)$. (*symmetry* property)
- If $i$ communicates with $j$ $(i \leftrightarrow j)$ and $j$ communicates with $k\,(j \leftrightarrow k)$, $i$ communicates with $k$ too $(i \leftrightarrow k)$. (*transitivity* property)

In a Markov chain, the accessibility relation divides states into some classes. All states communicates with each other within each class. If two states are in different classes, they do not communicate with each other even if a one-sided accessibility exists. That is, state $i$ and state $j$ are in the same class if and only if they communicate $(i \leftrightarrow j)$. If a Markov chain has only one class, in other words, all states are reachable with each other, the chain is said to be *irreducible.*

Let $f_i^{(n)}$ be the probability that when leaving state $i$, the first transition to itself occurs after $n$-steps. That is,

$$f_i^n = P[X_n = i, X_r \neq i \; for \; r = 1,2,\ldots,n-1 \mid X_0 = i]. \tag{2.7}$$

"For any state $i$ we let $f_i$ denote the probability that, starting in state $i$, the process will ever reenter state $i$" (Ross, 2007, p. 195). That is,

$$f_i = \sum_{n=1}^{\infty} f_i^{(n)} . \tag{2.8}$$

If $f_i < 1$, state $i$ is said to be *transient* and if $f_i = 1$, state $i$ is said to be *recurrent*. State $i$ is recurrent, if and only if turning back to state $i$ is certain. So, it is possible to find the mean recurrence time of state $i$. That is,

$$\mu_i = \sum_{n=1}^{\infty} n f_i^{(n)} . \tag{2.9}$$

It is said that $i$ is *recurrent null* if $\mu_i = \infty$ and *positive recurrent* or *recurrent nonnull* if $\mu_i < \infty$.

State $i$ is called *absorbing* if $P_{ii} = 1$. It means that if the process makes a transition to state $i$, it is certain to stay at state $i$. It does not make any transitions after making transition to state $i$. If state $i$ is an absorbing state, it is a class all by itself.

"Some Markov chain may exhibit periodicity in their behavior. The period of a state $i$ is defined as the greatest common divisor of all integer $n \geq 1$ for which $P_{ij}^{(n)} > 0$. If $P_{ii}^{(n)} = 0$ for all $n \geq 1$ define $d(i) = 0$. If $d(i) > 1$, state $i$ is said to be *periodic,* while if $d(i) = 1$, then $i$ is an *aperiodic* state." (Uzunoğlu, 2001, p. 24)

When a Markov chain is an aperiodic, irreducible and positive recurrent, it is known as *ergodic*.

*2.2.1.3 Limiting Probabilities*

Let $\boldsymbol{P}$ is a transition matrix of an aperiodic, irreducible and $m$-state Markov chain with state space $S$. The probability of being in state $j$ after a large number of transitions is defined as the *limiting probability.* In other words, when number of transitions $n \to \infty$, $P^{(n)}$ converges to a fixed value and these probabilities are independent of the initial state $(X_0 = i)$.

The probability of being in state $j$ after a large number of steps is

$$\lim_{n \to \infty} P_{ij}^{(n)} = \pi_j \tag{2.10}$$

It does not matter where the chain at the beginning. Let $\{X_n, n = 0,1,\ldots\}$ is a Markov chain with $S = \{1,2,\ldots s\}$ and $\pi_i$ be the probability of being in state $i$ in a long term. Then the following is the matrix of limiting probabilities,

$$\lim_{n\to\infty} P^{(n)} = \begin{bmatrix} \pi_1 & \pi_2 & \cdots \pi_s \\ \pi_1 & \pi_2 & \cdots \pi_s \\ \vdots & \vdots & \vdots \\ \pi_1 & \pi_2 & \cdots \pi_s \end{bmatrix}$$

and these probabilities can be shown with a $1\times s$ vector:

$$\pi = [\pi_1 \quad \pi_2 \ \cdots \pi_s]. \tag{2.11}$$

It is possible to compute these limiting probabilities. In an irreducible Markov chain with ergodic states, the limiting probabilities satisfies the following equations:

$$\pi_j = \sum_{i\in S} \pi_i P_{ij} \tag{2.12}$$

and

$$\sum_{j\in S} \pi_j = 1. \tag{2.13}$$

### *2.2.2 Continuous-Time Markov Chains (CTMC)*

A CTMC is a stochastic process that makes transitions between states in accordance with a Markov chain, but the distribution of time that the process spends in each state before moving to another state is *exponentially* distributed. The amount of time that the process spends in a state and the next transition must be independent random variables.

Let $\{X(t), t \geq 0\}$ be a continuous-time stochastic process which takes on values of a nonnegative-integer set. The process is a continuous-time Markov chain if it satisfies the Markov property for all $s, t \geq 0$ and nonnegative integers $i, j, x(u), 0 \leq u < s$, that is

$$\begin{aligned} &P\{X(t+s) = j \mid X(s) = i, X(u) = x(u), 0 \leq u < s\} \\ &= P\{X(t+s) = j \mid X(s) = i\} = P_{ij}(t) \end{aligned} \tag{2.14}$$

“In other words, a continuous-time Markov chain is a stochastic process having the Markovian property that the conditional distribution of the future $X(t+s)$ given the present $X(s)$ and the past $X(u)$, $0 \le u < s$, depends only on the present and is independent of the past." (Ross, 2007, p. 366)

Continuous-time Markov chain has *stationary* or *homogeneous* transition probabilities if,

$$P\{X(t+s) = j \mid X(s) = i\}$$

is independent of s.

Let $T_i$ be amount of time that process stays in state $i$ before moving to another state. Because the amount of time that the process spends in any state before making a transition follows exponential distribution, $T_i$ has the memoryless property. That is,

$$P\{T_i > s+t \mid T_i > s\} = P\{T_i > t\}. \tag{2.15}$$

**2.3 Semi-Markov Process**

Semi-Markov process is a stochastic process which makes transitions in accordance with a Markov chain, but the amount of time that spent in states before moving to another state is random. The process moves frome state $i$ to state $j$ with probability $P_{ij}$ and given that next transition is to state $j$ and the time spent until moving to state $j$ has a distribution $F_{ij}$. As it is said in chapter one, the time spent in a state before moving to an another state is called *sojourn time*. Transitions are related to a Markov chain like it is said before so, the state space $S$ is denoted by $S = \{0,1,2,3,...\}$.

Let the random variables $T_n$ and $S_n$ represents the sojourn times between transitions and the transition epochs for semi Markov process, respectively. So,

$$S_n = \sum_{i=0}^{n} T_n\ . \tag{2.16}$$

for $n \ge 0$ and $0 = S_0 \le S_1 \le S_2 \le \dots$.

$Q_{ij}(t)$ denotes the probability that after the process makes a transition to state $i$, next transition is to state $j$ in time less than or equal to $t$ (Ross, 1996).

$$Q_{ij}(t) = P[X_{n+1} = j, S_{n+1} - S_n \leq t \mid X_n = i] \tag{2.17}$$

and

$$\sum_{j=0}^{\infty} Q_{ij}(\infty) = 1, \qquad j = 0,1,2,... \tag{2.18}$$

and we must have,

$$Q_{ij}(t) \geq 0, \qquad Q_{ij}(0) = 0, \qquad i, j = 0,1,2,,\ldots, \qquad t \geq 0\,.$$

Then $Q = \{Q_{ij}(t),\ i, j \in S, t \geq 0\}$ is called *semi-Markov kernel* for the semi-Markov process.

There is an important relationship between the probabilities $P_{ij}$ and $Q_{ij}(t)$. If the time indicated with $t$ converges to infinity, $Q_{ij}(t)$ converges to $P_{ij}$ at the same time. That is,

$$\lim_{t\to\infty} Q_{ij}(t) = Q_{ij}(\infty) = P_{ij}\,. \tag{2.19}$$

When the process is in state $i$, it makes the next transition to state $j$ according to Markov chain with a probability $P_{ij}$. Besides, according to the definition of semi-Markov process, distribution of sojourn times are random so let "$F_{ij}^{(t)}$ represents the conditional probability that a transition will take place within an amount of time $t$, given that the process has just entered $i$ and will next enter $j$." (Ross, 1970, p. 85) That is,

$$F_{ij}(t) = P[S_{n+1} - S_n \leq t \mid X_n = i, X_{n+1} = j] \tag{2.20}$$

The cumulative distribution function of sojourn times $F_{ij}^{(t)}$ can be obtained by,

$$F_{ij}(t) = \frac{Q_{ij}(t)}{P_{ij}} \tag{2.21}$$

and it should not be forgotten that if $P_{ij} = 0$, let $F_{ij}^{(t)}$ be arbitrary.

Let $N_i(t)$ is the number of transitions in state $i$ that occur in time interval $(0,t]$ and $N(t)$ is the total number of transitions along the process. So,

$$N(t) = \sum_{i=0}^{\infty} N_i(t)\,. \tag{2.22}$$

Let $N(t) = (N_1(t), N_2(t), \ldots)$ be the vector process then, the stochastic process $\{N(t), t \geq 0\}$ is called *Markov renewal process*.

If $Z(t) = X_{N(t)}$ denotes the state of the process at time $t$, then $\{Z(t), t \geq 0\}$ is a stochastic process and it is called *semi-Markov process*.

When we consider the definitions of semi-Markov process and Markov renewal process, it can be said that "the semi-Markov process records the state of the process at each time point, while the Markov renewal process is a counting process which keeps track of the number of times each state has been visited." (Ross, 1970, p. 86)

The semi-Markov process waits for a time in a state before making a transition to another state. Since these sojourn times are random, they have a distribution. Let $H_i$ be the distribution of sojourn times in state $i$ before making a transition. In other words, when the semi-Markov process moves to state $i$, it spends some time before moving to another state and $H_i$ is the distribution of these sojourn times in state $i$. That is,

$$H_i(t) = \sum_{j=0}^{\infty} P_{ij} F_{ij}(t) = \sum_{j=0}^{\infty} Q_{ij}(t) \tag{2.23}$$

and it is supposed that $H_i(0) < 1$ for all $i$.

