


Spring 5-2020

**A PARTIAL LIKELIHOOD APPROACH TO LONGITUDINAL
CATEGORICAL DATA USING A CONTINUOUS TIME SEMI-MARKOV
CHAIN MODEL**

KUSHA A. MOHAMMADI

Follow this and additional works at: https://digitalcommons.library.tmc.edu/uthsph_dissertsopen


 Part of the [Community Psychology Commons](#), [Health Psychology Commons](#), and the [Public Health Commons](#)

A PARTIAL LIKELIHOOD APPROACH TO LONGITUDINAL
CATEGORICAL DATA USING A CONTINUOUS TIME
SEMI-MARKOV CHAIN MODEL

by

KUSHA A MOHAMMADI, BA, MS

APPROVED:



WENYAW CHAN, PhD



MICHAEL SWARTZ, PhD



R. SUE DAY, PhD



DEAN, THE UNIVERSITY OF
TEXAS SCHOOL OF PUBLIC
HEALTH

Copyright
by
KUSHA A MOHAMMADI, BA, MS
2020

DEDICATION

To Sara Beth

A PARTIAL LIKELIHOOD APPROACH TO LONGITUDINAL CATEGORICAL
DATA USING A CONTINUOUS TIME SEMI-MARKOV CHAIN MODEL

by

KUSHA A MOHAMMADI, BA, MS

Presented to the Faculty of The University of Texas

School of Public Health

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH

Houston, Texas

May 2020

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my dissertation advisor, Dr. Wenyaw Chan. His valuable experience, profound support, and patience created an enriching research environment for my dissertation journey. As my advisor, I am grateful for his assistance to find various data analysis projects to diversify my educational experience. Also, I'm deeply grateful for Dr. Michael Swartz who often shared valuable insight from his academic and research experience thus far. Dr. Sue Day for her extensive knowledge in critiquing current epidemiological research from her class and feedback in my dissertation. I am also grateful for Dr. Julia Benoit for serving on my committee and advice she has given me through the years.

I am very grateful to the the National Institute of General Medical Sciences (trainee grant program: T32GM074902) for the financial support throughout my academic endeavors. My thanks to the UT Health and the department of Biostatistics and data science for the opportunity to be apart of this program and pursue my goals.

Finally, I would also like to take the opportunity to thank my family and friends who have painfully listened to me talk about my statistical work over the years. Dr. Jack Mealy for his strong support, guidance, and mentorship from my undergraduate work to now. To all my friends who have been lifting me up through the years, I am truly appreciative for your encouragement through the challenging moments I experienced on this journey. I cannot express enough words of thankfulness to my dad, mom, and sister. My family has been a driving force throughout this academic journey and I am truly appreciative for their financial, emotional, and academic support. Finally, to my future wife, Sara Beth, I am truly grateful for your love, patience, and encouragement that helped me get through all the chaos of the dissertation process. Sara, through her own work ethic, inspired me to persevere through challenging moments and supported and loved me to the end. I am extremely grateful to you all.

A PARTIAL LIKELIHOOD APPROACH TO LONGITUDINAL CATEGORICAL
DATA USING A CONTINUOUS TIME SEMI-MARKOV CHAIN MODEL

Kusha A Mohammadi, BA, MS
The University of Texas
School of Public Health, 2020

Dissertation Chair: Wenyaw Chan, PhD

Longitudinal studies have been critical in understanding the characteristics of chronic diseases or interventions. Since many processes have natural multi-categorical responses over time, multi-state stochastic models have been used to estimate the transition rates between stages. Some multi-state models applied in practice assume the Markov property. The Markov property constrains the sojourn distribution to be exponentially distributed. While useful theoretical properties arise by the Markov assumption, we will consider a more flexible framework by allowing arbitrarily distributed waiting times. This describes a semi-Markov process which has already been applied to various fields in Public Health. Similar to Markov model developments, semi-Markov models have been extended to add covariate effects on each transition intensity for better estimation. Statistical inference methods for semi-Markov chains are still being developed for unique problems for efficient estimation and computational feasibility. Particularly, in this dissertation, we have developed a partial likelihood based approach under a semi-Markov framework. First, we will consider estimating parameters for a three to four stage process by a partial likelihood approach and examining the sensitivities of the transition intensity estimates with models that have a gamma or Weibull sojourn time. This approach will estimate the hazard rates between discrete stages. Secondly, we will extend the semi-Markov model to include covariate effects on the transition rates and again, analyze its results with models assuming the gamma or Weibull sojourn time. Two applications will

be considered to illustrate our method: A caregiver stress-level study from the Baylor's Alzheimer's Disease and Memory Disorders Center and a depression severity level study from the Hispanic Established Population for the Epidemiological Study of the Elderly (HEPESE).

Table of Contents

1	Introduction	1
1.1	Literature Review	1
1.1.1	Importance of Multi-State Models	1
1.1.2	The Underlying Idea of Semi-Markov Models	1
1.1.3	Applications & Developments for Semi-Markov Models	3
1.2	Data Description for the Alzheimer’s Disease Caregiver Stress Application	4
1.3	Data Description for the HEPese Application	5
1.4	Public Health Significance	6
1.5	Specific Aims	7
	References	9
2	Estimation Method of a Continuous-Time Semi-Markov Model for Longitudinal Categorical Outcomes: A Partial Likelihood Approach	12
2.1	Abstract	12
2.2	Introduction	13
2.3	Methods	15
2.3.1	Semi-Markov Model	15
2.3.2	Distributions of the Sojourn Time	17
2.3.3	Construction of the Partial Likelihood	18
2.4	Simulation	20
2.5	Longitudinal Application	22
2.6	Discussion	24
	References	43

3	A Continuous-Time Semi-Markov Model for Longitudinal Categorical Outcome with predictors: A Partial Likelihood Approach	46
3.1	Abstract	46
3.2	Introduction	47
3.3	Methods	48
3.3.1	The semi-Markov Model	48
3.3.2	Incorporation of Covariates	50
3.3.3	Distributions of the Sojourn Time	51
3.3.4	The Partial Likelihood by Adding Covariates	52
3.4	Simulation	54
3.5	Caregiver Stress Application	56
3.6	Discussion	58
	References	66
4	Trajectories in Depression Symptoms among Elderly Mexican Americans with Chronic Health Conditions: A Longitudinal Data Analysis	67
4.1	Abstract	67
4.2	Introduction	68
4.3	Methods	70
4.3.1	Elderly Hispanic Study Sample	70
4.3.2	Categorical Outcome Measure	71
4.3.3	Statistical Analysis	71
4.4	Results	72
4.5	Discussion	74
	References	84
5	Future Works	86
	Appendix A: Supplementary Materials for Chapter 4	89
A.1	Additional Tables & Figures to Describe the HEPSE Example	90

Appendix B: Code	92
B.1 Written in R 3.6.2 - "Dark and Stormy Night"	92
B.2 C++ Code written for Rcpp Package in R	121
Bibliography	129

List of Tables

2.1	Simulation Results for a three-state Semi-Markov Model	27
2.2	Simulation Results for a Four-State Semi-Markov Model (<i>continued</i>) . . .	29
2.5	Observed Transitions between 3-Levels of Caregiver Stress	32
2.6	Observed Transitions between 4-Levels of Caregiver Stress	32
2.7	Alzheimer’s Disease Caregiver 3-Level Stress Model Estimates	33
2.9	Estimated Transition Probabilities of the Embedded 3-State Markov Chain	40
2.10	Estimated Transition Probabilities of the Embedded 4-State Markov Chain	40
3.1	Observed Transitions between 3-Levels of Caregiver Stress	58
3.2	Simulation Results for a Three-State Semi-Markov Model with the Inclu- sion of Covariates (<i>continued</i>)	62
4.1	Characteristics of Elderly Hispanic Adults in HEPSE at Baseline	78
4.2	Observed Transitions between 4-Levels of Caregiver Stress	81
4.3	Probability of Moving to Another Depression Stage at Time of Transition	81
4.4	Elderly Mexican-American 4-Level Depression Model-Based Parameter Estimates assuming a Gamma Sojourn Distribution	82

List of Figures

2.1	Frequency of Transition between 3 Levels of Caregiver of Stress over Time	34
2.2	Frequency of Transition between 4 Levels of Caregiver of Stress over Time	35
2.3	Sojourn Time by Caregiver Stress-Level (i to j) while Overlaying each Semi-Markov Model for a 3-State Process. The red line denotes the exponential distribution, the orange line represents the Weibull distribution, and the green line represents the gamma distribution.	38
2.4	Sojourn Time by Caregiver Stress-Level (i to j) while Overlaying each Semi-Markov Model for a 4-State Process. The red line denotes the exponential distribution, the orange line represents the Weibull distribution, and the green line represents the gamma distribution.	39
2.5	Plot of the Hazard of the Continuous-Time semi-Markov Process assuming the gamma sojourn time distribution. State 1, 2, and 3 represent not/mild stressed, moderately depressed, and severely depressed, respectively. Time is in years.	41
2.6	Plot of the Hazard of the Continuous-Time semi-Markov Process assuming the gamma sojourn time distribution. State 1, 2, 3, and 4 represent not stressed, mildly stressed, moderately stressed, and severely stressed, respectively. Time is in years.	42
3.1	Frequency of Transition between 3 Levels of Caregiver Stress Over Time By Sex	63

3.2	Plot of the Hazard of the Continuous-Time Semi-Markov Process Assuming a Gamma Sojourn Distribution for an Alzheimer’s Disease Caregiver Application. The lines indicates hazard lines for the ages of 65 years, 70 years, 76 years, and 81 years old (green, red, orange and blue, respectively).	65
4.1	The timeline of the Hispanic Established Population for the Epidemiologic Study of the Elderly. Eight waves of data were collected from 1993 to 2013.	77
4.2	Frequency of Transition between 4 Levels of Depressive Symptoms over Time	80
4.3	Plot of the Hazard of the Continuous-Time semi-Markov Process assuming the gamma sojourn time distribution. State 1, 2, 3, and 4 represent not depressed, mildly depressed, moderately depressed, and severely depressed, respectively. Time is in years.	83
A.1	Sojourn Time by Depression-Level (i to j) while overlaying each Semi-Markov Model for a 4-State Process. The red line represents the exponential distribution, the orange line represents the Weibull distribution, and the green line represents the gamma distribution.	91

Chapter 1

Introduction

1.1 Literature Review

1.1.1 Importance of Multi-State Models

Multi-state models extend the classical survival model to analyze multiple transient stages or levels of disease [1]. Particularly, these type of models have been utilized in various fields to investigate the natural course of a variety of biological processes. There are a variety of longitudinal studies that focus on chronic diseases or interventions that have an observable multi-categorical response over time. Some common examples that can be modeled by a multi-state stochastic model are breast cancer [2, 3], HIV [4-6], Alzheimer's disease [7, 8], cirrhosis [9], asthma [10], and bipolar disease [11]. The multi-state model is an effective method in Public Health in estimating the transition rates between two discrete states or stages. By modeling the categorical disease or intervention stages, this can aid in improving prognosis, drug development, and clinical trial design.

1.1.2 The Underlying Idea of Semi-Markov Models

The Markov chain model has over 50 years of developed theory that allow the Markovian process to have a clear and convenient method to understanding various scientific insights [12]. In the statistical field, we define a Markov process as a stochastic

process that satisfies the Markov property. In simple terms, the Markov property holds if the probabilistic behavior of the future state depends only on the present state and disregards the past history of the chain [13]. Since the history of the chain does not affect the future evolution, many also have called this the memoryless property of the Markov process. Due to this fact, estimating the transition rates is easy and computationally feasible. However, this property also assumes the distribution of the sojourn time is exponentially distributed in the continuous case or geometrically distributed in the discrete case. In practice, these distributions may not realistically apply to all situations in Public Health. For instance, by assuming the waiting times are exponential, the time until a state change is likely to be instant or very short which may not be true of the underlying process [14]. For this reason, the Markov framework was then generalized to allow for arbitrary sojourn time distributions.

Semi-Markov processes were first introduced independently by Levy [15], and Smith [16] at the International Congress of Mathematics, and in the same year, Takács also characterized a stochastic process of the same type [17]. These papers detailed an inaugural class of stochastic processes that generalized the well-known Markov chain with finite state spaces. The motivation to develop this methodology was both theoretical and application based. Particularly, Levy was interested in understanding how the behavior of the sample paths would change if the sojourn time distribution was any general function (i.e. the sojourn time was not assumed to be exponentially distributed). For Smith, he sought to develop the theory of regenerative stochastic processes by applying it to a general form of the Markov chains. Alternatively, coincidences in particle counting problems provoked Takacás to study this type of a recurrent process, however, he never formally named the process. Rather, he classified the stochastic recurrent process as "of a Certain Kind". From 1954 on, these works started as the foundation to many other groundbreaking theoretical results and allowed semi-Markov processes to be constructed for a wide breadth of applications [18].

In the literature, the terms semi-Markov process (SMP) and Markov renewal process (MRP) are often used interchangeably, however, differentiate in definition. The MRP was first coined by Pyke and others in the 1960s who studied processes similar in nature to the semi-Markov processes [19–21]. In simple terms, a MRP is identical to the SMP except that each state is only defined at the jump time points whereas a SMP is defined for every given time. Because of this subtle mathematical difference, these terms over time in research have become synonyms for one another. Throughout this dissertation, we will refer to this general stochastic process as SMP and will formally define this process in Chapter 2.

1.1.3 Applications & Developments for Semi-Markov Models

In many applications, the semi-Markov model is a powerful tool because of its relaxed assumptions of the Markov property. Weiss and Zelen proposed a semi Markov model by assuming a gamma sojourn distribution to study right-censored observations in a clinical trial [14]. Foucher and others introduced a semi-Markov model to study HIV disease based on a Generalized Weibull distribution as the waiting time distribution [22]. Kang and Lagakos explored a semi-Markov process in a HPV study setting where they proposed a likelihood base approach for panel data [23]. Cao and others compared a Cox Markov model versus semi-Markov model in a study heart failure disease management to find sensible balance between model parsimonious and computational complexity [24]. Additionally, hidden semi-Markov models have been applied to a number of various areas such as speech recognition/synthesis, fMRI brain mapping, and handwriting recognition [25]. These examples illustrate the potential applications and use for semi-Markov models.

Semi-Markov models also have been extended to consider special types of data. Anderson and others proposed a Cox semi-Markov model to add covariate effects to each transition intensity for an application in bleeding episodes and mortality in liver cirrhosis

[26]. Titman presented a new statistical likelihood method to estimate transitions rates from panel data using phase-type approximations [27]. Shu and others utilized large sample theory to develop asymptotic theory for the Cox semi-Markov model to investigate the robustness and efficiency of semi-Markov estimators [28]. Aralis and Brookmeyer proposed a stochastic estimation procedure for panel observation data with back transitions while assuming a non-exponential distribution [29]. Yu has also extended the semi-Markov theory to consider misclassification in observed states called the hidden semi-Markov model (HSMM) [25]. Semi-Markov theory and estimation procedures continue to be developed for better estimation and unique practice situations.

1.2 Data Description for the Alzheimer’s Disease Caregiver Stress Application

In this longitudinal study, caregiver stress-level was recorded over a 21 year period by the Baylor Alzheimer’s Disease and Memory Disorders Center [30]. The primary aim of the study was to collection socio-demographic information as well as neuro-psychological information to evaluate probable Alzheimer’s Disease. As a secondary interest, a cohort of Alzheimer’s Disease (AD) caregivers, representing a family member or friend, were recruited to provide information on their health and well-being. The stress level was recorded on four levels: none, mild, moderate, and severe. Patients involved in the study were diagnosed with AD and cared for by family members and/or friends. Self-reported information on the caregiver stress-level over a 21 year period (1990 - 2011) was collected. The time between visits for each caregiver varied widely along with the number of recorded observations.

For this longitudinal analysis, we will use a continuous-time semi-Markov model to learn about that movements through the various caregiver stress-levels. We will include individuals with at least two recorded stress levels and complete covariate information.

Additionally, we could analyze how the caregiver covariate information affects the transition rates between stress-levels. Our proposed partial likelihood approach will be used to estimate the model parameters under an exponential, Weibull, and gamma wait time distribution. The final model will be determined through statistical measures and graphic overlays of the raw data. The transitions rates and sojourn times could be invaluable information to the literature of Alzheimer’s disease and help promote awareness in AD caregiver stress.

1.3 Data Description for the HEPSE Application

The longitudinal data analysis will be based on eight waves of data from the Hispanic Established Population for the Epidemiological Study of the Elderly (HEPESE). The HEPSE contains Mexican Americans aged 65 and older, who live in five southwestern states: Texas, New Mexico, Colorado, Arizona, and California [31]. The original study started in 1993 -1994 with 3050 subjects with a response rate of 83%. Additional follow-ups occurred every two years post baseline: Wave 2 in 1995 - 1996 ($M = 2438$), Wave 3 in 1998-1999 ($M = 1980$), Wave 4 in 2000 - 2001 ($M = 1682$), Wave 5 in 2004 - 2005 ($M = 2069$), Wave 6 in 2007 ($M = 1542$), Wave 7 in 2010 - 2011 ($M = 1078$), and Wave 8 in 2012 - 2013 ($M = 744$). Wave 5 added 905 new respondents that were aged 75 and older and followed up with the original cohort.

For this longitudinal analysis, we will include individuals who have more than 1 observation, are not missing depression information, and have depressive symptoms. We will categorize the Center for Epidemiological Studies Depression Scale [32] by the following criteria: not depressed (0 - 9 points), mildly depressed (10 - 15 points), moderately depressed (16 - 24 points), and severely depressed (more than 25 points)[33]. A continuous-time semi-Markov model is used to capture dynamic nature of the depression levels over time. The partial likelihood methodology will be utilized to estimate the

transition rates under an exponential, Weibull, and gamma wait time distribution. The appropriate model will be determined through statistical measures like AIC and goodness of fit test. Lastly, we will interpret the hazard model in the context of the HEPSE application and discuss its potential contribution to the mental health literature.

1.4 Public Health Significance

Statistical inference in the area of semi-Markov models continues to grow as more complex problems arise. Specifically, there is a continual need for the development of efficient estimators and computationally feasible methods to study dynamic disease/intervention behaviors in Public Health. By observing repeated outcome variables over time, a researcher can learn about an individual's trajectory dynamics in a continuous-time setting. This type of information is captured in longitudinal studies. Many times, longitudinal studies also collect explanatory variables which can be utilized in the model to potentially get better estimates on the transition rates. For this reason, longitudinal categorical data play an integral part in expanding the knowledge of multi-level disease/interventions. Accordingly, this research proposal is to develop a partial likelihood method considering the semi-Markov framework for categorical longitudinal data. This inaugural methodology will (1) estimate the transition rates between disease/intervention stages for a given sojourn distribution and (2) extend this approach to additionally account for subject covariate information. This proposed partial likelihood approach will greatly contribute to the semi-Markov literature. First, the structure of the partial likelihood method is familiar and simple as in the classical survival analysis. The redefined probabilistic statements in the partial likelihood allow for the complexity of the semi-Markov process to be analyzed. Secondly, by utilizing Rcpp package [34], and doParallel package [35] in R, we will develop an computationally efficient way to estimate the parameters from the semi-Markov process. The Rcpp package connects the C++ programming

language and R by allowing R to call C ++ functions easily into R code. This tool will help us improve computation time quickly and conveniently. Similarly, the doParallel package will improve computation speed by performing multi-core computing. Thirdly, we thoroughly analyzed two non-exponentially sojourn time distributions: Weibull and gamma distribution. Both of these distributions have been studied for unique survival problems because of the fact that they are generalizations of the exponential distribution. By assuming either the Weibull or gamma as the waiting time distribution, we have the flexibility in the CTSM to have multiple types of shapes for the hazard of the semi-Markov process which allows us to study a wide range of Public Health longitudinal applications.

1.5 Specific Aims

The specific aims of this proposed research are:

Specific Aim #1: To develop a partial likelihood estimation method for estimating parameters in a continuous time semi-Markov model with longitudinal data of three or four outcome categories and to compare its results with models that have a gamma or Weibull sojourn time.

In many research settings, a disease/intervention outcome had multiple levels measured over time. These multi-categories can be natural stages like no disease to pre-clinical disease to disease or scaled level states such as no pain to some pain to much pain. We will develop a partial likelihood approach to estimate the hazard rates between stages considering a non-exponential distribution. Specifically, we will compare an exponential sojourn distribution (i.e. Markov process) to (1) a Gamma sojourn distribution and (2) a Weibull sojourn distribution. This methodology will be applied to an Alzheimer's Disease caregiver stress level example.

Specific Aim #2: To extend the partial likelihood method in aim 1 to include covariate effects on the transition intensities while its outcome process is under a semi-Markov framework.

In addition to recording a categorical outcome over time, several studies many will collect patient information that can be used to further understand the covariate effects on the transition rates. These covariate hazard rates are helpful in interpreting the scientific associations. We will extend the partial likelihood approach to estimate the baseline hazard rates between stages and the covariate effects considering a non-exponential distribution. As before, we will compare an exponential sojourn distribution (i.e. Markov process) to (1) a Gamma sojourn distribution and (2) a Weibull sojourn distribution. To illustrate the proposed method, we will consider a longitudinal outcome as care-giver stress-level while incorporating some predictors.

Specific Aim #3: To examine the dynamic behavior in depression levels among older adults of Mexican descent from the Hispanic Established Population for the Epidemiological Study of the Elderly (HEPESE) by using a continuous-time semi-Markov model and applying the partial likelihood methodology in the first two aims.

The methodology developed in aim 1 and aim 2 will be utilized in a depression-level application among Mexican elderly to determine the baseline transition rates, and covariate effects while considering three semi-Markov models assumptions for the sojourn time distribution: 1) an exponential sojourn time distribution (i.e. Markov Model), 2) a gamma sojourn distribution and 3) a Weibull sojourn distribution. Each model will be compared to one another using appropriate statistical tests to find the most appropriate model for the data.

References

- [1] D. R. Cox, “Regression models and life-tables,” *Springer Series in Statistics Breakthroughs in Statistics*, pp. 527–541, 1992. DOI: 10.1007/978-1-4612-4380-9_37.
- [2] H. H. Chen, S. W. Duffy, and L. Tabar, “A markov chain method to estimate the tumour progression rate from preclinical to clinical phase, sensitivity and positive predictive value for mammography in breast cancer screening,” *The Statistician*, vol. 45, no. 3, p. 307, 1996. DOI: 10.2307/2988469.
- [3] R. Perez-Ocón, J. E. Ruiz-Castro, and M. L. Gámiz-Perez, “A multivariate model to measure the effect of treatments in survival to breast cancer,” *Biometrical Journal*, vol. 40, no. 6, pp. 703–715, 1998. DOI: 10.1002/(sici)1521-4036(199810)40:6<703::aid-bimj703>3.0.co;2-7.
- [4] I. M. Longini, W. S. Clark, R. H. Byers, J. W. Ward, W. W. Darrow, G. F. Lemp, and H. W. Hethcote, “Statistical analysis of the stages of hiv infection using a markov model,” *Statistics in Medicine*, vol. 8, no. 7, pp. 831–843, 1989. DOI: 10.1002/sim.4780080708.
- [5] A. Alioum, V. Leroy, D. Commenges, F. Dabis, and R. Salmon, “Effect of gender, age, transmission category, and antiretroviral therapy on the progression of human immunodeficiency virus infection using multistate markov models,” *Epidemiology*, vol. 9, no. 6, pp. 605–612, 1998. DOI: 10.1097/00001648-199811000-00007.
- [6] I. Kousignian, B. Autran, C. Chouquet, V. Calvez, E. Gomard, C. Katlama, Y. Rivière, and D. Costagliola, “Markov modelling of changes in hiv-specific cytotoxic t-lymphocyte responses with time in untreated hiv-1 infected patients,” *Statistics in Medicine*, vol. 22, no. 10, pp. 1675–1690, 2003. DOI: 10.1002/sim.1404.
- [7] R. J. Kryscio, F. A. Schmitt, J. C. Salazar, M. S. Mendiondo, and W. R. Markesbery, “Risk factors for transitions from normal to mild cognitive impairment and dementia,” *Neurology*, vol. 66, no. 6, pp. 828–832, 2006. DOI: 10.1212/01.wnl.0000203264.71880.45.
- [8] D. C. Ewbank, “A multistate model of the genetic risk of alzheimers disease,” *Experimental Aging Research*, vol. 28, no. 4, pp. 477–499, 2002. DOI: 10.1080/03610730290103096.
- [9] P. Jepsen, H. Vilstrup, and P. K. Andersen, “The clinical course of cirrhosis: The importance of multistate models and competing risks analysis,” *Hepatology*, vol. 62, no. 1, pp. 292–302, 2015. DOI: 10.1002/hep.27598.
- [10] P. Saint-Pierre, C. Combescure, J. Daurès, and P. Godard, “The analysis of asthma control under a markov assumption with use of covariates,” *Statistics in Medicine*, vol. 22, no. 24, pp. 3755–3770, Aug. 2003. DOI: 10.1002/sim.1680.
- [11] A. Duffy, J. Horrocks, S. Doucette, C. Keown-Stoneman, S. McCloskey, and P. Grof, “The developmental trajectory of bipolar disorder,” *British Journal of Psychiatry*, vol. 204, no. 2, pp. 122–128, 2014. DOI: 10.1192/bjp.bp.113.126706.

- [12] V. Barbu and N. Limnios, *Semi-Markov chains and hidden semi-Markov models toward applications: their use in reliability and DNA analysis*. Springer, 2008.
- [13] H. M. Taylor and S. Karlin, *An introduction to stochastic modeling*. Academic Press, 2014.
- [14] G. H. Weiss and M. Zelen, “A semi-markov model for clinical trials,” Jan. 1963. DOI: 10.21236/ad0407905.
- [15] P. Levy, “Processus semi-markoviens,” in *Processus semi-markoviens*. North-Holland Publishing Co., 1954, vol. 3, pp. 416–426.
- [16] S. L. W, “Regenerative stochastic processes,” in. Regenerative stochastic processes, 1955, vol. 232, pp. 6–31.
- [17] L. Takacs, “Some investigations concerning recurrent stochastic processes of a certain type,” in. Acta Mathematica Hungarica, 1954, vol. 3, pp. 115–128.
- [18] V. S. Korolyuk, S. M. Brodi, and A. F. Turbin, “Semi-markov processes and their applications,” *Journal of Soviet Mathematics*, vol. 4, no. 3, pp. 244–280, 1975, ISSN: 1573-8795. DOI: 10.1007-BF01097184.
- [19] R. Pyke, “Markov renewal processes: Definitions and preliminary properties,” *The Annals of Mathematical Statistics*, vol. 32, no. 4, pp. 1231–1242, 1961. DOI: 10.1214/aoms/1177704863.
- [20] R. Pyke and R. Schaufele, “Limit theorems for markov renewal processes,” *The Annals of Mathematical Statistics*, vol. 35, no. 4, pp. 1746–1764, 1964. DOI: 10.1214/aoms/1177700397.
- [21] R. Pyke and R. Schaufele, “The existence and uniqueness of stationary measures for markov renewal processes,” *The Annals of Mathematical Statistics*, vol. 37, no. 6, pp. 1439–1462, 1966. DOI: 10.1214/aoms/1177699138.
- [22] Y. Foucher, E. Mathieu, P. Saint-Pierre, J.-F. Durand, and J.-P. Daurès, “A semi-markov model based on generalized weibull distribution with an illustration for hiv disease,” *Biometrical Journal*, vol. 47, no. 6, pp. 825–833, 2005. DOI: 10.1002/bimj.200410170.
- [23] M. Kang and S. W. Lagakos, “Statistical methods for panel data from a semi-markov process, with application to hpv,” *Biostatistics*, vol. 8, no. 2, pp. 252–264, 2007. DOI: 10.1093/biostatistics/kxl006.
- [24] Q. Cao, E. Buskens, T. Feenstra, H. Hillege, and D. Postmus, “Continuous-time semi-markov models in health economic decision making: An illustrative example in heart failure disease management,” *Med Decis Making*, vol. 36, no. 1, pp. 59–71, 2015. DOI: 10.1177/0272989X15593080.
- [25] S.-Z. Yu, “General hidden semi-markov model,” *Hidden Semi-Markov Models*, pp. 27–58, 2016. DOI: 10.1016/b978-0-12-802767-7.00002-4.

- [26] P. K. Anderson, S. Esbjerg, and T. Sorensen, “Multi-state models for bleeding episodes and mortality in liver cirrhosis,” *Stat. Med.*, vol. 19, pp. 587–599, 2000. DOI: 10.1002/(sici)1097-0258(20000229)19:4<587::aid-sim358>3.0.co;2-0.
- [27] A. C. Titman, “Estimating parametric semi-markov models from panel data using phase-type approximations,” *Statistics and Computing*, vol. 24, no. 2, pp. 155–164, 2012. DOI: 10.1007/s11222-012-9360-6.
- [28] Y. Shu, J. P. Klein, and M.-J. Zhang, “Asymptotic theory for the cox semi-markov illness-death model,” *Lifetime Data Anal*, vol. 13, pp. 91–117, 2007. DOI: 10.1007/s10985-006-9018-9.
- [29] H. Aralis and R. Brookmeyer, “A stochastic estimation procedure for intermittently-observed semi-markov multistate models with back transitions,” *Statistical Methods in Medical Research*, pp. 1–18, 2017. DOI: 10.1177/0962280217736342.
- [30] J. S. Benoit, W. Chan, S. Luo, H.-W. Yeh, and R. Doody, “A hidden markov model approach to analyze longitudinal ternary outcomes when some observed states are possibly misclassified,” *Statistics in Medicine*, vol. 35, no. 9, pp. 1549–1557, 2016. DOI: 10.1002/sim.6861.
- [31] K. S. Markides, C. A. Stroup-Benham, J. S. Goodwin, L. C. Perkowski, M. Lichtenstein, and L. A. Ray, “The effect of medical conditions on the functional limitations of mexican-american elderly,” *Annals of Epidemiology*, vol. 6, no. 5, pp. 386–391, 1996. DOI: 10.1016/s1047-2797(96)00061-0.
- [32] L. S. Radloff, “The ces-d scale,” *Applied Psychological Measurement*, vol. 1, no. 3, pp. 385–401, 1977. DOI: 10.1177/014662167700100306.
- [33] J. R. Moon, J. Huh, J. Song, I.-S. Kang, S. W. Park, S.-A. Chang, J.-H. Yang, and T.-G. Jun, “The center for epidemiologic studies depression scale is an adequate screening instrument for depression and anxiety disorder in adults with congenial heart disease,” *Health and Quality of Life Outcomes*, vol. 15, no. 1, May 2017. DOI: 10.1186/s12955-017-0747-0.
- [34] D. Eddelbuettel and J. J. Balamuta, “Extending extitR with extitC++: A Brief Introduction to extitRcpp,” *PeerJ Preprints*, vol. 5, e3188v1, Aug. 2017, ISSN: 2167-9843. DOI: 10.7287/peerj.preprints.3188v1. [Online]. Available: <https://doi.org/10.7287/peerj.preprints.3188v1>.
- [35] M. Corporation and S. Weston, *Doparallel: Foreach parallel adaptor for the 'parallel' package*, R package version 1.0.15, 2019. [Online]. Available: <https://CRAN.R-project.org/package=doParallel>.

Chapter 2

Estimation Method of a Continuous-Time Semi-Markov Model for Longitudinal Categorical Outcomes: A Partial Likelihood Approach

Authors: Kusha A. Mohammadi, Wenyaw Chan, and Valory Pavlik

2.1 Abstract

In public health, longitudinal studies have been paramount to studying dynamic diseases and interventions. Many statistical developments through the years have contributed to improvements in modeling the dynamics of transitions among disease states. The multi-state Markov model, for example, has been most often utilized to estimate the transition rates between multi-categorical responses. Although, the Markov model may not be realistic in practice due to the Markov property, which assumes the sojourn time to be exponential distributed. The model proposed in this research considers the semi-Markov framework to analyze longitudinal categorical outcomes that allow for unspecified waiting time distributions. To estimate the parameters of the continuous-time semi-Markov model (CTSMM), we develop a partial likelihood approach for a three to four stage

process. We evaluated our method assuming the sojourn time follows a gamma, Weibull or exponential distribution and examined their sensitivities to our method. Simulations show relatively low bias and similar standard deviation and standard error calculations for both three and four state CTSMs. The coverage probabilities was lower than the expected 95%, however, the CTSM assuming a gamma wait time had the highest coverage across the rate parameters. A longitudinal application of Alzheimer’s disease care-giver stress level was used to illustrate the proposed partial likelihood approach.

Keywords: Semi-Markov Model, Longitudinal Data, Categorical Outcomes, Partial Likelihood Method

2.2 Introduction

In many public health settings, the Markov multi-state model has become an effective approach to analyze categorical events over a given time period. Particularly, it is a useful way to describe the natural course of a variety of biological processes by estimating the rates of transition between states. Some recent applications include breast cancer [2, 3], HIV [4–6], Alzheimer’s disease [7, 8], cirrhosis [9], asthma [10], and bipolar disease [11]. This convenient model, however, implies the Markov property which describes the probabilistic behavior of the future state depending only on the present state and disregarding the history of the chain [13]. Due to the Markov property, the distribution of the sojourn time is assumed to be exponentially distributed in the continuous-time case. This suggests the time until a state transition is likely to be instant or very short, which may not be realistic in practice [36]. It is preferred to have a framework that allows the sojourn time distribution to be unspecified.

