COUPLED HIDDEN MARKOV MODEL WITH BIVARIATE DISCRETE COPULA TO STUDY COMORBIDITY OF CHRONIC DISEASES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

ZARINA OFLAZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
STATISTICS

FEBRUARY 2022

Approval of the thesis:

COUPLED HIDDEN MARKOV MODEL WITH BIVARIATE DISCRETE COPULA TO STUDY COMORBIDITY OF CHRONIC DISEASES

submitted by **ZARINA OFLAZ** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Statistics Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** ______________

Prof. Dr. Özlem İlk Dağ
Head of Department, **Statistics** ______________

Prof. Dr. Ceylan Yozgatlıgil
Supervisor, **Department of Statistics, METU** ______________

Prof. Dr. Ayşe Sevtap Kestel
Co-supervisor, **Institute of Applied Mathematics, METU** ______________

Examining Committee Members:

Prof. Dr. Özlem İlk Dağ
Department of Statistics, METU ______________

Prof. Dr. Ceylan Yozgatlıgil
Department of Statistics, METU ______________

Assist. Prof. Dr. Uğur Karabey
Department of Actuarial Sciences, Hacettepe University ______________

Assoc. Prof. Dr. Könül Bayramoğlu Kavlak
Industrial Engineering, Boğaziçi University ______________

Assist. Prof. Dr. Mustafa Hilmi Pekalp
Actuarial Sciences, Ankara University ______________

Date: 09.02.2022

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Zarına Oflaz

Signature :

ABSTRACT

COUPLED HIDDEN MARKOV MODEL WITH BIVARIATE DISCRETE COPULA TO STUDY COMORBIDITY OF CHRONIC DISEASES

Oflaz, Zarına
Ph.D., Department of Statistics
Supervisor: Prof. Dr. Ceylan Yozgatlıgil
Co-Supervisor: Prof. Dr. Ayşe Sevtap Kestel

February 2022, 86 pages

A range of chronic diseases have a significant influence on each other and share common risk factors. Comorbidity, which shows the existence of two or more diseases interacting or triggering each other, is an important measure for actuarial valuations. The main proposal of the thesis is to model parallel interacting processes describing two or more chronic diseases by a combination of hidden Markov theory and copula function. This study introduces a novel coupled hidden Markov model with the bivariate discrete copula function in the hidden process. We use a novel discrete copula, namely the Binomial copula. We compute a complete data log-likelihood and develop an inference necessary to implement the model. To estimate the parameters of the model and deal with the numerical intractability of the log-likelihood, we propose a variational expectation maximization (VEM) algorithm. To perform the VEM algorithm, a lower bound of the model's log-likelihood is defined, and estimators of the parameters are computed in M-part. A possible numerical underflow occurring in the computation of forward-backward probabilities is solved.

The simulation study was conducted for two different odds ratios to assess the performance of the proposed model, resulting in satisfactory findings. Additionally, the proposed model was applied to hospital appointment data from a private hospital. The model defines the dependency structure of unobserved disease data and its dynamics. The application results demonstrate that the proposed model is useful for investigating disease comorbidity when only population dynamics over time are available and no clinical data are available.

Keywords: Coupled Hidden Markov Model, Discrete Copula, Dependency in Hidden Process, Comorbidity, Chronic Diseases

ÖZ

KRONİK HASTALIKLARIN KOMORBİDİTESİNİ İNCELEMEK İÇİN İKİ DEĞİŞKENLİ KESİKLİ KAPULA İLE BİRLEŞTİRİLMİŞ GİZLİ MARKOV MODELİ

Oflaz, Zarına
Doktora, İstatistik Bölümü
Tez Yöneticisi: Prof. Dr. Ceylan Yozgatlıgil
Ortak Tez Yöneticisi: Prof. Dr. Ayşe Sevtap Kestel

Şubat 2022 , 86 sayfa

Bir dizi kronik hastalığın birbirleri üzerinde önemli etkileri vardır ve ortak risk faktörlerini paylaşırlar. Birbiriyle etkileşen veya birbirini tetikleyen iki veya daha fazla hastalığın varlığını gösteren komorbidite, aktüeryal değerlemeler için önemli bir ölçüttür. Tezin ana önerisi, saklı Markov teorisi ve kapula fonksiyonunun bir kombinasyonu ile iki veya daha fazla kronik hastalığı tanımlayan paralel etkileşimli süreçleri modellemektir. Bu çalışmada, saklı süreçte iki değişkenli ayrık kapula fonksiyonu ile yeni bir birleşik saklı Markov modeli geliştirildi. Yeni bir ayrık kapula, yani Binom kapula kullanılmıştır. Tam-veri model olasılığını hesaplıyoruz ve modeli uygulamak için gerekli bir çıkarım geliştirilmiştir. Modelin parametrelerini tahmin etmek ve log-olasılığının sayısal zorluğu ele almak için, bir varyasyonel beklenti maksimizasyonu (VBM) algoritması önerilmiştir. VBM algoritmasını gerçekleştirmek için, modelin log-olabilirliğinin bir alt sınırı tanımlanır ve parametrelerin tahmin edicileri M-bölümünde hesaplanır. İleri-geri olasılıkların hesaplanmasında meydana gelen olası bir sayısal taşma çözülür.