Let $\eta_{ij}$ denote the expected amount of time spent in state $i$, given that the next transition is to state $j$. That is,

$$\eta_{ij} = \int_0^{\infty} t\, dF_{ij}(t) \tag{2.24}$$

and let $m_i$ denote the expected sojourn time in state $i$ during each visit. So, $m_i$ is the mean of the distribution $H_i$. Then,

$$m_i = \int_0^{\infty} t\, dH_i(t)\,. \tag{2.25}$$

It is possible to replace $H_i(t)$ in equation (2.25) with equation (2.23). So, $m_i$ can be written in a different form. That is,

$$\begin{aligned}
m_i &= \int_0^{\infty} t\, dH_i(t) \\
&= \int_0^{\infty} t\, d\left[\sum_{j=0}^{\infty} P_{ij} F_{ij}(t)\right] \\
&= \int_0^{\infty} t\, d\left[P_{i0}F_{i0}(t) + P_{i1}F_{i1}(t) + P_{i2}F_{i2}(t) + \ldots\right] \\
&= \int_0^{\infty} t\, d\left[P_{i0}F_{i0}(t)\right] + \int_0^{\infty} t\, d\left[P_{i1}F_{i1}(t)\right] + \int_0^{\infty} t\, d\left[P_{i2}F_{i2}(t)\right] + \ldots
\end{aligned}$$

and it is known that transition probabilities of the Markov chain are constants so,

$$m_i = P_{i0}\int_0^{\infty} t\, dF_{i0}(t) + P_{i1}\int_0^{\infty} t\, dF_{i1}(t) + P_{i2}\int_0^{\infty} t\, dF_{i2}(t) + \ldots.$$

by the equation (2.23),

$$m_i = P_{i0}\eta_{i0} + P_{i1}\eta_{i1} + P_{i2}\eta_{i2} + \ldots.$$

and it follows that,

$$m_i = \sum_{j=0}^{\infty} P_{ij}\eta_{ij}\,. \tag{2.26}$$

For a semi-semi Markov process, state $i$ is said to be *regular* if

$$P\{N(t) = \infty \mid X_0 = i\} = 0 \tag{2.27}$$

and when all states are regular, Markov renewal process is said to be regular. So, the number of transitions are finite, $N(t) < \infty$ for all $t \geq 0$.

The equation (2.20) can be written differently by cross-multiplying. That is,

$$Q_{ij}(t) = P_{ij}\, F_{ij}(t)\,. \tag{2.28}$$

Then, it is possible to show the relationship of discrete-time and continuous-time Markov chain with semi-Markov process. It is known that when all transition times are identically 1, the semi-Markov process is a discrete-time Markov chain and when the sojourn times of the semi-Markov process follow exponential distribution, the semi-Markov process is a continuous-time Markov process. That is,

$$Q_{ij}(t) = P_{ij}\, I(t \geq 1) \tag{2.29}$$

for discrete-time Markov chain and,

$$Q_{ij}(t) = P_{ij}\,(1 - e^{-\lambda_i t}) \tag{2.30}$$

for continuous-time Markov chain.

### *2.3.1 Classification of States*

Let us define $G_{ij}(\cdot)$ which is the distribution function of time until first transition occurs into $j$, given that the stochastic process starts in state $i$. However, let it be known that $G_{ij}(\infty)$ denotes the probability that a Markov renewal process started in state $i$ will ever move into state $j$. Then,

$$G_{ij}(t) = P\{N_j(t) > 0 \mid Z(0) = i\} \tag{2.31}$$

and

$$P_{ij}(t) = P\{Z(t) = j \mid Z(0) = i\} \tag{2.32}$$

for all $i, j$ and $t \geq 0$.

$\mu_{ij}$ denotes the first moments of the distribution $G_{ij}(t)$ and it is obtained by

$$\mu_{ij} = \begin{cases} \infty & \text{if } G_{ij}(\infty) < 1 \\ \int_0^\infty t\, dG_{ij}(t) & \text{otherwise.} \end{cases} \tag{2.33}$$

If we have the moments, it is possible to obtain $\mu_{ii}$ which denotes the *mean recurrence time* of state $i$.

- State $i$ and state $j$ are to *communicate* when $i = j$ or $G_{ij}(\infty) \cdot G_{ij}(\infty) > 0$.
- There is an equivalence communication between states and the disjoint equivalence classes are called *classes.*

- A semi-Markov process is said to be *irreducible* if there is only one class in the process.
- If $G_{ii}(\infty)=1$, $i$ is said to be *recurrent* and *transient* for other $G_{ii}(\infty)$. Besides, state $i$ is recurrent, if and only if it is recurrent in Markov chain.
- When $i$ is recurrent and $\mu_{ii}<\infty$, it is said to be *positive recurrent*. For a finite state space, state $i$ is positive recurrent, if and only if it is positive recurrent in the Markov chain and $m_i<\infty$.
- When $i$ is recurrent and $\mu_{ii}=\infty$, it is said to be *null recurrent*.

There is an other way to show the recurrence of $i$. When $m_i<\infty$, $i$ is recurrent if and only if, either

$$\int_0^\infty P_{ii}(t)dt = m_i + m_i E[N_i(\infty)] = \infty \tag{2.34}$$

or

$$E[N_i(\infty)] = \infty . \tag{2.35}$$

If state $i$ is said to be transient, the random variable $N_i(\infty)$ follows geometric distribution with mean

$$\frac{G_{ii}(\infty)}{1-G_{ii}(\infty)} . \tag{2.36}$$

So, from (2.33)

$$\int_0^\infty P_{ii}(t)dt = m_i + m_i \frac{G_{ii}(\infty)}{1-G_{ii}(\infty)} \tag{2.37}$$

and if we look at (2.37), it is clear that when $m_i=\infty$ or state $i$ is recurrent, (2.37) is infinite.

### *2.3.2 Stationary Distribution Probabilities*

For an irreducible semi-Markov process if $\mu_{jj}<\infty$, then,

$$P_j = \lim_{t\to\infty} P_{ij}(t) = \lim_{t\to\infty} P\{Z(t)=j \mid Z(0)=i\}$$

exists and as we know from the limiting probabilities of Markov chain, it is independent of the initial state of the process, and

$$P_j = \frac{m_j}{\mu_{jj}}. \tag{2.38}$$

Furthermore, again for an irreducible semi-Markov process if $\mu_{jj} < \infty$, then

$$\frac{m_j}{\mu_{jj}} = \lim_{t\to\infty} \frac{\text{amount of time in } j \text{ during } [0,t]}{t} \tag{2.39}$$

with probability 1. So, $m_j / \mu_{jj}$ is said to be equal to the long-run proportion of time in $i$.

Suppose that we have an irreducible and positive recurrent Markov chain. Let $\pi_i, i \geq 0$ be the limiting probabilities of the Markov chain. It is known that $\pi_i$ is the proportion of transitions into state $i$ and $m_i$ is the expected sojourn time in state $i$ during each visit. Stationary distribution probabilities are proportional to $\pi_j m_j$ for the semi-Markov process. Then,

$$P_j = \frac{\pi_j m_j}{\sum_{i=0}^{\infty} \pi_i m_i}. \tag{2.40}$$

It seems that the stationary distribution probabilities are dependent to transition probabilities $P_{ij}$ and expected sojourn times $m_i$ for $i, j \geq 0$.

# CHAPTER THREE
# METHODS USED FOR THE ESTIMATION OF SOJOURN TIMES FOR SEMI-MARKOV PROCESS

Semi-Markov process is introduced in detail in chapter two and from now on we have all necessary informations about the process especially distribution of sojourn times. In this chapter, two methods is explained for the estimation of sojourn times. As we know, estimation methods may differ according to the type of data set. The first method used is EM-algorithm. It is used for the estimation of parameters based on likelihood function when there exists some of missing, censored or inappropriate values in data set. The second method used is an empirical estimator for sojourn times. Then, these two methods is applied to a real life data set in the next chapter.

## 3.1 Expectation-Maximization (EM) Algorithm

EM-algorithm is a general iterative algorithm for parameter estimation using maximum likelihood approach when some of the variables are not observed or inappropriate such as missing or censored. It is performed two steps in each iteration. Algorithm starts with *E-step* which finds the expectation of the log-likelihood using the current estimates of the parameters. It follows with *M-step* which finds parameters by maximizing the expectation of log-likelihood found at the E-step. These two steps are repeated until convergence.

If the related data set is an incomplete data, this means that there exists two sample spaces. One of them is *observed data* and the other is *unobserved data*.

Let $\underline{y} = (y_1, \ldots, y_p)$ be a random vector and $f(y;\theta)$ is the joint probability density function of $\underline{y}$ where $\theta = (\theta_1, \ldots, \theta_p)$ is the vector of unknown parameters. If complete-data vector $\underline{y}$ is observed, then it is enough to compute the maximum likelihood estimate of $\theta$ based on the distribution of $\underline{y}$. The log-likelihood function

$$\log L(\theta; y) = \ell(\theta; y) = \log f(y;\theta) \tag{3.1}$$

is maximized.

If there are some missing values, only a function of the complete-data $\underline{y}$ is observed. The observed data is denoted by $y_{obs}$ but it should not be forgotten that it is an observed but an incomplete data. The unobserved or missing data is denoted by $y_{mis}$ and it is assumed that missing data are missing at random. So, it is possible to express $\underline{y}$ as $(y_{obs}, y_{mis})$ and then,

$$\begin{aligned} f(y;\theta) &= f(y_{obs}, y_{mis};\theta) \\ &= f_1(y_{obs};\theta)\,.\,f_2(y_{mis} \mid y_{obs};\theta), \end{aligned} \tag{3.2}$$

where $f_1$ and $f_2$ are the joint densities of $y_{obs}$ and $y_{mis}$, respectively. As it is seen from the $f_2$, $y_{mis}$ has a conditional probability density given the data $y_{obs}$.

Then, the log-likelihood function of complete-data $\underline{y}$ is,

$$\ell(\theta; y) = \ell_{obs}(\theta; y_{obs}) + \log f_2(y_{mis} \mid y_{obs};\theta) \tag{3.3}$$

and so, the log-likelihood function of the observed-data $y_{obs}$ is,

$$\ell_{obs}(\theta; y_{obs}) = \ell(\theta; y) - \log f_2(y_{mis} \mid y_{obs};\theta)\,. \tag{3.4}$$

If maximizing $\ell_{obs}$ is difficult but maximizing the complete-data log-likelihood $\ell(\theta; y)$ is easier, EM algorithm is very useful and easy to apply. But, we know that $\underline{y}$ contains unobserved values so, it is not possible to evaluate and to maximize $\ell$. EM algorithm is used to maximize this $\ell(\theta; y)$ in accordance with its intuitive idea. $\ell(\theta; y)$ is replaced with its conditional expectation given $y_{obs}$. It should not be forgotten that this conditional expectation is computed according to the distribution of $\underline{y}$ updated at the current estimate of $\theta$.