Semi Markov models have become a flexible alternative to the Markov framework because it allows for arbitrary waiting time distributions. Weiss and Zelen utilized the semi-Markov framework by investigating right-censored observations in a clinical trial

assuming a gamma sojourn distribution [14]. Foucher and others assumed a generalized Weibull distribution for the wait time to analyze a HIV application [22]. To find a sensible balance between model parsimony and computational complexity, Cao and others compared a cox Markov model versus a cox semi-Markov model to comprehend heart disease failure [24]. These examples spotlight the potential applications in medical research for semi-Markov models.

Approaches to analyze longitudinal categorical outcome data under a semi-Markov model continue to grow as more intricate applications arise. Ouhbi and Limnios considered a non-parametric estimation method for semi-Markov kernels and its hazard function [37, 38]. To account for interval censoring and truncation, Sternberg and Satten proposed a discrete-time non-parametric estimation procedure for a semi-Markov model to HIV applications [39]. Damerджи presented a maximum likelihood estimation approach to calculate the transition rates of the generalized semi-Markov process [40]. To estimate the transition intensity and survival function for a three state semi-Markov model, Joly and Commenges described a penalized likelihood approach with censor and truncated data [41].

In this paper, we develop an alternative estimation method to analyze longitudinal categorical outcome data with three to four stages. We propose a partial likelihood estimation method for estimating parameters in a continuous-time semi-Markov model with longitudinal data. Specifically, we will assume semi-Markov models with exponential, Weibull, and Gamma sojourn time distributions and examine their sensitivities with our method. The proposed estimation method provides a more flexible and realistic tool than the Markov model and extends to biological processes with three to four stages. To illustrate our method, we will apply the method to Alzheimer’s care-giver stress level application.

The remainder of the paper is organized as follows. Section 2.3 defines a semi-Markov process, the sojourn time distributions, and the partial likelihood function. A

simulation study will be used to evaluate the statistical properties of the partial likelihood method in section 2.4 and applied to an Alzheimer’s care-giver stress level example in section 2.5. Lastly, the paper concludes with a discussion in section 2.6

2.3 Methods

2.3.1 Semi-Markov Model

We will consider a random process that is a continuous-time multi-state stochastic process with a finite state space, $\Phi = \{1, 2, \dots, b\}$. For $n = 1, \dots, D$, let $T = (T_n)_{n \in \mathbb{N}}$ denote the consecutive states transition time points where D is the total number of transitions. $T_0 = 0$ is defined as the time of the origin for the stochastic process. Let $S_n = T_n - T_{n-1}$ denote the sojourn times where we set $S_0 = 0$. Then, let $S = (S_n)_{n \in \mathbb{N}}$ be the successive sojourn times in the visited states. Also, let $X = (X_n)_{n \in \mathbb{N}}$ be the sequence of observed states for the n^{th} transition where the state $X_n(t)$ is defined for $t \in [T_n, T_{n+1}]$ and has an initial distribution $\omega_i = P(X_0 = i), i \in \Phi$. This sequence forms an embedded homogeneous Markov chain. Then (X, T) is a homogeneous semi-Markov process if the two assumptions hold true. First, as a subject enters state i , we assume the next state the subject enters is state j with probability, $p_{ij}, i, j \in \Phi$. Second, given that the following state is j , the time until the next transition from i to j has distribution F_{ij} (i.e. an arbitrary sojourn distribution).

The continuous-time semi-Markov kernel, $Q_{ij}(t)$, corresponds to the probability of jumping toward state j between time t and $t + \Delta t$ after being in state i :

$$\begin{aligned} Q_{ij}(t) &= P(X_{n+1} = j, t \leq S_{n+1} | \Lambda_{n-1}) \\ &= P(X_{n+1} = j, t \leq S_{n+1} | X_n = i) \end{aligned} \tag{2.1}$$

where $\Lambda_n = \{(X_0, T_0); \dots; (X_n, T_n)\}$ denotes the history of the semi-Markov chain, $i, j \in \Phi$, and $t \in [T_n, T_{n+1}]$.

The transition probabilities, p_{ij} , from state i to j is formally defined as

$$p_{ij} = \lim_{t \rightarrow \infty} Q_{ij}(t) = P(X_{n+1} = j | X_n = i) \quad (2.2)$$

where $0 \leq p_{ij} \leq 1 \forall i, j \in \Phi$, and $\sum_j p_{ij} = 1$.

Lastly, the distribution function of the waiting time determines the amount of time t that a subject stays in state i before transitioning to state j :

$$F_{ij}(t) = \lim_{\Delta t \rightarrow 0^+} \frac{P[t \leq S_{n+1} < t + \Delta t | X_{n+1} = j, X_n = i]}{\Delta t} \quad (2.3)$$

where $i, j \in \Phi$, and $t \in [T_n, T_{n+1}]$. The marginal probability distribution of the sojourn time is derived from equation 2.2 and 2.3 and written in the following way

$$F_i(t) = \sum_{i \neq j} p_{ij} F_{ij}(t) \quad (2.4)$$

By these relations, we have the following model:

$$Q_{ij}(t) = p_{ij} F_{ij}(t) \quad (2.5)$$

where $i, j \in \Phi$. From distribution function, $F_{ij}(t)$, we can easily derive the probability density function ($f_{ij}(t)$), survival function ($S_{ij}(t)$), and hazard function, $\nu_{ij}(t)$. The hazard of the semi-Markov process is then defined as the probability of transitioning to a state j between time t and $t + \Delta t$, given the previous state is i for a duration t ,

$$h_{ij}(t) = \lim_{\Delta t \rightarrow 0^+} \frac{P(X_{n+1} = j, t \leq S_{n+1} < t + \Delta t | S_{n+1} \geq t, X_n = i)}{\Delta t} \quad (2.6)$$

Using all these relations, we can relate $h_{ij}(t)$ to the hazard function of the sojourn time, survival function of the sojourn time, and transition probabilities,

$$h_{ij}(t) = \frac{p_{ij} \nu_{ij}(t) S_{ij}(t)}{S_i(t)} \quad (2.7)$$

2.3.2 Distributions of the Sojourn Time

By using classical survival relations, we can deduce the hazard function of the sojourn time, $\nu_{ij}(t)$, from equation 2.3:

$$\nu_{ij}(t) = \lim_{\Delta t \rightarrow 0^+} \frac{P[t \leq S_{n+1} < t + \Delta t | S_{n+1} \geq t, X_{n+1} = j, X_n = i]}{\Delta t} \quad (2.8)$$

where t represents the time in a particular state. We will consider three different sojourn distributions in continuous-time semi-Markov model. By assuming these various distributions, the semi-Markov model can be applied to a wide set of problems within Public Health.

Exponential Distribution (λ_{ij})

The hazard function for the exponential distribution with rate parameter, λ_{ij} , is constant over time and is given by

$$\nu_{ij}(t) = \lambda_{ij} \quad (2.9)$$

where $t > 0$, $\forall \lambda_{ij} > 0$, $i, j \in \Phi$. By assuming the waiting time is exponentially distributed, the processes reduces to the well-known Markov model.

Gamma Distribution (ξ, λ_{ij})

The hazard function for the 2-parameter gamma distribution can be viewed as a generalization of the exponential distribution. The hazard function with rate parameter, λ_{ij} , and shape parameter, ξ , is defined as

$$\nu_{ij}(t) = \frac{\lambda_{ij}^\xi t^{\xi-1} e^{-\lambda_{ij}t}}{\Gamma(\xi) - \Gamma(\xi, \lambda_{ij}t)} \quad (2.10)$$

where $t > 0$, $\forall \lambda_{ij} > 0$, $i, j \in \Phi$, $\xi > 0$,

$$\Gamma(a, x) = \int_0^x z^{a-1} e^{-z} dz$$

for $a > 0$, is the incomplete gamma functions, and

$$\Gamma(a) = \int_0^\infty z^{a-1} e^{-z} dz$$

is the gamma function.

For simplicity, we will assume the shape parameter, ξ , is constant across all transitions from i to j .

Weibull Distribution (k, λ_{ij})

The hazard function for the 2-parameter Weibull distribution can also be viewed as a general form of the exponential distribution. The hazard function with rate parameter, λ_{ij} , and shape parameter, k , is defined as follows

$$\nu_{ij}(t) = k\lambda_{ij}t^{k-1} \tag{2.11}$$

where $k > 0$, $\forall \lambda_{ij} > 0$, $i, j \in \Phi$, and $t > 0$.

Additionally, for simplicity, we will assume that the shape parameter, k , is constant across all transitions from i to j .

2.3.3 Construction of the Partial Likelihood

In this section, we outline the construction of the partial likelihood that will allow us to estimate $\mathbf{\Omega} = \{\lambda_{ij}, \xi^*, k^*\}$ where $*$ represents if the shape parameter needs to be estimated based on the sojourn distribution. For m subjects, $m = 1, \dots, M$, we have a longitudinal data based on the jump times, T^{n_m} and respective state transition, X^{n_m} . We will order all the data for M individuals based on the transition times and will be represented by $(T^{(n)}, X^{(n)})$ for the n^{th} transition, $n = 1, \dots, D$. We will define the risk set, $\mathcal{R}(\tau-)$, as all the subjects who are still being observed prior to time, τ , where for

simplicity of notation, we let $\tau = T^{(n)}$. Let $I_{X_l(\tau-)}(u)$ denote the current state, u , for subject l prior to time τ . Let $S^{(n)}$ be the ordered time the individual is in a particular state, i , before transitioning to state j (i.e. sojourn time). Let φ be the time already spent in a particular state, u , for a subject l . Thus, the probability that an subject transitions at time τ given one individual l transitions in $\mathcal{R}_l(\tau-)$ at time τ is

$$\frac{h_{ij}(S^{(n)}|\mathbf{\Omega})}{\sum_{l \in \mathcal{R}(\tau-)} \sum_{u \in \Phi} h_u(\varphi|\mathbf{\Omega}) I_{X_l(\tau-)}(u)} \quad (2.12)$$

where $i, j, u \in \Phi$, and $h_{ij}(\cdot|\cdot)$ is the hazard parametric model for the transition i to j depending on the sojourn distribution chosen. For hazard functions, refer back to section 2.2. Then the partial likelihood is formed by multiplying all the conditional probabilities over all the transitions D . This is given by

$$L(\mathbf{\Omega}) = \prod_{n=1}^D \frac{h_{ij}(S^{(n)}|\mathbf{\Omega})}{\sum_{l \in \mathcal{R}(\tau-)} \sum_{u \in \Phi} h_u(\varphi|\mathbf{\Omega}) I_{X_l(\tau-)}(u)} \quad (2.13)$$

The partial likelihood is analogous to the classical Cox partial likelihood developed in 1972 [42]. In some applications, we may encounter ties in the set of transition times. While there are various constructions to take into account event ties, we will consider Breslow's ties method [43]. With this modification, the partial likelihood is as follows

$$L(\mathbf{\Omega}) = \prod_{n=1}^D \frac{\prod_{g \in d_n} h_{i_g j_g}(S_g^{(n)}|\mathbf{\Omega})}{\left[\sum_{l \in \mathcal{R}(\tau-)} \sum_{u \in \Phi} h_u(\varphi|\mathbf{\Omega}) I_{X_l(\tau-)}(u) \right]^{d_n}} \quad (2.14)$$

where d_n is the number of events at a given transition time, τ and g represents one of the d_n transitions ($i \rightarrow j$) at time τ . We can estimate the parameters $\mathbf{\Omega}$ by maximizing equation 2.14 or by using its logarithmic transformation as shown

$$\begin{aligned}
l(\boldsymbol{\Omega}) = & \sum_{n=1}^D \sum_{g \in d_n} \log [h_{i_g j_g}(S_g^{(n)} | \boldsymbol{\Omega})] \\
& - \sum_{n=1}^D d_n \log \left[\sum_{l \in \mathcal{R}(\tau-)} \sum_{u \in \Phi} h_u(\varphi | \boldsymbol{\Omega}) I_{X_l(\tau-)}(u) \right]
\end{aligned} \tag{2.15}$$

where $h_u(t) = \sum_{r \in \Phi} h_{ur}(t|-)$. $h_{ij}(t)$ is the hazard of the semi-Markov process and can have an exponential, gamma, or Weibull sojourn time distribution outlined in Section 2.3.2.

In a typical setting, the maximum likelihood estimates are attained by taking the log of the likelihood function and setting the first derivative to zero. However, the first derivative of the partial likelihood is not possible to derive with respect to $\boldsymbol{\Omega}$ due to the complexity of the hazard functions. We utilized a non-linear optimization method that is derivative free to maximize the log partial likelihood function. Since the second derivative is not available in closed form, we attained non-parametric bootstrap samples to estimate the standard errors of each parameter of interest. In each bootstrap sample, we re-sampled M subject's with replacement, applied the likelihood function defined by equation 2.15, and calculated each standard error by using all the bootstrap samples. All analysis used R 3.6.2 [44], Rcpp package [34], and doParallel package [35].

2.4 Simulation

Simulation studies were conducted to assess the partial likelihood method outlined in section 2.3. Specifically, we simulate two semi-Markov process cases: first, a three-state continuous-time process assuming an exponential, Weibull, and gamma sojourn time distribution and second, a four-state continuous-time process assuming an exponential, Weibull, and gamma sojourn time distribution. For each model, we obtained 1000 simulations and 50 non-parametric bootstrap samples to calculate the standard errors. For

comparability, we determined a common mean between the gamma and Weibull distribution and set the exponential mean to this common mean. Nelder-Mead’s non-linear optimization was used to estimate parameters of the log partial likelihood in equation 2.15. We evaluated our proposed estimation procedure by the bias, standard deviation, standard error, mean square error (MSE), and 95% coverage probability. Statistical bias is a measure of distance between the expected value and underlying true value. In simulations, we would expect the standard deviation and standard error to be relatively close to indicate the standard error represents the true sampling variation. The mean square error measures the average of the square of the errors. We also view it as a combination of the variation and bias squared. Lastly, coverage probability will help us measure the proportion of the time that the confidence interval contains the true value. For our simulations, we would expect the coverage probability to be close to 95%.

The simulation results for the semi-Markov three-state process and four-state process are summarized for each model in tables 2.1 and 2.2, respectively. By choosing a non-exponential sojourn distribution (Weibull or gamma), the 3-state semi-Markov model assuming a Weibull wait time distribution had the lowest mean square error across all the rate parameters (Table 2.1). In terms of 95% coverage probability, all 3-state models under-performed in terms of capturing the true parameter 95% of the time. Using parallel computing, the full simulation studies for the exponential, Weibull, and gamma case on 20 cores required 10.81, 27.17, and 25.31 hours to run, respectively. Table 2.2 refers to the summary of simulation results for the four-state semi-Markov model. Similar to the 3-state model, the 4-state Weibull simulation shows the lowest mean square error compared to the semi-Markov model assuming a gamma sojourn distribution. While the 95% coverage probability was not met, the gamma case had the highest overall coverage across the rate parameters. In terms of computation time, the model assuming an exponential distribution on 20 cores took 13.60 hours, assuming a Weibull distribution on 30 cores took 25.86 hours, and assuming a gamma distribution on 30 cores took 24.24

hours. In the 3-state and 4-state simulation, the exponential sojourn case (i.e. the Markov Model) performed computationally faster with low bias and variance. In tables 2.3 and 2.4, Markov chain simulations for a full likelihood approach using the multi-state model (msm) [45] in R and our proposed likelihood approach are analyzed. For bias and mean square error, both the Markov likelihood methods are compatible. For the standard error and standard deviation, the partial likelihood method is slightly better than the full likelihood approach but the coverage probability is worse for the partial likelihood method. The full likelihood approach had around 95% coverage probability using the asymptotic standard errors (i.e. from the Hessian matrix) for the confidence interval.

2.5 Longitudinal Application

We applied the partial likelihood approach to a caregiver stress-level application that was recorded over a 21 year period by the Baylor Alzheimer’s Disease and Memory Disorders Center [30]. The primary aim of the study was to collect socio-demographic information as well as neuro-psychological information to assess probable Alzheimer’s Disease. As a secondary interest, a cohort of Alzheimer’s Disease (AD) caregivers, representing a family member or friend, were recruited to provide information on their health and well-being. The care-giver stress level was recorded on four levels: none (state 1), mild (state 2), moderate (state 3), and severe (state 4). The time between visits for each caregiver varied widely along with the number of recorded observations. We included individuals with at least two recorded stress levels ($M = 681$ subjects). The longest observation time was 13.78 years. We re-categorized the stress level as follows: none or mild (state 1), moderate (state 2), and severe (state 3). We utilized a continuous-time semi-Markov model to learn about that movements through the various caregiver stress-levels. Employing various sojourn time distributions, we analyzed the best model using a likelihood ratio test that follows a chi square distribution.

Table 2.5 and 2.6 shows the frequencies of transition between 3-level and 4-level caregiver stress. The most transitions occurred from none/mild stress level to moderate stress levels (303 transitions) in the 3-level case. In the 4-stress level example, 255 transitions went from mild to moderate stress level. Figure 2.1 and Figure 2.2 summarize the dynamic behavior of transitions between caregiver stress levels for 3 states and 4 states over time, respectively. Over the 14 year period, we observed a decline in the number in the none/mild stress level and an increase in the other levels (Figure 2.1). Similarly, in the four state analysis, we observe that the moderate and severe states increase steadily over time while the other two lower states steadily decrease.

To understand the behavior of caregiver stress level over time, we applied a continuous-time semi Markov model and estimated the parameters of the model by the partial likelihood function for a three and four state process. The results are presented in table 2.7 and 2.8 for each process, respectively. Before we analyzed the final results, we carefully found which model closely follows the data. In figure 2.1, the sojourn time within each transition (i to j) is plotted for the 3-stress level application. Using the parameter estimates, the exponential distribution (red line), Weibull distribution (orange line), and gamma distribution (green line) are overlaid onto the data density curve. Overall, it suggests the semi-Markov model assuming a gamma sojourn distribution closely fits the caregiver stress-level data. From figure 2.2, we find a similar trend where the semi-Markov model assuming the gamma wait time distribution closely resembles the data curves across all transitions. For a more quantitative comparison, we calculated the Akaike's information criterion (AIC) for each model. For the three state process assuming exponential, Weibull, and gamma sojourn time, the AIC was 12494.77, 12368.73, and 12290.37, respectively. The AIC for the 4 state-process for each model was 15077.69, 14888.21, and 14757.50, respectively. The smallest AIC value indicates the better model which suggests the semi-Markov model with a gamma sojourn time is the preferred model in both cases. Based on these results, we can utilize the semi-Markov model defined in

equation 2.5 or 2.7 to find the hazard rate for each transition from one stress level to the next under a gamma sojourn time distribution. Figure 2.5 and 2.6 gives an visual representation of the hazard of the semi-Markov process using the model estimates for a 3-state process and 4-state process, respectively. Lastly, we summarized the estimated probability transition matrix in table 2.9 and 2.10 for the 3-state and 4-state applications.

2.6 Discussion

In this paper, we detailed and assessed the partial likelihood approach to analyzing longitudinal categorical outcomes using a continuous-time semi-Markov model. We explored various sojourn distributions including the exponential sojourn time distribution which reduces the semi-Markov model to a Markov Model. Simulations demonstrate relatively low bias and variance across each model considered under a 3 or 4 state process. Overall, the mean square error was marginally higher in the semi-Markov model that assumes a gamma sojourn distribution compared to the others. Computationally, the complexity of the hazard function of the semi-Markov process presented some difficulties in performing the simulations efficiently. By utilizing *Rcpp* package and *doParallel* package in R, we found the total time greatly reduced to run 1000 simulations and 50 non-parametric bootstrap samples. Bootstrap samples were calculated to obtain the standard errors of the estimate since the fisher's information matrix was unable. These bootstrap standard errors were used to calculate the 95% coverage probability for all the parameters. Nearly all the 95% coverage probabilities were observed to be in the range of 78% - 90%. This indicates the 95% confidence interval may be too narrow to capture the true underlying estimate. To explain this result, we compared Markov models using the full likelihood and the partial likelihood (Table 2.3 and 2.4). We observed a slightly lower standard deviation and standard error from the partial likelihood, although, poor coverage probability for both the 3-state and 4-state Markov chain process. This may be due to the

unstable estimates in the partial likelihood from possibly outlying datasets that were generated and using the non-parametric bootstrap approach to compute the confidence intervals. We reason these are why the partial likelihood approach for the semi-Markov model had under-performing coverage probabilities for the exponential, Weibull, and gamma sojourn times.

Our results are not without limitations. The derivative of the log partial likelihood is not available in closed and instead, we had to consider a derivative free numerical optimization approach. This proves to be a difficult optimization problem. The Nelder-Mead (NM) non-linear optimizer needs to satisfy certain properties in order to converge to the maximizing point (i.e. the optimal parameters). While we observed convergence (convergence code = 0), there is the possibility the NM optimization determined a local maximum rather than a global maximum. Other optimizations methods may need to be considered to understand this non-linear problem. Secondly, we observed the 95% coverage probabilities to be less than what was expected. The non-parametric bootstrap procedure may not be appropriate for this approach. We re-sampled from the subjects with replacement and re-estimated the parameters to get a bootstrap distribution. We used 50 bootstrap samples to obtain the standard errors which seemed reasonable. However, the marginally low 95% coverage suggests we obtained a narrow 95% Confidence Interval. Some potential solutions are to consider a parametric bootstrap sample such as in the R package called, 'msm' [45], or find an optimization method to estimate the hessian matrix.

We applied the partial likelihood approach to an Alzheimer's Disease caregiver stress level application done by the Baylor Alzheimer's Disease and Memory Disorders Center [30]. This application provides an excellent example where the time spent before transition may not be exponentially distributed. Three semi-Markov models were considered for this example while considering the data as a 3-state process and 4-state process. The results suggested the semi-Markov model assuming a gamma wait time

distribution is the most appropriate mode for both cases. We can use this information to understand the how the stress-level behaves in a cohort of Alzheimer’s Disease caregivers. To build on these results, further investigation into other sojourn time distributions are needed to find other models that can consider bi-modal distributions like those seen in figure 2.1 and 2.2. For illustration purposes, we graphed the hazard of the semi-Markov process for a 3-state and 4 state process in figure 2.5 and 2.6, respectively. We observed that a participant who is caring for an Alzheimer’s disease patient is at higher risk of transitioning to a higher stress level than progressing back.

While we used a stress-level example in this paper, the partial likelihood approach can be applied to any longitudinal categorical outcome data. Our method can handle 3 or 4 state processes with the ability to use a continuous-time semi-Markov model. The model and approach can assume three different parametric distributions: exponential, Weibull, or Gamma. This allows the partial likelihood approach to be applicable to many different public health areas.

Table 2.1: Simulation Results for a three-state Semi-Markov Model

	True	Estimate	Bias	SD	SE	MSE	95% CP
<i>Exponential Sojourn Time¹</i>							
λ_{12}	0.47	0.4716	0.0016	0.0302	0.0225	0.0009	0.841
λ_{13}	0.68	0.6835	0.0035	0.0463	0.0356	0.0022	0.854
λ_{21}	0.49	0.4914	0.0014	0.0289	0.0225	0.0008	0.863
λ_{23}	0.63	0.6316	0.0016	0.0288	0.0215	0.0008	0.851
λ_{31}	0.52	0.5237	0.0037	0.0394	0.0311	0.0016	0.878
λ_{32}	0.63	0.6312	0.0012	0.0271	0.0205	0.0007	0.857
<i>Weibull Sojourn Time²</i>							
λ_{12}	1.9	1.9334	0.0334	0.0666	0.0702	0.0056	0.857
λ_{13}	1.3	1.3412	0.0412	0.0574	0.0672	0.0050	0.803
λ_{21}	1.8	1.8340	0.0340	0.0638	0.0656	0.0052	0.830
λ_{23}	1.4	1.4378	0.0378	0.0440	0.0549	0.0034	0.786
λ_{31}	1.7	1.7359	0.0359	0.0714	0.0767	0.0064	0.862
λ_{32}	1.4	1.4367	0.0367	0.0411	0.0544	0.0030	0.786
k	2.0	2.0012	0.0012	0.0396	0.0333	0.0016	0.895
<i>Gamma Sojourn Time³</i>							
λ_{12}	1.90	1.9631	0.0631	0.1642	0.1638	0.0309	0.874
λ_{13}	1.30	1.3618	0.0618	0.1288	0.1282	0.0204	0.849
λ_{21}	1.80	1.8638	0.0638	0.1616	0.1482	0.0302	0.830
λ_{23}	1.40	1.4556	0.0556	0.0966	0.1052	0.0124	0.849
λ_{31}	1.70	1.7701	0.0701	0.1794	0.1780	0.0371	0.882
λ_{32}	1.40	1.4567	0.0567	0.0932	0.1044	0.0119	0.843
ψ	0.89	0.8926	0.0026	0.0324	0.0272	0.0011	0.875

¹ 300 subjects in each simulation for 10 time units long; On 20 cores, computation time was 10.81 hours.

² 200 subjects in each simulation for 5 time units long; On 20 cores, computation time was 27.17 hours.

³ 150 subjects in each simulation for 5 time units long; On 20 cores, computation time was 25.31 hours.

Table 2.2: Simulation Results for a Four-State Semi-Markov Model

	True	Estimate	Bias	SD	SE	MSE	95% CP
<i>Exponential Sojourn Time¹</i>							
λ_{12}	0.47	0.4756	0.0056	0.0425	0.0343	0.0018	0.858
λ_{13}	0.68	0.6816	0.0016	0.0476	0.0380	0.0023	0.870
λ_{14}	0.81	0.8124	0.0024	0.0573	0.0421	0.0033	0.857
λ_{21}	0.49	0.4935	0.0035	0.0452	0.0342	0.0021	0.864
λ_{23}	0.63	0.6352	0.0052	0.0505	0.0384	0.0026	0.868
λ_{24}	0.74	0.7461	0.0061	0.0565	0.0437	0.0032	0.851
λ_{31}	0.52	0.5241	0.0041	0.0397	0.0308	0.0016	0.871
λ_{32}	0.63	0.6329	0.0029	0.0442	0.0337	0.0020	0.863
λ_{34}	0.39	0.3935	0.0035	0.0362	0.0279	0.0013	0.857
λ_{41}	0.74	0.7435	0.0035	0.0475	0.0385	0.0023	0.875
λ_{42}	0.44	0.4430	0.0030	0.0375	0.0290	0.0014	0.863
λ_{43}	0.49	0.4925	0.0025	0.0388	0.0298	0.0015	0.856
<i>Weibull Sojourn Time²</i>							
λ_{12}	1.9	1.9512	0.0512	0.1094	0.1105	0.0146	0.846
λ_{13}	1.3	1.3402	0.0402	0.0598	0.0699	0.0052	0.836
λ_{14}	1.1	1.1390	0.0390	0.0506	0.0638	0.0041	0.823
λ_{21}	1.8	1.8527	0.0527	0.0992	0.1060	0.0126	0.862
λ_{23}	1.4	1.4518	0.0518	0.0679	0.0811	0.0073	0.815
λ_{24}	1.2	1.2486	0.0486	0.0577	0.0751	0.0057	0.810
λ_{31}	1.7	1.7626	0.0626	0.0842	0.0977	0.0110	0.797
λ_{32}	1.4	1.4554	0.0554	0.0663	0.0843	0.0075	0.800
λ_{34}	2.3	2.3532	0.0532	0.1314	0.1303	0.0201	0.864
λ_{41}	1.2	1.2487	0.0487	0.0552	0.0717	0.0054	0.815
λ_{42}	2.0	2.0550	0.0550	0.1022	0.1100	0.0135	0.866
λ_{43}	1.8	1.8587	0.0587	0.0907	0.0987	0.0117	0.817

Table 2.2: Simulation Results for a Four-State Semi-Markov Model (*continued*)

	True	Estimate	Bias	SD	SE	MSE	95% CP
k	2.0	1.9950	-0.0050	0.0423	0.0305	0.0018	0.887
<i>Gamma Sojourn Time</i> ³							
λ_{12}	1.90	2.0232	0.1232	0.3160	0.2879	0.1150	0.874
λ_{13}	1.30	1.4129	0.1129	0.1780	0.1914	0.0444	0.850
λ_{14}	1.10	1.1994	0.0994	0.1421	0.1668	0.0301	0.834
λ_{21}	1.80	1.9116	0.1116	0.2808	0.2723	0.0913	0.890
λ_{23}	1.40	1.5187	0.1187	0.1995	0.2070	0.0539	0.866
λ_{24}	1.20	1.3080	0.1080	0.1568	0.1846	0.0362	0.853
λ_{31}	1.70	1.8304	0.1304	0.2371	0.2419	0.0732	0.866
λ_{32}	1.40	1.5219	0.1219	0.1813	0.2043	0.0477	0.863
λ_{34}	2.30	2.4513	0.1513	0.3752	0.3543	0.1637	0.889
λ_{41}	1.20	1.3069	0.1069	0.1571	0.1734	0.0361	0.837
λ_{42}	2.00	2.1240	0.1240	0.3123	0.3005	0.1129	0.900
λ_{43}	1.80	1.9201	0.1201	0.2524	0.2585	0.0781	0.869
ψ	0.89	0.8979	0.0079	0.0416	0.0369	0.0018	0.886

¹ 300 subjects in each simulation for 10 time units long; On 20 cores, computation time was 13.60 hours.

² 100 subjects in each simulation for 10 time units long; On 30 cores, computation time was 25.86 hours.

³ 100 subjects in each simulation for 5 time units long; On 30 cores, computation time was 24.24 hours.