Önerilen modelin performansını değerlendirmek için iki farklı olasılık oranının simülasyon çalışması yapıldı ve tatmin edici bulgular elde edildi. Ayrıca önerilen model özel bir hastaneden hastane randevu verilerine uygulanmıştır. Model, gözlemlenmemiş hastalık verilerinin bağımlılık yapısını ve dinamiklerini tanımlar. Uygulama sonuçları, önerilen modelin yalnızca zaman içindeki popülasyon dinamikleri mevcut olduğunda ve hiçbir klinik veri mevcut olmadığında hastalık komorbiditesini araştırmak için yararlı olduğunu göstermektedir.

Anahtar Kelimeler: Birleştirilmiş Gizli Markov Modeli, Kesikli Kapula, Gizli Süreçte Bağımlılık, Komorbidite, Kronik Hastalıkları

To My Lovely Family

And

My Grandmother Mathematician Zoya Zhanatkizi

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisors, Prof. Dr. Ceylan Yozgatlıgil and Prof. Dr. Ayşe Sevtap Kestel, for your patience, guidance and support. I have benefited greatly from your wealth of knowledge and meticulous editing. I am extremely grateful that you took me on as a student and continued to have faith in me over the years.

I am grateful to Prof. Dr. Özlem İlk Dağ, the Head of Department of Statistics, and Assist. Prof. Dr. Uğur Karabey for the valuable suggestions, ever encouraging and motivating guidance you provided me as part of my dissertation committee. Thank you to my committee members, Assoc. Prof. Dr. Könül Bayramoğlu Kavlak and Assist. Prof. Dr. Mustafa Hilmi Pekalp. Your detailed feedback have been very important to me. I also want to thank my dissertation committee for letting my defense be an enjoyable moment, and for your brilliant comments and suggestions.

Thank you to Prof. Dr. Bayram Sade, the Rector of KTO Karatay University, for all of the kind words and support you have provided. My big appreciation to Assist. Prof. Dr. Ahmet Atılgan, the Head of Department of Insurance and Social Security, for your unconditional and endless support during my PhD study. Also, I extend my thanks to all my colleagues at Karatay University and Middle East Technical University for their continuous encouragement and support, for their time, advice and moral support.

I would like to express my gratitude to my students at KTO Karatay University. Their eagerness to learn and discover new things, as well as their ambition to grow themselves, made my academic and teaching path more meaningful.

Thank you to my parents and parents-in-law for your endless support. You have always stood behind me, and this was no exception. Mom, thank you for all of your love and for always reminding me of the end goal. Thank you to my grandmother Zoya Zhanatkizi, the talented mathematician, whose enthusiasm for mathematics has always encouraged me. If I had a chance, I would like to discuss with you Science and Mathematics.

Last, but not least, my warm and heartfelt thanks go to my lovely husband and son, for their tremendous support and hope they had given to me. Without that hope, this thesis would not have been possible. Thank you all for the strength you gave me.

TABLE OF CONTENTS

CHAPTERS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

LIST OF ABBREVIATIONS

$(X_{i,t})$	the state-dependent process
$(\mathbf{S}_i)$	the hidden process
$S_{i,t}$	the hidden state for disease i at time t
$X_{i,t}$	the observation for disease i at time t
$(S_t)_t$ or $(\mathbf{S}_t)$	the joint hidden process
$\Phi_q(X_{i,t})$	an emission distribution
m_q	an initial distribution of the states $1 \leq q \leq Q$
$\pi_{q,r}$	a transition probability
$\bar{\mathbf{p}}(\mathbf{S}_i, \mathbf{S}_j)$	a bivariate copula probability math function
$\log P(\mathbf{X}, \mathbf{S})$	the complete data log-likelihood
$J(\mathbf{X}, \theta, \tilde{P}$	the lower bound of the complete data log-likelihood
$\tilde{P}(\mathbf{S})$	an approximation of hidden process
$\tau^r_{i,t}, \Lambda^{qr}_{i,t}$	the conditional expectations given the observations $\mathbf{X}_{i,t}$
$h^l_{i,t}$	a variational parameter for a Markov chain
$F^q_{i,t}$	the forward probabilities
$G^r_{i,t+1}$	the backward probabilities
p_{xy}	joint probability for $x, y \in \{0, 1\}$
HMM	hidden Markov model
CHMM	coupled hidden Markov model
CDLL	complete data log-likelihood
EM	expectation maximization algorithm
VEM	variational expectation maximization algorithm