Let $\theta^{(0)}$ be an initial value for $\theta$. It is required in the first iteration of the algorithm to compute the conditional expectation of $\ell(\theta; y)$. That is,

$$Q(\theta;\theta^{(0)}) = E_{\theta^{(0)}}\left[\ell(\theta; y) \mid y_{obs}\right]. \tag{3.5}$$

After computing this expectation, $Q(\theta;\theta^{(0)})$ is maximized with respect to the parameters $\theta$ and updated estimates are found such that

$$Q(\theta^{(1)};\theta^{(0)}) \geq Q(\theta;\theta^{(0)})\,. \tag{3.6}$$

for all $\theta$.

So, two steps follow each other iteratively until convergence. It should be computed $Q(\theta;\theta^{(t)})$ in the E-step. That is,

$$Q(\theta;\theta^{(t)}) = E_{\theta^{(t)}}\left[\ell(\theta; y \mid y_{obs}\right]. \tag{3.7}$$

Then, it follows by the M-step which $\theta^{(t+1)}$ is found such that

$$Q(\theta^{(t+1)};\theta^{(t)}) \geq Q(\theta;\theta^{(t)}) \tag{3.8}$$

for all $\theta$. These two steps are repeated until the difference of estimates $|\hat{\theta}^{(t+1)} - \hat{\theta}^{(t)}|$ is less than or equal to $\varepsilon$, where $\varepsilon$ is a proper small quantity.

If the complete-data $\underline{y}$ is belonged to the exponential family, E-step and M-step are simplified and become easier to use. That means because of belonging to the exponential family, the log-likelihood function is linear in the sufficient statistic for $\theta$. Then, it is sufficient for us to compute the expectation of the complete-data sufficient statistic given $y_{obs}$. Besides, instead of maximizing the expected log-likelihood, when $\underline{y}$ is from exponential family, the conditional expectations of the sufficient statistics can be directly substituted with the sufficient statistics which is in the expressions obtained for the complete-data maximum likelihood estimators of $\theta$ to obtain the next iterate.

Some properties of EM algorithm:

- It is easily applied because it relies on complete-data computations. That is, E-step involves taking expectations over complete-data conditional distributions and M-step only requires complete-data maximum likelihood estimation.
- EM algorithm is a numerically stable algorithm.

- Each iteration needs to increase the log-likelihood $\ell(\theta; y_{obs})$ and if $\ell(\theta; y_{obs})$ is bounded, $\ell(\theta^{(t)}; y_{obs})$ converges to a stationary point.
- If $\theta^{(t)}$ converges, it does so to a local maximum of $\ell(\theta; y_{obs})$ and to the unique maximum likelihood estimators if $\ell(\theta; y_{obs})$ is unimodal.
- If there are lots of missing data, the convergence rate of EM algorithm can be extremely slow.

### *3.1.1 Log-normal Example*

Let the complete-data $y$ be a random sample from log-$N(\mu, \sigma^2)$. The probability density function of log-normal distribution is

$$f(y \mid \mu, \sigma^2) = \frac{1}{y\sigma\sqrt{2\pi}} e^{-\frac{(\ln y - \mu)^2}{2\sigma^2}} \quad ; \quad y > 0. \tag{3.9}$$

Then, the likelihood function is

$$\begin{aligned} L(\mu, \sigma^2; y_1, y_2, \ldots, y_n) &= \prod_{i=1}^{n} \frac{1}{y_i\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln y_i - \mu}{\sigma}\right)^2} \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \left(\prod_{i=1}^{n} \frac{1}{y_i}\right) e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(\ln y_i - \mu)^2} \end{aligned} \tag{3.10}$$

and the log-likelihood function is

$$\begin{aligned} &\ln L(\mu, \sigma^2; y_1, y_2, \ldots, y_n) \\ &= -\frac{n}{2}\left(\ln 2\pi + \ln \sigma^2\right) - \sum_{i=1}^{n} \ln y_i - \frac{1}{2\sigma^2} \sum_{i=1}^{n} \left[(\ln y_i)^2 - 2\mu \ln y_i + \mu^2\right] \end{aligned} \tag{3.11}$$

To obtain the maximum likelihood estimates, we need to differentiate (3.11) with respect to the parameters $\mu$ and $\sigma^2$, then equal these derivations to zero.

$$\frac{\partial \ln L(\mu, \sigma^2; y_1, y_2, \ldots, y_n)}{\partial \mu} = \frac{\sum_{i=1}^{n} (\ln y_i - \mu)^2}{\sigma^2} = 0$$

$$\Rightarrow \hat{\mu} = \frac{\sum_{i=1}^{n} \ln y_i}{n} \tag{3.12}$$

and

$$\frac{\partial \ln L(\mu,\sigma^2; y_1, y_2, \ldots, y_n)}{\partial \sigma^2} = -\sigma^2 n + \sum_{i=1}^{n} (\ln y_i - \mu)^2 = 0$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (\ln y_i - \hat{\mu})^2}{n} = \frac{\sum_{i=1}^{n} (\ln y_i)^2}{n} - \left[ \frac{\sum_{i=1}^{n} \ln y_i}{n} \right]^2. \tag{3.13}$$

"Let $X_1 = k_1, \ldots, X_n = k_n$ be a random sample of size $n$ from $p_X(k;\theta)$. The statistic $\hat{\theta} = h(X_1, \ldots, X_n)$ is *sufficient* for $\theta$ if the likelihood function, $L(\theta)$, factors into the product of the pdf for $\hat{\theta}$ and a constant that does not involve $\theta$-that is, if

$$L(\theta) = \prod_{i=1}^{n} p_X(k_i;\theta) = p_{\hat{\theta}}(\theta_e;\theta) b(k_1, \ldots, k_n).$$

A similar statement holds if the data consist of a random sample $Y_1 = y_1, \ldots, Y_n = y_n$ drawn from a continuous pdf $f_Y(y;\theta)$." (Larsen, & Marx, 2012, p. 326)

For log-normal distribution, let us show the sufficient statistics for $\theta = (\mu, \sigma^2)$. We know that the likelihood function for log-normal distribution is

$$L(\mu,\sigma^2;y_1,y_2,\ldots,y_n) = \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \left( \prod_{i=1}^{n} \frac{1}{y_i} \right) e^{-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\ln y_i - \mu)^2}$$

and the expression in the exponential of likelihood function can be written as

$$-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (\ln y_i)^2 + \frac{\mu}{\sigma^2} \sum_{i=1}^{n} \ln y_i - \frac{n\mu^2}{2\sigma^2}.$$

It is seen that first factor involves parameters but no $y_i$'s, second factor involves $y_i$'s but no parameters so, the sufficient statistics must come from the exponential term in the third factor. Besides, the second and third terms are functions that depend on $y_i$ only through a function. In other words, $L(\theta)$ is factored into the product of the pdf for $\hat{\theta}$ and a constant that does not involve $\theta$. If we look at the likelihood function, it is seen that $\sum_{i=1}^{n} \ln y_i$ and $\sum_{i=1}^{n} (\ln y_i)^2$ are the sufficient statistics for

$\theta = (\mu, \sigma^2)$ and the log-likelihood function of complete-data is linear in sufficient statistics.

Suppose that $y_i, i = 1, \ldots m$ are the observed values and $y_i, i = m+1, \ldots, n$ are the random missing values and it should not be forgotten that $y_i$ are assumed to be i.i.d. $\log - N(\mu, \sigma^2)$. The observed-data is denoted by $y_{obs} = (y_1, y_2, \ldots, y_n)$.

The E-step requires the computations of conditional expectations of sufficient statistics given $y_{obs}$ because the complete-data $\underline{y}$ is from exponential family. That is,

$$E_\theta\left(\sum_{i=1}^{n} \ln y_i \mid y_{obs}\right) \text{ and } E_\theta\left(\sum_{i=1}^{n} (\ln y_i)^2 \mid y_{obs}\right).$$

So, at the $t^{\text{th}}$ iteration of the E-step, we need to compute

$$\begin{aligned} h_1^{(t)} &= E_{\mu^{(t)}, \sigma^{2(t)}}\left(\sum_{i=1}^{n} \ln y_i \mid y_{obs}\right) \\ &= \sum_{i=1}^{m} \ln y_i + (n-m)\mu^{(t)} \end{aligned} \tag{3.14}$$

because $E_{\mu^{(t)}, \sigma^{2(t)}}(\ln y_i) = \mu^{(t)}$ where $\mu^{(t)}$ and $\sigma^{2(t)}$ are the current estimators for $\mu$ and $\sigma^2$, and

$$\begin{aligned} h_2^{(t)} &= E_{\mu^{(t)}, \sigma^{2(t)}}\left(\sum_{i=1}^{n} (\ln y_i)^2 \mid y_{obs}\right) \\ &= \sum_{i=1}^{m} (\ln y_i)^2 + (n-m)\left[{\sigma^{(t)}}^2 + {\mu^{(t)}}^2\right] \end{aligned} \tag{3.15}$$

because $E_{\mu^{(t)}, \sigma^{2(t)}}[(\ln y_i)^2] = \sigma^{2(t)} + {\mu^{(t)}}^2$.

The M-step requires to substitute the conditional expectations of complete-data sufficient statistics computed in the E-step with the sufficient statistics which are in the maximum likelihood estimators (3.12) and (3.13) to obtain the updated estimators for new iterates. That is,

$$\mu^{(t+1)} = \frac{h_1^{(t)}}{n} \tag{3.16}$$

and

$$\sigma^{2(t+1)} = \frac{h_2^{(t)}}{n} - \mu^{(t+1)2}. \tag{3.17}$$

Let $\mu^{(0)}$ and $\sigma^{2^{(0)}}$ be starting values to the algorithm. Thus, the EM-algorithm iterates properly between (3.14) and (3.15) and (3.16) and (3.17) to obtain new estimates such that $(\mu^{(1)}, \sigma^{2^{(1)}}), (\mu^{(2)}, \sigma^{2^{(2)}}), \ldots$. Iterations are carried on until the difference of estimates $|\hat{\theta}^{(t+1)} - \hat{\theta}^{(t)}|$ is less than or equal to $\varepsilon$, where $\varepsilon$ is a predetermined small quantity.