Table 2.3: Simulation Comparison between Estimating 3-State Markov Chain using Full Likelihood Approach and Partial Likelihood Approach

	<i>Full Likelihood Approach</i> ¹						<i>Partial Likelihood Approach</i>						
	True	Estimate	Bias	SD	SE²	MSE	95% CP	Estimate	Bias	SD	SE³	MSE	95% CP
λ_{12}	0.47	0.4741	0.0041	0.0666	0.0661	0.0044	0.9464	0.4716	0.0016	0.0302	0.0225	0.0009	0.841
λ_{13}	0.68	0.6795	-0.0005	0.0760	0.0749	0.0058	0.9503	0.6835	0.0035	0.0463	0.0356	0.0022	0.854
λ_{21}	0.49	0.4943	0.0043	0.0629	0.0623	0.0040	0.9513	0.4914	0.0014	0.0289	0.0225	0.0008	0.863
λ_{23}	0.63	0.6287	-0.0013	0.0712	0.0708	0.0051	0.9593	0.6316	0.0016	0.0288	0.0215	0.0008	0.851
λ_{31}	0.52	0.5230	0.0030	0.0630	0.0627	0.0040	0.9603	0.5237	0.0037	0.0394	0.0311	0.0016	0.878
λ_{32}	0.63	0.6253	-0.0047	0.0601	0.0660	0.0036	0.9672	0.6312	0.0012	0.0271	0.0205	0.0007	0.857

¹ The full likelihood approach uses the multi-state model (msm) in R to obtain the estimates

² The asymptotic standard errors of the full likelihood was calculated from the Hessian

³ The standard error of the partial likelihood was calculated from non-parametric bootstrap samples

Table 2.4: Simulation Comparison between Estimating 4-State Markov Chain using Full Likelihood Approach and Partial Likelihood Approach

	<i>Full Likelihood Approach</i> ¹							<i>Partial Likelihood Approach</i>						
	True	Estimate	Bias	SD	SE²	MSE	95% CP	Estimate	Bias	SD	SE³	MSE	95% CP	
λ_{12}	0.47	0.4789	0.0089	0.1553	0.1615	0.0242	0.9611	0.4756	0.0056	0.0425	0.0343	0.0018	0.858	
λ_{13}	0.68	0.6838	0.0038	0.1536	0.1575	0.0236	0.9750	0.6816	0.0016	0.0476	0.0380	0.0023	0.870	
λ_{14}	0.81	0.7940	-0.0160	0.1683	0.1771	0.0286	0.9750	0.8124	0.0024	0.0573	0.0421	0.0033	0.857	
λ_{21}	0.49	0.4978	0.0078	0.1651	0.1708	0.0273	0.9661	0.4935	0.0035	0.0452	0.0342	0.0021	0.864	
λ_{23}	0.63	0.6237	-0.0063	0.1458	0.1490	0.0213	0.9780	0.6352	0.0052	0.0505	0.0384	0.0026	0.868	
λ_{24}	0.74	0.7416	0.0016	0.1588	0.1660	0.0252	0.9741	0.7461	0.0061	0.0565	0.0437	0.0032	0.851	
λ_{31}	0.52	0.5306	0.0106	0.1280	0.1321	0.0165	0.9731	0.5241	0.0041	0.0397	0.0308	0.0016	0.871	
λ_{32}	0.63	0.6226	-0.0074	0.1204	0.1241	0.0145	0.9750	0.6329	0.0029	0.0442	0.0337	0.0020	0.863	
λ_{34}	0.39	0.3865	-0.0035	0.1196	0.1211	0.0143	0.9571	0.3935	0.0035	0.0362	0.0279	0.0013	0.857	
λ_{41}	0.74	0.7171	-0.0229	0.1439	0.1522	0.0212	0.9711	0.7435	0.0035	0.0475	0.0385	0.0023	0.875	
λ_{42}	0.44	0.4480	0.0080	0.1237	0.1289	0.0154	0.9750	0.4430	0.0030	0.0375	0.0290	0.0014	0.863	
λ_{43}	0.49	0.4933	0.0033	0.1227	0.1232	0.0151	0.9591	0.4925	0.0025	0.0388	0.0298	0.0015	0.856	

¹ The full likelihood approach uses the multi-state model (msm) in R to obtain the estimates

² The asymptotic standard errors of the full likelihood was calculated from the Hessian

³ The standard error of the partial likelihood was calculated from non-parametric bootstrap samples

Table 2.5: Observed Transitions between 3-Levels of Caregiver Stress

From State	Stress Level	To State		
		<i>1</i>	<i>2</i>	<i>3</i>
<i>1</i>	<i>None/Mild</i>	0	303	49
<i>2</i>	<i>Moderate</i>	234	0	205
<i>3</i>	<i>Severe</i>	61	170	0

Table 2.6: Observed Transitions between 4-Levels of Caregiver Stress

From State	Stress Level	To State			
		<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
<i>1</i>	<i>None</i>	0	78	48	8
<i>2</i>	<i>Mild</i>	63	0	255	41
<i>3</i>	<i>Moderate</i>	40	194	0	205
<i>4</i>	<i>Severe</i>	18	42	170	0

Table 2.7: Alzheimer's Disease Caregiver 3-Level Stress Model Estimates

	Estimate	Sojourn Time	95% Bootstrap CI
<i>Exponential Sojourn Time</i>			
λ_{12}	0.4625	2.1620	(0.4235, 0.5015)
λ_{13}	0.5463	1.8209	(0.452, 0.6406)
λ_{21}	0.5701	1.7540	(0.5187, 0.6216)
λ_{23}	0.6001	1.6664	(0.5478, 0.6524)
λ_{31}	0.5921	1.6888	(0.5249, 0.6593)
λ_{32}	0.4643	2.1539	(0.406, 0.5225)
<i>Weibull Sojourn Time</i>			
λ_{12}	0.7497	1.2076	(0.6303, 0.8692)
λ_{13}	0.8195	1.1094	(0.6562, 0.9827)
λ_{21}	0.9241	0.9798	(0.7672, 1.081)
λ_{23}	0.9433	0.9598	(0.7865, 1.1001)
λ_{31}	0.8481	1.0675	(0.7136, 0.9827)
λ_{32}	0.7946	1.1394	(0.6525, 0.9368)
k	1.4658	-	(1.3811,1.5504)
<i>Gamma Sojourn Time</i>			
λ_{12}	1.7269	1.6235	(1.5136, 1.9402)
λ_{13}	1.9081	1.4693	(1.5782, 2.238)
λ_{21}	2.1542	1.3015	(1.9226, 2.3858)
λ_{23}	2.2574	1.2420	(2.0052, 2.5096)
λ_{31}	2.0888	1.3422	(1.8161, 2.3615)
λ_{32}	1.9538	1.4350	(1.6852, 2.2224)
ψ	2.8036	-	(2.5828,3.0244)

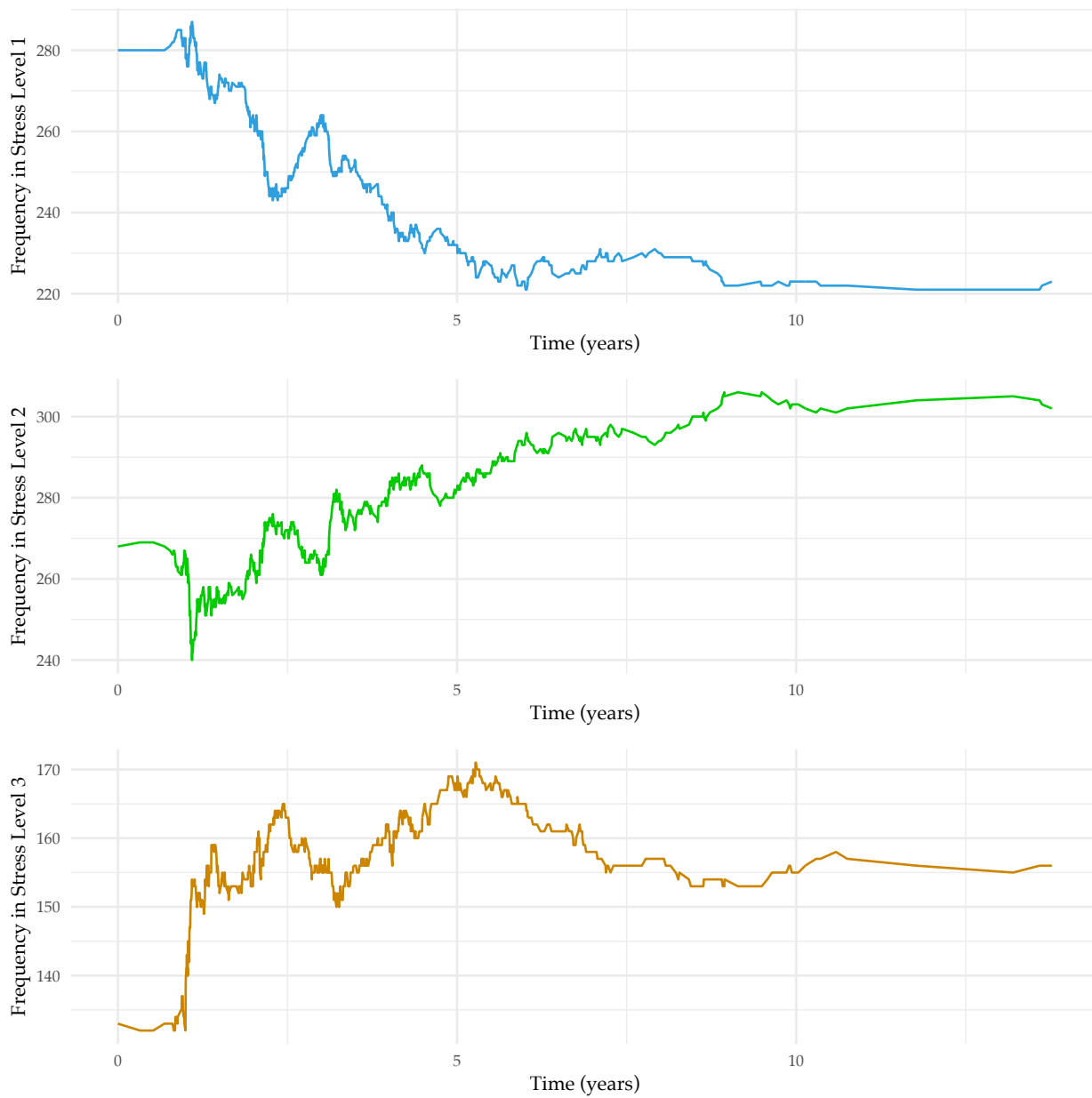


Figure 2.1: Frequency of Transition between 3 Levels of Caregiver of Stress over Time

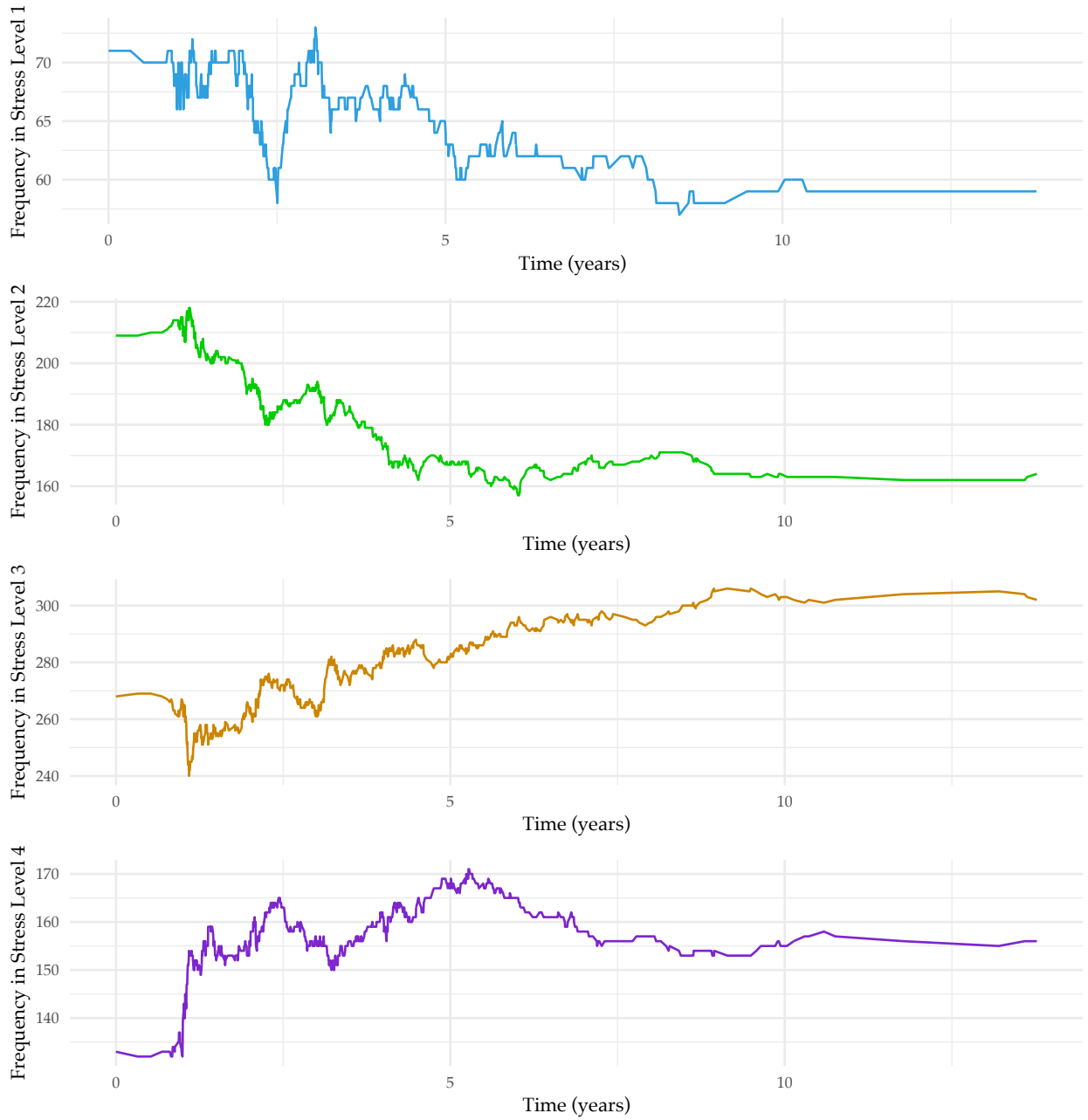


Figure 2.2: Frequency of Transition between 4 Levels of Caregiver of Stress over Time

Table 2.8: Alzheimer’s Disease Caregiver 4-Level Stress Model Estimates

	Estimate	Sojourn Time	95% Bootstrap CI
<i>Exponential Sojourn Time</i>			
λ_{12}	0.5414	1.8471	(0.4741, 0.6087)
λ_{13}	0.475	2.1051	(0.3917, 0.5584)
λ_{14}	0.4195	2.3840	(0.2454, 0.5935)
λ_{21}	0.6371	1.5696	(0.561, 0.7133)
λ_{23}	0.4952	2.0196	(0.4566, 0.5337)
λ_{24}	0.5956	1.6789	(0.5061, 0.6852)
λ_{31}	0.5347	1.8703	(0.4637, 0.6056)
λ_{32}	0.5382	1.8581	(0.4825, 0.5938)
λ_{34}	0.5675	1.7620	(0.5145, 0.6206)
λ_{41}	0.5691	1.7570	(0.4716, 0.6667)
λ_{42}	0.5554	1.8006	(0.4791, 0.6317)
λ_{43}	0.4359	2.294	(0.3866, 0.4853)
<i>Weibull Sojourn Time</i>			
λ_{12}	0.9958	0.9162	(0.7498, 1.2417)
λ_{13}	0.9780	0.9328	(0.7321, 1.2239)
λ_{14}	1.0542	0.8654	(0.6241, 1.4842)
λ_{21}	1.0617	0.8593	(0.8531, 1.2703)
λ_{23}	1.0014	0.9111	(0.7888, 1.214)
λ_{24}	1.0420	0.8755	(0.7743, 1.3098)
λ_{31}	1.0486	0.8701	(0.8568, 1.2403)
λ_{32}	1.0406	0.8767	(0.8716, 1.2096)
λ_{34}	1.0684	0.8539	(0.8802, 1.2565)
λ_{41}	1.0400	0.8772	(0.5748, 1.5052)
λ_{42}	1.1199	0.8146	(0.7745, 1.4654)
λ_{43}	1.0044	0.9083	(0.7702, 1.2386)

Table 2.8: Alzheimer's Disease Caregiver 4-Level Stress Model Estimates (*continued*)

	Estimate	Sojourn Time	95% Bootstrap CI
k	1.3913	-	(1.1955, 1.5872)
<i>Gamma Sojourn Time</i>			
λ_{12}	2.1626	1.4567	(1.7498, 2.5753)
λ_{13}	1.9590	1.6081	(1.5037, 2.4144)
λ_{14}	1.7527	1.7974	(-7.8255, 11.3308)
λ_{21}	2.4813	1.2696	(2.008, 2.9546)
λ_{23}	2.1800	1.4451	(1.7945, 2.5654)
λ_{24}	2.3765	1.3256	(1.8259, 2.9272)
λ_{31}	2.1908	1.4380	(1.7468, 2.6349)
λ_{32}	2.2356	1.4092	(1.7974, 2.6737)
λ_{34}	2.3162	1.3601	(1.8788, 2.7537)
λ_{41}	2.2922	1.3744	(1.5894, 2.995)
λ_{42}	2.3248	1.3551	(1.9415, 2.708)
λ_{43}	2.1246	1.4828	(1.7509, 2.4984)
ψ	3.1503	-	(2.7782, 3.5224)

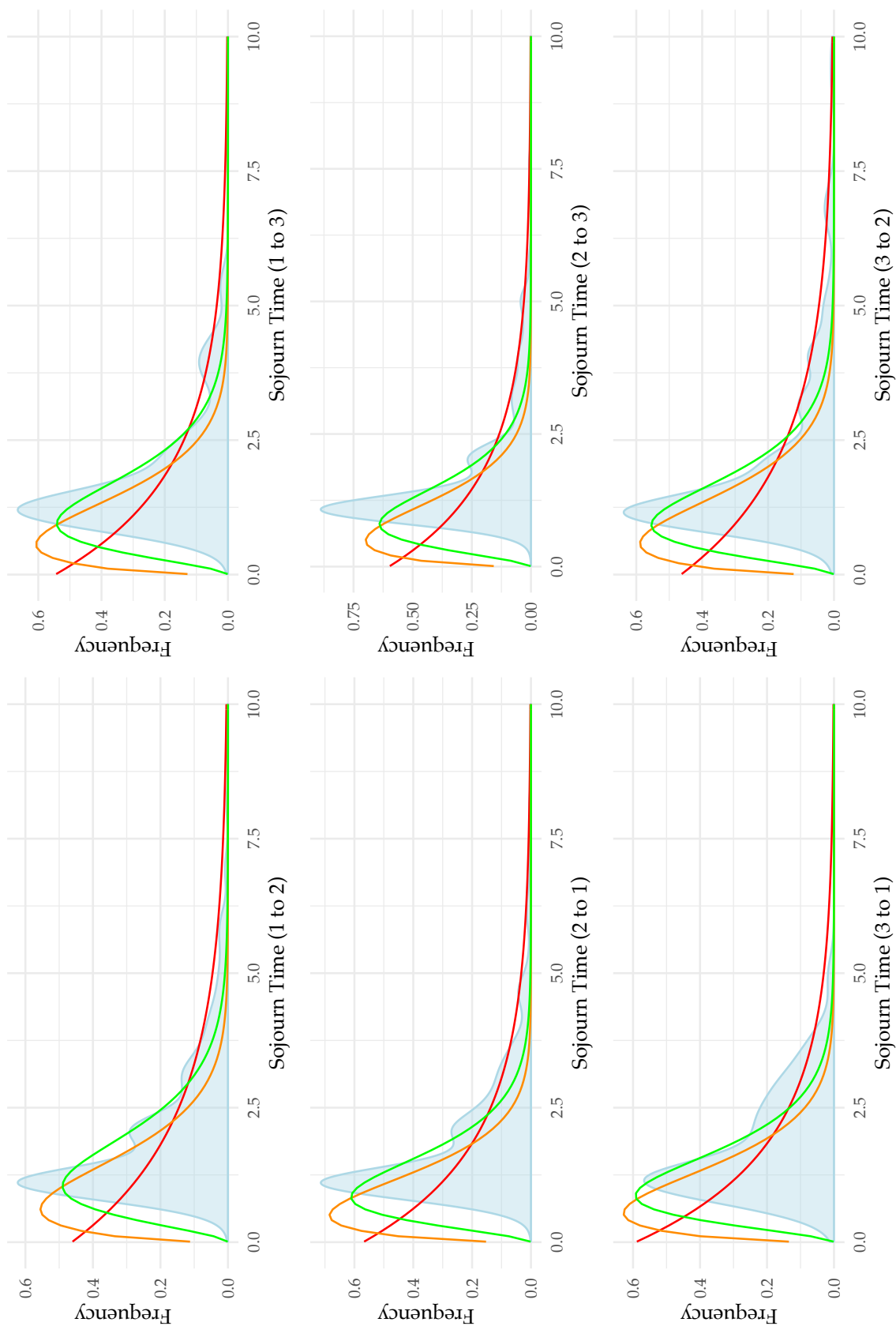


Figure 2.3: Sojourn Time by Caregiver Stress-Level (i to j) while Overlaying each Semi-Markov Model for a 3-State Process. The red line denotes the exponential distribution, the orange line represents the Weibull distribution, and the green line represents the gamma distribution.

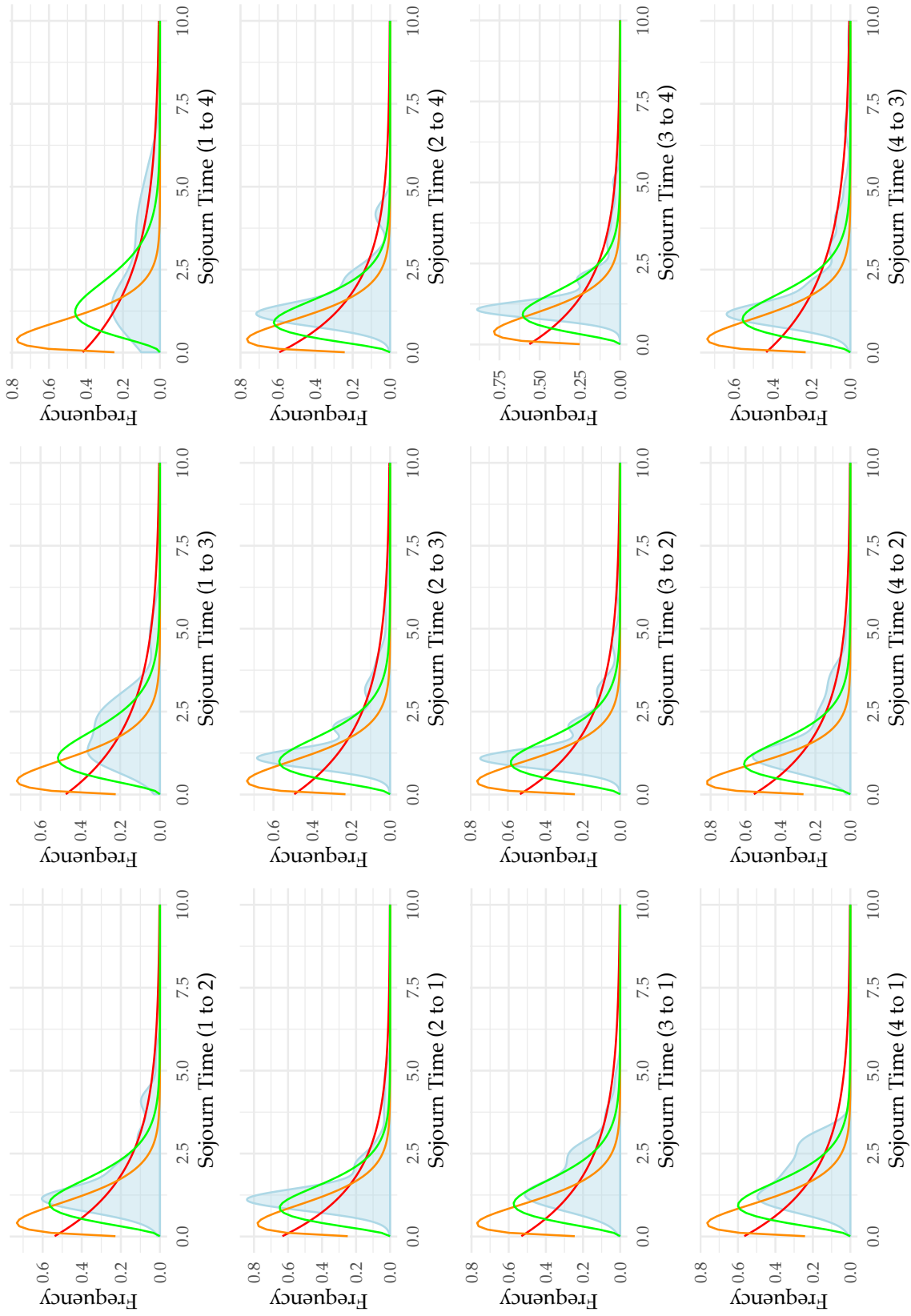


Figure 2.4: Sojourn Time by Caregiver Stress-Level (i to j) while Overlaying each Semi-Markov Model for a 4-State Process. The red line denotes the exponential distribution, the orange line represents the Weibull distribution, and the green line represents the gamma distribution.

Table 2.9: Estimated Transition Probabilities of the Embedded 3-State Markov Chain

From State	Stress Level	To State		
		<i>1</i>	<i>2</i>	<i>3</i>
<i>1</i>	<i>None/Mild</i>	0.000	0.861	0.139
<i>2</i>	<i>Moderate</i>	0.533	0.000	0.467
<i>3</i>	<i>Severe</i>	0.264	0.736	0.000

Table 2.10: Estimated Transition Probabilities of the Embedded 4-State Markov Chain

From State	Stress Level	To State			
		<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
<i>1</i>	<i>None</i>	0.000	0.582	0.358	0.060
<i>2</i>	<i>Mild</i>	0.175	0.000	0.710	0.114
<i>3</i>	<i>Moderate</i>	0.091	0.442	0.000	0.467
<i>4</i>	<i>Severe</i>	0.082	0.182	0.736	0.000

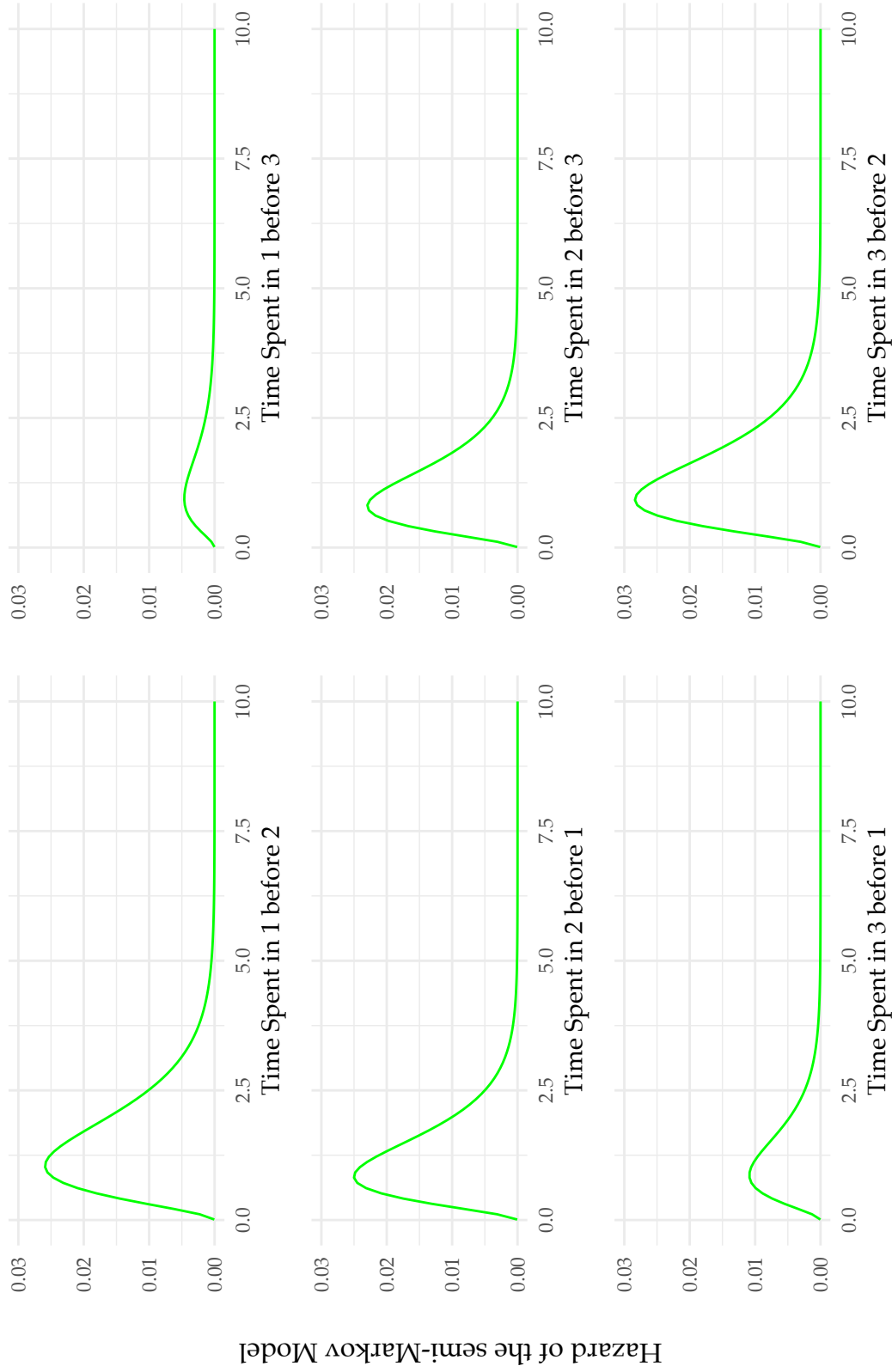


Figure 2.5: Plot of the Hazard of the Continuous-Time semi-Markov Process assuming the gamma sojourn time distribution. State 1, 2, and 3 represent not/mild stressed, moderately depressed, and severely depressed, respectively. Time is in years.

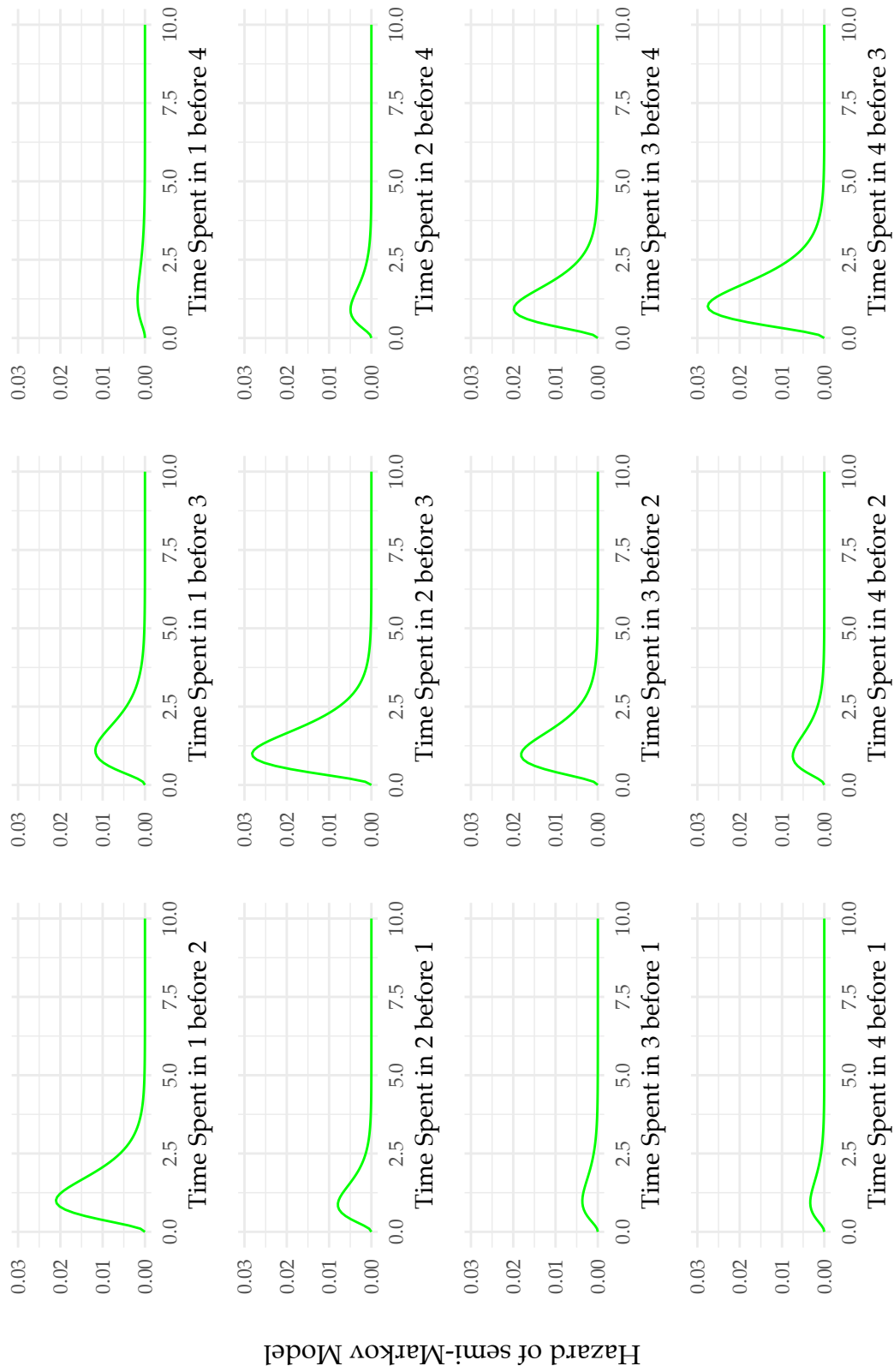


Figure 2.6: Plot of the Hazard of the Continuous-Time semi-Markov Process assuming the gamma sojourn time distribution. State 1, 2, 3, and 4 represent not stressed, mildly stressed, moderately stressed, and severely stressed, respectively. Time is in years.

References

- [2] H. H. Chen, S. W. Duffy, and L. Tabar, “A markov chain method to estimate the tumour progression rate from preclinical to clinical phase, sensitivity and positive predictive value for mammography in breast cancer screening,” *The Statistician*, vol. 45, no. 3, p. 307, 1996. DOI: 10.2307/2988469.
- [3] R. Perez-Ocón, J. E. Ruiz-Castro, and M. L. Gámiz-Perez, “A multivariate model to measure the effect of treatments in survival to breast cancer,” *Biometrical Journal*, vol. 40, no. 6, pp. 703–715, 1998. DOI: 10.1002/(sici)1521-4036(199810)40:6<703::aid-bimj703>3.0.co;2-7.
- [4] I. M. Longini, W. S. Clark, R. H. Byers, J. W. Ward, W. W. Darrow, G. F. Lemp, and H. W. Hethcote, “Statistical analysis of the stages of hiv infection using a markov model,” *Statistics in Medicine*, vol. 8, no. 7, pp. 831–843, 1989. DOI: 10.1002/sim.4780080708.
- [5] A. Alioum, V. Leroy, D. Commenges, F. Dabis, and R. Salmon, “Effect of gender, age, transmission category, and antiretroviral therapy on the progression of human immunodeficiency virus infection using multistate markov models,” *Epidemiology*, vol. 9, no. 6, pp. 605–612, 1998. DOI: 10.1097/00001648-199811000-00007.
- [6] I. Kousignian, B. Autran, C. Chouquet, V. Calvez, E. Gomard, C. Katlama, Y. Rivière, and D. Costagliola, “Markov modelling of changes in hiv-specific cytotoxic t-lymphocyte responses with time in untreated hiv-1 infected patients,” *Statistics in Medicine*, vol. 22, no. 10, pp. 1675–1690, 2003. DOI: 10.1002/sim.1404.
- [7] R. J. Kryscio, F. A. Schmitt, J. C. Salazar, M. S. Mendiondo, and W. R. Markesbery, “Risk factors for transitions from normal to mild cognitive impairment and dementia,” *Neurology*, vol. 66, no. 6, pp. 828–832, 2006. DOI: 10.1212/01.wnl.0000203264.71880.45.
- [8] D. C. Ewbank, “A multistate model of the genetic risk of alzheimers disease,” *Experimental Aging Research*, vol. 28, no. 4, pp. 477–499, 2002. DOI: 10.1080/03610730290103096.
- [9] P. Jepsen, H. Vilstrup, and P. K. Andersen, “The clinical course of cirrhosis: The importance of multistate models and competing risks analysis,” *Hepatology*, vol. 62, no. 1, pp. 292–302, 2015. DOI: 10.1002/hep.27598.
- [10] P. Saint-Pierre, C. Combescure, J. Daurès, and P. Godard, “The analysis of asthma control under a markov assumption with use of covariates,” *Statistics in Medicine*, vol. 22, no. 24, pp. 3755–3770, Aug. 2003. DOI: 10.1002/sim.1680.
- [11] A. Duffy, J. Horrocks, S. Doucette, C. Keown-Stoneman, S. Mccloskey, and P. Grof, “The developmental trajectory of bipolar disorder,” *British Journal of Psychiatry*, vol. 204, no. 2, pp. 122–128, 2014. DOI: 10.1192/bjp.bp.113.126706.
- [13] H. M. Taylor and S. Karlin, *An introduction to stochastic modeling*. Academic Press, 2014.