CHAPTER 1

INTRODUCTION

The precise estimation of mortality and morbidity rates is critical for optimal pricing of life insurance and health insurance products. Chronic diseases have a significant impact on how insurance reserves and premiums are managed, as chronic diseases account for the majority of claimed patient illnesses. Especially, unexpectedly high costs regarding the critical illnesses require understanding the factors in the likelihood of such an event.

Comorbidity is a medical term that refers to when a patient has two or more diseases concurrently. A range of chronic diseases have a significant influence on each other and share common risk factors. Our aim is to investigate these unknown shared factors that contribute to the development of these diseases. We are particularly interested in estimating the probability of chronic disease comorbidity, or how the presence of one disease may affect the likelihood of being exposed to another. For example, a person with diabetes is at a greater risk of developing cardiovascular disease, such as ischemic heart disease, than a person without diabetes.

The theoretical set up in the thesis is mainly to combine hidden Markov and copula theories describing the probabilistic joint behavior of two or more chronic diseases. We use a coupled hidden Markov model (CHMM) as the fundamental model in this study to capture the comorbidity phenomenon by combining the hidden processes underlying chronic diseases. The causal nature of disease-disease interactions can be represented using dynamic networks, in which the strength of the edges connecting nodes varies over time in accordance to the joint copula distribution. The underlying network dynamics are frequently unknown, and what we perceive are sequences of observed events propagating across the network. To infer latent network dynamics from observed sequences, one must consider both when and what events occurred in the past, as both provide insight into the mechanisms behind disease generation and progression.

We propose a novel CHMM with copula accounting for interaction between diseases in the latent space. Since the hidden process of the CHMM is defined on a discrete state space, the probability of joint hidden states is modeled by bivariate discrete copula proposed by Geenens, 2020. Sklar's theorem states that the copula of a discrete random vector is not completely identifiable, resulting in serious inconsistencies. Therefore, Geenens, 2020 develops the rejuvenating approach of copula modeling for discrete data based on Yule's (Yule, 1912), Goodman and Kruskal's (Goodman and Kruskal, 1979), and Mosteller's (Mosteller, 1968) conceptions.

Several research have used copula theory to examine competing risks and comorbidity. While competing risks are investigated using a variety of copula-based approaches (Y. Wang and Pham, 2011,Kaishev et al., 2007,Chen, 2010, Escarela and Carriere, 2003, Lo and Wilke, 2010), there is only one study examining chronic disease comorbidity using a copula-based approach (Stöber et al., 2015). To our knowledge, no study has been conducted that uses a combination of any type of HMM and copula function to investigate comorbidity or competing risks.

The proposed model is an incomplete data model for which the expectation maximization (EM) algorithm (Dempster et al., 1977) is the most often used probability maximization technique. However, the model's exact inference creates a number of computing problems. In this study, we define a probabilistic model in general form when the number of underlying hidden chains exceeds two, resulting in a large number of parameters. We employ variational approximation to make the expectation (E) step of the EM algorithm tractable in terms of computation. The resulting variational EM (VEM) seeks to maximize the lower bound on the log-likelihood. While VEM is initially formally described in machine learning applications such as (L. K. Saul et al., 1996), it is now frequently deployed and generalized in a variety of ways. Comprehensive summary of studies can be found in in the works of Jaakkola, 2000, Wainwright and Jordan, 2008, Blei et al., 2017. Ormerod and Wand, 2010 gives also an explanation of VEM in statistical terms.

In this study, we present theoretical advances for defining a probabilistic model and developing the necessary inference to implement the model. In particular, we compute the complete data log-likelihood (CDLL) and its lower bound, derive forward-backward probabilities and conditional expectations required for the E-step, and derive estimators for the model parameters required for the maximization (M) step. We propose an approximate inference algorithm based on a variational approach. A simulation study is performed to assess the performance of the proposed method and an application to the detection of comorbidity levels in heart diseases and hypertension is presented.