### 3.2 Empirical Estimators

In this chapter, some empirical estimators proposed by Limnios, Ouhbi and Sadek (2005) for stationary distribution and mean sojourn times of semi-Markov process is studied. These estimators is easy to use on data sets because calculations makes it very easy to obtain estimates for parameters.

Assume that the Markov chain $\{X_n\}$ is irreducible. Let $\pi_i, i \geq 0$ be the limiting probabilities and $m_i < \infty$ be the mean sojourn time into state $i$.

Suppose that we have an irreducible and positive recurrent Markov chain. Let $\pi_i, i \geq 0$ be the limiting probabilities of the Markov chain. It is known that $\pi_i$ is the proportion of transitions into state $i$ and $m_i$ is the expected sojourn time in state $i$. Then, we know that the stationary distribution probabilities for semi-Markov process are

$$P_j = \frac{\pi_j m_j}{\sum_{i=0}^{\infty} \pi_i m_i}.$$

The estimator proposed for the stationary distribution of semi-Markov process is

$$\hat{P}_j = \frac{\hat{\pi}_j \hat{m}_j}{\sum_{i=0}^{\infty} \hat{\pi}_i \hat{m}_i} \tag{3.18}$$

where $\hat{\pi}_j$ is an estimator for the limiting probabilities of Markov chain and $\hat{m}_j$ is an estimator for the mean sojourn time into state $j$ (Limnios, Ouhbi, & Sadek, 2005). That is,

$$\hat{\pi}_j = \frac{N_j(t)}{N(t)} \tag{3.19}$$

for all $i$, where $N_i(t)$ is the number of transitions in state $i$ that occur in time interval $(0,t]$ and $N(t)$ is the number of transitions along the process and

$$\hat{m}_j = \frac{1}{N_j(t)} \sum_{k=1}^{N_j(t)} T_{jk} \tag{3.20}$$

where $T_{jk}$ is the *k*th sojourn time in state $j$. Note that if $N_i(t) = 0$, $\hat{m}_j$ and $\hat{\pi}_j$ are equals to zero.

Asymptotic properties of these two estimators are given in the paper (Limnios, Ouhbi, & Sadek, 2005). It is shown that they are strongly consistent and unbiased estimators, as *t* tends to infinity.

# CHAPTER FOUR
# APPLICATION

In this chapter, a real life data set is characterized with semi-Markov process. The sojourn time distribution is estimated by two approaches. Mean sojourn times and stationary distributions for the data set is presented.

## 4.1 Problem Definition and Assumptions

Winning, drawing and losing are the possible results in a football match and a team plays in a certain amount of matches every year. Number of matches may change according to number of teams in the league and the organizations participated. We interested in the scores of the official football matches that Besiktas Gymnastics Club played. Data set consists of all the official match results and dates of professional football team Besiktas Gymnastics Club from 1959 to 2014. Data is obtained from the official website of Turkish Football Federation (n.d).

After playing a match, a team gets a score and waits for a length of time until next match. There are many factors that affects the success of a team in a football match. However, we assume that the result a team obtain from a football match is affected substantially by the result of the most previous match and how many days ago this result is obtained. The dates of the matches that a football club plays are not known in advance.

$Z(t) = X_{N(t)}$ denotes the result of the match played at time $t$, then $\{Z(t), t \geq 0\}$ is a semi-Markov process.

## 4.2 States of the Markov Chain

We know that there are three possible results at the end of a football match. So, Markov chain for the process has three states. States of the Markov chain are denoted by numbers. Losing a match is represented by *state 0,* drawing a match is represented by *state 1* and winning a match is represented by *state 2.* Then, the state space of the process is $S = \{0,1,2\}$.

After defining the states of the process, it is now time for explaining the number of transitions and the transition probabilities of the Markov chain for the semi-Markov process the process.

In this study, playing matches successively means making transitions from one state into another state. For instance, suppose that Besiktas wins a match and draws the next match they played. This means that the Markov chain makes transition from state 2 into state 1 or drawing after losing a match means making a transition from state 0 into state 1.

### 4.3 Transitions and Transition Probabilities

Besiktas has played 2373 official football matches from 1959 to 2014. They win 1230, draw 634 and lose 509 of these matches. So, there are 2372 transitions along the process. The transition matrix consists of transitions numbers from state $i$ into state $j$. That is,

$$\boldsymbol{N} = \begin{bmatrix} N_{00} & N_{01} & N_{02} \\ N_{10} & N_{11} & N_{12} \\ N_{20} & N_{21} & N_{22} \end{bmatrix} = \begin{bmatrix} 124 & 130 & 255 \\ 134 & 177 & 323 \\ 251 & 327 & 651 \end{bmatrix}.$$

Transition matrix gives us information about behavior of the process in successive matches. Here, $N_{ij}$ is the number of transitions when the process is in state *i* and the next transition is into state *j*. For instance, $N_{21} = 327$ means after Besiktas win a match, they draw at the next match for 327 times.

We know that $P(X_t = i)$ is the probability of being in state $i$ at time $t$ and

$$P_{ij} = P\{X_{t+1} = j \mid \mathrm{X}_{\mathrm{t}} = i\}$$

is the probability of making a transition into state $j$ at time $t+1$ given that the process is in state $i$ at time $t$. We can estimate one-step transition probabilities of the Markov chain by the following formula:

$$P_{ij} = \frac{N_{ij}}{\sum_j N_{ij}}. \tag{4.1}$$

Thus, one-step transition probability matrix of the Markov chain for semi-Markov process can be obtained. It should not be forgotten that row totals of the transition probability matrix must be 1. That is,

$$\boldsymbol{P} = \begin{bmatrix} P_{00} & P_{01} & P_{02} \\ P_{10} & P_{11} & P_{12} \\ P_{20} & P_{21} & P_{22} \end{bmatrix} = \begin{bmatrix} 0.2436 & 0.2554 & 0.5010 \\ 0.2113 & 0.2792 & 0.5095 \\ 0.2042 & 0.2661 & 0.5297 \end{bmatrix}.$$

### *4.3.1 Limiting Probabilities*

We know that the probability of being in state $i$ in the long term is called limiting probabilities and it is denoted by $\pi_i$. We have three states 0,1 and 2 so vector of probabilities is $\pi = [\pi_0, \pi_1, \pi_2]$. We also know that these limiting probabilities satisfy the equations (2.11) and (2.12). Then,

$$\begin{gathered} \pi_0 = 0.2436\pi_0 + 0.2113\pi_1 + 0.2042\pi_2 \\ \pi_1 = 0.2554\pi_0 + 0.2792\pi_1 + 0.2661\pi_2 \\ \pi_2 = 0.5010\pi_0 + 0.5095\pi_1 + 0.5297\pi_2 \\ \pi_0 + \pi_1 + \pi_2 = 1 \end{gathered}$$

is solved and limiting probabilities fo the Markov chain are

$$\pi = [0.2146, 0.2673, 0.5181].$$

## 4.4 Sojourn Times for Semi-Markov Process

Time spent in a state before making a transition to another state is called sojourn time. In this study, after getting a result $i$, $i \in S$, Besiktas waits for a time until the next result $j$, $j \in S$ which is called sojourn time. Sojourn times are examined in days.

Think about a season that matches are played. In general, teams start playing their matches at the end of summer and finish playing matches at the beginning of next years summer every season. So, they do not play any official matches along a summer, roundly 3 or 4 months. While sojourn time between two successive matches is usually about a week, not playing a match about 3 or 4 months is not a proper situation. Sojourn times between last match of a season and first match of next season is too higher than between two matches in league season. Besides, at the end of a season teams generally change their squads, sell and buy players. Maybe their goals change at the beginning of a season such that championship or title race. So, these sojourn times are assumed to be inappropriate values.

We can not say that these data sets of sojourn times follow a certain distribution because there are inappropriate values and these values are not considered in statistical analysis in other words assumed to be missing. But, histogram of *observed-data* help us identifying the distribution of complete-data so we can make comments about the distributions of sojourn times.

The following Figure 4.1, Figure 4.2 and Figure 4.3 are the histograms of observed-data sojourn times $(\mathrm{ST_i})$ when the process is in state $i$, for $i = 0,1,2$.

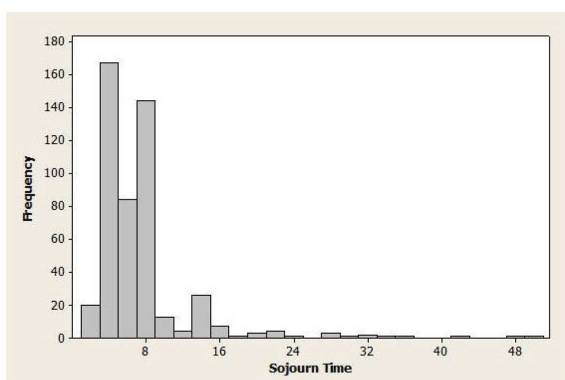


Figure 4.1 Histogram of observed-data $\mathrm{ST}_0$

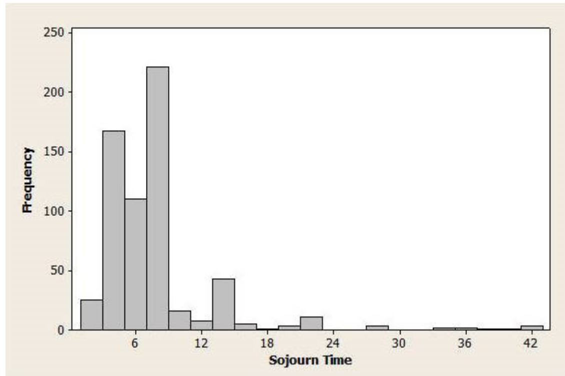


Figure 4.2 Histogram of observed-data $\mathrm{ST}_1$

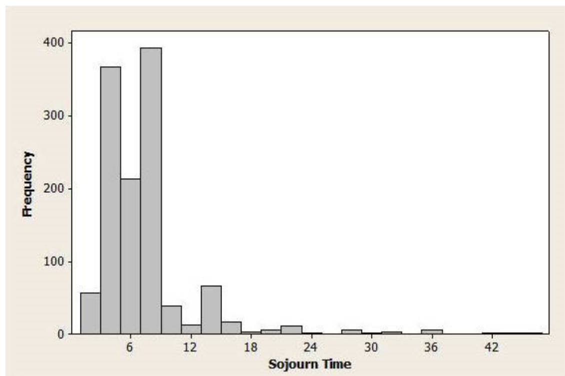


Figure 4.3 Histogram of observed-data $\mathrm{ST}_2$

The following Figure 4.4- 4.12 are the histograms of observed-data sojourn times given that the process has just entered $i$ and will next enter $j$ for $i, j = 0,1,2$ $(\mathrm{ST}_{ij})$.