- [14] G. H. Weiss and M. Zelen, “A semi-markov model for clinical trials,” Jan. 1963. DOI: 10.21236/ad0407905.
- [22] Y. Foucher, E. Mathieu, P. Saint-Pierre, J.-F. Durand, and J.-P. Daurès, “A semi-markov model based on generalized weibull distribution with an illustration for hiv disease,” *Biometrical Journal*, vol. 47, no. 6, pp. 825–833, 2005. DOI: 10.1002/bimj.200410170.
- [24] Q. Cao, E. Buskens, T. Feenstra, H. Hillege, and D. Postmus, “Continuous-time semi-markov models in health economic decision making: An illustrative example in heart failure disease management,” *Med Decis Making*, vol. 36, no. 1, pp. 59–71, 2015. DOI: 10.1177/0272989X15593080.
- [30] J. S. Benoit, W. Chan, S. Luo, H.-W. Yeh, and R. Doody, “A hidden markov model approach to analyze longitudinal ternary outcomes when some observed states are possibly misclassified,” *Statistics in Medicine*, vol. 35, no. 9, pp. 1549–1557, 2016. DOI: 10.1002/sim.6861.
- [34] D. Eddelbuettel and J. J. Balamuta, “Extending extitR with extitC++: A Brief Introduction to extitRcpp,” *PeerJ Preprints*, vol. 5, e3188v1, Aug. 2017, ISSN: 2167-9843. DOI: 10.7287/peerj.preprints.3188v1. [Online]. Available: <https://doi.org/10.7287/peerj.preprints.3188v1>.
- [35] M. Corporation and S. Weston, *Doparallel: Foreach parallel adaptor for the 'parallel' package*, R package version 1.0.15, 2019. [Online]. Available: <https://CRAN.R-project.org/package=doParallel>.
- [36] R. B. Fetter and J. D. Thompson, “A decision model for the design and operation of a progressive patient care hospital,” *Medical Care*, vol. 7, no. 6, pp. 450–462, 1969. DOI: 10.1097/00005650-196911000-00004.
- [37] B. Ouhbi and N. Limnios, “Non-parametric estimation for semi-markov kernels with application to reliability analysis,” *Applied Stochastic Models and Data Analysis*, vol. 12, no. 4, pp. 209–220, 1996. DOI: 10.1002/(sici)1099-0747(199612)12:4<209::aid-asm284>3.0.co;2-t.
- [38] B. Ouhbi and N. Limnios, “Nonparametric estimation for semi-markov processes based on its hazard rate functions,” *Statistical Inference for Stochastic Processes*, vol. 2, pp. 151–173, 1999.
- [39] M. R. Sternberg and G. A. Satten, “Discrete-time nonparametric estimation for semi-markov models of chain-of-events data subject to interval censoring and truncation,” *Biometrics*, vol. 55, no. 2, pp. 514–522, 1999. DOI: 10.1111/j.0006-341x.1999.00514.x.
- [40] H. Damerджи, “Maximum likelihood estimation for generalized semi-markov processes,” *Discrete Event Dynamic Systems*, vol. 6, no. 1, pp. 73–104, 1996. DOI: 10.1007/bf01796784.

- [41] P. Joly and D. Commenges, “A penalized likelihood approach for a progressive three-state model with censored and truncated data: Application to aids,” *Biometrics*, vol. 55, no. 3, pp. 887–890, 1999. DOI: 10.1111/j.0006-341x.1999.00887.x.
- [42] D. R. Cox, “Regression models and life-tables,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972. DOI: 10.1111/j.2517-6161.1972.tb00899.x.
- [43] N. Breslow, “Covariance analysis of censored survival data,” *Biometrics*, vol. 30, no. 1, p. 89, 1974. DOI: 10.2307/2529620.
- [44] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019. [Online]. Available: <https://www.R-project.org/>.
- [45] C. H. Jackson, “Multi-state models for panel data: The msm package for R,” *Journal of Statistical Software*, vol. 38, no. 8, pp. 1–29, 2011. [Online]. Available: <http://www.jstatsoft.org/v38/i08/>.

Chapter 3

A Continuous-Time Semi-Markov Model for Longitudinal Categorical Outcome with predictors: A Partial Likelihood Approach

Authors: Kusha A. Mohammadi, Wenyaw Chan, and Valory Pavlik

3.1 Abstract

A continuous-time semi-Markov models (CTSMM) can be utilized as an alternative to studying longitudinal categorical outcomes to the classic transition model in cases where the Markov assumption is too restrictive or unrealistic. Often longitudinal studies collect subject covariate information to potentially better explain the outcome distributional changes over time. However, when we consider a three-state semi-Markov processes (SMP), we are limited to the statistical approaches to estimate the transition covariate effects under a semi-Markov model. To address this issue, we develop a partial likelihood approach to incorporate predictors to evaluate the transition covariate effects while considering various sojourn-time distributions: exponential, gamma, and Weibull. This method contributes to statistical inference in the area of semi-Markov models and provides a computationally feasible approach to study a breadth of longitudinal appli-

cations. We assessed the proposed method through extensive simulation studies and examined their sensitivities. The simulation results suggest accurate estimation with low bias of the transition effects for a CTSM and coverage probability close to the expected 95%. We applied our partial likelihood approach to a longitudinal example in which the care-giver stress-level over time are used as outcome while incorporating some predictors.

Keywords: CTSM, Covariate Transition Effects, Longitudinal Categorical Outcomes, Partial Likelihood

3.2 Introduction

In many applications, continuous-time semi-Markov model (CTSM) is a powerful tool because it relaxes the strict Markov assumption. The Markov property implies that the holding time has an exponential distribution. Fetter and Thompson revealed the movement through various health states for an individual may not be Markovian for many diseases [36]. For this reason, the semi-Markov framework is often considered to allow for arbitrary sojourn time distributions.

Statistical inference in the area of semi-Markov models continues to grow as more complex problems arise. Anderson and others proposed a Cox semi-Markov model to add covariate effects to each transition intensity for an application in bleeding episodes and mortality in liver cirrhosis [26]. Titman presented a new statistical likelihood method to estimate transition rates from panel data using phase-type approximations [27]. Shu and others utilized large sample theory to develop asymptotic theory for the Cox semi-Markov model to investigate the robustness and efficiency of semi-Markov estimators [28]. Aralis and Brookmeyer proposed a stochastic estimation procedure for panel observation data with back transitions while assuming a non-exponential distribution [29]. Yu has also extended the semi-Markov theory to consider misclassification in observed states called the hidden semi-Markov model (HSMM) [25]. While all these examples contributed

greatly to the stochastic literature, there is still a continual need for the development of efficient estimators and computationally feasible methods to study multi-state semi-Markov processes while incorporating covariate information.

In this article, we propose partial likelihood method to estimate the transition covariate effects under a semi-Markov model. Specifically, we will analyze a three categorical disease-state process while adjusting for some covariates over time. For each CTSM, we will assume three wait time distributions: exponential (i.e. Markov model), gamma, and Weibull. The gamma and Weibull distributions are widely used in classical survival analysis because they generalize the exponential distribution. By using these distributions, we have a more flexible model to analyze a variety of longitudinal categorical disease problems. Additionally, our proposed estimation procedure will provide critical information that can be used in studying dynamic disease/interventions in medical research. To highlight our method, we will apply the partial likelihood approach to an Alzheimer’s caregiver stress-level example after controlling for some covariates.

The remainder of the paper is organized as follows. Section 3.3 defines the semi-Markov process and outlines the partial likelihood method that incorporates subject covariate information. Extensive simulation studies are summarized in section 3.4 and applied to 3-level outcome of care-giver stress in section 3.5. The paper concludes with a discussion in section 3.6.

3.3 Methods

3.3.1 The semi-Markov Model

First, let’s consider a Markov renewal process (X_n, T_n) where $0 = T_0 < T_1 < \dots < T_D$ are consecutive state transition time points to states X_n for D total number of transitions. $S = (S_n)_{n \in \mathbb{N}}$ is defined as the successive holding times in the visited states. The sequence $X = (\{X_n\})$ forms an embedded discrete-time homogeneous Markov chain for a discrete

state space, $\Phi = \{1, 2, \dots, b\}$. Given the initial distribution, $\omega_i = P(X_0 = i)$, $i \in \Phi$, the probability of moving from a state i to state j is $p_{ij} = P(X_{n+1} = j | X_n = i)$, $p_{ij} > 0$, for $i \neq j$ and $p_{ij} = 0$ for $i = j$. The semi-Markov kernel, Q_{ij} , satisfies the following

$$\begin{aligned} Q_{ij}(t) &= P(X_{n+1} = j, t \leq S_{n+1} | \Lambda_{n-1}) \\ &= P(X_{n+1} = j, t \leq S_{n+1} | X_n = i) \end{aligned} \quad (3.1)$$

where $\Lambda_n = \{(X_0, T_0); \dots; (X_n, T_n)\}$ denotes the history of the semi-Markov chain, $i, j \in \Phi$, and $t \in \mathbb{R}^+$.

The distribution function of the sojourn time, F_{ij} , determines the amount of time $t \in \mathbb{R}^+$ an individual stays in state i before transitioning to state j :

$$F_{ij}(t) = \lim_{\Delta t \rightarrow 0^+} \frac{P[t \leq S_{n+1} < t + \Delta t | X_{n+1} = j, X_n = i]}{\Delta t} \quad (3.2)$$

where $i, j \in \Phi$, and $t \in \mathbb{R}^+$. We can relate the semi-Markov kernel to the distribution function of the holding time through the transition probabilities:

$$Q_{ij}(t) = p_{ij} F_{ij}(t) \quad (3.3)$$

where $i, j \in \Phi$. Using the classical survival relations, we can deduce the hazard of the semi-Markov process which is the probability of moving to a state j between time t and $t + \Delta t$, given the previous state is i for a duration t ,

$$\begin{aligned} h_{ij}(t) &= \lim_{\Delta t \rightarrow 0^+} \frac{P(X_{n+1} = j, t \leq S_{n+1} < t + \Delta t | S_{n+1} \geq t, X_n = i)}{\Delta t} \\ &= \frac{p_{ij} S_{ij}(t) \nu_{ij}(t)}{S_i(t)} \end{aligned} \quad (3.4)$$

where $i \neq j$, $i, j \in \Phi$, $h_{ii}(t) = -\sum_{j \neq i} h_{ij}(t)$. The survival and hazard of the sojourn time is denoted by $S_{ij}(t)$ and $\nu_{ij}(t)$. Additionally, the survival function of the

wait time, $S_i(t)$, is defined as $\sum_{j \in \Phi} p_{ij}(1 - F_{ij}(t))$.

3.3.2 Incorporation of Covariates

To add the influence of covariates on the holding time distributions, we utilize a cox proportional hazard model for p ($p = 1, \dots, k$) explanatory variables with some known functional form of the covariates, $\psi(\cdot)$ [42]. Let $T_{(1)} < \dots < T_{(D)}$ be the ordered transition times of all the m individuals, $m = 1, \dots, M$. Then, let $\mathbf{Z}_{ij}^{(n)}(t)$ be a vector representing the individual's covariate information at n^{th} transition from i to j at time t , $i, j \in \Phi$. The general form for the hazard rates function while accounting for covariates is as follows,

$$\nu_{ij}(t|\mathbf{Z}_{ij}^{(n)}(t)) = \nu_{0,ij}(t)\psi(\mathbf{Z}_{ij}^{(n)}(t))$$

From this general form, we will make few assumptions:

1. The proportionality of hazards holds within each i to j state transition but does not hold between.
2. The vector of covariate effects, $\boldsymbol{\beta}$, is the same across all $i \rightarrow j$ transitions (i.e. $\boldsymbol{\beta}_{ij} = \boldsymbol{\beta}$; $\mathbf{Z}_{ij}^{(n)}(t) = \mathbf{Z}^{(n)}(t)$)
3. The covariates are independent of the transition time, t (i.e. $\mathbf{Z}^{(n)}(t) = \mathbf{Z}^{(n)}$).
4. The same baseline intensity distribution, $\nu_{0,ij}(t)$, is assumed for each state transition from i to j (e.g. Weibull distribution for $1 \rightarrow 2$, $1 \rightarrow 3$, etc.).

Under these assumptions, we have the following simplified hazard model that integrate p predictors,

$$\nu_{ij}(t|\mathbf{Z}^{(n)}) = \nu_{0,ij}(t)\exp(\boldsymbol{\beta}^t \mathbf{Z}^{(n)}) \tag{3.5}$$

From equation 3.5, the regression coefficients have the well-known interpretation of relative risk given the assumptions hold. However, the covariate effects on the semi-Markov hazard function (i.e. equation 3.4) will be interpreted graphically due to its intricacy.

3.3.3 Distributions of the Sojourn Time

We studied three different sojourn distributions in continuous-time semi-Markov model. For simplicity, we will assume that the shape parameters (i.e. k, ψ) is constant across all transitions from i to j . The simplest distribution is the exponential distribution which has constant hazard over time (equivalent to the Markov model) with a positive rate parameter, $\lambda_{ij}, i, j \in \Phi$.

$$\nu_{ij}(t) = \lambda_{ij} \tag{3.6}$$

Secondly, the gamma distribution has the flexibility of dealing with many different distribution shapes in practice. This generalized form of the exponential distribution defines the hazard function of the waiting time with positive rate parameter, $\lambda_{ij}, i, j \in \Phi$, and positive shape parameter, ξ as

$$\nu_{ij}(t) = \frac{\lambda_{ij}^{\xi} t^{\xi-1} e^{-\lambda_{ij} t}}{\Gamma(\xi) - \Gamma(\xi, \lambda_{ij} t)} \tag{3.7}$$

where $\Gamma(a)$ is the gamma function, and $\Gamma(a, x)$ is the incomplete gamma function, $a > 0$.

Lastly, the Weibull distribution generalizes exponential case by allowing a second parameter to alter the shape of the distribution. This adaptable feature has allowed this distribution to be used in many practical applications. The hazard of a two-parameter Weibull distribution with positive rate parameter, $\lambda_{ij}, i, j \in \Phi$, and positive shape parameter, k is as follows

$$\nu_{ij}(t) = k\lambda_{ij}t^{k-1} \quad (3.8)$$

All these distributions allow the semi-Markov model to be adaptable to a broad set of longitudinal categorical studies.

3.3.4 The Partial Likelihood by Adding Covariates

We collect data based on the triple $(T^{n_m}, X^{n_m}, \mathbf{Z}^{n_m})$ for n_m transitions for the m^{th} subject, $m = 1, \dots, M$. Let $(T^{(n)}, X^{(n)}, \mathbf{Z}^{(n)})$ be the ordered data based on the transition times, t , where $T^{(n)}$ is the n^{th} transition time, $X^{(n)}$ is the transitioning state (i.e. state j), and $\mathbf{Z}^{(n)}$ is the vector of covariate information at the n^{th} transition. The ordered data combines all M subjects by their transition time. For convenience, we will let $\tau = T^{(n)}$. We define the risk set, $\mathcal{R}(\tau-)$, as the set of all individuals who are still under study at a time prior to τ . We denote, $I_{X_l(\tau-)}(u)$, to identify the current state, u , of subject l prior to transition time τ . Let $S^{(n)}$ be the time spent in a particular state, i , before transitioning to state j (i.e. sojourn time) for the n^{th} transition. Let φ be the time already spent in a particular state, u , for a subject l . For each possible n transitions, the probability that there is a n^{th} transition ($i \rightarrow j$) at time τ with covariates $\mathbf{Z}^{(n)}$ given that one subject transition in the risk set at that time is

$$\frac{h_{ij}(S^{(n)}|\mathbf{Z}^{(n)})}{\sum_{l \in \mathcal{R}(\tau-)} \sum_{u \in \Phi} h_u(\varphi|\mathbf{Z}^{(l)}) I_{X_l(\tau-)}(u)} \quad (3.9)$$

where $h_{ij}(\cdot|\cdot)$ is transition rate function defined by equation 3.4.

Let the parameters of interest be defined by $\Theta = \{\boldsymbol{\beta}, \lambda_{ij}, \xi^*, k^*\}$ where $*$ indicates the parameter that needs to be estimated depending on the sojourn distribution (defined in Section 3.3.3). Then the partial likelihood is formed by multiplying all the conditional probabilities over all the transitions D . This is given by

$$L(\Theta) = \prod_{n=1}^D \frac{h_{ij}(S^{(n)}|\mathbf{Z}^{(n)}, \Theta)}{\sum_{l \in \mathcal{R}(\tau-)} \sum_{u \in \Phi} h_u(\varphi|\mathbf{Z}^{(l)}, \Theta) I_{X_l(\tau-)}(u)} \quad (3.10)$$

where $i, j, u \in \Phi$, $h_u(t) = -\sum_{j \neq i} h_{ij}(t)$. The numerator of the likelihood function depends only on the individual's explanatory variables who is currently transitioning from state $i \rightarrow j$. The denominator of the likelihood includes all the information of the subjects who are still at risk prior to time τ .

A special case to consider is when there are ties present. In practice, it is common to collect many subject's information at a common calendar time (e.g. every year). We will consider one method constructed by Breslow [43] where there are ties among the events. The partial likelihood can be expressed as

$$L(\Theta) = \prod_{n=1}^D \frac{\prod_{g \in d_n} h_{i_g j_g}(S_g^{(n)}|\mathbf{Z}^{(n)}, \Theta)}{\left[\sum_{l \in \mathcal{R}(\tau-)} \sum_{u \in \Phi} h_u(\varphi|\mathbf{Z}^{(l)}, \Theta) I_{X_l(\tau-)}(u) \right]^{d_n}} \quad (3.11)$$

where d_n is the number of events at a given transition time, τ , for the n^{th} transition.

Parameter estimation can be carried out by optimizing the likelihood function or equivalently, the log likelihood, $l(\Theta)$,

$$l(\Theta) = \sum_{n=1}^D \sum_{g \in d_n} \log \left[h_{i_g j_g}(S_g^{(n)}|\mathbf{Z}^{(n)}, \Theta) \right] - \sum_{n=1}^D d_n \log \left[\sum_{l \in \mathcal{R}(\tau-)} \sum_{u \in \Phi} h_u(\varphi|\mathbf{Z}^{(l)}, \Theta) I_{X_l(\tau-)}(u) \right] \quad (3.12)$$

Ordinarily, the parameters of interest are acquired by deriving the first derivative of the likelihood function and setting it to zero. Due to the complex structure of the semi-Markov hazard function, the first derivative is not available in closed form and the

fisher’s information matrix is not easily computed. This presents a difficult optimization problem where we need a derivative-free numerical optimization approach to obtain the estimates for the CTSMM. Non-parametric bootstrap samples are used to estimate the standard errors for the parameters in Θ . For each bootstrap sample, M individuals were re-sampled with replacement and new estimates for the CTSMM were collected. All analysis used R 3.6.2 [44], Rcpp package [34], and doParallel package [35].

3.4 Simulation

To evaluate the performance of the proposed partial likelihood estimator, we describe a simulation study to examine a three-state CTSMM in this section. Three semi-Markov processes are simulated to represent the hold time distributions, $F_{i,j}$ outlined in subsection 3.3.3 (i.e. exponential, gamma, and Weibull). A general simulating algorithm for a semi-Markov process up to a time $t = T$ is given [25]:

1. Supply p_{ij} and F_{ij}
2. Choose an initial state, i_0 . Set $t = 0$.
3. Set $i = i_0$
4. Generate the following state, $j \sim p_{i_0,k}$
5. If $t < T$, then
 - a. Generate a sojourn time, $S_{i,j} \sim F_{i,j}$
 - b. Set $t = t + S_{i,j}$.
 - c. Set $i = j$;
 - d. Generate a new state $j \sim p_{i,k}$.
6. Else, stop.

In words, the semi-Markov process randomly determines the following state j based on the transition probabilities, p_{ij} , after entering a state i . Then it randomly determines the amount of time spent, S_{ij} , in a state i before transitioning to a state j

based on the sojourn time distribution, F_{ij} . The process continues until we reach a maximum observe time, T . For each simulation, the algorithm was run for 400 subjects and 1000 simulations. The standard errors were approximated by using 50 non-parametric bootstrap samples and used to calculate the 95% bootstrap confidence intervals. The hazard of semi-Markov process were dependent on one continuous (β_1) and one binary covariate, β_2 as described in subsection 3.3.2. The Nelder-Mead method was used to maximize the partial log likelihood function in equation 3.12. We assess our proposed method by the bias, standard deviation, standard error, mean-square error (MSE), and 95% probability coverage. Briefly, the bias describes the distance between the estimated value and the true value, and the standard deviation represents how close the numbers are to the mean. The standard error will explain how far the sample statistic deviates from the actual population parameter. In simulations, we would expect the standard deviation and standard errors to be relatively close to one another. The mean square error combines these two components of bias and variability to imply the mean difference between the estimated and observed parameters. Generally, a relatively low MSE value indicates a well-fitted model. Lastly, the 95% coverage probability refers to the number of times the true parameter is in the confidence interval. It is desirable to have coverage probabilities near 95% to indicate an efficient estimation method.

The simulations results for the three semi-Markov models were summarized in table 3.2. Among all three models, the bias remained relatively low (< 0.09) with the exception of one rate parameter, λ_{31} , in the SMM assuming a gamma sojourn time. The highest variability was observed in the estimation of the beta coefficients (β_1 and β_2) for the exponential and Weibull sojourn time. Most rates, shape, and beta coefficient estimates had relatively low MSE with the variance (i.e. standard deviation squared) driving higher values in some numbers. In the last column, the 95% bootstrap coverage probability is given. By assuming a Weibull or gamma wait time distribution, the results showed most of the estimates hovering around the expected 0.95 range. The Markov

model (i.e. exponential sojourn) exhibited some low coverage probabilities in 80 percent range and one in the 70 percent range. Low coverage indicate either biased estimates or anti-conservative standard errors and should be considered when assuming an exponential holding time distribution. Some estimates showed standard deviations much larger than the standard errors in the exponential and weibull sojourn time models.

3.5 Caregiver Stress Application

In this section, we describe how our method can be implemented in a caregiver stress-level example. The Baylor Alzheimer’s Disease and Memory Disorders Center recruited individuals to evaluate probable Alzheimer’s Disease using the criteria from the National Institute of Neurological and Communicative Disorders and Stroke [30]. Over a 21 year period, socio-demographic and neuro-psychological information was collected from participants to better understand the progression from a non-clinical neurological state to an Alzheimer’s Disease state. A second interest of the study was to focus on the health and well-being of the family members or friends who cared for the Alzheimer’s Disease patients. Caregivers were asked to fill out a questionnaire where self-reported stress-level was documented. No stress, mildly stressed, moderately stressed, and severely stressed were the four possible categories given to self-evaluate their stress level. Demographic characteristics were also gathered to describe and highlight possible differences between certain components. Using the data, we will conduct a longitudinal study where we treat the self-reported stress-level as a semi-Markov chain with three potential outcomes (i.e. None/mildly stressed, moderately stressed, and severely stressed). Additionally, patient sex and age at baseline were incorporated in the model to potentially better explain the behavior of caregiver stress level over time. Caregivers who have at least one transition and complete demographic information were included in the study. To find the best fit for the data, we calculate the Akaike’s information criterion (AIC) for each model and

choose the the model with the lowest AIC value.

Table 3.1 shows the number of transitions, $i \rightarrow j$, within the Alzheimer's Disease dataset. The lowest number of transitions occurred between severe to None/Mild and None/Mild to Severe (61 and 49 transitions, respectively). The most transitions seem to occur from moderate to none/mild and back to moderate (234 and 303 counts, respectively). To illustrate the changes in each stress level, we graphed the frequency of transitions by sex in figure 3.1. In the initial years of observation, females tend to be in the transition out of the lower two levels and move into the higher stress levels. For men, there was steady increase in the two highest levels as time progresses forward indicating some form of higher stress levels while caring for an Alzheimer's Disease patient.

In table 3.3, we presented the results of the continuous-time semi-Markov model that incorporated covariate effects. Three different sojourn time distributions were considered for the semi-Markov model. To find the most best model, the AIC were calculated for each model. The lowest AIC (11884.7) was found for the SMM assuming a gamma sojourn time compared to the exponential and Weibull sojourn (12330 and 12197.75, respectively). Regardless of the model, all suggest that the sex covariate effect (β_1) was statistically insignificant due to the fact the 95% bootstrap confidence interval contained zero. This suggests that the risk between males and females are the same and can be excluded from the model. The second coefficient, the age (β_2) effect, was found to be statistically significant. Since we standardized age, we would interpret the transition from a moderate to severe state in the following way: One standard deviation increase in age increases the risk of transitioning from a moderate to severe stress level by 1.59 (i.e. $\exp(0.4664)$). Since we assumed the beta coefficient to be the same across all levels, the interpretation is similar. To highlight the differences across age, figure 3.2 gives a visual context to the hazard of the semi-Markov process for various ages. Within each transition, we can conclude the risk of transitioning to a higher stress level is greater for older caregivers than younger ones within the first few years (i.e. compare the blue line

to green line).

Table 3.1: Observed Transitions between 3-Levels of Caregiver Stress

From State	Stress Level	To State		
		1	2	3
1	<i>None/Mild</i>	0	303	49
2	<i>Moderate</i>	234	0	205
3	<i>Severe</i>	61	170	0

3.6 Discussion

In this paper, we proposed a partial likelihood approach for a continuous-time semi-Markov chain model that includes covariate effects on the hazard function. The CTSMM helps us model the transitions between discrete disease states and allows us to examine how demographic or environmental factors affect the transition rates. An Alzheimer’s disease caregiver stress application was a natural example to exemplify the use of a CTSMM. Additionally, by using a CTSMM, we have the flexibility to specifying an arbitrary sojourn distribution. We explored three sojourn distributions: exponential (i.e. Markov model), Weibull, and gamma distributions. Both the Weibull and gamma are generalization of the exponential distribution with the added benefit of allowing a shape parameter to alter the distribution. By utilizing a cox hazard model, we can incorporate covariate effects on the hazard of transition. The partial likelihood method was then constructed and evaluated through some simulations. The simulation performance of the partial likelihood suggested relatively efficient estimation in models assuming a Weibull or gamma sojourn time distribution. Both models exhibited relatively low bias and mean square error. However, for some rates parameters, the standard deviations were much larger than the standard errors. From Chapter 2, one result suggested that the

partial likelihood approach may be unstable for some rate estimates due to some outlying datasets in the simulations. We reason that this is the source of the discrepancy between standard deviation and standard errors for some transition rates in these simulations. Additionally, all estimates showed coverage probabilities around the expected 95% range. All in all, the partial likelihood approach seems to be a reasonable method to estimate the parameters in a 3-level CTSM incorporating covariate effects on the hazard rates.

The longitudinal caregiver example demonstrates the type of analysis that can be conducted with a multi-level categorical outcome. We modeled 3 stress-levels (none/mild, moderate, and severe) as semi-Markov chain and investigated how the patients age and sex affect the transition rates. By explored three different holding time distributions, we were able to find the most appropriate model to fit the caregiver stress data. The AIC values suggested that the time until transitioning to another state was modeled best when assuming a gamma distribution than a Weibull or exponential distribution. This serves as an example where Markov model may not be realistic. As in any analysis, we can interpret the significant coefficients as relative risk by exponentiating the beta estimates and graph the hazard of the semi-Markov process as in figure 3.2. Our longitudinal analysis suggests that the age of a caregiver affects the transitioning rate through the three levels of stress.

The work presented are not without some limitations. First, we had a complex optimization problem such that we needed to find a non-linear optimizer for the log partial likelihood. A Nelder-Mead optimization was used for its ability to find the optimal parameters without the derivative. However, depending on the initial parameters, the partial log likelihood may converge to a local minima or not at all. Additional non-linear optimization methods need to be consider to understand how reliable the Nelder-Mead estimates are for this approach. Secondly, we limited our research to three parameter distributions. We used the language of 'most appropriate' model because it is not clear whether it is the best model. From our simulation results, we find non-exponential dis-

tribution performing relatively well with this partial likelihood approach. More research is needed to expand the number of distributions we can have to explore unique data distributions. For instance, the raw sojourn times in the caregiver stress example for each transition suggested a bi-modal shape than uni-modal. Although, we would argue this phenomenon occurred in this example because of the varied interview times and inability to capture the time spent in each stress level. Further research into these matters will develop the robustness of the partial likelihood approach.

In this paper, we proposed a partial likelihood method that incorporates covariate information on the the transition rates of a continuous-time semi-Markov chain. A natural extension of the proposed approach is to consider a four state process and explore other parametric and non-parametric sojourn times.

Table 3.2: Simulation Results for a Three-State Semi-Markov Model with the Inclusion of Covariates¹

	True	Estimate	Bias	SD	SE	MSE	95% Coverage
<i>Exponential</i>							
λ_{12}	0.47	0.4896	0.0196	0.1019	0.0728	0.0108	0.914
λ_{13}	0.68	0.6925	0.0125	0.1370	0.1075	0.0189	0.879
λ_{21}	0.49	0.5082	0.0182	0.1053	0.0775	0.0114	0.908
λ_{23}	0.63	0.6495	0.0195	0.3120	0.0706	0.0978	0.892
λ_{31}	0.52	0.5444	0.0244	0.1218	0.1045	0.0154	0.926
λ_{32}	0.63	0.6379	0.0079	0.0998	0.0642	0.0100	0.904
β_1	0.50	0.5941	0.0941	0.7790	0.0819	0.6157	0.769
β_2	1.00	1.0092	0.0092	0.3415	0.0828	0.1167	0.877
<i>Weibull</i>							
λ_{12}	0.47	0.4736	0.0036	0.0787	0.0549	0.0062	0.942
λ_{13}	0.68	0.6865	0.0065	0.1877	0.0838	0.0353	0.933
λ_{21}	0.49	0.5032	0.0132	0.3488	0.0594	0.1218	0.933
λ_{23}	0.63	0.6352	0.0052	0.1338	0.0632	0.0179	0.921
λ_{31}	0.52	0.5288	0.0088	0.0917	0.0768	0.0085	0.953
λ_{32}	0.63	0.6305	0.0005	0.0735	0.0544	0.0054	0.932
k	2.00	2.0298	0.0298	0.1292	0.1278	0.0176	0.911
β_1	0.50	0.5671	0.0671	1.1238	0.0916	1.2675	0.972
β_2	1.00	1.0045	0.0045	0.4144	0.0824	0.1717	0.961
<i>Gamma</i>							
λ_{12}	1.9	1.9856	0.0856	0.2881	0.3613	0.0903	0.931
λ_{13}	1.3	1.3660	0.0660	0.2239	0.2722	0.0545	0.907
λ_{21}	1.8	1.8843	0.0843	0.2917	0.3443	0.0922	0.940
λ_{23}	1.4	1.4334	0.0334	0.1595	0.2070	0.0265	0.927
λ_{31}	1.7	1.8336	0.1336	0.3760	0.4691	0.1592	0.946

Table 3.2: Simulation Results for a Three-State Semi-Markov Model with the Inclusion of Covariates (*continued*)

	True	Estimate	Bias	SD	SE	MSE	95% Coverage
λ_{32}	1.4	1.4327	0.0327	0.1607	0.1913	0.0269	0.926
ψ	2.0	2.0745	0.0745	0.1066	0.1799	0.0169	0.902
β_1	0.5	0.5088	0.0088	0.0459	0.0687	0.0022	0.974
β_2	-1.0	-0.9935	0.0065	0.0592	0.0778	0.0036	0.919

¹ For each CTSM, we simulated 400 subjects for 4 time units long.