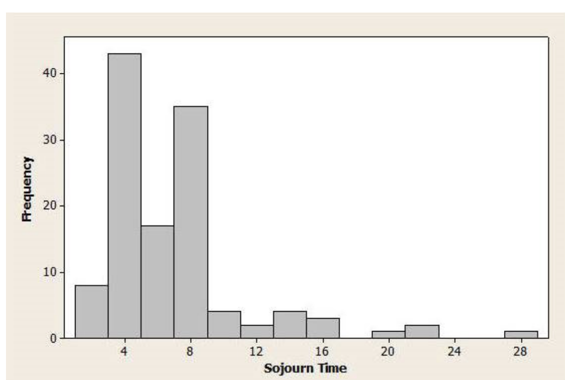


Figure 4.4 Histogram of observed-data $\mathrm{ST}_{00}$

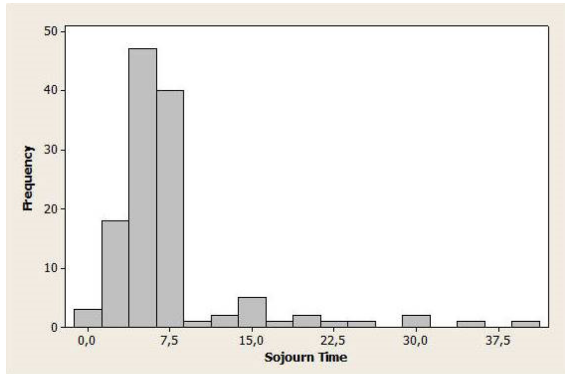


Figure 4.5 Histogram of observed-data $ST_{01}$

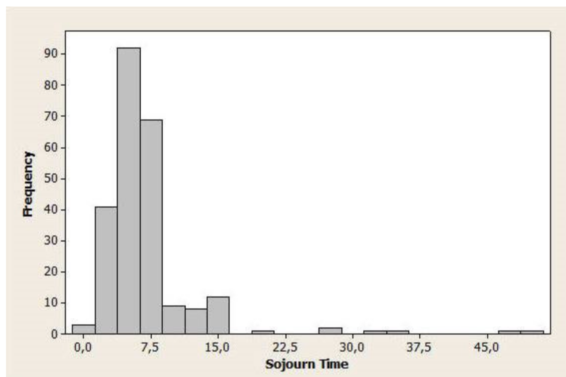


Figure 4.6 Histogram of observed-data $ST_{02}$

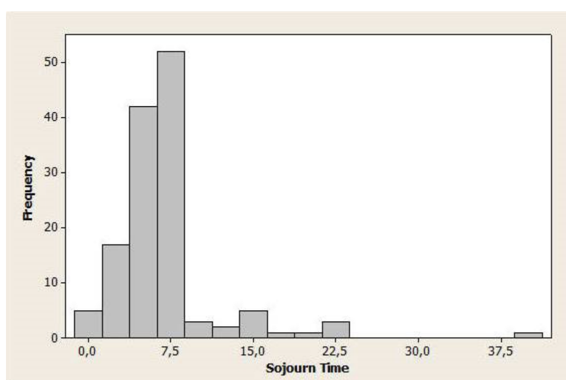


Figure 4.7 Histogram of observed-data $ST_{10}$

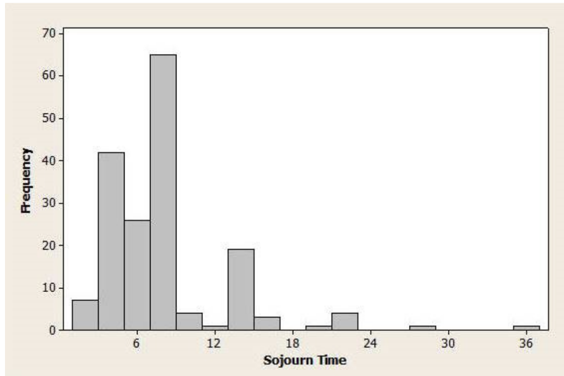


Figure 4.8 Histogram of observed-data $ST_{11}$

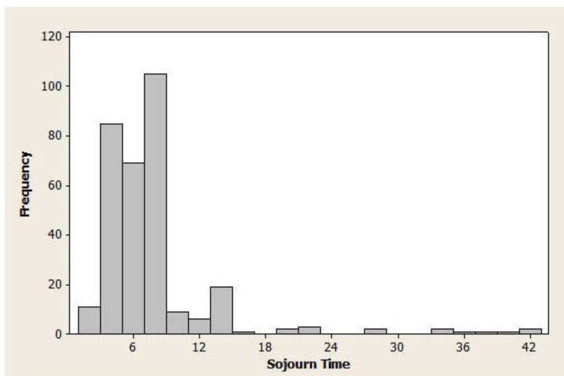


Figure 4.9 Histogram of observed-data $ST_{12}$

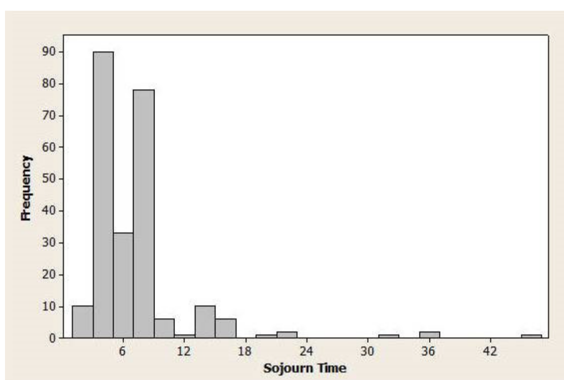


Figure 4.10 Histogram of observed-data $ST_{20}$

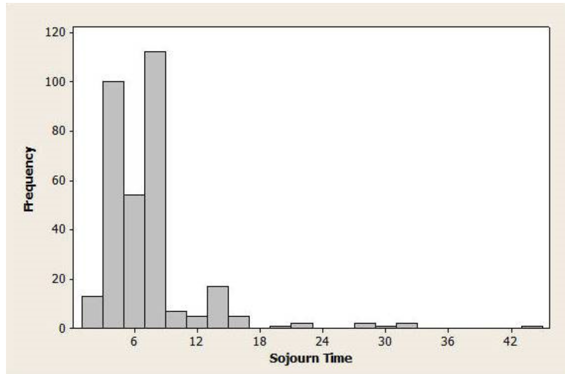


Figure 4.11 Histogram of observed-data $\mathrm{ST}_{21}$

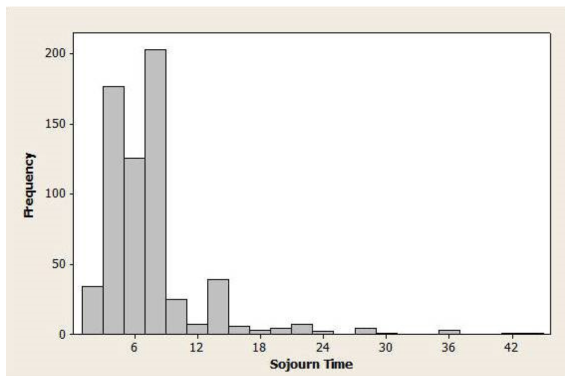


Figure 4.12 Histogram of observed-data $\mathrm{ST}_{22}$

It should not be forgotten that the figures are the histograms of observed-data sojourn times and they give us information about the complete-data and their distributions. It is seen from the figures that sojourn times for semi-Markov process most probably follow a heavy-tailed distribution. When we apply the goodness of fit tests, log-normal distribution is the best distribution for the sojourn times in the states. So, we assumed that the sojourn times in the states are log-normally distributed.

### 4.5 Parameter Estimation Using the EM-Algorithm

The complete-data includes missing values and then, parameters are estimated using EM-algorithm. Log-normal distribution has two parameters, $\mu$ and $\sigma^2$. EM-algorithm for log-normal distribution is given in chapter four. It should be noted that complete-data sojourn times are log-normally distributed and the estimation values calculated with EM-algorithm are estimates for the distribution parameters of sojourn times.

Let $\mu^{(0)} = 2$ and $\sigma^{2^{(0)}} = 1$ be the starting values. The EM-algorithm starts with the initial values and continue iterating until the differences $|\mu^{(t+1)} - \mu^{(t)}|$ and $|\sigma^{2^{(t+1)}} - \sigma^{2^{(t)}}|$ are smaller than or equal to predetermined value $\varepsilon = 10^{-4}$.

The iterations of EM-algorithm and the distribution parameters of sojourn times are shown at the following Tables 4.1-4.12. The values emphasized with $(*)$ are the differences which are smaller than or equal to $\varepsilon$ and so, the values emphasized with (**) are the estimates for the distribution parameters for sojourn times.

It should be noted that complete-data sojourn times are log-normally distributed and the estimation values calculated with EM-algorithm are estimates for the distribution parameters of sojourn times.

Table 4.1 Parameter estimation using EM-algorithm for $\mathrm{ST}_{00}$

| $t$ | $\mu^{(t)}$ | $\mu^{(t+1)}$ | $\lvert\mu^{(t+1)} - \mu^{(t)}\rvert$ | $\sigma^{2^{(t)}}$ | $\sigma^{2^{(t+1)}}$ | $\lvert\sigma^{2^{(t+1)}} - \sigma^{2^{(t)}}\rvert$ |
|---|---|---|---|---|---|---|
| 0 | 2 | 1.675894 | 0.324106 | 1 | 0.406047 | 0.593953 |
| 1 | 1.675894 | 1.665439 | 0.010455 | 0.406047 | 0.383390 | 0.022657 |
| 2 | 1.665439 | 1.665102 | 0.000337 | 0.383390 | 0.382655 | 0.000735 |
| 3 | 1.665102 | 1.665091** | 0.000011* | 0.382655 | 0.382632** | 0.000023* |

Then, $\mathrm{ST}_{00}$ follows log-$N(1.665091, 0.382632)$

Table 4.2 Parameter estimation using EM-algorithm for $\mathrm{ST}_{01}$

| $t$ | $\mu^{(t)}$ | $\mu^{(t+1)}$ | $\lvert\mu^{(t+1)} - \mu^{(t)}\rvert$ | $\sigma^{2^{(t)}}$ | $\sigma^{2^{(t+1)}}$ | $\lvert\sigma^{2^{(t+1)}} - \sigma^{2^{(t)}}\rvert$ |
|---|---|---|---|---|---|---|
| 0 | 2 | 1.777022 | 0.222978 | 1 | 0.417866 | 0.582134 |
| 1 | 1.777022 | 1.768446 | 0.008576 | 0.417866 | 0.393491 | 0.024375 |
| 2 | 1.768446 | 1.768116 | 0.000330 | 0.393491 | 0.392550 | 0.000941 |
| 3 | 1.768116 | 1.768103** | 0.000013* | 0.392550 | 0.392514** | 0.000036* |

Then, $\mathrm{ST}_{01}$ follows log- $N(1.768103, 0.392514)$

Table 4.3 Parameter estimation using EM-algorithm for $\mathrm{ST}_{02}$

| $t$ | $\mu^{(t)}$ | $\mu^{(t+1)}$ | $\lvert \mu^{(t+1)} - \mu^{(t)} \rvert$ | $\sigma^{2(t)}$ | $\sigma^{2(t+1)}$ | $\lvert \sigma^{2(t+1)} - \sigma^{2(t)} \rvert$ |
|---|---|---|---|---|---|---|
| 0 | 2 | 1.740673 | 0.259327 | 1 | 0.369954 | 0.630046 |
| 1 | 1.740673 | 1.726435 | 0.014238 | 0.369954 | 0.331468 | 0.038486 |
| 2 | 1.726435 | 1.725653 | 0.000782 | 0.331468 | 0.329343 | 0.002125 |
| 3 | 1.725653 | 1.725611 | 0.000042 | 0.329343 | 0.329227 | 0.000116 |
| 4 | 1.725611 | 1.725608** | 0.000003* | 0.329227 | 0.329220** | 0.000007* |

Then, $\mathrm{ST}_{02}$ follows log- $N(1.725608, 0.329220)$

Table 4.4 Parameter estimation using EM-algorithm for $\mathrm{ST}_{10}$

| $t$ | $\mu^{(t)}$ | $\mu^{(t+1)}$ | $\lvert \mu^{(t+1)} - \mu^{(t)} \rvert$ | $\sigma^{2(t)}$ | $\sigma^{2(t+1)}$ | $\lvert \sigma^{2(t+1)} - \sigma^{2(t)} \rvert$ |
|---|---|---|---|---|---|---|
| 0 | 2 | 1.734776 | 0.265224 | 1 | 0.377215 | 0.622785 |
| 1 | 1.734776 | 1.730817 | 0.003959 | 0.377215 | 0.366854 | 0.010361 |
| 2 | 1.730817 | 1.730758 | 0.000059 | 0.366854 | 0.366699 | 0.000155 |
| 3 | 1.730758 | 1.730757** | 0.000001* | 0.366699 | 0.366697** | 0.000002* |

Then, $\mathrm{ST}_{10}$ follows log- $N(1.730757, 0.366697)$

Table 4.5 Parameter estimation using EM-algorithm for $\mathrm{ST}_{11}$

| $t$ | $\mu^{(t)}$ | $\mu^{(t+1)}$ | $\lvert \mu^{(t+1)} - \mu^{(t)} \rvert$ | $\sigma^{2(t)}$ | $\sigma^{2(t+1)}$ | $\lvert \sigma^{2(t+1)} - \sigma^{2(t)} \rvert$ |
|---|---|---|---|---|---|---|
| 0 | 2 | 1.834350 | 0.165650 | 1 | 0.389785 | 0.610215 |
| 1 | 1.834350 | 1.831542 | 0.002808 | 0.389785 | 0.378970 | 0.010815 |
| 2 | 1.831542 | 1.831495 | 0.000047 | 0.378970 | 0.378786 | 0.000184 |
| 3 | 1.831495 | 1.831494** | 0.000001* | 0.378786 | 0.378783** | 0.000003* |

Then, $\mathrm{ST}_{11}$ follows log- $N(1.831494, 0.378783)$

Table 4.6 Parameter estimation using EM-algorithm for $\mathrm{ST}_{12}$

| $t$ | $\mu^{(t)}$ | $\mu^{(t+1)}$ | $\lvert \mu^{(t+1)} - \mu^{(t)} \rvert$ | $\sigma^{2(t)}$ | $\sigma^{2(t+1)}$ | $\lvert \sigma^{2(t+1)} - \sigma^{2(t)} \rvert$ |
|---|---|---|---|---|---|---|
| 0 | 2 | 1.779318 | 0.220682 | 1 | 0.393015 | 0.606985 |
| 1 | 1.779318 | 1.776585 | 0.002733 | 0.393015 | 0.384888 | 0.008127 |
| 2 | 1.776585 | 1.776551 | 0.000034 | 0.384888 | 0.384787 | 0.000101 |
| 3 | 1.776551 | 1.776551** | 0.000000* | 0.384787 | 0.384786** | 0.000001* |

Then, $\mathrm{ST}_{12}$ follows log-$N(1.776551, 0.384786)$

Table 4.7 Parameter estimation using EM-algorithm for $\mathrm{ST}_{20}$

| $t$ | $\mu^{(t)}$ | $\mu^{(t+1)}$ | $\lvert \mu^{(t+1)} - \mu^{(t)} \rvert$ | $\sigma^{2(t)}$ | $\sigma^{2(t+1)}$ | $\lvert \sigma^{2(t+1)} - \sigma^{2(t)} \rvert$ |
|---|---|---|---|---|---|---|
| 0 | 2 | 1.722984 | 0.277016 | 1 | 0.386126 | 0.613874 |
| 1 | 1.722984 | 1.711948 | 0.011036 | 0.386126 | 0.358490 | 0.027636 |
| 2 | 1.711948 | 1.711508 | 0.000440 | 0.358490 | 0.357383 | 0.001107 |
| 3 | 1.711508 | 1.711490** | 0.000018* | 0.357383 | 0.357339** | 0.000044* |

Then, $\mathrm{ST}_{20}$ follows log-$N(1.711490, 0.357339)$

Table 4.8 Parameter estimation using EM-algorithm for $\mathrm{ST}_{21}$

| $t$ | $\mu^{(t)}$ | $\mu^{(t+1)}$ | $\lvert \mu^{(t+1)} - \mu^{(t)} \rvert$ | $\sigma^{2(t)}$ | $\sigma^{2(t+1)}$ | $\lvert \sigma^{2(t+1)} - \sigma^{2(t)} \rvert$ |
|---|---|---|---|---|---|---|
| 0 | 2 | 1.732096 | 0.267904 | 1 | 0.373472 | 0.626528 |
| 1 | 1.732096 | 1.728000 | 0.004096 | 0.373472 | 0.362778 | 0.010694 |
| 2 | 1.728000 | 1.727937 | 0.000063 | 0.362778 | 0.362614 | 0.000164 |
| 3 | 1.727937 | 1.727936** | 0.000001* | 0.362614 | 0.362611** | 0.000003* |

Then, $\mathrm{ST}_{21}$ follows log-$N(1.727936, 0.362611)$

Table 4.9 Parameter estimation using EM-algorithm for $\mathrm{ST}_{22}$

| $t$ | $\mu^{(t)}$ | $\mu^{(t+1)}$ | $\lvert \mu^{(t+1)} - \mu^{(t)} \rvert$ | $\sigma^{2(t)}$ | $\sigma^{2(t+1)}$ | $\lvert \sigma^{2(t+1)} - \sigma^{2(t)} \rvert$ |
|---|---|---|---|---|---|---|
| 0 | 2 | 1.739865 | 0.260135 | 1 | 0.409003 | 0.590997 |
| 1 | 1.739865 | 1.736668 | 0.003197 | 0.409003 | 0.400898 | 0.008105 |
| 2 | 1.736668 | 1.736629** | 0.000039* | 0.400898 | 0.400798** | 0.000100* |

Then, $\mathrm{ST}_{22}$ follows log-$N(1.736629, 0.400798)$

The estimation values of $\mathrm{ST}_{ij}$, for $i, j = 0,1,2$ for log-normal distribution are presented in the following matrices.

$$\hat{\mu}_{ij} = \begin{bmatrix} 1.665091 & 1.768103 & 1.725608 \\ 1.730757 & 1.831494 & 1.776551 \\ 1.711490 & 1.727936 & 1.736629 \end{bmatrix}$$

and

$$\hat{\sigma}_{ij}^{2}=\begin{bmatrix}0.382632 & 0.392514 & 0.329220\\ 0.366697 & 0.378783 & 0.384786\\ 0.357339 & 0.362611 & 0.400798\end{bmatrix}.$$

Table 4.10 Parameter estimation using EM-algorithm for $\mathrm{ST}_0$

| $t$ | $\mu^{(t)}$ | $\mu^{(t+1)}$ | $\lvert\mu^{(t+1)}-\mu^{(t)}\rvert$ | $\sigma^{2(t)}$ | $\sigma^{2(t+1)}$ | $\lvert\sigma^{2(t+1)}-\sigma^{2(t)}\rvert$ |
|---|---|---|---|---|---|---|
| 0 | 2 | 1.741955 | 0.258045 | 1 | 0.390397 | 0.609603 |
| 1 | 1.741955 | 1.730272 | 0.011683 | 0.390397 | 0.359646 | 0.030751 |
| 2 | 1.730272 | 1.729743 | 0.000529 | 0.359646 | 0.358247 | 0.001399 |
| 3 | 1.729743 | 1.729719** | 0.000024* | 0.358247 | 0.358184** | 0.000063* |