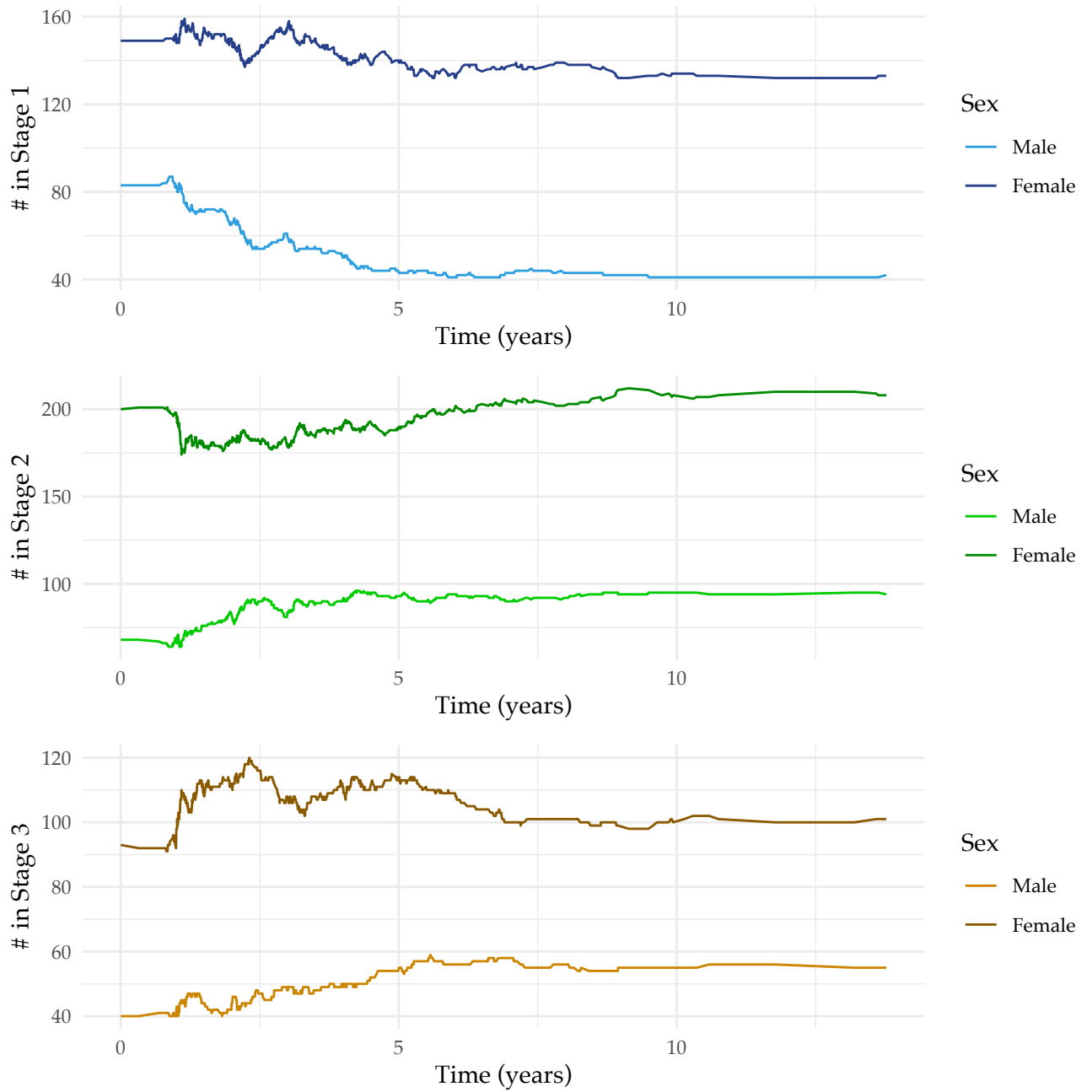


Figure 3.1: Frequency of Transition between 3 Levels of Caregiver Stress Over Time By Sex

Table 3.3: Alzheimer's Disease Caregiver 3-Level Stress Model Parameter Estimates

	<i>Exponential Sojourn</i>		<i>Weibull Sojourn</i>		<i>Gamma Sojourn</i>	
	Estimate	95% Bootstrap CI	Estimate	95% Bootstrap CI	Estimate	95% Bootstrap CI
λ_{12}	0.5387	(0.4898, 0.5876)	0.4854	(0.4391, 0.5316)	1.8351	(1.5433, 2.127)
λ_{13}	0.4957	(0.3262, 0.6653)	0.3970	(0.2161, 0.5780)	2.1437	(0.5572, 3.7302)
λ_{21}	0.5001	(0.4258, 0.5745)	0.4295	(0.3395, 0.5195)	1.7839	(1.3878, 2.1799)
λ_{23}	0.6021	(0.5470, 0.6573)	0.5510	(0.4947, 0.6074)	2.0334	(1.6312, 2.4356)
λ_{31}	0.8222	(0.6832, 0.9612)	0.7823	(0.6280, 0.9366)	4.1514	(2.9490, 5.3538)
λ_{32}	0.4215	(0.3388, 0.5041)	0.3721	(0.2757, 0.4685)	1.1598	(0.6765, 1.6430)
β_1	-0.0621	(-0.1475, 0.0233)	-0.0653	(-0.1658, 0.0353)	-0.0204	(-0.1173, 0.0765)
β_2	0.1260	(0.0634, 0.1886)	0.2338	(0.1147, 0.3528)	0.4664	(0.1879, 0.7449)
Shape ¹	-	-	1.8387	(1.7526, 1.9248)	3.4867	(3.1951, 3.7784)

¹ The shape parameters refer to k and ψ for the Weibull and gamma distributions, respectively.

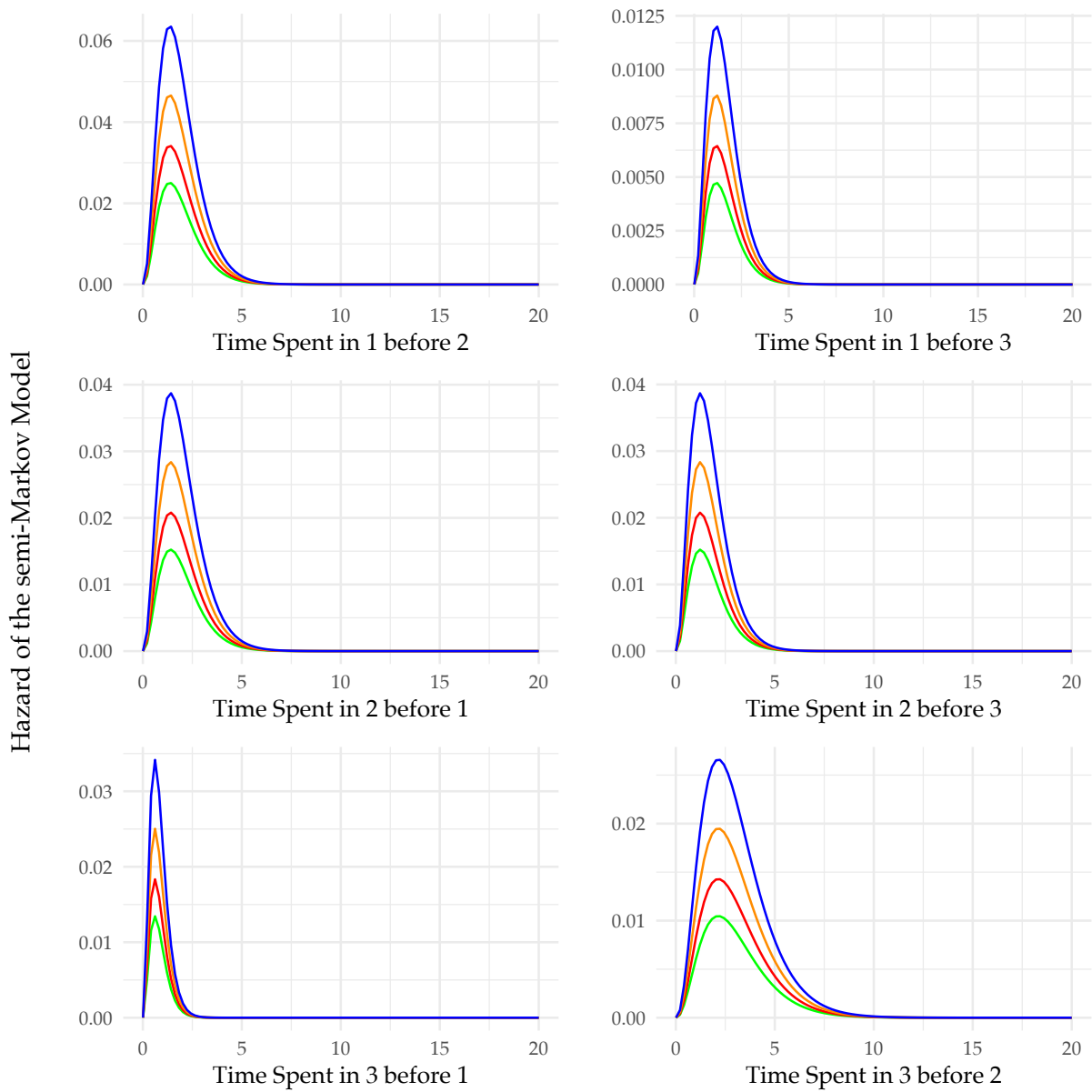


Figure 3.2: Plot of the Hazard of the Continuous-Time Semi-Markov Process Assuming a Gamma Sojourn Distribution for an Alzheimer’s Disease Caregiver Application. The lines indicates hazard lines for the ages of 65 years, 70 years, 76 years, and 81 years old (green, red, orange and blue, respectively).

References

- [25] S.-Z. Yu, “General hidden semi-markov model,” *Hidden Semi-Markov Models*, pp. 27–58, 2016. DOI: 10.1016/b978-0-12-802767-7.00002-4.
- [26] P. K. Anderson, S. Esbjerg, and T. Sorensen, “Multi-state models for bleeding episodes and mortality in liver cirrhosis,” *Stat. Med.*, vol. 19, pp. 587–599, 2000. DOI: 10.1002/(sici)1097-0258(20000229)19:4<587::aid-sim358>3.0.co;2-0.
- [27] A. C. Titman, “Estimating parametric semi-markov models from panel data using phase-type approximations,” *Statistics and Computing*, vol. 24, no. 2, pp. 155–164, 2012. DOI: 10.1007/s11222-012-9360-6.
- [28] Y. Shu, J. P. Klein, and M.-J. Zhang, “Asymptotic theory for the cox semi-markov illness-death model,” *Lifetime Data Anal*, vol. 13, pp. 91–117, 2007. DOI: 10.1007/s10985-006-9018-9.
- [29] H. Aralis and R. Brookmeyer, “A stochastic estimation procedure for intermittently-observed semi-markov multistate models with back transitions,” *Statistical Methods in Medical Research*, pp. 1–18, 2017. DOI: 10.1177/0962280217736342.
- [30] J. S. Benoit, W. Chan, S. Luo, H.-W. Yeh, and R. Doody, “A hidden markov model approach to analyze longitudinal ternary outcomes when some observed states are possibly misclassified,” *Statistics in Medicine*, vol. 35, no. 9, pp. 1549–1557, 2016. DOI: 10.1002/sim.6861.
- [34] D. Eddelbuettel and J. J. Balamuta, “Extending extitR with extitC++: A Brief Introduction to extitRcpp,” *PeerJ Preprints*, vol. 5, e3188v1, Aug. 2017, ISSN: 2167-9843. DOI: 10.7287/peerj.preprints.3188v1. [Online]. Available: <https://doi.org/10.7287/peerj.preprints.3188v1>.
- [35] M. Corporation and S. Weston, *Doparallel: Foreach parallel adaptor for the 'parallel' package*, R package version 1.0.15, 2019. [Online]. Available: <https://CRAN.R-project.org/package=doParallel>.
- [36] R. B. Fetter and J. D. Thompson, “A decision model for the design and operation of a progressive patient care hospital,” *Medical Care*, vol. 7, no. 6, pp. 450–462, 1969. DOI: 10.1097/00005650-196911000-00004.
- [42] D. R. Cox, “Regression models and life-tables,” *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972. DOI: 10.1111/j.2517-6161.1972.tb00899.x.
- [43] N. Breslow, “Covariance analysis of censored survival data,” *Biometrics*, vol. 30, no. 1, p. 89, 1974. DOI: 10.2307/2529620.
- [44] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019. [Online]. Available: <https://www.R-project.org/>.

Chapter 4

Trajectories in Depression Symptoms among Elderly Mexican Americans with Chronic Health Conditions: A Longitudinal Data Analysis

Authors: Kusha A. Mohammadi, and Wenyaw Chan

4.1 Abstract

Background:

Depression is one of the most prevalent mental health issues among older Mexican-Americans populations. Hispanic Americans are facing a mental health crisis where research is needed to understand the behaviors of depression symptoms while coping with one or multiple chronic illnesses such as heart disease, cancer, diabetes mellitus, stroke, hypertension, and kidney disease.

Methods:

Eight waves of data from the Hispanic Established Population for the Epidemiologic Study of the Elderly (HEPESE) which spans 20 years (1993 - 2013) was studied. We categorized the CES-D score into four categories to describe the level of severity of depres-

sion: Not depressed, mildly depressed, moderately depressed, and severely depressed. A continuous-time semi-Markov model was used to describe the dynamic severity depression level over two decades and a partial likelihood approach obtained the parameter estimates.

Results:

Respondents ($n = 3,079$) were shown to naturally progress toward higher depression levels after beginning in the not depressed state within the first 7 years. From the semi-Markov model, we identified that an elderly Hispanic person with any of the six chronic illnesses will spend about 15 years in the severely depressed level which is about 3-4 year longer than the other depression levels. We also report that there is a high risk of transitioning from non-depressed level back to a higher depressed level (mild, moderate, and severe) upon entering.

Conclusions:

Our current study indicates elderly Hispanics coping with one or multiple of the six chronic illnesses are likely to spend the most time in mild to severe depressed levels and have a higher risk of transitioning to a more severe depression level from a non-depressed level upon entering.

Keywords: Depressive Symptoms, HEPSE, CTSM, Partial Likelihood Method

4.2 Introduction

Many recent longitudinal studies have investigating factors associated with risk of depressive symptoms among Hispanics. One study used multivariate logistic regression to determine the associations between depressive symptoms and sociodemographics, chronic health conditions, disability, and cultural factors [46]. Another research team found that social network characteristics have a direct link between depressive symptoms and chronic health conditions [47]. There is additional evidence that social support and

church attendance were protective factors against increase depressive symptoms during pre-widowhood [48]. Further, age-adjusted odds of depressive symptoms in Hispanic women was 2.11 times the odds of non-Hispanic women while the men did not have a significant odds ratio [49]. Oh and others illustrated a relationship between depression and negative family interaction among cancer Hispanic individuals whom also experienced depression [50].

In 2010, the elderly Hispanic population (65 years and older) made up about seven percent of the United States population and is projected to rise to about twenty percent by 2050 [51]. Because of the rise in the elderly Hispanic population, it has been of interest to expand the knowledge of changes in depressive symptoms over time. Katon indicated a higher incidence and prevalence of major depression in individuals with chronic medical illness [52]. The odds of increased depressive symptoms for those living with specific chronic illness like heart attack was significantly higher (OR = 1.86; p-value = 0.03) than those with low depressive symptoms among Mexican-American adults aged 65 and older [53]. Another investigation reported that Mexican Americans had significantly earlier onset major depressive disorder as compared with African Americans [54]. Monserud and Markides highlighted church attendance was associated with a slower increase in depressive symptoms and greater social support was related to more depressive symptoms in the context of widowhood [48]. Changes in depression symptoms among older Mexican-Americans continues to be at the forefront of research.

In this paper, we investigate the dynamic changes of four levels of depression among elderly Mexican Americans with chronic health illnesses over time. We analyze eight time points of data from the Hispanic Established Population for the Epidemiologic Study of the Elderly (HEPESE) on adults 65 years and older. The present study contributes to Hispanic mental health literature in the following ways. First, we will be able to better describe the natural course between depressive level symptoms over time using a continuous-time semi-Markov model. This type of longitudinal data anal-

ysis will provide a more complex illustration of depression symptom patterns over time among older adults of Mexican descent by offering the rate of transition from one level to another, estimating the time spent in each depression severity level, and graphically depicting the hazard of transition. Secondly, the study addresses the mental health crisis for Mexican-American who experience changes in depressive symptoms while coping with certain chronic conditions. These chronic illnesses include heart attack, diabetes mellitus, cancer, stroke, hypertension, and kidney disease. Lastly, the study has the potential to inform future policy to develop depression programs to aid elderly Mexican-Americans coping with the chronic illness.

4.3 Methods

4.3.1 Elderly Hispanic Study Sample

We based our analysis on the Hispanic Established Population for the Epidemiologic Study of the Elderly (HEPESE). The HEPESE contains Mexican Americans aged 65 and older, who live in five southwestern states: Texas, New Mexico, Colorado, Arizona, and California [31]. The original study started in 1993 -1994 with 3050 subjects with a response rate of 83% [55] (Figure 4.1). Additional follow-ups occurred every two years post baseline: Wave 2 in 1995 - 1996 (M = 2438) [56], Wave 3 in 1998-1999 (M = 1980) [57], Wave 4 in 2000 - 2001 (M = 1682) [58], Wave 5 in 2004 - 2005 (M = 2069) [59], Wave 6 in 2006 - 2007 (M = 1542) [60], Wave 7 in 2010 - 2011 (M = 1078) [61], and Wave 8 in 2012 - 2013 (M = 744) [62]. Wave 5 added 905 new respondents that were aged 75 and older and followed up with the original cohort. The interviews took place inside the respondent's home in both Spanish or English based on their preference. The survey consisted of questionnaire elements of self-reported sociodemographic, cultural, and health-related measures. We used six chronic illness items to determine if the respondent was diagnosed with one or more chronic illnesses (Heart attack, diabetes mellitus, cancer,

stroke, hypertension, and kidney disease). We included those from the original cohort who had more than 1 observation (CES-D measure), and those who at least on of the six chronic illnesses. Additionally, we made use of the new respondents in Wave 5 and included them with the same criteria. Overall, we had a sample size of 3,079 subjects that were analyzed.

4.3.2 Categorical Outcome Measure

We based the depression symptoms on the Center for Epidemiological Studies Depression Scale (CES-D) [32] and categorized by the following criteria[33]: not depressed (0 - 9 points), mildly depressed (10 - 15 points), moderately depressed (16 - 24 points), and severely depressed (more than 25 points). CES-D score is comprised of 20 questions experienced during the past week. For each item, the answers vary in score from 0 (none/rarely) to 3 (most of the time). Respondents with a CES-D of 16 imply more psychological distress [63]. For every respondent's observations, we classified the CES-D score within four possible categorical outcomes to obtain a full trajectory of depression level over time.

4.3.3 Statistical Analysis

A continuous-time semi-Markov model (CTSMM) was used to capture dynamic nature of the depression levels over the duration of the HEPSE. The models for this analysis were un-adjusted models, meaning they did not consider any subject information on the transition rates. We used a partial likelihood approach to estimate the parameters from the CTSMM. To find the final model, the changes in depression severity over time were examined from models with different waiting time assumptions. Based on the Akaike information criterion (AIC), we proceed with the model with the lowest value. We interpreted the hazard of the CTSMM in the context of the elderly Hispanic Americans with one or more multiple chronic illnesses and discussed its potential contribution to

the mental health literature. All analysis were carried out in R 3.6.2.

4.4 Results

The baseline characteristics of the 3,079 respondents overall and by depression level are presented in Table 4.1. Overall, the elderly Mexican-American's are about 69% women, 74 years old, 54% married, and 40% very satisfied with life. The vast majority of individuals were diagnosed with hypertension (2372 individuals) compared to all other chronic illnesses. At baseline, most of the participants are classified as not depressed (1168 subjects) whereas mildly depressed, moderately depressed, and severely depressed are close in sample size (785, 623, 503, respectively). Among severely depressed, 35% of them are widowed - not divorced, 28.5% are somewhat to not at all satisfied with life, 74.2% had hypertension, and 40.8% are diabetic. Within the moderately depressed stage, 74% are women, 8.7% had a stroke, 77.7% were hypertensive, and 14.4% had a heart attack. The lower depression levels (not depressed and mildly depressed) had similar percentages across all the characteristics in Table 4.1.

In table 4.2, we find the total transitions over the observed time period within this dataset. For instance, there was 205 elderly Mexican-American respondents who transitioned from a moderate depression level to a severe depression level. The other observed transitions can be interpreted similarly. From figure 4.2, we observe how the counts in each depression level over the 20 year changed over the study duration. Overall, we examined a steep decrease in the number of respondents transitioning out of the not depressed stage to the higher stress level categories. After 15 years, we still notice a steady increase in the moderately depressed (CES-D score 16 - 24 points) level. This indicates that the elderly Mexican-American respondents tend to eventually go towards to third stress level after a number of years.

Table 4.3 refers to the probability changing to another depression severity state

at the time of transition. The estimated probability of transitioning from not depressed to mildly depressed and mildly depressed to moderately depressed are 46.5% and 27.0%, respectively. Additionally, we find there is a 26.2% chance of moving from a not depressed state to a severely depressed level and 24.4% chance of moving from a severely depressed level to a moderately depressed level.

From the CTSM, we investigated three sojourn distributions: exponential, Weibull, and gamma distribution. After optimizing the partial likelihood functions, the AIC was calculated to find the most appropriate model. The AIC was 28318.47, 27844.53, and 27617.36 for the CTSM which assumed the waiting time distributions to be exponential, Weibull, and gamma, respectively. This indicates that the model that specifies the gamma sojourn time is the most appropriate model for the elderly Hispanic data. Refer to the appendix A to view the raw sojourn times for each transition while overlaying each distribution over and model based estimates for the other models. Table 4.4 represents the parameter transition rate estimates of the CTSM that describes the changes in depressive symptoms among chronically ill elderly Mexican-Americans. Using the mean of the gamma distribution, we can find the estimated time we would expect to spend in each depressive state. The expected time spent in the not depressed, mildly depressed, moderately depressed, and severely depressed levels were 10.79, 12.61, 12.53, and 15.16 years, respectively. This suggests an elderly Mexican American living with chronic disease spent the longest time in the severely depressed state before transitioning back to a less severe level which is about 3 to 4 years longer than the other depression levels. Figure 4.3 illustrates the hazard of the semi-Markov process assuming a gamma sojourn time transitioning from one state to another. All hazards reflect a negatively (left) skewed shape. A higher curve indicates a greater risk of transitioning out of that depression level, whereas a flatter curve indicates a lower risk of moving out of that depression level. Particularly, the risk of transitioning from a not depressed level to a more severe state (2, 3, or 4) was high soon after entering the level. Additionally, the risk of

a respondent moving from a severely depressed level back to any of the lower states was relatively low upon entering that level.

4.5 Discussion

In this paper, we described the trajectory of multi-categorical depressive process among elderly Mexican-Americans coping with one or multiple chronic illnesses. We constructed a semi-Markov chain model to analyze the behavior through a series of states and to give some quality insight to risk of transitioning to a worse depression level. The 3079 respondents included in this study showed a high percentage ($> 39\%$) of life satisfaction (i.e. very satisfied) across all classification of depressive symptoms. This could indicate two explanations for the HEPese Data. First, there is a presence of respondent bias where the respondents are not truthfully answering the questions. Second, the CES-D score may not be a representative indicator for depressive symptoms. In the latter case, we would need to consider a different measure for depressive symptoms. However, in 2017, Moon and others determined CES-D to be an adequate screening instrument for depression in adults with high predictive power (area under the ROC: 0.92) [33]. It is also important to note that the outcome measure for depressive symptoms were self-reported and not from medical exams or records.

Based on the eight-waves of the HEPese longitudinal study, participants were shown to naturally progress toward the higher depression levels after beginning in the not depressed level within the first 7 years (figure 4.2). Thereafter, participants tended to stabilize in each depression state, however, there was still a tendency to transition to the moderately depressed state as time progressed on. The most appropriate model was found by calculating the AIC for each model. The gamma sojourn distribution closely resembled the HEPese time until transition for each state movement. Although, the time spent in a particular depression state from the HEPese dataset was revealed to

have had a bi-modal shape rather than uni-modal shape (see appendix A). Recall, the HEPSE follow-ups were collected bi-yearly from baseline. Further, some waves were not consistently observed in two year increments, rather, the participants were followed up with 3 or 4 years between interviews. Because the study procedure was collected in this way, the exact duration a respondent spent in each state may not be known or may vary greatly. We strongly believe this explains the bi-modal distributions present in the HEPSE example. Despite the finding, the gamma distribution does closely capture the natural distribution of the data. From the semi-Markov model, we identified that an elderly Hispanic person with some chronic illness will spend most of their time in the severely depressed state upon transitioning to that state (table 4.4). Additionally, we observed that an individual from the HEPSE is at higher risk of transitioning to a higher depression level (i.e. mildly depressed, moderately depressed, or severely depressed) from being classified as not depressed (figure 4.3). We also report that there is a lower risk of transitioning from severely depressed state back to a not depressed level. This indicates elderly Hispanics coping with any of these six chronic illnesses are likely to exhibit mild to severe depression symptoms as time progresses forward.

Our study is not without its limitations. First, the HEPSE study collected 8 waves of data every two to three years. If observations were collected more frequently, we could better understand the expected time a particular respondent would spend in any given depressive categorical state. Second, the results are not generalizable to a general population of elderly adults. In this current study, we did not weight the respondent's outcome measure at each of the eight waves. To make population inferences, we would need to consider a weighted analysis to make general conclusions about the pattern of depression severity over time. Lastly, we did not adjust for any confounders on the transition rates of the CTSM. There are potentially many risk factors for depression symptoms that were also collected from each subject. Therefore, it is not clear if the association in this analysis are indeed accurate due to the lack of confounders and

potential covariates.

The current research is a first look at the dynamic changes of depressive symptoms among Mexican-Americans who are chronically ill utilizing the HEPSESE. Additionally, it's one of the first studies to utilize a semi-Markov model to examine the changes through a series of depressive symptom states over eight waves of observations in HEPSESE.

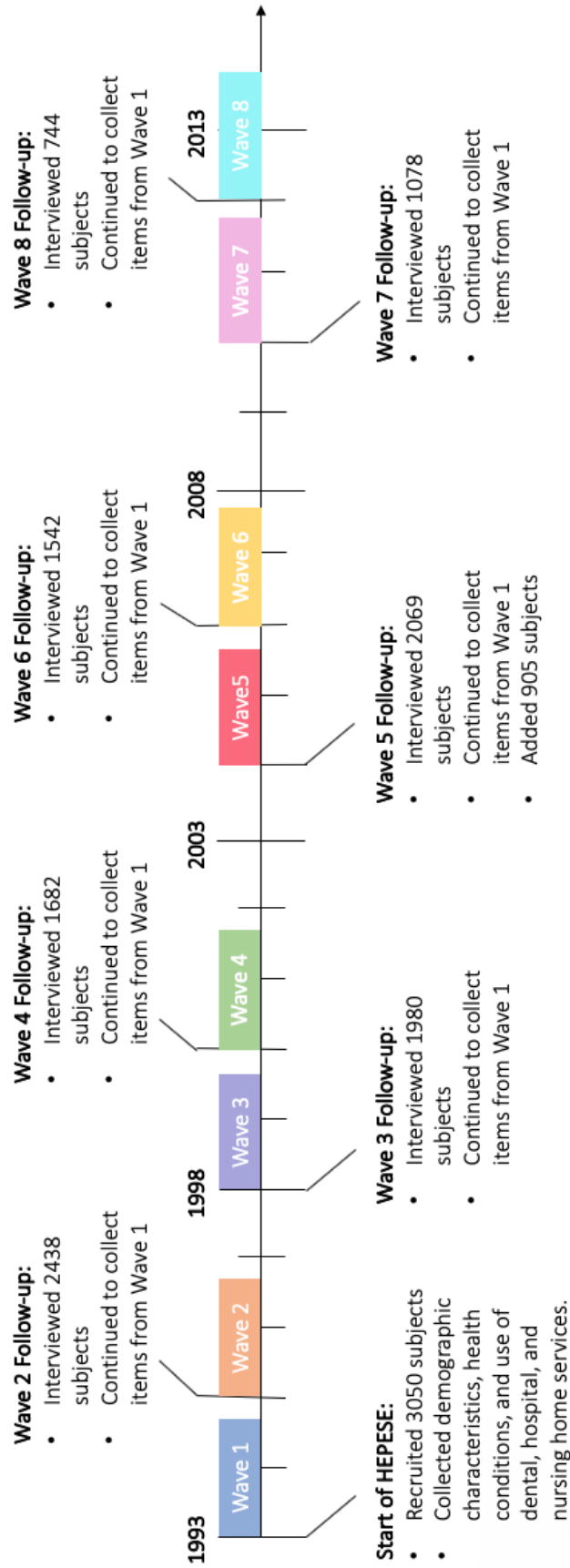


Figure 4.1: The timeline of the Hispanic Established Population for the Epidemiologic Study of the Elderly. Eight waves of data were collected from 1993 to 2013.

Table 4.1: Characteristics of Elderly Hispanic Adults in HEPSE at Baseline

	Overall	Not Depressed	Mildly Depressed	Moderately Depressed	Severely Depressed
N	3079	1168	785	623	503
Age (<i>Mean(SD)</i>)	73.91 (6.38)	73.69 (6.32)	74.00 (6.26)	74.51 (6.68)	73.56 (6.29)
Sex (<i>N (%)</i>)					
Female	2129 (69.1)	765 (65.5)	539 (68.7)	461 (74.0)	364 (72.4)
Male	950 (30.9)	403 (34.5)	246 (31.3)	162 (26.0)	139 (27.6)
Marriage Status (<i>N (%)</i>)					
Married	1658 (53.8)	659 (56.4)	421 (53.6)	318 (51.0)	260 (51.7)
Divorced	177 (5.7)	65 (5.6)	43 (5.5)	38 (6.1)	31 (6.2)
Separated	97 (3.2)	33 (2.8)	24 (3.1)	22 (3.5)	18 (3.6)
Never married	111 (3.6)	44 (3.8)	29 (3.7)	20 (3.2)	18 (3.6)
Widowed	1031 (33.5)	366 (31.3)	266 (33.9)	223 (35.8)	176 (35.0)
Don't know	3 (0.1)	0 (0.0)	1 (0.1)	2 (0.3)	0 (0.0)
Missing	2 (0.1)	1 (0.1)	1 (0.1)	0 (0.0)	0 (0.0)
Life Satisfaction (<i>N (%)</i>)					
Very satisfied	1240 (40.3)	461 (39.5)	335 (42.7)	268 (43.0)	176 (35.0)
Completely satisfied	1064 (34.6)	461 (39.5)	263 (33.5)	179 (28.7)	161 (32.0)
Somewhat satisfied	589 (19.1)	186 (15.9)	151 (19.2)	133 (21.3)	119 (23.7)
Not at all satisfied	70 (2.3)	20 (1.7)	12 (1.5)	14 (2.2)	24 (4.8)
Don't know	10 (0.3)	2 (0.2)	3 (0.4)	4 (0.6)	1 (0.2)
Missing	106 (3.4)	38 (3.3)	21 (2.7)	25 (4.0)	22 (4.4)
Heart Attack (<i>N (%)</i>)					
Yes	456 (14.8)	170 (14.6)	104 (13.2)	90 (14.4)	92 (18.3)
No	2565 (83.3)	979 (83.8)	666 (84.8)	520 (83.5)	400 (79.5)
Suspect or possible	52 (1.7)	16 (1.4)	15 (1.9)	11 (1.8)	10 (2.0)
Don't know	4 (0.1)	2 (0.2)	0 (0.0)	1 (0.2)	1 (0.2)
Missing	2 (0.1)	1 (0.1)	0 (0.0)	1 (0.2)	0 (0.0)
Diabetes Mellitus (<i>N (%)</i>)					
Yes, definitely	1174 (38.1)	435 (37.2)	301 (38.4)	233 (37.4)	205 (40.8)
Yes, borderline	75 (2.4)	33 (2.8)	18 (2.3)	14 (2.2)	10 (2.0)

Table 4.1: Characteristics of Elderly Hispanic Adults in HEPSE at Baseline (*continued*)

	Overall	Not Depressed	Mildly Depressed	Moderately Depressed	Severely Depressed
No	1821 (59.1)	698 (59.8)	463 (59.0)	373 (59.9)	287 (57.1)
Don't Know	9 (0.3)	2 (0.2)	3 (0.4)	4 (0.5)	1 (0.2)
Cancer (<i>N (%)</i>)					
Yes	254 (8.2)	97 (8.3)	68 (8.7)	46 (7.4)	43 (8.5)
No	2818 (91.5)	1069 (91.5)	716 (91.2)	576 (92.5)	457 (90.9)
Suspect or possible	3 (0.1)	2 (0.2)	0 (0.0)	0 (0.0)	1 (0.2)
Don't know	4 (0.1)	0 (0.0)	1 (0.1)	1 (0.2)	2 (0.4)
Stroke (<i>N (%)</i>)					
Yes	257 (8.3)	99 (8.5)	64 (8.2)	54 (8.7)	40 (8.0)
No	2798 (90.9)	1062 (90.9)	714 (91.0)	567 (91.0)	455 (90.5)
Suspect or possible	22 (0.7)	6 (0.5)	7 (0.9)	1 (0.2)	8 (1.6)
Missing	2 (0.1)	1 (0.1)	0 (0.0)	1 (0.2)	0 (0.0)
Hypertensive (<i>N (%)</i>)					
Yes	2372 (77.0)	892 (76.4)	623 (79.4)	484 (77.7)	373 (74.2)
No	674 (21.9)	263 (22.5)	154 (19.6)	133 (21.3)	124 (24.7)
Suspect or possible	29 (0.9)	12 (1.1)	7 (0.9)	5 (0.8)	5 (1.0)
Don't know	4 (0.1)	1 (0.1)	1 (0.1)	1 (0.2)	1 (0.2)
Kidney Disease (<i>N (%)</i>)					
Yes	126 (4.1)	39 (3.3)	25 (3.2)	31 (5.0)	31 (6.2)
No	1018 (33.1)	387 (33.1)	265 (33.8)	197 (31.6)	169 (33.6)
Don't know	30 (1.0)	9 (0.8)	11 (1.4)	5 (0.8)	5 (1.0)
Missing	1905 (61.9)	733 (62.8)	484 (61.7)	390 (62.6)	298 (59.2)