Then, $\mathrm{ST}_0$ follows log-$N(1.729719, 0.358184)$

Table 4.11 Parameter estimation using EM-algorithm for $\mathrm{ST}_1$

| $t$ | $\mu^{(t)}$ | $\mu^{(t+1)}$ | $\lvert\mu^{(t+1)}-\mu^{(t)}\rvert$ | $\sigma^{2(t)}$ | $\sigma^{2(t+1)}$ | $\lvert\sigma^{2(t+1)}-\sigma^{2(t)}\rvert$ |
|---|---|---|---|---|---|---|
| 0 | 2 | 1.784803 | 0.215197 | 1 | 0.392001 | 0.607998 |
| 1 | 1.784803 | 1.781743 | 0.003060 | 0.392001 | 0.382688 | 0.009313 |
| 2 | 1.781743 | 1.781699 | 0.000044 | 0.382688 | 0.382556 | 0.000132 |
| 3 | 1.781699 | 1.781699** | 0.000000* | 0.382556 | 0.382554** | 0.000002* |

Then, $\mathrm{ST}_1$ follows log-$N(1.781699, 0.382554)$

Table 4.12 Parameter estimation using EM-algorithm for $\mathrm{ST}_2$

| $t$ | $\mu^{(t)}$ | $\mu^{(t+1)}$ | $\lvert\mu^{(t+1)}-\mu^{(t)}\rvert$ | $\sigma^{2(t)}$ | $\sigma^{2(t+1)}$ | $\lvert\sigma^{2(t+1)}-\sigma^{2(t)}\rvert$ |
|---|---|---|---|---|---|---|
| 0 | 2 | 1.734290 | 0.265710 | 1 | 0.394816 | 0.605183 |
| 1 | 1.734290 | 1.729317 | 0.004973 | 0.394816 | 0.382144 | 0.012671 |
| 2 | 1.729317 | 1.729224 | 0.000093 | 0.382144 | 0.381907 | 0.000237 |
| 3 | 1.729224 | 1.729223** | 0.000001* | 0.381907 | 0.381902** | 0.000005* |

Then, $\mathrm{ST}_2$ follows log-$N(1.729223, 0.381902)$

So, the estimate vectors of $\mathrm{ST}_i$, for $i = 0,1,2$ for log-normal distribution,

$$\hat{\mu}_i = \begin{bmatrix} 1.729719 \\ 1.781699 \\ 1.729223 \end{bmatrix}$$

and

$$\hat{\sigma}_{\mathrm{i}}^2 = \begin{bmatrix} 0.358184 \\ 0.382554 \\ 0.381902 \end{bmatrix}.$$

### 4.6 Characterization by Semi-Markov Process

The expectation and the variance of a random variable $Y$ that follows log-normal distribution are

$$E(Y) = e^{\left(\mu + \frac{\sigma^2}{2}\right)} \tag{4.2}$$

and

$$V(Y) = e^{\left(2\mu + 2\sigma^2\right)} - e^{\left(2\mu + \sigma^2\right)} \tag{4.3}$$

We know that $\mathrm{ST}_{ij}$ are log-normally distributed and the parameter estimates are found. So, the expectations of log-normal distribution using (4.2) with respect to $\hat{\mu}_{ij}$ and $\hat{\sigma}_{ij}^2$ matrices are the mean sojourn times given that the process has just entered $i$ and will next enter $j$ $(\eta_{ij})$ and the variances of $\mathrm{ST}_{ij}$ $(\upsilon_{ij})$ are found using (4.3). The mean $\mathrm{ST}_{ij}$ are

$$\eta_{ij} = \begin{bmatrix} 6.400702 & 7.130357 & 6.620819 \\ 6.780878 & 7.545008 & 7.163109 \\ 6.620433 & 6.747976 & 6.938104 \end{bmatrix}$$

and variances of $\mathrm{ST}_{ij}$ are

$$\upsilon_{ij} = \begin{bmatrix} 19.097227 & 24.439731 & 17.090697 \\ 20.367496 & 26.215366 & 24.079865 \\ 18.825994 & 19.902426 & 23.732510 \end{bmatrix}.$$

Cumulative distribution function of log-normal distribution is

$$F(Y) = \frac{1}{2} + \frac{1}{2} erf\left( \frac{\ln y - \mu}{\sigma\sqrt{2}} \right) \tag{4.4}$$

where $erf$ is the error function,

$$erf\,(y) = \frac{2}{\sqrt{\pi}} \int_0^y e^{-x^2}\,dx \tag{4.5}$$

So, the conditional probability given $\mathrm{ST}_{ij}$ is

$$F_{ij}(t) = \frac{1}{2} + \frac{1}{2} erf\left( \frac{\ln(t) - \mu_{ij}}{\sigma_{ij}\sqrt{2}} \right). \tag{4.6}$$

Then,

$$F_{00}(t) = \frac{1}{2} + \frac{1}{2} erf\left( \frac{\ln(t) - 1.665091}{\sqrt{2}.(0.382632)} \right) \tag{4.7}$$

$$F_{01}(t) = \frac{1}{2} + \frac{1}{2} erf\left( \frac{\ln(t) - 1.768103}{\sqrt{2}.(0.392514)} \right) \tag{4.8}$$

$$F_{02}(t) = \frac{1}{2} + \frac{1}{2} erf\left( \frac{\ln(t) - 1.725608}{\sqrt{2}.(0.329220)} \right) \tag{4.9}$$

$$F_{10}(t) = \frac{1}{2} + \frac{1}{2} erf\left( \frac{\ln(t) - 1.730757}{\sqrt{2}.(0.366697)} \right) \tag{4.10}$$

$$F_{11}(t) = \frac{1}{2} + \frac{1}{2} erf\left( \frac{\ln(t) - 1.831494}{\sqrt{2}.(0.378783)} \right) \tag{4.11}$$

$$F_{12}(t) = \frac{1}{2} + \frac{1}{2} erf\left( \frac{\ln(t) - 1.776551}{\sqrt{2}.(0.384786)} \right) \tag{4.12}$$

$$F_{20}(t) = \frac{1}{2} + \frac{1}{2} erf\left(\frac{\ln(t) - 1.711490}{\sqrt{2.(0.357339)}}\right) \tag{4.13}$$

$$F_{21}(t) = \frac{1}{2} + \frac{1}{2} erf\left(\frac{\ln(t) - 1.727936}{\sqrt{2.(0.362611)}}\right) \tag{4.14}$$

$$F_{22}(t) = \frac{1}{2} + \frac{1}{2} erf\left(\frac{\ln(t) - 1.736629}{\sqrt{2.(0.400798)}}\right). \tag{4.15}$$

All $F_{ij}(t)$ are found and they are shown as a matrix

$$F_{ij}(t) = \begin{bmatrix} F_{00}(t) & F_{01}(t) & F_{02}(t) \\ F_{10}(t) & F_{11}(t) & F_{12}(t) \\ F_{20}(t) & F_{21}(t) & F_{22}(t) \end{bmatrix}.$$

For instance, given that Besiktas has just draw a match and will next lose a match, the probability of occurring this transition in less than or equal to 7 days is

$$F_{10}(7) = P[S_{n+1} - S_n \leq 7 \mid X_n = 1, X_{n+1} = 0] = 0.6388$$

By the equation (3.6), it is possible to construct the matrix $Q_{ij}(t)$.

$$Q_{00}(t) = (0.2436)\left[\frac{1}{2} + \frac{1}{2} erf\left(\frac{\ln(t) - 1.665091}{\sqrt{2.(0.382632)}}\right)\right] \tag{4.16}$$

$$Q_{01}(t) = (0.2554)\left[\frac{1}{2} + \frac{1}{2} erf\left(\frac{\ln(t) - 1.768103}{\sqrt{2.(0.392514)}}\right)\right] \tag{4.17}$$

$$Q_{02}(t) = (0.5010)\left[\frac{1}{2} + \frac{1}{2} erf\left(\frac{\ln(t) - 1.725608}{\sqrt{2.(0.329220)}}\right)\right] \tag{4.18}$$

$$Q_{10}(t) = (0.2113)\left[\frac{1}{2} + \frac{1}{2} erf\left(\frac{\ln(t) - 1.730757}{\sqrt{2.(0.366697)}}\right)\right] \tag{4.19}$$

$$Q_{11}(t) = (0.2792)\left[\frac{1}{2} + \frac{1}{2} erf\left(\frac{\ln(t) - 1.831494}{\sqrt{2.(0.378783)}}\right)\right] \tag{4.20}$$

$$Q_{12}(t) = (0.5095)\left[\frac{1}{2} + \frac{1}{2} erf\left(\frac{\ln(t) - 1.776551}{\sqrt{2.(0.384786)}}\right)\right] \tag{4.21}$$

$$Q_{20}(t) = (0.2042)\left[\frac{1}{2} + \frac{1}{2} erf\left(\frac{\ln(t) - 1.711490}{\sqrt{2.(0.357339)}}\right)\right] \tag{4.22}$$

$$Q_{21}(t) = (0.2661)\left[\frac{1}{2} + \frac{1}{2} erf\left(\frac{\ln(t) - 1.727936}{\sqrt{2.(0.362611)}}\right)\right] \tag{4.23}$$

$$Q_{22}(t) = (0.5297)\left[\frac{1}{2} + \frac{1}{2} erf\left(\frac{\ln(t) - 1.736629}{\sqrt{2.(0.400798)}}\right)\right] \tag{4.24}$$

and then,

$$Q_{ij}(t) == \begin{bmatrix} Q_{00}(t) & Q_{01}(t) & Q_{02}(t) \\ Q_{10}(t) & Q_{11}(t) & Q_{12}(t) \\ Q_{20}(t) & Q_{21}(t) & Q_{22}(t) \end{bmatrix}$$

The elements of $Q_{ij}(t)$ gives the probability that after the process makes a transition to state $i$, next transition is to state $j$ in time less than or equal to $t$.