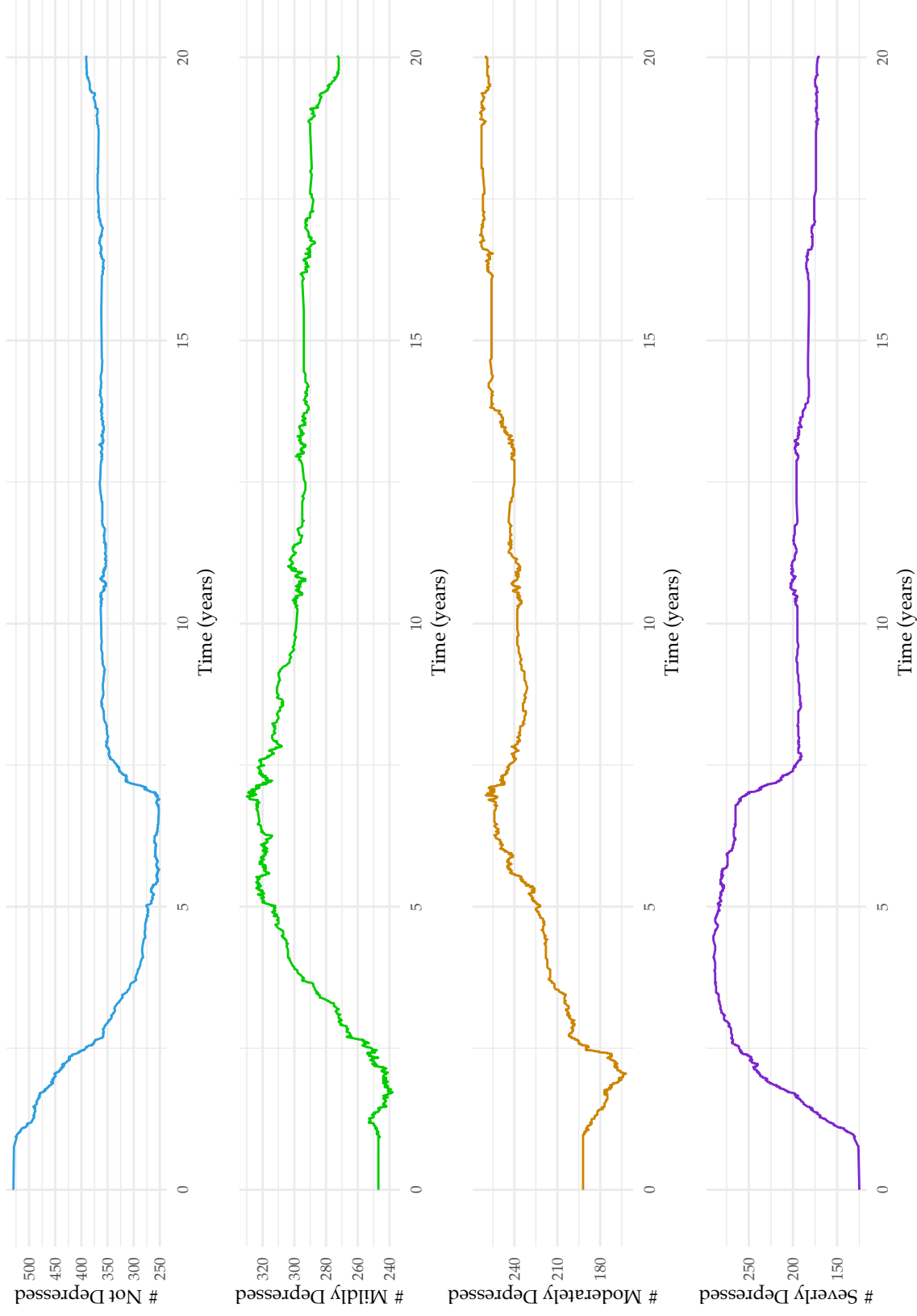


Figure 4.2: Frequency of Transition between 4 Levels of Depressive Symptoms over Time

Table 4.2: Observed Transitions between 4-Levels of Caregiver Stress

From State	Depression Level	To State			
		1	2	3	4
1	<i>None</i>	0	362	212	204
2	<i>Mild</i>	296	0	138	78
3	<i>Moderate</i>	169	98	0	96
4	<i>Severe</i>	173	78	81	0

Table 4.3: Probability of Moving to Another Depression Stage at Time of Transition

From State	Depression Level	To State			
		1	2	3	4
1	<i>Not Depressed</i>	-	0.465	0.272	0.262
2	<i>Mildly Depressed</i>	0.578	-	0.270	0.152
3	<i>Moderately Depressed</i>	0.466	0.270	-	0.264
4	<i>Severely Depressed</i>	0.521	0.235	0.244	-

Table 4.4: Elderly Mexican-American 4-Level Depression Model-Based Parameter Estimates assuming a Gamma Sojourn Distribution

	Estimate	Sojourn Time¹	95% Bootstrap CI
λ_{12}	0.8307	4.05	(0.7619, 0.8996)
λ_{13}	0.8475	3.97	(0.7696, 0.9253)
λ_{14}	1.2132	2.77	(1.1135, 1.3129)
λ_{21}	0.5959	5.65	(0.5396, 0.6523)
λ_{23}	0.9307	3.61	(0.8452, 1.0161)
λ_{24}	1.0038	3.35	(0.8485, 1.159)
λ_{31}	0.6346	5.30	(0.5545, 0.7148)
λ_{32}	0.9380	3.59	(0.8287, 1.0473)
λ_{34}	0.9244	3.64	(0.7728, 1.0759)
λ_{41}	0.5135	6.55	(0.4443, 0.5827)
λ_{42}	0.7643	4.40	(0.6227, 0.9059)
λ_{43}	0.8000	4.21	(0.7010, 0.8989)
ψ	3.3643	-	(3.1788, 3.5497)

¹ The time spent (in years) in state i before transitioning to a state j

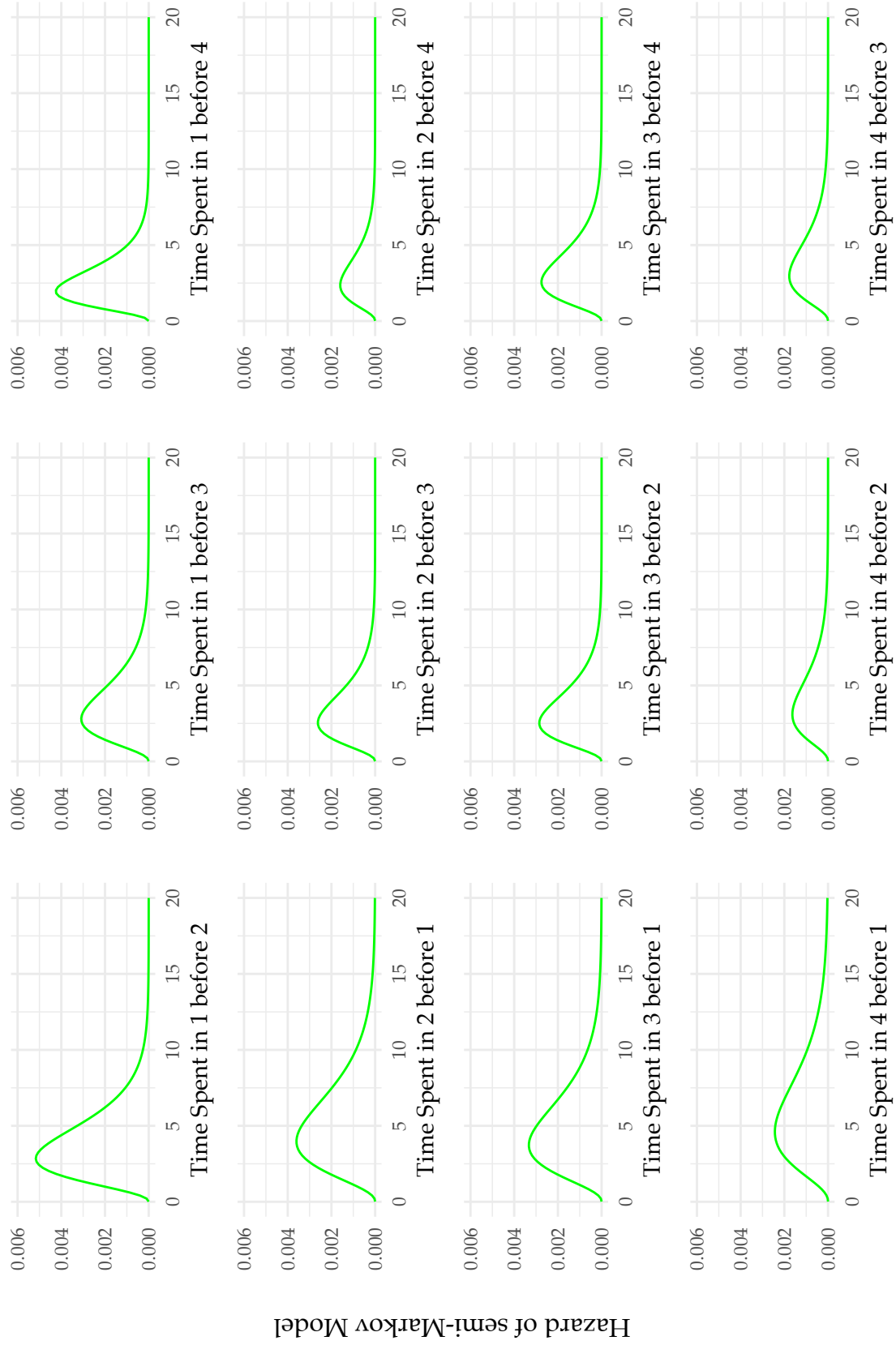


Figure 4.3: Plot of the Hazard of the Continuous-Time semi-Markov Process assuming the gamma sojourn time distribution. State 1, 2, 3, and 4 represent not depressed, mildly depressed, moderately depressed, and severely depressed, respectively. Time is in years.

References

- [31] K. S. Markides, C. A. Stroup-Benham, J. S. Goodwin, L. C. Perkowski, M. Lichtenstein, and L. A. Ray, “The effect of medical conditions on the functional limitations of mexican-american elderly,” *Annals of Epidemiology*, vol. 6, no. 5, pp. 386–391, 1996. DOI: 10.1016/s1047-2797(96)00061-0.
- [32] L. S. Radloff, “The ces-d scale,” *Applied Psychological Measurement*, vol. 1, no. 3, pp. 385–401, 1977. DOI: 10.1177/014662167700100306.
- [33] J. R. Moon, J. Huh, J. Song, I.-S. Kang, S. W. Park, S.-A. Chang, J.-H. Yang, and T.-G. Jun, “The center for epidemiologic studies depression scale is an adequate screening instrument for depression and anxiety disorder in adults with congenital heart disease,” *Health and Quality of Life Outcomes*, vol. 15, no. 1, May 2017. DOI: 10.1186/s12955-017-0747-0.
- [46] S. A. Black, K. S. Markides, and T. Q. Miller, “Correlates of depressive symptomatology among older community-dwelling mexican americans: The hispanic epese,” *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, vol. 53B, no. 4, Jan. 1998. DOI: 10.1093/geronb/53b.4.s198.
- [47] S. Soto, E. M. Arredondo, M. T. Villodas, J. P. Elder, E. Quintanar, and H. Madanat, “Depression and chronic health conditions among latinos: The role of social networks,” *Journal of Immigrant and Minority Health*, vol. 18, no. 6, pp. 1292–1300, 2016. DOI: 10.1007/s10903-016-0378-2.
- [48] M. A. Monserud and K. S. Markides, “Changes in depressive symptoms during widowhood among older mexican americans: The role of financial strain, social support, and church attendance,” *Aging & Mental Health*, vol. 21, no. 6, pp. 586–594, Jul. 2016. DOI: 10.1080/13607863.2015.1132676.
- [49] C. J. Swenson, J. Baxter, S. M. Shetterly, S. L. Scarbro, and R. F. Hamman, “Depressive symptoms in hispanic and non-hispanic white rural elderly the san luis valley health and aging study,” *American Journal of Epidemiology*, vol. 152, no. 11, pp. 1048–1055, Jan. 2000. DOI: 10.1093/aje/152.11.1048.
- [50] H. Oh, K. Ell, and A. Subica, “Depression and family interaction among low-income, predominantly hispanic cancer patients: A longitudinal analysis,” *Supportive Care in Cancer*, vol. 22, no. 2, pp. 427–434, Apr. 2013. DOI: 10.1007/s00520-013-1993-2.
- [51] *The next four decades: The older population in the united states: 2010 to 2050*. [Online]. Available: <https://www.census.gov/prod/2010pubs/p25-1138.pdf>.
- [52] W. J. Katon, “Clinical and health services relationships between major depression, depressive symptoms, and general medical illness,” *Biological Psychiatry*, vol. 54, no. 3, pp. 216–226, 2003. DOI: 10.1016/s0006-3223(03)00273-7.

- [53] S. Rote, N.-W. Chen, and K. Markides, “Trajectories of depressive symptoms in elderly mexican americans,” *Journal of the American Geriatrics Society*, vol. 63, no. 7, pp. 1324–1330, 2015. DOI: 10.1111/jgs.13480.
- [54] S. A. Riolo, T. A. Nguyen, J. F. Greden, and C. A. King, “Prevalence of depression by race/ethnicity: Findings from the national health and nutrition examination survey iii,” *American Journal of Public Health*, vol. 95, no. 6, pp. 998–1000, 2005. DOI: 10.2105/ajph.2004.047225.
- [55] K. S. Markides, “Hispanic established populations for the epidemiologic studies of the elderly, 1993-1994: [arizona, california, colorado, new mexico, and texas],” *ICPSR Data Holdings*, Apr. 2000. DOI: 10.3886/icpsr02851.v2.
- [56] K. S. Markides, “Hispanic established populations for epidemiologic studies of the elderly, wave ii, 1995-1996: [arizona, california, colorado, new mexico, and texas],” *ICPSR Data Holdings*, 2002. DOI: 10.3886/icpsr03385.v2.
- [57] K. S. Markides, “Hispanic established populations for epidemiologic studies of the elderly, wave iii, 1998-1999: [arizona, california, colorado, new mexico, and texas],” *ICPSR Data Holdings*, May 2004. DOI: 10.3886/icpsr04102.v2.
- [58] K. S. Markides and L. A. Ray, “Hispanic established populations for epidemiologic studies of the elderly, wave iv, 2000-2001 [arizona, california, colorado, new mexico, and texas],” *ICPSR Data Holdings*, Apr. 2005. DOI: 10.3886/icpsr04314.v2.
- [59] K. S. Markides, L. A. Ray, R. Angel, and D. V. Espino, “Hispanic established populations for the epidemiologic study of the elderly (hepese) wave 5, 2004-2005 [arizona, california, colorado, new mexico, and texas],” *ICPSR Data Holdings*, 2009. DOI: 10.3886/icpsr25041.
- [60] K. S. Markides, L. A. Ray, R. Angel, and D. V. Espino, “Hispanic established populations for the epidemiologic study of the elderly (hepese) wave 6, 2006-2007 [arizona, california, colorado, new mexico, and texas],” *ICPSR Data Holdings*, 2012. DOI: 10.3886/icpsr29654.v1.
- [61] K. S. Markides, N.-W. Chen, R. Angel, R. Palmer, and J. Graham, “Hispanic established populations for the epidemiologic study of the elderly (hepese) wave 7, 2010-2011,” *ICPSR Data Holdings*, Dec. 2016. DOI: 10.3886/ICPSR36537.v2.
- [62] K. S. Markides, N.-W. Chen, R. Angel, and R. Palmer, “Hispanic established populations for the epidemiologic study of the elderly (hepese) wave 8, 2012-2013,” *ICPSR Data Holdings*, Nov. 2016. DOI: 10.3886/ICPSR36578.v2.
- [63] J. H. Boyd, M. M. Weissman, W. D. Thompson, and J. K. Myers, “Screening for depression in a community sample: Understanding the discrepancies between depression symptom and diagnostic scales,” *Archives of General Psychiatry*, vol. 39, no. 10, pp. 1195–1200, Oct. 1982, ISSN: 0003-990X. DOI: 10.1001/archpsyc.1982.04290100059010.

Chapter 5

Future Works

When studying longitudinal categorical outcomes, many studies tend to use a multi-state Markov approach to analyze the behavior through a series of discrete states. Although, the Markov property may not be realistic since it imposes the holding time to be exponentially distributed. By this reasoning, a semi-Markov model seems more applicable because of its flexibility for an arbitrary sojourn time distribution. However, depending on the sojourn time distribution, the estimation method may be computationally difficult and inefficient for higher state semi-Markov processes. This problem motivates the work presented in this dissertation. We have constructed a partial likelihood method that has the ability to study the dynamics of a process as a semi-Markov chain which includes incorporating covariate effects on the transition rates. The partial likelihood approach utilized familiar survival analysis properties to develop an analogous form to estimate the parameter estimates of a semi-Markov process. This approach has a couple of advantages. First, the structure of the partial likelihood method is familiar and simple as in the classical survival analysis. The redefined probabilistic statements in the partial likelihood allow for the complexity of the semi-Markov process to be analyzed. Provided that the simulation were acceptable, we have shown our method to apply to three to four state semi-Markov processes. Secondly, by utilizing Rcpp package [34], and doParallel package [35] in R, we developed an computationally efficient way

to estimate the parameters from the semi-Markov process. The Rcpp package connects the C++ programming language and R by allowing R to call C++ functions easily into R code. This tool helped us improve computation time quickly and conveniently. Similarly, the doParallel package helped us improve computation speed by performing multi-core computing. Thirdly, we thoroughly analyzed two non-exponentially sojourn time distributions: Weibull and gamma distribution. Both of these distributions have been studied for unique survival problems because of the fact that they are generalizations of the exponential distribution. By assuming either the Weibull or gamma as the waiting time distribution, we have the flexibility in the CTSM to have multiple types of shapes for the hazard of the semi-Markov process. In both our applications, we found the time until transition to be non-exponential (i.e. Markov model is not appropriate). Lastly, our approach is applicable to a multitude of longitudinal categorical settings in public health.

The limitations outlined in the dissertation will be considered for our future research directions. First, due to the complexity of the hazard of the semi-Markov process, the derivative of the partial likelihood was not available in closed form. Typically, with the first derivative, we are able to derive the maximum likelihood estimates and standard errors from the Fisher's information matrix. In the absence of the first derivative, this proved to be a complicated optimization problem where we needed a derivative free optimization approach. For all three aims, we considered the Nelder-Mead (NM) non-linear optimization to find the optimal estimates. While we observed convergence (i.e. convergence code = 0), there is still a possibility that we may be in a local maximum rather than a global maximum. For future research, we need to explore other optimization methods to further investigate the convergence we observed. Secondly, we used a non-parametric bootstrap method to estimate the standard errors for each parameter of the CTSM. In chapter 3, we observed the coverage probabilities to be lower than the expected 95% which indicated that the confidence interval was not capturing the true value

95% of the time. This might suggest we need to reconsider the approach to estimating the standard errors. However, the results in aim 2 suggested the coverage probabilities to be closer to the expected 95%. For future research, we would compare and contrast (1) a parametric bootstrap sampling procedure and (2) an optimization method to approximate the hessian matrix. Lastly, we recognize our proposed method was limited to two sojourn distributions: Weibull and gamma. For future work, we would extend this to consider other parametric distributions such as double exponential, normal, and Pareto. We also desire to investigate non-parametric distributions for the holding time of the CTSMM.

Appendix A: Supplementary Materials for Chapter 4

Supplementary Materials for

**Trajectories in Depression Symptoms among Elderly Mexican
Americans with Chronic Health Conditions: A Longitudinal
Data Analysis**

A.1 Additional Tables & Figures to Describe the HEPSE Example

Table A.1: Elderly Mexican-American 4-Level Depression CTSMM assuming an exponential and Weibull distribution

	<i>Exponential Sojourn</i>			<i>Weibull Sojourn</i>		
	Estimate	95% Bootstrap CI	Estimate	Estimate	95% Bootstrap CI	Estimate
λ_{12}	0.2701	(0.2553,0.2850)	0.3518	0.3518	(0.3233,0.3804)	
λ_{13}	0.2690	(0.2478,0.2902)	0.3520	0.3520	(0.322,0.382)	
λ_{14}	0.3934	(0.3525,0.4343)	0.4348	0.4348	(0.4003,0.4693)	
λ_{21}	0.1626	(0.1496,0.1757)	0.2621	0.2621	(0.2327,0.2916)	
λ_{23}	0.2496	(0.2287,0.2706)	0.334	0.334	(0.3017,0.3662)	
λ_{24}	0.3102	(0.2664,0.3541)	0.3717	0.3717	(0.3268,0.4166)	
λ_{31}	0.1825	(0.1676,0.1975)	0.2889	0.2889	(0.2547,0.3231)	
λ_{32}	0.2546	(0.2253,0.2839)	0.3518	0.3518	(0.316,0.3875)	
λ_{34}	0.2886	(0.2468,0.3304)	0.3463	0.3463	(0.3083,0.3842)	
λ_{41}	0.1380	(0.1254,0.1506)	0.2202	0.2202	(0.1939,0.2465)	
λ_{42}	0.2164	(0.1935,0.2394)	0.2714	0.2714	(0.2401,0.3027)	
λ_{43}	0.2236	(0.1934,0.2538)	0.2855	0.2855	(0.2468,0.3241)	
k			1.5564	1.5564	(1.478,1.6349)	

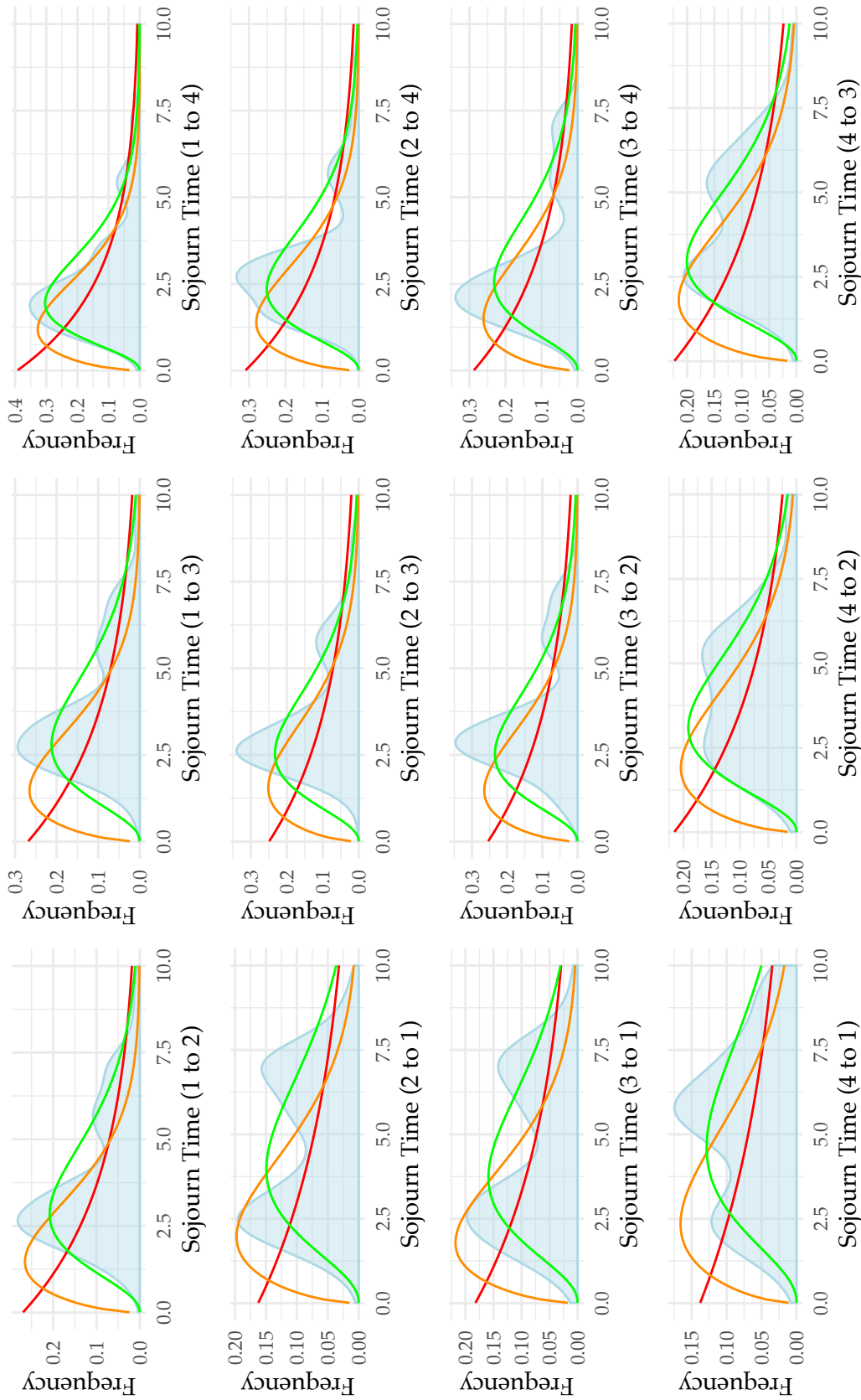


Figure A.1: Sojourn Time by Depression-Level (i to j) while overlaying each Semi-Markov Model for a 4-State Process. The red line represents the exponential distribution, the orange line represents the Weibull distribution, and the green line represents the gamma distribution.

Appendix B: Code

B.1 Written in R 3.6.2 - "Dark and Stormy Night"

Listing B.1: The Functions used to Implement the Partial Likelihood Approach, Simulation, and Output

```
#' Simulate Dataset of a Semi Markov Process
#'
#' @description
#' To simulate a three or four stage semi-markov process for a
#' given time distribution.
#' Sojourn distributions are exponential or gamma. Simulation
#' inputs are p0 (initial
#' distribution), pij (probability matrix), and Fij (sojourn
#' distribution).
#'
#' @param pij Probability matrix (i X j)
#' @param Rij Matrix (i X j) of rate parameters
#' @param p0 Initial State (if one not supplied, a random one
#' will be chosen)
#' @param Tmax Max time to observe patients (10 years)
#' @param nsubj Number of simulated subjects (defaulted to 700
#' subjects)
#' @param distn Sojourn Distribution ("exp" = exponential, "gamma
#' " = gamma, "weibull" = Weibull)
#' @param covar the covariates are needed set to TRUE (defaulted
#' to FALSE)
#' @param beta vector of true parameters for covariates (
#' defaulted to 0 for no covariates)
#' @param psi true gamma shape parameter constant for all
#' transition i to j
#' @param k true weibull shape parameter constant for all
#' transition i to j
#' @param Tmin Minimum start time for each subject (defaulted to
#' 0)
#' @param binprob Simulating binomial random draw for dictomous
#' variable (default p = 0.5)
#' @param umin Simulating uniform random draw for continuous
```

```

    variable (lower bound); (default min = 10)
#' @param umax Simulating uniform random draw for continuous
    variable (upper bound) (default max = 20)
#' @param nobs If there are panel observations, give number of
    observations (defaulted to NULL)
#' @param censorTmax If the last observation time needs to be set
    to Tmax
#' @return DF A simulated dataset of n subjects
#' @importFrom stats rexp rgamma rweibull runif rbinom
#' @export
SemiMarkovSim <- function(pij, Rij, p0 = NA, Tmax = 5, nsubj =
    700, distn = "exp", covar = FALSE, beta = 0, psi = NULL, k =
    NULL, nobs = NULL, Tmin = 0, binprob = 0.5, umin = -1, umax =
    1, censorTmax = FALSE) {

    # Create dataframe to save all subject data
    DF <- data.frame()

    # If covar is TRUE, then simulate a dictomous and continuous
    variable
    if (covar == TRUE) {
        X1 <- runif(n = nsubj, min = umin, max = umax) # The
            continuous variable
        X2 <- rbinom(n = nsubj, size = 1, prob = binprob) # The
            binary variable
        cov.l <- cbind(X1, X2)
        bcovs <- beta %*% t(cov.l)
        #rates <- lapply(1:nsubj, function(c) Rij*exp(aij + bcovs[,c
            ]))
        rates <- lapply(1:nsubj, function(c) Rij*exp(bcovs[,c]))
    } else {
        rates <- lapply(1:nsubj, function(g) Rij)
        cov.l <- replicate(nsubj, cbind(0,0))
    }

    # If start state is given, then start with supplied state
    # else start with random state
    if(is.numeric(p0)) {
        init.state <- rep(p0, nsubj)
    } else {
        init.state <- replicate(nsubj, sample(1:nrow(pij), size = 1))
    }

    df <- lapply(1:nsubj, function(l) SMC(l, pij, rates[[l]], init.
        state[l], distn, psi, k, Tmax, covar, cov.l[l,], Tmin, nobs,
        censorTmax))

    DF <- do.call(rbind, df)

    DF <- as.matrix(DF)

```

```

    return(DF)
}

#' Simulate Panel Observations
#'
#' @description
#' Using simulate times and states, create a new vector of panel
  observe times and states.
#' For example, observations are to be once every year (e.g. 1,
  2, 3).
#'
#' @param states vector of transition states
#' @param time vector of transition times
#' @param nobs Number of panel observations
#' @return i.state vector of observe state (from i) for panel
  times
#' @return j.state vector of observe state (to i) for panel times
#' @return int.time vector of panel observed times
#'
SimPanObs <- function(states, time, nobs) {
  int.time <- 1:nobs
  obs.state <- c(states[1])
  for(k in 1:nobs) {
    new.time <- time[which(time <= int.time[k])]
    obs.state <- c(obs.state, states[which(time == max(new.time))
    ])
  }
  int.time <- c(0, int.time)
  j.state <- obs.state

  # Reduce data non-repeated states
  ReduceData <- Reduce(j.state, int.time)
  j.state <- ReduceData$State
  int.time <- ReduceData$Time
  int.stime <- c(0, diff(int.time))
  i.state <- c(j.state[1], j.state[1:(length(j.state) -1)])

  return(list(i.state = i.state, j.state = j.state, int.time =
    int.time, int.stime = int.stime))
}

#' Eliminate Repeated Adjacent Observe States
#'
#' @description
#' Function to deleting repeated observe states for each
  individual, however, maintaining
  the visit time in each state (e.g. time = c(1,2,4); state = c
  (3,3,4) -> time = c(1,4);

```

```

#' state = c(3,4)).
#'
#' @param state Subject observed state's with adjacent repeated
states
#' @param time Subject visit time for each corresponding observed
state
#' @return State Subject observed state's without adjacent
repeated states
#' @return Time Subject visit time for each corresponding
observed state
#'
Reduce <- function(state, time){
  # Count the number of consecutive runs in the State vector
  runs <- rle(state)

  # Find positions in vector where runs are more than 1
  myruns <- which(runs$lengths >= 1)

  # Cumulative Sums of consecutive runs
  runs.len <- cumsum(runs$lengths)

  # Keep positions (indicies) that have single runs
  end <- runs.len[myruns]

  # Get observed state's without adjacent repeated states
  new_state <- state[end]

  # Get corresponding visit time
  # If run on is at beginning of process, re-index
  if(state[1] == state[2]) {end <- c(1, end[-length(end)]]}
  new_time <- time[end]

  return(list(State = new_state, Time = new_time))
}

#' Simulate a Simple Semi-Markov Chain Assuming A Sojourn
Distribution
#'
#' @description
#' To simulate one semi-markov chain assuming a exponential,
weibull, or gamma sojourn
#' distribution by supplying an initial distribution (p0),
probability matrix (pij), and
#' Fij (sojourn distribution).
#'
#' @param l indicator for the kth subject
#' @param pij Probability matrix (i X j)
#' @param Rij Matrix (i X j) of rate parameters
#' @param init.state Initial State for the Markov Chain
#' @param distn the specified distribution ("exp", "gamma", "
weibull")

```

```

#' @param psi shape parameter constant for all transition i to j
#' @param k shape parameter constant for all transition i to j
#' @param Tmax Max time to observe patients (10 years)
#' @param covar Boolean statement if covariates are included
#' @param covs the simulated covariate vector
#' @param Tmin Minimum start time for each subject (defaulted to
  0)
#' @param nobs If there are panel observations, give number of
  observations (defaulted to NULL)
#' @param censorTmax If the last observation time needs to be set
  to Tmax

#' @importFrom stats rexp rgamma rweibull runif rbinom
#' @return df A simulated dataset for the kth subject
SMC <- function(l, pij, Rij, init.state, distn, psi = NULL, k =
  NULL, Tmax = 10, covar = FALSE, covs = NULL, Tmin = 0, nobs =
  NULL, censorTmax = TRUE) {

  i.vec <- c(init.state)           # Store initial state in "To
    state" vector, i
  j.vec <- c(init.state)           # Store initial state in "From
    state" vector, j
  stime <- c(Tmin)                 # Set the initial
    sojourn time to be 0
  time <- c(Tmin)                  # Set the t to be Tmin

  t <- Tmin                        # Set t value to Tmin
  curr.state <- init.state          # Set current state to initial
    state

  while(t < Tmax) {

    # Sample a next state
    next.state <- sample(1:nrow(pij), size = 1, prob = pij[curr.
      state,])

    if (distn == "exp") {
      # Draw a Exp R.V. for the current state
      wait.time <- rexp(n = 1, rate = Rij[curr.state, next.state
        ])

    } else if (distn == "gamma") {
      # Draw a gamma R.V. for the current state
      wait.time <- rgamma(n = 1, shape = psi, rate = Rij[curr.
        state, next.state])

    } else if (distn == "weibull") {
      # Draw a Weibull R.V. for the current state
      wait.time <- rweibull(n = 1, shape = k, scale = 1/Rij[curr.
        state, next.state])
    } else {

      stop(paste("The Distribution", distn, "is not an option"))
    }
  }
}

```

```

}

ctime <- t + wait.time           # Cumulative time
i.vec <- c(i.vec, curr.state)    # Store current state in i.
    vec
j.vec <- c(j.vec, next.state)    # Store next state in j.vec
time <- c(time, ctime)          # Store time in a vector,
    time
stime <- c(stime, wait.time)     # Store sojourn time in a
    vector, stime

curr.state <- next.state         # Set next state to current
    state
t <- ctime                       # update time, t, to
    current time
}

# If the last time observation is > observed time, then censor
    time to max time
if (max(t) > Tmax & censorTmax == TRUE) {
  nvec <- length(time)
  time[nvec] <- Tmax
  stime[nvec] <- Tmax - time[nvec - 1]
}

# If there are set interval time supplied, then resimulate
    states & time
if (!is.null(nobs)){
  obs <- SimPanObs(j.vec, time, nobs) # Obtain these states &
    times
  stime <- obs$int.stime
  time <- obs$int.time              # Save the new observed
    times
  i.vec <- obs$i.state              # Save the i states for
    each time t
  j.vec <- obs$j.state              # Save the j states for
    each time t
}

ID <- rep(1, length(time))        # Create ID number for
    subject k

# Save the data in a temporary subject dataframe
df <- data.frame(ID = ID, time = time, i = i.vec, j = j.vec,
  stime = stime)

# If there is covariates, save the covariates as a dataframe
if (covar == TRUE) {
  # Create a dataframe to attach to temporary subject dataframe
  dfcovs <- data.frame(t(replicate(n = length(time), covs)))
}

```

```

    # Save covariates to the temporary subject dataframe
    df <- cbind(df, dfcovs)
  }

  return(df)
}

#' Simulate a Markov Chain and Estimate Transition Rates
#'
#' @description
#' Using msm package, we will simulate a markov chain and
  estimate the transition rates
#' using the full likelihood approach. The model uses the hessian
  matrix to calculate the
#' asymptotic standard errors.
#'
#' @param Qij Matrix (i X j) of transition rate parameters
#' @param nsim The number of simulations to be run
#' @param n.subj The number of subjects in the simulation study
#' @param n.obs The number of observations for each subject
#' @param seed Use any integer to set the seed for reproducible
  results (default= 1234)
#' @return est a vector of estimates, standard errors, lower and
  upper confidence interval
#' values
#' @import msm
#' @import doParallel
#' @import doRNG
#' @import suMisc
#' @export
sim_mc <- function(Qij, nsim, n.subj = 300, n.obs = 11, seed =
  1234){

  registerDoRNG(seed = seed)
  est <- foreach(b = 1:nsim, .combine = 'cbind') %dopar% {
    temp <- data.frame(subject = rep(1:n.subj, rep(n.obs ,n.subj)
      ), time = rep(seq(0, (n.obs - 1), 1), n.subj))
    DF <- simmulti.msm(temp, qmatrix = Qij, start = sample(1:ncol
      (Qij), n.subj, replace = T))

    if(ncol(Qij) == 3) {
      Q.init <- matrix(c(0.0, 0.6, 0.6, 0.6, 0.0, 0.6, 0.6, 0.6,
        0.0), ncol = 3, nrow = 3, byrow = T)
    } else {
      Q.init <- matrix(c(0.0, 0.6, 0.6, 0.6,
        0.6, 0.0, 0.6, 0.6,
        0.6, 0.6, 0.0, 0.6,
        0.6, 0.6, 0.6, 0.0), ncol = 4, nrow = 4,
        byrow = T)
    }
  }
}

```



```

tryCatch({
  # MSM model
  ad.msm <- msm(state ~ time, subject, data = DF, qmatrix = Q.
    init)

  qmat <- qmatrix.msm(ad.msm)
  qb.est <- t(qmat$estimates)[t(Qij) > 0]
  qb.se <- t(qmat$SE)[t(Qij) > 0]
  qb.lwr <- t(qmat$L)[t(Qij) > 0]
  qb.upr <- t(qmat$U)[t(Qij) > 0]

  c(qb.est, qb.se, qb.lwr, qb.upr)
}, error=function(e){})
}

return(est)
}

#' Summarize the Markov Chain simulation results in a LaTeX table
#'
#' To take simulation results in matrix form and output quick
#' latex code
#'
#' @param sim Matrix of Markov Chain (from sim.mc) simulation
#' results
#' @param true Vector of true values (i.e. transition rate
#' parameters)
#' @param nstate Number of states
#' @return Latex code for summary table
#' @importFrom stats var
#' @import kableExtra
#' @export
MakeTableMC <- function(sim, true, nstate = 3){

  # Extract the raw simulation parameter results from the
  # confidence intervals
  npar <- nstate*(nstate - 1)
  simD <- sim[1:npar,]
  se <- sim[(npar + 1):(2*npar),]
  lwr <- sim[(2*npar + 1):(3*npar),]
  upr <- sim[(3*npar + 1):(4*npar),]

  # Evaluate the simulation results from mean, bias, var, mse,
  # and coverage probability
  mean.s <- apply(simD, 1, mean)
  bias.s <- sapply(1:nrow(simD), function(i) mean(simD[i,] - true

```

```

    [i]))
sd.s <- apply(simD, 1, sd)
mse.s <- sd.s^2 + bias.s^2
cov.s <- sapply(1:nrow(lwr), function(i) mean(lwr[i,] <= true[i,] & upr[i,] >= true[i]))
se.s <- apply(se, 1, mean)

# Collect results into a new matrix
sum.tab <- cbind(true, mean.s, bias.s, sd.s, se.s, mse.s, cov.s)
sum.tab <- round(sum.tab, 4)
colnames(sum.tab) <- c("True", "Estimate", "Bias", "SD", "SE", "MSE", "95\\% Coverage")
rownames(sum.tab) <- GreekLabels(nstate, "exp", 0, covar = FALSE)

kTab <- kable(sum.tab, "latex", vline = "", escape = F, caption = "Full Likelihood Results from Markov Chain", linesep = "", align = rep('c',7), position = "!ht")
return(kTab)
}

#' Mean of F Distribution
#'
#' Calculate the mean of the distribution
#'
#' @param rate Rate parameter
#' @param shape Shape parameter (if applicable)
#' @param dist Distribution to calculate mean (options "exp", "weibull", "gamma")
#' @return mean value Returns mean value for the specified distribution
#' @examples
#' mean_F(2, dist = "exp")
#' mean_F(2, 3, "weibull")
#' mean_F(2, 2, "gamma")
#' @export
mean_F <- function(rate, shape = NULL, dist) {
  if (dist == "exp") {
    R <- 1/rate
  } else if (dist == "gamma") {
    R <- shape/rate
  } else if (dist == "weibull") {
    R <- gamma(1 + 1/shape)/rate
  }

  if(is.matrix(R) == TRUE) {diag(R) <- 0}
  return(round(R, 2))
}

```

```

#' Find Shape Parameters to Find Common Mean
#'
#' @description
#' To find gamma shape parameter for common mean between the
  weibull and gamma
#' distribution.
#'
#' @param wei_shape weibull shape parameter
#' @return Gamma Shape parameter
#' @export
ShapeFind <- function(wei_shape) {
  return(gamma(1 + 1/wei_shape))
}

#' Hazard Function for Computation
#'
#' @description
#' To calculate the hazard values using either the exponential,
  weibull, or gamma
#' distribution. The hazard function for the semi-markov process
  is uses the probability
#' matrix (pij), sojourn pdf (fij), and sojourn cdf (Fij).
#'
#' @param ri Vector of rate parameters for state i
#' @param pij Probability transition matrix (i X j)
#' @param i state from
#' @param j state to
#' @param s sojourn time spent in the previous state i
#' @param distn Sojourn Distribution ("exp" = exponential, "gamma
  " = gamma, "weibull" = Weibull)
#' @param beta beta coefficient parameters
#' @param covs covariate information vector for transitioning
  subject
#' @param psi shape parameter for the gamma distribution
#' @param k shape parameter for the weibull distribution
#' @param covar boolean statement if covariates are included
#' @return hazard value from respective hazard function
#' @importFrom stats dexp pexp dgamma pgamma pweibull dweibull
#'
hfunction <- function(ri, pij, i, j, s, distn, beta, covs, psi, k
, covar) {

  # Hazard function = pij * fij / sum_j( pij (1 - Fij) )

  if (distn == "exp") {
    # Exponential Hazard function using exponential pdf, cdf, and
      pij
    num <- pij[i, j] * dexp(s, ri[j])
    den <- sapply(which(1:nrow(pij) != i), function(u) pij[i,u]*

```

```

        (1 - pexp(s, ri[u]))
    h <- num/sum(den)
} else if (distn == "gamma") {
    # Gamma Hazard function using exponential pdf, cdf, and pij

    num <- pij[i, j] * dgamma(s, psi, ri[j])
    den <- sapply(which(1:nrow(pij) != i), function(u) pij[i,u]*
        (1 - pgamma(s, psi, ri[u])))
    h <- num/sum(den)
} else if (distn == "weibull") {
    # Weibull Hazard function using exponential pdf, cdf, and pij
    num <- pij[i, j] * dweibull(s, k, 1/ri[j])
    den <- sapply(which(1:nrow(pij) != i), function(u) pij[i,u]*
        (1 - pweibull(s, k, 1/ri[u])))
    h <- num/sum(den)
} else {
    # Send warning message if the distribution is not one of
    # these
    stop(paste("The Distribution", distn , "is not an option"))
}

# If covariates (i.e. TRUE), modify hazard function by cox
# model
#if (covar == TRUE) {h <- h*exp(as.numeric(aij[i,j] + beta %*%
#    covs))}
if (covar == TRUE) {h <- h*exp(as.numeric(beta %*% covs))}

return(h)
}

#' Probability Transition Matrix from Data
#'
#' To calculate the transition counts using the data.
#'
#' @param state the observed state transition
#' @param ID The subject's identification tag
#' @param data the data frame with these variables
#' @return matrix A matrix of observed probability transition
# counts
#' @export
#' @importFrom msm statetable.msm
PijCount <- function(data) {
    P <- statetable.msm(j, ID, data)
    return(P/apply(P, 1, sum))
}

```

```

#' Parameter Preparation
#'
#' To separate the rate parameters from the shape parameter
#'
#' @param par vector of parameter estimates
#' @param nstate number of states in the process (defaulted to 4)
#' @param nBeta number of covariates
#' @param distn Specify the distribution ("exp", "gamma", "
  weibull")
#' @param covar boolean statement if covariates are included
#' @return List of the parameters by matrix of rates, shape
  parameter, beta coefficients, transition specific constants aij
  .
#'
ParPrep <- function(par, nstate, nBeta, distn, covar) {

  if (distn == "exp") {
    nr <- nstate*(nstate - 1)
    if (covar == TRUE) {
      beta <- par[(nr+1):(nr+nBeta)]
      #aij <- par[(nr+nBeta+1):length(par)]
      #aMat <- Vec2Mat(aij, nstate)
    } else {
      beta <- rep(0, 2)
      #aMat <- matrix(0, ncol = nstate, nrow = nstate)
    }
    par <- par[1:nr]
    shape = 0
  } else if (distn == "gamma" | distn == "weibull") {
    nr <- nstate*(nstate - 1)
    shape <- par[(nr+1)]
    if (covar == TRUE) {
      beta <- par[(nr+2):(nr+ nBeta + 1)]
      #aij <- par[(nr+nBeta+2):length(par)]
      #aMat <- Vec2Mat(aij, nstate)
    } else {
      beta <- rep(0, 2)
      #aMat <- matrix(0, ncol = nstate, nrow = nstate)
    }
    par <- par[1:nr]
  } else {
    stop(paste("The Distribution", distn , "is not an option"))
  }

  rMat <- Vec2Mat(par, nstate)

  #return(list(parM = rMat, shape = shape, beta = beta, aijM =
    aMat))

```

```

return(list(parM = rMat, shape = shape, beta = beta))
}

#' Vector to Matrix
#'
#' Take the parameter estimates to nstate X nstate matrix
#'
#' @param par vector of parameter estimates
#' @param nstate number of states in the process (defaulted to 4)
#' @return A matrix from the vector form (i X j)
#' @export
Vec2Mat <- function(par, nstate = 4) {
  W <- matrix(1, nrow = nstate, ncol = nstate) # Create a Matrix
  (nstate X nstate)
  diag(W) <- 0 # Fill the
  diagonals with 0s
  W[W > 0] <- par # Fill parameters
  estimates in Matrix
  q <- t(W) # Transpose matrix
  to realign elements
  return(q)
}

#' List of Prepared Data
#'
#' To organize and separate data for partial likelihood
#'
#' @param data A data frame in this order (ID, time, i, j, s,
  covs..)
#' @param covar boolean statement if covariates are included
#' @return A list of data needed for optimization; numData,
  denData, denTies, denTime, cov
#'
PrepData <- function(data, covar = FALSE) {
  # Index for variables in data
  ID <- 1; time <- 2; i = 3; j = 4; s = 5
  pij <- PijCount(data)

  df1 <- data[order(data[,time]),] # Order data by
  transition time
  df2 <- df1[df1[,time] > 0, ] # Data with transition
  times greater than 0
  Ttime <- df2[,time] # Obtain the ordered
  transition times
  UTime <- unique(Ttime) # Unique Transition times

  # Count the number of ties in the data
  dn <- sapply(UTime, function(t) sum(t == Ttime))

  # Matrix of information (state i, state j, and sojourn time s)

```

```

    for the numerator
numData <- df2[,c(i, j, s)]

# List of data by subject for the denominator (transition time
  t, state i, and state j)
denData <- sapply(unique(df1[,ID]), function(w) df1[which(df1[,
  ID] == w), c(time,i,j)])

if (covar == TRUE) {
  # Matrix of Covariates for the numerator
  ncov <- ncol(df2)
  numCovs <- df2[, (s+1):ncov]

  # List of Covariates for the denominator
  denCovs <- sapply(unique(df1[,ID]), function(w) df1[which(df1
    [,ID] == w), ][1, (s+1):ncov])
} else {
  numCovs <- matrix(0, ncol = 2)
  denCovs <- matrix(0, ncol = 2)
}

return(list(numData = numData, denData = denData, denTies = dn,
  denTime = UTime, numCovs = numCovs, denCovs = t(denCovs),
  pij = pij))
}

#' PARTIAL LOG LIKELIHOOD
#'
#' @description
#' To estimate the maximum likelihood estimates by maximizing the
  partial log likelihood,
#' PLL.
#'
#' @param par parameters needed for estimation
#' @param data A dataset in matrix form and ordered format (e.g.
  ID, time, i, j, sojourn time)
#' @param distn Specify the distribution ("exp", "gamma", "
  weibull")
#' @param nstate Number of states in the semi-Markov process (
  defaulted to 4)
#' @param nBeta number of covariates
#' @param covar boolean statement if covariates are included
#' @return The partial log likelihood value for a given set
  parameters
#' @import Rcpp
#' @export
PLL <- function(par, data, distn = "exp", nstate = 4, nBeta = 2,
  covar = FALSE) {

```

```

# Before optimization prepare data by separating into multiple
  matrices
## List arguments: numData, denData, denTies, denTime
df <- PrepData(data, covar)

# Obtain parameters for PLL
Lpar <- ParPrep(par, nstate, nBeta, distn, covar)
rates <- Lpar$parM
shape <- Lpar$shape
beta <- Lpar$beta
#aij <- Lpar$aijM
#aij <- matrix(0, nstate, nstate)

# Probability transition matrix
pij <- df$pij

# Calculation of the partial likelihood
## Calculation of the numerator
## numData structure: (state) i, (state) j, (sojourn time) s
lnum <- apply(df$numData, 1, function(v) hfunction(rates[v
  [1],], pij, v[1], v[2], v[3], distn, beta, df$numCovs[which(
  v[3] == v[3]),], shape, shape, covar))

## Calculation of the denominator
## denData Structure: time (of transition), (state) i, (state)
  j
## denTime: Vector of unique transition times
## denCovs: Matrix of covariate information
A <- df$denData
B <- df$denTime
C <- as.matrix(df$denCovs)
ldenom <- RiskSet(A, B, rates, pij, distn, beta, C, shape,
  shape, covar)

## Calculate the total of the partial log likelihood
l <- sum(log(lnum)) - sum(df$denTies * log(ldenom))

if(is.infinite(l)){l <- -1e6}
if(is.nan(l)){l <- -1e6}
return(-l)
}

#' Simulation Optimization
#'
#' @description
#' To evaluate the partial log likelihood by nonlinear

```



```

optimization using Lagrange method.
#' Each simulation will also have bootstrap samples to collect
the standard errors.
#'
#' @param pij Probability matrix (i X j)
#' @param Rij Matrix (i X j) of rate parameters
#' @param p0 Initial State (if one not supplied, a random one
will be choosen)
#' @param Tmax Max time to observe patients (10 years)
#' @param nsubj Number of simulated subjects (defaulted to 700
subjects)
#' @param nsim Number of Simulations
#' @param nboot Number of bootstrap samples
#' @param nstate Number of states
#' @param nBeta Number of coefficents
#' @param distn Sojourn Distribution ("exp" = exponential, "gamma
" = gamma, "weibull" = Weibull)
#' @param covar If covariates are needed (defaulted to FALSE)
#' @param beta vector of true parameters for covariates (
defaulted to 0 for no covariates)
#' @param psi true gamma shape parameter constant for all
transition i to j
#' @param k true weibull shape parameter constant for all
transition i to j
#' @param aij Transition specific intercepts matrix (i X j)
#' @param Tmin Minimum start time for each subject (defaulted to
0)
#' @param binprob Simulating binomial random draw for dictomous
variable (default p = 0.5)
#' @param umin Simulating uniform random draw for continuous
variable (lower bound); (default min = 10)
#' @param umax Simulating uniform random draw for continuous
variable (upper bound) (default max = 20)
#' @param nobs If there are panel observations, give number of
observations (defaulted to NULL)
#' @param censorTmax If the last observation time needs to be set
to Tmax
#' @param control Control options for optim r
#' @return estMat matrix of optimization estimates
#' @return SeList A list of bootstrap estimates for each
simulation
#' @importFrom stats optim sd
#' @useDynLib PLSMM, .registration = TRUE
#' @export
SimOptim <- function(pij, Rij, p0 = NA, Tmax = 5, nsubj = 700,
nsim = 500, nboot = 10, nstate = 4, nBeta = 2, distn = "exp",
covar = FALSE, beta = 0, psi = NULL, k = NULL, aij = NULL,
nobs = NULL, Tmin = 0, binprob = 0.5, umin = -1, umax = 1,
censorTmax = FALSE, control = list(reltol = 1e-4)) {

LenPar <- nstate*(nstate - 1) + 1*(!is.null(psi) == T)*(length(
psi)) + 1*(!is.null(k) == T)*(length(k)) + 1*(all(beta != 0)

```

```

)*(length(beta))
  est <- matrix(NA, nrow = nsim, ncol = LenPar)
  upr <- matrix(NA, nrow = nsim, ncol = LenPar)
  lwr <- matrix(NA, nrow = nsim, ncol = LenPar)

  for (b in 1:nsim) {
    data <- SemiMarkovSim(pij, Rij, p0, Tmax, nsubj,
      distn, covar, beta, psi, k, aij, nob, Tmin,
      binprob, umin, umax, censorTmax)

    par <- init.par(data, distn, nstate, nBeta, covar
      )

    est[b, ] <- optim(par = par, fun = PLL, data =
      data, distn = distn, nstate = nstate, nBeta =
      nBeta, covar = covar, method = "Nelder-Mead",
      control = control)$par

    res2 <- matrix(NA, nrow = nboot, ncol = length(
      par))

    for (v in 1:nboot){
      data2 <- data[which(data[,1] %in% sample(
        unique(data[,1]), size = nsubj,
        replace = TRUE)),]

      par2 <- init.par(data2, distn, nstate,
        nBeta, covar)

      res2[v,] <- optim(par = par2, fun = PLL,
        data = data2, distn = distn, nstate =
        nstate, nBeta = nBeta, covar = covar,
        method = "Nelder-Mead", control =
        control)$par

    }

    lwr[b,] <- sapply(1:length(par), function(a) est[
      b,a] - 1.96* sd(res2[,a]))
    upr[b,] <- sapply(1:length(par), function(a) est[
      b,a] + 1.96* sd(res2[,a]))

  }

  return(list(est = est, lwr = lwr, upr = upr))
}

#' Simulation Optimization by paralleling loops
#'
```

```

#' @description
#' To evaluate the partial log likelihood by Nelder-Mead method
  optimization.
#' Each simulation will also have bootstrap samples to collect
  the standard errors.
#'
#' @param pij Probability matrix (i X j)
#' @param Rij Matrix (i X j) of rate parameters
#' @param p0 Initial State (if one not supplied, a random one
  will be choosen)
#' @param Tmax Max time to observe patients (10 years)
#' @param nsubj Number of simulated subjects (defaulted to 700
  subjects)
#' @param nsim Number of Simulations
#' @param nboot Number of bootstrap samples
#' @param nstate Number of states
#' @param nBeta Number of coefficents
#' @param distn Sojourn Distribution ("exp" = exponential, "gamma
  " = gamma, "weibull" = Weibull)
#' @param covar the covariates are needed set to TRUE (defaulted
  to FALSE)
#' @param beta vector of true parameters for covariates (
  defaulted to 0 for no covariates)
#' @param psi true gamma shape parameter constant for all
  transition i to j
#' @param k true weibull shape parameter constant for all
  transition i to j
#' @param aij Transition specific intercepts matrix (i X j)
#' @param Tmin Minimum start time for each subject (defaulted to
  0)
#' @param binprob Simulating binomial random draw for dictomous
  variable (default p = 0.5)
#' @param umin Simulating uniform random draw for continuous
  variable (lower bound); (default min = 10)
#' @param umax Simulating uniform random draw for continuous
  variable (upper bound) (default max = 20)
#' @param nobobs If there are panel observations, give number of
  observations (defaulted to NULL)
#' @param censorTmax If the last observation time needs to be set
  to Tmax
#' @param control Controls for optim
#' @param seed Set the seed for reproducibility
#' @return est A matrix of optimization estimates and confidence
  intervals
#' @importFrom stats optim sd
#' @import doParallel
#' @import doRNG
#' @useDynLib PLSMM, .registration = TRUE
#' @export
SimOptimPar <- function(pij, Rij, p0 = NA, Tmax = 5, nsubj = 700,
  nsim = 500, nboot = 10, nstate = 4, nBeta = 2, distn = "exp",
  covar = FALSE, beta = 0, psi = NULL, k = NULL, aij = NULL,

```

```

nobs = NULL, Tmin = 0, binprob = 0.5, umin = -1, umax = 1,
censorTmax = FALSE, control = list(reltol = 1e-4), seed =
1234) {

registerDoRNG(seed = seed)
  est <- foreach(b = 1:nsim, .combine = 'cbind', .packages
= 'PLSMM') %dopar% {

    data <- SemiMarkovSim(pij, Rij, p0, Tmax, nsubj, distn,
covar, beta, psi, k, nobs, Tmin, binprob, umin,
umax, censorTmax)

    par <- init.par(data, distn, nstate, nBeta, covar
)

    res <- optim(par = par, fn = PLL, data = data,
distn = distn, nstate = nstate, nBeta = nBeta,
covar = covar, method = "Nelder-Mead",
control = control)$par

    bootres <- BootCI(data, res, distn, nstate, nsubj
, nBeta, nboot, covar)

    c(res, bootres)
  }

return(est)
}

#' Bootstrap samples to find the 95\% Confidence Intervals
#'
#' @description
#' To find the 95\% Confidence intervals by bootstrap samples
#'
#' @param data Simulated Semi-Markov Dataset
#' @param res Best results from Nelder-Mead Optimization
#' @param distn Sojourn Distribution ("exp" = exponential, "gamma
" = gamma, "weibull" = Weibull)
#' @param nstate Number of states
#' @param nsubj Number of subjects
#' @param nBeta Number of coefficients
#' @param nboot Number of bootstrap samples
#' @param covar the covariates are needed set to TRUE (defaulted
to FALSE)
#' @param control Control options for optim
#' @return A vector of confidence intervals for each parameter in
order
#' @importFrom stats optim sd

```

```

#' @import doParallel
#' @export
BootCI <- function(data, res, distn, nstate, nsubj, nBeta, nboot,
  covar, control) {

  res2 <- foreach(v = 1:nboot, .combine = 'cbind', .
    packages = 'PLSMM') %do% {

    data2 <- data[which(data[,1] %in% sample(unique(data
      [,1]), size = nsubj, replace = TRUE)),]

    par2 <- init.par(data2, distn, nstate, nBeta,
      covar)

    optim(par = par2, fn = PLL, data = data2, distn =
      distn, nstate = nstate, nBeta = nBeta, covar =
      covar, method = "Nelder-Mead", control =
      control)$par
  }

  CI <- sapply(1:length(res), function(a) res[a] + c(-1,1)
    * 1.96* sd(res2[a,]))

  return(c(CI[1,], CI[2,]))
}

#' Summarize the simulation results in a LaTeX table
#'
#' To take simulation results in matrix form and output quick
  latex code
#'
#' @param sim Matrix of simulation results
#' @param true Vector of true values (rate parameters, shape
  parameters, coefficients, etc.)
#' @param nstate Number of states
#' @param distn The sojourn distribution (e.g. exp, weibull,
  gamma)
#' @param nBeta Number of covariate parameters
#' @param covar Boolean statement if covariates are present
#' @return Latex code for summary table
#' @importFrom stats var
#' @import kableExtra
#' @export
MakeTable <- function(sim, true, nstate = 4, distn = "exp", nBeta
  = 2, covar = FALSE){

  # Extract the raw simulation parameter results from the
    confidence intervals
  npar <- nstate*(nstate - 1)

```

```

if(distn != "exp") {npar = npar + 1}
if(covar == TRUE) {npar = npar + nBeta}
simD <- sim[1:npar,]
lwr <- sim[(npar+1):(npar+npar),]
upr <- sim[(npar + npar + 1):(npar + npar + npar),]

# Evaluate the simulation results from mean, bias, var, mse,
  and coverage probability
mean.s <- apply(simD, 1, mean)
bias.s <- sapply(1:nrow(simD), function(i) mean(simD[i,] - true
  [i]))
sd.s <- apply(simD, 1, sd)
mse.s <- sd.s^2 + bias.s^2
cov.s <- sapply(1:nrow(lwr), function(i) mean(lwr[i,] <= true[i]
  ] & upr[i,] >= true[i]))
se.s <- sapply(1:nrow(upr), function(i) median((upr[i,] - true[
  i]) / 1.96))

# Collect results into a new matrix
sum.tab <- cbind(true, mean.s, bias.s, sd.s, se.s, mse.s, cov.s
  )
sum.tab <- round(sum.tab, 4)
colnames(sum.tab) <- c("True", "Estimate", "Bias", "Variance",
  "SE", "MSE", "95\\% Coverage")
rownames(sum.tab) <- GreekLabels(nstate, distn, nBeta, covar)

kTab <- kable(sum.tab, "latex", vline = "", escape = F, caption
  = "Simulation Results Assuming an Exponential Sojourn Time
  Distribution", linesep = "", align = rep('c',7), position =
  "!ht")
return(kTab)
}

#' Make Greek Labels for Tables
#'
#' To make row labels formatted for latex output
#'
#' @param nstate Number of States in the System
#' @param distn Sojourn Distribution ("exp" = exponential, "gamma
  " = gamma, "weibull" = Weibull)
#' @param nBeta Number of covariate parameters
#' @param covar Boolean statement if covariates are present
#' @return Vector of parameters in latex format
GreekLabels <- function(nstate, distn, nBeta, covar = FALSE) {
  Vec <- c()
  for (i in 1:nstate) {
    for (j in which(1:nstate != i)) {
      Vec <- c(Vec, paste("$\\lambda_{", i, j, "}$", sep = ""))
    }
  }
}

```

```

}
if (distn == "weibull") {Vec <- c(Vec, paste("$k$"))}
if (distn == "gamma") {Vec <- c(Vec, paste("$\\psi$"))}
if(covar == TRUE) {
  for (p in 1:nBeta){
    Vec <- c(Vec, paste("$\\beta_", p, "$", sep = ""))
  }
}
return(Vec)
}

#' Real Data Optimization by paralleling loops
#'
#' @description
#' To evaluate the partial log likelihood by Nelder-Mead method
  optimization.
#' The real data example will have bootstrap samples to calculate
  standard errors.
#'
#' @param data Real longitudinal data in the format of (ID, time,
  i, j, sojourn, X1,..,Xn)
#' @param nboot Number of bootstrap samples
#' @param nstate Number of states
#' @param nBeta Number of coefficients
#' @param distn Sojourn Distribution ("exp" = exponential, "gamma"
  = gamma, "weibull" = Weibull)
#' @param covar the covariates are needed set to TRUE (defaulted
  to FALSE)
#' @param control Control options for optim
#' @param seed Set the seed for reproducibility
#' @return est A matrix of optimization estimates and confidence
  intervals
#' @importFrom stats optim sd
#' @useDynLib PLSMM, .registration = TRUE
#' @export
PLLSMM <- function(data, nboot = 10, nstate = 4, nBeta = 2, distn
  = "exp", covar = FALSE, control = list(reltol = 1e-4), seed =
  1234) {

  nsubj <- length(unique(data[,1]))

  par <- init.par(data, distn, nstate, nBeta, covar)

  res <- optim(par = par, fn = PLL, data = data, distn =
    distn, nstate = nstate, nBeta = nBeta, covar = covar,
    method = "Nelder-Mead", control = control)$par

  bootres <- BootCIPar(data, res, distn, nstate, nsubj,
    nBeta, nboot, covar, control, seed)

```

```

    est <- list(res = res, boot = bootres)

    return(est)
}

#' Bootstrap samples to find the 95\% Confidence Intervals by
#' Paralleling Loops
#'
#' @description
#' To find the 95\% Confidence intervals by bootstrap samples
#' using Parallel loops
#'
#' @param data Simulated Semi-Markov Dataset
#' @param res Best results from Nelder-Mead Optimization
#' @param distn Sojourn Distribution ("exp" = exponential, "gamma"
#'   " = gamma, "weibull" = Weibull)
#' @param nstate Number of states
#' @param nsubj Number of subjects
#' @param nBeta Number of coefficients
#' @param nboot Number of bootstrap samples
#' @param covar the covariates are needed set to TRUE (defaulted
#'   to FALSE)
#' @param control Control options for optim
#' @param seed Set the seed for reproducibility
#' @return A vector of confidence intervals for each parameter in
#'   order
#' @importFrom stats optim sd
#' @import doParallel
#' @import doRNG
#' @export
BootCIPar <- function(data, res, distn, nstate, nsubj, nBeta,
  nboot, covar, control, seed) {

  registerDoRNG(seed = seed)
  res2 <- foreach(v = 1:nboot, .combine = 'cbind', .
    packages = 'PLSMM') %dopar% {

    tryCatch({
      data2 <- data[which(data[,1] %in% sample(unique(data
        [,1]), size = nsubj, replace = TRUE)),]

      par2 <- init.par(data2, distn, nstate, nBeta,
        covar)

      optim(par = par2, fn = PLL, data = data2, distn =
        distn, nstate = nstate, nBeta = nBeta, covar =
        covar, method = "Nelder-Mead", control =

```



```

        control)$par
      }, error=function(e){})
    }

    CI <- sapply(1:length(res), function(a) res[a] + c(-1,1)
      * 1.96* sd(res2[a,]))

    return(CI)
  }

#' Summarize the real data example results in a LaTeX table
#'
#' To take each Semi-Markov model results in matrix form and
#' output quick latex code
#'
#' @param Results The Semi-Markov Model Results
#' @param nstate The number of states
#' @param distn The distribution of the sojourn time ("exp", "
#' weibull", "gamma")
#' @param nBeta Number of covariate parameters
#' @param covar Boolean statement if covariates are present
#' @return Latex code for summary table
#' @import kableExtra
#' @export
ResTable <- function(Results, nstate = 3, distn, nBeta, covar =
  FALSE){

  # Extract the parameter estimates from the model
  nrate <- nstate*(nstate - 1)
  if(distn == "exp") {stime <- mean_F(Results$res[1:nrate],0,
    dist = "exp")}
  if(distn != "exp") {stime <- c(mean_F(Results$res[1:nrate],
    Results$res[nrate+1], dist = distn), NA)}
  sum.tab <- cbind(round(Results$res, 4), round(stime, 4), paste(
    "(", round(Results$boot[1,], 4), ", ", round(Results$boot
    [2,],4), ")", sep = ""))
  colnames(sum.tab) <- c("Estimate", "Sojourn Time", "95\\%
    Confidence Inteval")
  rownames(sum.tab) <- GreekLabels(nstate, distn, nBeta, covar)

  kTab <- kable(sum.tab, "latex", vline = "", escape = F, caption
    = "Semi-Markov Model Assuming an Exponential Sojourn Time
    Distribution", linesep = "", align = rep('c',3), position =
    "!ht")

  return(kTab)
}

```

```

#’ Calculate the AIC for each model
#’
#’ Using the partial loglikelihood, we calculate the
#’ corresponding AIC for the model
#’
#’ @param res The Semi-Markov Model parameter estimates
#’ @param data The real application data in the format (ID, time,
#’ i, j, stime)
#’ @param distn The distribution of the sojourn time ("exp", "
#’ weibull", "gamma")
#’ @param nstate The number of states
#’ @param nBeta Number of covariate parameters
#’ @param covar Boolean statement if covariates are present
#’ @return aic The quality statistical measure to compare to
#’ other models
#’ @export

AIC_PLL <- function(res, data, distn, nstate = 4, nBeta = 2,
  covar = F) {

  q <- nstate*(nstate - 1)
  if(distn != "exp") {q = q + 1}
  if(covar) {q = q + nBeta}
  ll <- PLL(res, data, distn, nstate, nBeta, covar)
  aic <- 2*ll + (2*q)
  return(aic)
}

#’ Hazard Function for Graphical representation
#’
#’ @description
#’ To calculate the hazard values using either the exponential,
#’ weibull, or gamma
#’ distribution. The hazard function for the semi-markov process
#’ uses the probability
#’ matrix (pij), sojourn pdf (fij), and sojourn cdf (Fij).
#’
#’ @param rij Matrix of rate parameters for all states i, j
#’ @param pij Probability transition matrix (i X j)
#’ @param i state from
#’ @param j state to
#’ @param s sojourn time spent in the previous state i
#’ @param distn Sojourn Distribution ("exp" = exponential, "gamma
#’ " = gamma, "weibull" = Weibull)
#’ @param beta beta coefficient parameters
#’ @param covs covariate information vector for transitioning
#’ subject
#’ @param psi shape parameter for the gamma distribution
#’ @param k shape parameter for the weibull distribution