For instance, suppose that Besiktas plays and loses a match. The probability of winning next match in less then or equal to 5 days is

$$Q_{02}(5) = P[X_{n+1} = 2, S_{n+1} - S_n \leq 5 \mid X_n = 0] = 0.2372$$

and we know that $Q_{ij}(\infty) = P_{ij}$. That is,

$$Q_{02}(\infty) = P[X_{n+1} = 2, S_{n+1} - S_n \leq \infty \mid X_n = 0] = 0.5010 = P_{02}$$

$m_i$ denotes the expected sojourn time in state $i$ during each visit and by the equation (3.11)

$$\begin{aligned} m_0 &= P_{00}\eta_{00} + P_{01}\eta_{01} + P_{02}\eta_{02} = 6.697332 \\ m_1 &= P_{10}\eta_{10} + P_{11}\eta_{11} + P_{12}\eta_{12} = 7.188949 \\ m_2 &= P_{20}\eta_{20} + P_{21}\eta_{21} + P_{22}\eta_{22} = 6.822639 \end{aligned}$$

Then, we obtain

$$m = [m_0, m_1, m_2] = [6.697332, 7.188949, 6.822639] \tag{4.25}$$

***4.6.1 Stationary Distribution Probabilities of the Semi-Markov Process***

Limiting probabilities of the Markov chain are

$$\pi = [0.2146, 0.2673, 0.5181]$$

So, by the equation (3.24) stationary distribution probabilities of the semi-Markov process are

$$P_0 = \frac{(0.2146)(6.697332)}{(0.2146)(6.697332) + (0.2673)(7.188949) + (0.5181)(6.822639)} = 0.2085$$

$$P_1 = \frac{(0.2673)(7.188949)}{(0.2146)(6.697332) + (0.2673)(7.188949) + (0.5181)(6.822639)} = 0.2787$$

$$P_2 = \frac{(0.5181)(6.822639)}{(0.2146)(6.697332) + (0.2673)(7.188949) + (0.5181)(6.822639)} = 0.5128$$

Then,

$$P = [0.2085, 0.2787, 0.5128]. \tag{4.26}$$

### 4.7 Estimation with the Empirical Estimator

In this section, sojourn times that are defined as inappropriate in EM-algorithm are removed from the data set. So, there are 2317 transitions in the process. There are three empirical estimators explained in chapter three. First one is the empirical estimator for limiting probabilities of Markov chain. That is, by equation (3.19)

$$\hat{\pi}_0 = \frac{N_0(t)}{N(t)} = \frac{493}{2317} = 0.2128$$

$$\hat{\pi}_1 = \frac{N_1(t)}{N(t)} = \frac{621}{2317} = 0.2680$$

$$\hat{\pi}_2 = \frac{N_2(t)}{N(t)} = \frac{1203}{2317} = 0.5192$$

where, $N_i(t)$ is the number of transitions in state $i$ that occur in time interval $(0, t]$ and $N(t)$ is the total number of transitions in the process.

$$\Rightarrow \hat{\pi} = [0.2128, 0.2680, 0.5192]. \tag{4.27}$$

The second one is the empirical estimator for the mean sojourn times. That is, by equation (3.20)

$$\hat{m}_0 = \frac{1}{493}\sum_{k=1}^{493} T_{0k} = 6.8763$$

$$\hat{m}_1 = \frac{1}{621}\sum_{k=1}^{621} T_{1k} = 7.2099$$

$$\hat{m}_2 = \frac{1}{1203}\sum_{k=1}^{1203} T_{2k} = 6.8176$$

where $T_{jk}$ is the $k$th sojourn time in state $j$.

$$\Rightarrow \hat{m} = [6.8763, 7.2099, 6.8176]. \tag{4.28}$$

The third one is the empirical estimator for the stationary distribution probabilities of the semi-Markov process. That is, by equation (3.18)

$$\hat{P}_0 = \frac{(0.2128)(6.8763)}{(0.2128)(6.8763)+(0.2680)(7.2099)+(0.5192)(6.8176)} = 0.2110$$

$$\hat{P}_1 = \frac{(0.2680)(7.2099)}{(0.2128)(6.8763)+(0.2680)(7.2099)+(0.5192)(6.8176)} = 0.2786$$

$$\hat{P}_2 = \frac{(0.5192)(6.8176)}{(0.2128)(6.8763)+(0.2680)(7.2099)+(0.5192)(6.8176)} = 0.5104$$

and so,

$$\hat{P} = [0.2110, 0.2786, 0.5104]. \tag{4.29}$$

### 4.8 Comparison of the Results

The distribution parameters of sojourn times are estimated by using EM-algorithm. Then, the mean sojourn times and stationary distribution probabilities for semi-Markov process are estimated with EM-algorithm and an empirical estimator. Table 5.13 contains the estimates obtained in this study.

Table 4.13 Estimation values for semi-Markov process

| $i$ | $\pi_i$ | $\hat{\pi}_i$ | $m_i$ | $\hat{m}_i$ | $P_i$ | $\hat{P}_i$ |
|---|---|---|---|---|---|---|
| 0 | 0.2146 | 0.2128 | 6.6973 | 6.8763 | 0.2085 | 0.2110 |
| 1 | 0.2673 | 0.2680 | 7.1889 | 7.2099 | 0.2787 | 0.2786 |
| 2 | 0.5181 | 0.5192 | 6.8226 | 6.8176 | 0.5128 | 0.5104 |

The empirical estimators studied in chapter three are unbiased and strongly consistent estimators. Then, it can be said that they are reliable estimators for population parameters. The Table 4.13 shows that the empirical estimates are close to the values found using EM-algorithm. So, EM-algorithm is useful for this data set which has random missing values. The assumption of log-normal distribution is a sound decision. So, we can say that sojourn times are log-normally distributed.

# CHAPTER FIVE
# CONCLUSION

In this study, methods used for the estimation of sojourn times that exists in the literature are reviewed. The concept semi-Markov process is explained in detail and applied to a real life data set. The data contains some of inappropriate values therefore, EM-algorithm is used for the estimation of distribution parameters of sojourn times. Then stationary distribution probabilities and mean sojourn times of semi-Markov process are estimated using the technique defined in the literature and empirical estimators. Finally, a comparison is made between the estimates.

Because of the existence of inappropriate values which are assumed to be missing, we can not say that these sojourn times follow a certain distribution. We decided that the most probable distribution is log-normal for sojourn times by inferring from the observed-data. Then, we estimated the parameters of log-normally distributed sojourn times with EM-algorithm. Stationary distribution probabilities and mean sojourn times are estimated for semi-Markov process. Besides, the empirical estimators used for the estimation in this thesis are unbiased and strongly consistent estimators. So, it can be said that they are good and reliable estimators for the related parameters. Stationary distribution probabilities and mean sojourn times are estimated again with the empirical estimators. Thus, we have two estimation values for stationary distribution probabilities and mean sojourn times of semi-Markov process. When we look at the estimation values, it is seen that these values are close to each other. Then,

- According to the assumptions of semi-Markov process, the match results and dates of a football club can be characterized with semi-Markov process.
- EM-algorithm is useful for the estimation of distribution parameters of sojourn times which contains missing values.
- We said that the empirical estimators are good and reliable estimators and the estimates obtained using EM-algorithm are close to the empirical estimates. So, the assumption of log-normal distribution for sojourn times sounds good and

we can say that the sojourn times for semi-Markov process are log-normally distributed.

It is possible to calculate the probability that after the process makes a transition to state $i$, next transition is to state $j$ in time less than or equal to $t$ with $Q_{ij}(t)$ and the probability of making a transition within an amount of time $t$, given that the process has just entered $i$ and will next enter $j$ with $F_{ij}^{(t)}$.

The empirical estimators we used in this thesis are proposed before and already exist in the literature. For further study, new studies are possible about the estimation of sojourn times and stationary distributions for semi-Markov process. Different methods can be developed according to the type of related data set.

## REFERENCES


Altinok, Y., & Kolcak, D. (1999). An application of the semi-Markov model for earthquake occurences in North Anatolia, Turkey. *Journal of the Balkan Geophysical Society, 2* (4), 90-99.

Atherya, K. B., McDonald, D., & Ney, P. (1978). Limit theorems for semi-Markov processes and renewal theory for Markov chains. *The Annals of Probability, 6* (5), 788-797.

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

Ezhov, I., & Korolyuk, V. S. (1967). Semi-Markov processes and their applications. *Kibernetika, 3* (5), 58-65.

Larsen, J., & Marx, L. (2012). *An introduction to mathematical statistics and its applications* (5th ed.). United States of America: Pearson Education.

Levy, P. (1954). Process semi-markoviens. *Proceedings of the International Congress on Mathematics, Amsterdam, 2*, 416-426.

Limnios, N., Ouhbi, B., & Sadek, A. (2005). Empirical estimator of stationary distribution for semi-Markov processes. *Communications is Statistics - Theory and Methods, 34* (4), 987-995.

Masala, G. (2014). Rainfall derivatives pricing with an underlying semi-Markov model for precipitation occurences. *Stochastic Environmental Research and Risk Assessment, 28*, 717-727.

Ouhbi, B., & Limnios, N. (1999). Nonparametric estimation for semi-Markov processes based on its hazard rate functions. *Statistical Inference for Stochastic Processes, 2*, 151-173.

Perman, M., Senagacnik, A., & Tuma, M. (1997). Semi-Markov models with an application to power-plant reliability analysis. *IEEE Transactions on Reliability, 46* (4), 526-532.

Pons, O. (2007). Estimation for the distribution function of one- and tho-dimensional censored variables or sojourn times of Markov-renewal processes. *Communications in Statistics – Theory and Methods, 36* (14), 2527-2542.

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

Ross, S. M. (1970). *Applied probability models with optimization applications*. New York: Dover.

Ross, S. M. (1996). *Stochastic processes* (2nd ed.). New York: John Wiley & Sons.

Ross, S. M. (2007). *Introduction to probability models* (9th ed.). United States of America: Elsevier.

Smith, L. (1954). Regenerative stochastic processes. *Proceedings of the International Congress on Mathematics, Amsterdam, 2*, 304-305.

Takács, L. (1954). Some investigations concerning recurrent stochastic processes of a certain type. *Magyar Tudományos Akadémia Matematikai Kutató Intézetének Közleményei, 3,* 115-128.

Turkish Football Federation, (n.d). *Futbol bilgi bankası.* Retrieved January, 2015, from http://www.tff.org/Default.aspx?pageID=328.

Ugwuowo, F. I., & Udoumoh, E. F. (2011). Multi-stage semi-Markov models for recovery process. *Communications in Statistics – Theory and Methods, 40* (16), 3011-3019.

Uzunoğlu, U. (2001). *Markov decision processes and an application*. M.Sc. Thesis, Dokuz Eylül University, İzmir.

Voelkel, J. G., & Crowley, J. (1984). Nonparametric inference for a class of semi-Markov processes with censored observations. *The Annals of Statistics, 12* (1), 142-160.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics, 11* (1), 95-103.

Yadavalli, V. S. S., & Natarajan, R. (2001). A semi-Markov model of a manpower system. *Stochastic Analysis and Applications, 19* (6), 1077-1086.