```

```

#’ @param covar boolean statement if covariates are included
#’ @return hazard value from respective hazard function
#’ @importFrom stats dexp pexp dgamma pgamma pweibull dweibull
#’ @export
hf <- function(rij, pij, i, j, s, distn, beta = 0, covs = 0, psi
  = 0, k = 0, covar = FALSE) {

  # Hazard function = pij * fij / sum_j( pij (1 - Fij) )

  if (distn == "exp") {
    # Exponential Hazard function using exponential pdf, cdf, and
    # pij
    num <- pij[i, j] * dexp(s, rij[i,j])
    den <- sapply(which(1:nrow(pij) != i), function(u) pij[i,u]*
      (1 - pexp(s, rij[i,u])))
    h <- num/sum(den)

  } else if (distn == "gamma") {
    # Gamma Hazard function using exponential pdf, cdf, and pij

    num <- pij[i, j] * dgamma(s, psi, rij[i,j])
    den <- sapply(which(1:nrow(pij) != i), function(u) pij[i,u]*
      (1 - pgamma(s, psi, rij[i,u])))
    h <- num/sum(den)

  } else if (distn == "weibull") {
    # Weibull Hazard function using exponential pdf, cdf, and pij
    num <- pij[i, j] * dweibull(s, k, 1/rij[i,j])
    den <- sapply(which(1:nrow(pij) != i), function(u) pij[i,u]*
      (1 - pweibull(s, k, 1/rij[i,u])))
    h <- num/sum(den)

  } else {
    # Send warning message if the distribution is not one of
    # these
    stop(paste("The Distribution", distn , "is not an option"))
  }

  # If covariates (i.e. TRUE), modify hazard function by cox
  # model
  if (covar == TRUE) {h <- h*exp(as.numeric(beta %*% covs))}

  return(h)
}

#’ Density plot of the sojourn time data application and various
#’ sojourn ditributions
#’

```

```

#' @description
#' To create a graphical representation of the raw data sojourn
  time and overlay it with
#' the semi-Markov model parameter estimates
#'
#' @param data The real application data in the format (ID, time,
  i, j, stime)
#' @param r state from
#' @param s state to
#' @param rate_e The matrix of rate estimates for the exponential
  case
#' @param rate_w The matrix of rate estimates for the weibull
  case
#' @param shape_w The shape parameter for the weibull case
#' @param rate_g The matrix of rate estimates for the gamma case
#' @param shape_g The shape parameter for the gamma case
#' @return p1 A plot for the density for sojourn i to j
#' @import ggplot2
#' @importFrom stats dexp pexp dgamma pgamma pweibull dweibull
#' @export
Density_Plot <- function(data, r, s, rate_e, rate_w, shape_w,
  rate_g, shape_g){
  df <- data %>% filter(i == r & j == s)
  p1 <- ggplot(data = df, aes(x = stime)) +
    geom_density(fill = "lightblue", color = "
      lightblue", alpha = 0.4) +
    xlim(c(0.01,10)) +
    stat_function(fun = dexp, args = list(rate = rate
      _e[r,s]), colour = "red") +
    stat_function(fun = dweibull, args = list(shape =
      shape_w, scale = 1/rate_w[r,s]), colour = "
      darkorange") +
    stat_function(fun = dgamma, args = list(shape =
      shape_g, rate = rate_g[r,s]), colour = "green"
    ) +
    labs(x = paste("Sojourn Time ", "(", r, " to ", s,
      ")"), sep = ""), y = "Frequency") +
    theme_minimal() +
    theme(text = element_text(family = "Palatino"))
  return(p1)
}

```

```

#' Plot of the hazard of the semi-Markov Model
#'
#' @description
#' To create a graphical representation of the hazard of the semi
  -Markov model

```

```

#'
#' @param i state from
#' @param j state to
#' @param Rije The matrix of rate estimates for the exponential
  case
#' @param Rijw The matrix of rate estimates for the weibull case
#' @param k The shape parameter for the weibull case
#' @param Rijg The matrix of rate estimates for the gamma case
#' @param psi The shape parameter for the gamma case
#' @param betae beta coefficient parameters
#' @param covs covariate information vector for transitioning
  subject
#' @param covar boolean statement if covariates are included
#' @param realData boolean statement if real data example
#' @return p1 A plot for the hazard of the semi-markov model from
  i to j
#' @import ggplot2
#' @export
PlotHSMM <- function(i, j, s, pij, Rije = 0, Rijw = 0, k = 0,
  Rijg = 0, psi = 0, betae = 0, betaw = 0, betag = 0, covs = 0,
  cseq, covar = FALSE, realData = FALSE) {

  if (realData == FALSE) {
    df <- data.frame(s = s,
                      he = hf(Rije, pij, i, j,
                              s, "exp", betae, covs,
                              psi, k, covar),
                      hw = hf(Rijw, pij, i, j,
                              s, "weibull", betaw,
                              covs, psi, k, covar),
                      hg = hf(Rijg, pij, i, j,
                              s, "gamma", betag,
                              covs, psi, k, covar))

    p1 <- ggplot(df, aes(x = s, y = he)) + geom_line(
      colour = "red") +
      geom_line(aes(x = s, y =
                    hw), colour = "
                    darkorange") +
      geom_line(aes(x = s, y =
                    hg), colour = "green")
      +

      labs(x = paste("Time
                      Spent in ", i, "
                      before ", j, sep=""),
           y = "") +
      theme_minimal() +
      theme(text = element_text
            (family = "Palatino"))
  } else if (realData == "Age") {
    covs = cbind(rep(0, length(cseq)), cseq)
  }
}

```

```

df <- data.frame(s = s,
                 hg = hf(Rijg, pij, i, j,
                        s, "gamma", betag,
                        covs[1,], psi, k,
                        covar),
                 hg2 = hf(Rijg, pij, i, j,
                          s, "gamma", betag,
                          covs[2,], psi, k,
                          covar),
                 hg3 = hf(Rijg, pij, i, j,
                          s, "gamma", betag,
                          covs[3,], psi, k,
                          covar),
                 hg4 = hf(Rijg, pij, i, j,
                          s, "gamma", betag,
                          covs[4,], psi, k,
                          covar))

p1 <- ggplot(df, aes(x = s, y = hg)) + geom_line(
  colour = "green") +
  geom_
  line(aes(x = s, y = hg2), colour = "red") +
  geom_line(aes(x = s, y =
                hg3), colour = "
                darkorange") +
  geom_line(aes(x = s, y =
                hg4), colour = "blue")
  +
  labs(x = paste("Time
                  Spent in ", i, "
                  before ", j, sep=""),
        y = "") +
  theme_minimal() +
  theme(text = element_text
        (family = "Palatino"))
} else {
df <- data.frame(s = s,
                 hg = hf(Rijg, pij, i, j,
                        s, "gamma", betag,
                        cov1, psi, k, covar))
p1 <- ggplot(df, aes(x = s, y = hg)) + geom_line(
  colour = "green") +
  scale_y_continuous(limits
                     = c(0,0.006)) +
  labs(x = paste("Time
                  Spent in ", i, "
                  before ", j, sep=""),
        y = "") +
  theme_minimal() +
  theme(text = element_text

```

```

}
return(p1)
}

```

B.2 C++ Code written for Rcpp Package in R

Listing B.2: The Risk Set Function Translated into C++

```

#include <Rcpp.h>
using namespace Rcpp;

//' @title
//' Find the index of the next lowest number
//'
//' @description Obtains the next lowest index in a vector
//'
//' @param temp a vector of numbers
//' @param tstar a number value for which to evaluate the vector
//' @return an index or position in the vector for the next
//' lowest value
int NextLow(NumericVector temp,
            double tstar) {
    int nV = temp.size(); /* Find the size of the
                           temp vector*/
    NumericVector V(nV); /* Create a new vector of
                           size nV*/

    for (int x = 0; x < nV; x++) { /* Fill the vector V with
                                    index values*/
        V[x] = x;
    }

    NumericVector Y = V[temp < tstar]; /* Find the indices which
                                        is next lowest*/
    int K = max(Y); /* Index for the next
                    lowest time*/

    return K;
}

//' @title
//' The hazard function defined by the sojourn distribution
//'
//' @description Gives a hazard value for a the current state of
//' a subject
//'
//' @param ri A vector of rates at the current state i

```

```

//' @param pij The probability matrix of the system
//' @param state The current state of the subject at the
    transition time
//' @param sojourn The time spent at the current state
//' @param distn The distribution specified for the system
//' @param beta A vector of coefficients
//' @param covs A vector of covariates for the subject
//' @param psi Shape parameter for the gamma distribution
//' @param k Shape parameter for the psi distribution
//' @param covar A boolean statement to denote if covariates are
    present
//' @return a hazard value for the current state for subject m
double hfunction(NumericVector ri,
                 NumericMatrix pij,
                 double state,
                 double sojourn,
                 String distn,
                 NumericVector beta,
                 NumericVector covs,
                 double psi,
                 double k,
                 bool covar) {

    /* Hazard function = pij * fij / sum_j (pij(1 - Fij))*/

    if (distn == "exp") {
        /* Exponential hazard function using exponential pdf, cdf,
            and pij */
        int u = ri.length(); /* Obtain the
            number of states */
        NumericVector haz(u); /* Initiate hazard
            total */

        for (int e = 0; e < u; e++) {
            double tempP = pij( state , e ); /*
                Probability transition matrix value*/
            double rval = ri[e]; /* rate
                value for the state e */
            double scale = 1/rval; /* change
                rate to scale value */
            double d = R::dexp(sojourn, scale, false); /* pdf
                for exponential distribution*/
            double num = tempP*d; /*
                numerator calculation*/

            NumericVector hden(u); /*
                initiate hazard denominator sum*/
            for (int w = 0; w < u; w++) {
                double tempP2 = pij( state , w ); /*
                    Probability Matrix value */

```



```

    double r2 = ri[w]; /*
        rate value for the state e */
    double s2 = 1/r2; /*
        change rate to scale value */
    double p = R::pexp(sojourn, s2, true, false); /*
        CDF for exponential distribution*/
    double den = tempP2*(1-p); /*
        one value for denominator sum*/
    hden[w] = den; /*
        sum the denominator sums */

}

double hij = num/sum(hden); /*
    hazard from state i to state j*/

if (covar == true) {
    double hijtemp = hij;
    /*double aval = aij( state , e ); */
    /* Obtain Transition specific constant */
    double zb = sum(beta * covs); /*
        Obtain the regression sum*/
    double expz = exp(zb); /* Sum over
        the aij and ZB*/
    hij = hijtemp*expz; /*
        Add covariates to hazard calculation*/
}

haz[e] = hij; /*
    Add to total hazard*/

}

double haztot = sum(haz);
return haztot;
} else if (distn == "gamma") {
    /* Gamma hazard function using exponential pdf, cdf, and pij
    */
    int u = ri.length(); /* Obtain the
        number of states */
    NumericVector haz(u); /* Initiate hazard
        total */

    for (int e = 0; e<u; e++) {
        double tempP = pij( state , e ); /*
            Probability transition matrix value*/
        double rval = ri[e]; /*
            rate value for the state e */
        double scale = 1/rval; /*
            change rate to scale value */
        double d = R::dgamma( sojourn, psi, scale, false); /*
            PDF for the gamma distribution*/
    }
}

```

```

double num = tempP*d; /*
    numerator calculation */

NumericVector hden(u); /*
    initiate hazard denominator sum*/
for (int w = 0; w<u; w++) {
    double tempP2 = pij( state , w );
        /* Probability Matrix value */
    double r2 = ri[w];
        /* rate value for the state e */
    double s2 = 1/r2;
        /* change rate to scale value */
    double p = R::pgamma( sojourn, psi, s2, true, false);
        /* CDF for the gamma distribution*/
    double den = tempP2*(1-p);
        /* one value for denominator sum*/
    hden[w] = den;
        /* sum the denominator sums */
}

double hij = num/sum(hden);
    /* hazard from state i to state j*/

if (covar == true) {
    double hijtemp = hij;
        /*double aval = aij( state , e ); */
        /* Obtain Transition specific
            constant */
    double zb = sum(beta * covs);
        /* Obtain the regression sum*/
    double expz = exp(zb); /* Sum
        over the aij and ZB*/
    hij = hijtemp*expz;
        /* Add covariates to hazard calculation*/
}

haz[e] = hij; /*
    Add to total hazard*/

}

double haztot = sum(haz);
return haztot;

} else if (distn == "weibull") {
    /* Weibull hazard function using exponential pdf, cdf, and
        pij */
    int u = ri.length(); /* Obtain the
        number of states */
    NumericVector haz(u); /* Initiate hazard
        total */

```

```

for (int e = 0; e<u; e++) {
    double tempP = pij( state , e );           /*
        Probability transition matrix value*/
    double rval = ri[e];                       /*
        rate value for the state e */
    double scale = 1/rval;                     /*
        change rate to scale value */
    double d = R::dweibull( sojourn , k , scale , false); /*
        PDF for the gamma distribution*/
    double num = tempP*d;                       /*
        numerator calculation*/

    NumericVector hden(u);                       /*
        initiate hazard denominator sum*/
    for (int w = 0; w<u; w++) {
        double tempP2 = pij( state , w );
            /* Probability Matrix value */
        double r2 = ri[w];
            /* rate value for the state e */
        double s2 = 1/r2;
            /* change rate to scale value */
        double p = R::pweibull( sojourn , k , s2 , true , false);
            /* CDF for the gamma distribution*/
        double den = tempP2*(1-p);
            /* one value for denominator sum*/
        hden[w] = den;
            /* sum the denominator sums */
    }

    double hij = num/sum(hden);
        /* hazard from state i to state j*/

    if (covar == true) {
        double hijtemp = hij;
        /*double aval = aij( state , e ); */
            /* Obtain Transition specific
            constant */
        double zb = sum(beta * covs);
            /* Obtain the regression sum*/
        double expz = exp(zb);                       /* Sum
            over the aij and ZB*/
        hij = hijtemp*expz;
            /* Add covariates to hazard calculation*/
    }

    haz[e] = hij;                                   /*
        Add to total hazard*/

}
double haztot = sum(haz);
return haztot;

```

```

    } else {
        Rprintf("Warning: The Distribution given is not available")
        ;
        double haztot = -99;
        return haztot;
    }
}

/** @title Calculate the total hazard in the risk set
**/
**/ @description Using the data, the function will calculate the
total hazard for all the subejct currently at Risk
**/
**/ @param denData A list of by subject information
**/ @param denTime A vector of ordered transitioning times
**/ @param rij A matrix of rate parameters for each state i and
state j
**/ @param pij The probability matrix for the system
**/ @param distn The distribution specified for the system
**/ @param beta A vector of coefficents
**/ @param covs A vector of covariates for the subject
**/ @param psi Shape parameter for the gamma distribution
**/ @param k Shape parameter for the psi distribution
**/ @param covar A boolean statement to denote if covariates are
present
**/ @return The total hazard at the nth transition
**/ [[Rcpp::export]]
NumericVector RiskSet(List denData,
                      NumericVector denTime,
                      NumericMatrix rij,
                      NumericMatrix pij,
                      String distn,
                      NumericVector beta,
                      NumericMatrix covs,
                      double psi = 0,
                      double k = 0,
                      bool covar = false) {

    int Ti = 0; /* Index for the time column*/
    int I = 1; /* Index for the i state column
    */
    int J = 2; /* Index for the j state column
    */

    int nList = denData.size(); /* Size of the subject list*/
    int nT = denTime.size(); /* Size of the transition time
    vector*/

```

```

NumericVector lden(nT);          /* Initiallize the likelihood
    denominator*/

for (int a = 0; a < nT; a++) {

    double tstar = denTime[a];    /* Obtain the current
        transition time*/
    double lt = 0;                /* Temporary storage for
        the denominator of the likelihood*/

    for (int m = 0; m < nList; m++) {

        NumericMatrix subjM = denData[m];    /* Obtain the
            information for the mth subject*/
        NumericVector time = subjM( _ , Ti ); /* Obtain the
            time for the mth subject*/
        double tmax = max(time);            /* Find the max
            observation time for the mth subject*/

        if (tmax >= tstar) {

            int A = NextLow(time, tstar);    /* Get
                the index for the next lowest time*/
            double JumpT = time[A];        /* Last
                jump time for subject m*/

            int curstate = subjM( A , J ); /*
                Obtain the current subject during transition*/
            if (JumpT == tstar) {curstate = subjM( A , I );} /* If
                the trandistion time is the last obs, get j state*/
            int state = curstate - 1;      /* Re-
                Index state to match c++ index*/

            double sojourn = tstar - JumpT; /*
                Calculate the time already in current state*/
            NumericVector ri = rij( state , _ );
                /* Obatin the vector rates for the current state */
            NumericVector covV = covs( 0 , _ );
            if (covar == true) {NumericVector covV = covs( m , _ );}
                /* Obtain the covariate information for
                subject m */

            /* Calculate the hazard for specified distribution,
                current state of subject m*/
            double ltemp = hfunction(ri, pij, state, sojourn, distn,
                beta, covV, psi, k, covar);
            lt = lt + ltemp;                /* Add
                subject m hazard to total */
        }
    }
}

```

```
    }  
    lden[a] = lt;          /* Store the total hazard for this  
                           transition in a vector */  
  }  
  return lden;           /* Return the vector subject likelihood  
                           contribution*/  
}
```

Bibliography

- [1] D. R. Cox, “Regression models and life-tables,” *Springer Series in Statistics Breakthroughs in Statistics*, pp. 527–541, 1992. DOI: 10.1007/978-1-4612-4380-9_37.
- [2] H. H. Chen, S. W. Duffy, and L. Tabar, “A markov chain method to estimate the tumour progression rate from preclinical to clinical phase, sensitivity and positive predictive value for mammography in breast cancer screening,” *The Statistician*, vol. 45, no. 3, p. 307, 1996. DOI: 10.2307/2988469.
- [3] R. Perez-Ocón, J. E. Ruiz-Castro, and M. L. Gámiz-Perez, “A multivariate model to measure the effect of treatments in survival to breast cancer,” *Biometrical Journal*, vol. 40, no. 6, pp. 703–715, 1998. DOI: 10.1002/(sici)1521-4036(199810)40:6<703::aid-bimj703>3.0.co;2-7.
- [4] I. M. Longini, W. S. Clark, R. H. Byers, J. W. Ward, W. W. Darrow, G. F. Lemp, and H. W. Hethcote, “Statistical analysis of the stages of hiv infection using a markov model,” *Statistics in Medicine*, vol. 8, no. 7, pp. 831–843, 1989. DOI: 10.1002/sim.4780080708.
- [5] A. Alioum, V. Leroy, D. Commenges, F. Dabis, and R. Salmon, “Effect of gender, age, transmission category, and antiretroviral therapy on the progression of human immunodeficiency virus infection using multistate markov models,” *Epidemiology*, vol. 9, no. 6, pp. 605–612, 1998. DOI: 10.1097/00001648-199811000-00007.
- [6] I. Kousignian, B. Autran, C. Chouquet, V. Calvez, E. Gomard, C. Katlama, Y. Rivière, and D. Costagliola, “Markov modelling of changes in hiv-specific cytotoxic t-lymphocyte responses with time in untreated hiv-1 infected patients,” *Statistics in Medicine*, vol. 22, no. 10, pp. 1675–1690, 2003. DOI: 10.1002/sim.1404.
- [7] R. J. Kryscio, F. A. Schmitt, J. C. Salazar, M. S. Mendiondo, and W. R. Markesbery, “Risk factors for transitions from normal to mild cognitive impairment and dementia,” *Neurology*, vol. 66, no. 6, pp. 828–832, 2006. DOI: 10.1212/01.wnl.0000203264.71880.45.
- [8] D. C. Ewbank, “A multistate model of the genetic risk of alzheimers disease,” *Experimental Aging Research*, vol. 28, no. 4, pp. 477–499, 2002. DOI: 10.1080/03610730290103096.

- [9] P. Jepsen, H. Vilstrup, and P. K. Andersen, “The clinical course of cirrhosis: The importance of multistate models and competing risks analysis,” *Hepatology*, vol. 62, no. 1, pp. 292–302, 2015. DOI: 10.1002/hep.27598.
- [10] P. Saint-Pierre, C. Combescure, J. Daurès, and P. Godard, “The analysis of asthma control under a markov assumption with use of covariates,” *Statistics in Medicine*, vol. 22, no. 24, pp. 3755–3770, Aug. 2003. DOI: 10.1002/sim.1680.
- [11] A. Duffy, J. Horrocks, S. Doucette, C. Keown-Stoneman, S. McCloskey, and P. Grof, “The developmental trajectory of bipolar disorder,” *British Journal of Psychiatry*, vol. 204, no. 2, pp. 122–128, 2014. DOI: 10.1192/bjp.bp.113.126706.
- [12] V. Barbu and N. Limnios, *Semi-Markov chains and hidden semi-Markov models toward applications: their use in reliability and DNA analysis*. Springer, 2008.
- [13] H. M. Taylor and S. Karlin, *An introduction to stochastic modeling*. Academic Press, 2014.
- [14] G. H. Weiss and M. Zelen, “A semi-markov model for clinical trials,” Jan. 1963. DOI: 10.21236/ad0407905.
- [15] P. Levy, “Processus semi-markoviens,” in *Processus semi-markoviens*. North-Holland Publishing Co., 1954, vol. 3, pp. 416–426.
- [16] S. L. W, “Regenerative stochastic processes,” in. Regenerative stochastic processes, 1955, vol. 232, pp. 6–31.
- [17] L. Takacs, “Some investigations concerning recurrent stochastic processes of a certain type,” in. Acta Mathematica Hungarica, 1954, vol. 3, pp. 115–128.
- [18] V. S. Korolyuk, S. M. Brodi, and A. F. Turbin, “Semi-markov processes and their applications,” *Journal of Soviet Mathematics*, vol. 4, no. 3, pp. 244–280, 1975, ISSN: 1573-8795. DOI: 10.1007-BF01097184.
- [19] R. Pyke, “Markov renewal processes: Definitions and preliminary properties,” *The Annals of Mathematical Statistics*, vol. 32, no. 4, pp. 1231–1242, 1961. DOI: 10.1214/aoms/1177704863.
- [20] R. Pyke and R. Schaufele, “Limit theorems for markov renewal processes,” *The Annals of Mathematical Statistics*, vol. 35, no. 4, pp. 1746–1764, 1964. DOI: 10.1214/aoms/1177700397.
- [21] R. Pyke and R. Schaufele, “The existence and uniqueness of stationary measures for markov renewal processes,” *The Annals of Mathematical Statistics*, vol. 37, no. 6, pp. 1439–1462, 1966. DOI: 10.1214/aoms/1177699138.
- [22] Y. Foucher, E. Mathieu, P. Saint-Pierre, J.-F. Durand, and J.-P. Daurès, “A semi-markov model based on generalized weibull distribution with an illustration for hiv disease,” *Biometrical Journal*, vol. 47, no. 6, pp. 825–833, 2005. DOI: 10.1002/bimj.200410170.

- [23] M. Kang and S. W. Lagakos, “Statistical methods for panel data from a semi-markov process, with application to hpv,” *Biostatistics*, vol. 8, no. 2, pp. 252–264, 2007. DOI: 10.1093/biostatistics/kxl006.
- [24] Q. Cao, E. Buskens, T. Feenstra, H. Hillege, and D. Postmus, “Continuous-time semi-markov models in health economic decision making: An illustrative example in heart failure disease management,” *Med Decis Making*, vol. 36, no. 1, pp. 59–71, 2015. DOI: 10.1177/0272989X15593080.
- [25] S.-Z. Yu, “General hidden semi-markov model,” *Hidden Semi-Markov Models*, pp. 27–58, 2016. DOI: 10.1016/b978-0-12-802767-7.00002-4.
- [26] P. K. Anderson, S. Esbjerg, and T. Sorensen, “Multi-state models for bleeding episodes and mortality in liver cirrhosis,” *Stat. Med.*, vol. 19, pp. 587–599, 2000. DOI: 10.1002/(sici)1097-0258(20000229)19:4<587::aid-sim358>3.0.co;2-0.
- [27] A. C. Titman, “Estimating parametric semi-markov models from panel data using phase-type approximations,” *Statistics and Computing*, vol. 24, no. 2, pp. 155–164, 2012. DOI: 10.1007/s11222-012-9360-6.
- [28] Y. Shu, J. P. Klein, and M.-J. Zhang, “Asymptotic theory for the cox semi-markov illness-death model,” *Lifetime Data Anal*, vol. 13, pp. 91–117, 2007. DOI: 10.1007/s10985-006-9018-9.
- [29] H. Aralis and R. Brookmeyer, “A stochastic estimation procedure for intermittently-observed semi-markov multistate models with back transitions,” *Statistical Methods in Medical Research*, pp. 1–18, 2017. DOI: 10.1177/0962280217736342.
- [30] J. S. Benoit, W. Chan, S. Luo, H.-W. Yeh, and R. Doody, “A hidden markov model approach to analyze longitudinal ternary outcomes when some observed states are possibly misclassified,” *Statistics in Medicine*, vol. 35, no. 9, pp. 1549–1557, 2016. DOI: 10.1002/sim.6861.
- [31] K. S. Markides, C. A. Stroup-Benham, J. S. Goodwin, L. C. Perkowski, M. Lichtenstein, and L. A. Ray, “The effect of medical conditions on the functional limitations of mexican-american elderly,” *Annals of Epidemiology*, vol. 6, no. 5, pp. 386–391, 1996. DOI: 10.1016/s1047-2797(96)00061-0.
- [32] L. S. Radloff, “The ces-d scale,” *Applied Psychological Measurement*, vol. 1, no. 3, pp. 385–401, 1977. DOI: 10.1177/014662167700100306.
- [33] J. R. Moon, J. Huh, J. Song, I.-S. Kang, S. W. Park, S.-A. Chang, J.-H. Yang, and T.-G. Jun, “The center for epidemiologic studies depression scale is an adequate screening instrument for depression and anxiety disorder in adults with congenital heart disease,” *Health and Quality of Life Outcomes*, vol. 15, no. 1, May 2017. DOI: 10.1186/s12955-017-0747-0.
- [34] D. Eddelbuettel and J. J. Balamuta, “Extending extitR with extitC++: A Brief Introduction to extitRcpp,” *PeerJ Preprints*, vol. 5, e3188v1, Aug. 2017, ISSN: 2167-9843. DOI: 10.7287/peerj.preprints.3188v1. [Online]. Available: <https://doi.org/10.7287/peerj.preprints.3188v1>.

- [35] M. Corporation and S. Weston, *Doparallel: Foreach parallel adaptor for the 'parallel' package*, R package version 1.0.15, 2019. [Online]. Available: <https://CRAN.R-project.org/package=doParallel>.
- [36] R. B. Fetter and J. D. Thompson, "A decision model for the design and operation of a progressive patient care hospital," *Medical Care*, vol. 7, no. 6, pp. 450–462, 1969. DOI: 10.1097/00005650-196911000-00004.
- [37] B. Ouhbi and N. Limnios, "Non-parametric estimation for semi-markov kernels with application to reliability analysis," *Applied Stochastic Models and Data Analysis*, vol. 12, no. 4, pp. 209–220, 1996. DOI: 10.1002/(sici)1099-0747(199612)12:4<209::aid-asm284>3.0.co;2-t.
- [38] B. Ouhbi and N. Limnios, "Nonparametric estimation for semi-markov processes based on its hazard rate functions," *Statistical Inference for Stochastic Processes*, vol. 2, pp. 151–173, 1999.
- [39] M. R. Sternberg and G. A. Satten, "Discrete-time nonparametric estimation for semi-markov models of chain-of-events data subject to interval censoring and truncation," *Biometrics*, vol. 55, no. 2, pp. 514–522, 1999. DOI: 10.1111/j.0006-341x.1999.00514.x.
- [40] H. Damerdj, "Maximum likelihood estimation for generalized semi-markov processes," *Discrete Event Dynamic Systems*, vol. 6, no. 1, pp. 73–104, 1996. DOI: 10.1007/bf01796784.
- [41] P. Joly and D. Commenges, "A penalized likelihood approach for a progressive three-state model with censored and truncated data: Application to aids," *Biometrics*, vol. 55, no. 3, pp. 887–890, 1999. DOI: 10.1111/j.0006-341x.1999.00887.x.
- [42] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972. DOI: 10.1111/j.2517-6161.1972.tb00899.x.
- [43] N. Breslow, "Covariance analysis of censored survival data," *Biometrics*, vol. 30, no. 1, p. 89, 1974. DOI: 10.2307/2529620.
- [44] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2019. [Online]. Available: <https://www.R-project.org/>.
- [45] C. H. Jackson, "Multi-state models for panel data: The msm package for R," *Journal of Statistical Software*, vol. 38, no. 8, pp. 1–29, 2011. [Online]. Available: <http://www.jstatsoft.org/v38/i08/>.
- [46] S. A. Black, K. S. Markides, and T. Q. Miller, "Correlates of depressive symptomatology among older community-dwelling mexican americans: The hispanic epese," *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, vol. 53B, no. 4, Jan. 1998. DOI: 10.1093/geronb/53b.4.s198.

- [47] S. Soto, E. M. Arredondo, M. T. Villodas, J. P. Elder, E. Quintanar, and H. Madanat, “Depression and chronic health conditions among latinos: The role of social networks,” *Journal of Immigrant and Minority Health*, vol. 18, no. 6, pp. 1292–1300, 2016. DOI: 10.1007/s10903-016-0378-2.
- [48] M. A. Monserud and K. S. Markides, “Changes in depressive symptoms during widowhood among older mexican americans: The role of financial strain, social support, and church attendance,” *Aging & Mental Health*, vol. 21, no. 6, pp. 586–594, Jul. 2016. DOI: 10.1080/13607863.2015.1132676.
- [49] C. J. Swenson, J. Baxter, S. M. Shetterly, S. L. Scarbro, and R. F. Hamman, “Depressive symptoms in hispanic and non-hispanic white rural elderly the san luis valley health and aging study,” *American Journal of Epidemiology*, vol. 152, no. 11, pp. 1048–1055, Jan. 2000. DOI: 10.1093/aje/152.11.1048.
- [50] H. Oh, K. Ell, and A. Subica, “Depression and family interaction among low-income, predominantly hispanic cancer patients: A longitudinal analysis,” *Supportive Care in Cancer*, vol. 22, no. 2, pp. 427–434, Apr. 2013. DOI: 10.1007/s00520-013-1993-2.
- [51] *The next four decades: The older population in the united states: 2010 to 2050*. [Online]. Available: <https://www.census.gov/prod/2010pubs/p25-1138.pdf>.
- [52] W. J. Katon, “Clinical and health services relationships between major depression, depressive symptoms, and general medical illness,” *Biological Psychiatry*, vol. 54, no. 3, pp. 216–226, 2003. DOI: 10.1016/s0006-3223(03)00273-7.
- [53] S. Rote, N.-W. Chen, and K. Markides, “Trajectories of depressive symptoms in elderly mexican americans,” *Journal of the American Geriatrics Society*, vol. 63, no. 7, pp. 1324–1330, 2015. DOI: 10.1111/jgs.13480.
- [54] S. A. Riolo, T. A. Nguyen, J. F. Greden, and C. A. King, “Prevalence of depression by race/ethnicity: Findings from the national health and nutrition examination survey iii,” *American Journal of Public Health*, vol. 95, no. 6, pp. 998–1000, 2005. DOI: 10.2105/ajph.2004.047225.
- [55] K. S. Markides, “Hispanic established populations for the epidemiologic studies of the elderly, 1993-1994: [arizona, california, colorado, new mexico, and texas],” *ICPSR Data Holdings*, Apr. 2000. DOI: 10.3886/icpsr02851.v2.
- [56] K. S. Markides, “Hispanic established populations for epidemiologic studies of the elderly, wave ii, 1995-1996: [arizona, california, colorado, new mexico, and texas],” *ICPSR Data Holdings*, 2002. DOI: 10.3886/icpsr03385.v2.
- [57] K. S. Markides, “Hispanic established populations for epidemiologic studies of the elderly, wave iii, 1998-1999: [arizona, california, colorado, new mexico, and texas],” *ICPSR Data Holdings*, May 2004. DOI: 10.3886/icpsr04102.v2.
- [58] K. S. Markides and L. A. Ray, “Hispanic established populations for epidemiologic studies of the elderly, wave iv, 2000-2001 [arizona, california, colorado, new mexico, and texas],” *ICPSR Data Holdings*, Apr. 2005. DOI: 10.3886/icpsr04314.v2.

- [59] K. S. Markides, L. A. Ray, R. Angel, and D. V. Espino, “Hispanic established populations for the epidemiologic study of the elderly (hepese) wave 5, 2004-2005 [arizona, california, colorado, new mexico, and texas],” *ICPSR Data Holdings*, 2009. DOI: 10.3886/icpsr25041.
- [60] K. S. Markides, L. A. Ray, R. Angel, and D. V. Espino, “Hispanic established populations for the epidemiologic study of the elderly (hepese) wave 6, 2006-2007 [arizona, california, colorado, new mexico, and texas],” *ICPSR Data Holdings*, 2012. DOI: 10.3886/icpsr29654.v1.
- [61] K. S. Markides, N.-W. Chen, R. Angel, R. Palmer, and J. Graham, “Hispanic established populations for the epidemiologic study of the elderly (hepese) wave 7, 2010-2011,” *ICPSR Data Holdings*, Dec. 2016. DOI: 10.3886/ICPSR36537.v2.
- [62] K. S. Markides, N.-W. Chen, R. Angel, and R. Palmer, “Hispanic established populations for the epidemiologic study of the elderly (hepese) wave 8, 2012-2013,” *ICPSR Data Holdings*, Nov. 2016. DOI: 10.3886/ICPSR36578.v2.
- [63] J. H. Boyd, M. M. Weissman, W. D. Thompson, and J. K. Myers, “Screening for depression in a community sample: Understanding the discrepancies between depression symptom and diagnostic scales,” *Archives of General Psychiatry*, vol. 39, no. 10, pp. 1195–1200, Oct. 1982, ISSN: 0003-990X. DOI: 10.1001/archpsyc.1982.04290100059010.