The following section contains a review of the literature on the use of HMM with covariates in latent processes to determine the morbidity and mortality rate of IHD. Additionally, we provide a review of studies that have used classical statistical approaches to investigate the comorbidity and competing risks problem. Additionally, studies utilizing multi-state models, specifically HMMs, are examined, as well as HMMs in conjunction with the copula function. We compare the proposed model's strength to conventional models of comorbidity or competing risks in this section. Additionally, the first chapter discusses extended hidden Markov models (HMMs), which combine two or more HMMs to model the interaction of multiple processes. We discuss the proposed model's advantages in comparison to other coupled HMMs in the literature. The second chapter discusses the theoretical foundations of CHMM. The third chapter provides theoretical framework for CHMM and contemporaneous copula theory, as well as foundation for a newly developed copula function defined on discrete space. The fourth chapter defines the proposed model's structure and assumptions, the VEM algorithm, and the proposals necessary to conduct statistical inferences on the model and its derivations. Additionally, the fourth chapter contains an algorithm developed for estimating the model's parameters. The fifth chapter describes the simulation design and the results of the simulation study. The sixth chapter includes data analysis of a real-world data set containing patients with heart disease and hypertension. Additionally, the application of the proposed model to the data set and interpretation of the results are provided.

1.1 Literature Review

Numerous statistical models have been developed to elucidate the morbidity and mortality of chronic diseases. Mixed-effects regression models to estimate mortality and morbidity in extremely preterm infants (Schmidt et al., 2019). Machine learning models and statistical models (logistic regression and the k-nearest neighbor method) are implied to predict pneumonia mortality (Cooper et al., 1997).

The Kaplan-Meier estimates of survival curves and the univariate and multivariate Cox's proportional hazards models are being the most well-known ones (Salles et al., 2004,Souza et al., 2015). Extensive application of these methods, such as the hazard ratio of ischemic bowel (W.-S. Hu and Lin, 2017), chronic renal (Valdez-Ortiz et al., 2018), chronic Chagas' (Gonçalves et al., 2010), and ischemic heart diseases (Yen and Chen, 2013) using univariate and multivariate Cox's model can be found in the literature. These studies aim mostly to determine the risk factors associated with mortality due to these chronic diseases. The Kaplan-Meier for survival and logistic regression modeling are conducted to predict the mortality of patients who undergo vascular surgery (Esteban et al., 2019). Also, a composite endpoint with a Kaplan–Meier type analysis is used to evaluate patients' disease progression (Wong et al., 2007). Additionally, Poisson regression model is applied to determine the effect of fine particulate matter on daily mortality (Maté et al., 2010). Z.-y. Huang et al., 2020 presents a K-means-based Multiple locally weighted linear regression model to predict new local COPD hospitalizations number per week. Artificial Neural Network is used as a predictive model for diagnosing hypertension (Tengnah et al., 2019). Morbidity and life expectancy of patients with hypertension is studied using the Markov prediction model (Suciu et al., 2019).

Chronic diseases typically have a multi-state nature that dynamically progress from early to late stages and are influenced by a variety of internal and external risk factors. To investigate the structure of disease progression, multi-state models are increasingly utilized. For instance, breast cancer progression is examined with non-homogeneous exponential regression Markov models (Hsieh et al., 2002). Meenaxi and Singh, 2018 presented the use of the Markov process to demonstrate its efficacy in providing a survival analysis of a patient with chronic heart failure due to a reduced ejection fraction. The progression rate of type 2 diabetes is quantified using the Markov model (Shih et al., 2007). Collaborative topic modeling and the Gaussian mixture method have been employed to study the distribution and progression of chronic disease in a population using information on human mobility patterns (Y. Wang et al., 2018). Luo et al., 2021 proposed applying the continuous-time HMM to explore the progression of chronic obstructive pulmonary disease using longitudinal health records. A multi-state continuous time non-homogeneous Markov model is employed to the study disease progression of patients with decreased renal function (Begun et al., 2013). Visual analytics with HMMs have been employed to investigate disease progression pathways of chronic diseases (B. C. Kwon et al., 2021). The three-state Markov model and the Phase Type Law have been employed to investigate the morbidity and mortality rates of a chronic disease (Akat et al., 2019).

HMM is a promising statistical dynamic model applied in different fields as engineering, data analysis, genetics, energy, and medicine. HMM, is used as image recognition model, for example, in the analysis of transrectal ultrasound images to detect prostate cancer (Llobet et al., 2007), to predict ovarian cancer using 8 tumor suppressor genes (Saif et al., 2018). HMM is applied to cluster patient treatment pathways (Najjar et al., 2018), and 2-state HMM is used to link unobserved sepsis state and observed clinical predictors (Parente et al., 2018).

Furthermore, HMM is employed to study chronic lifelong diseases. HMM is used for analyzing heart sounds, e.g. fuzzy HMM is applied to Doppler ultrasound results to detect heart valve diseases (Uğuz et al., 2008), discrete HMM is used for the classification of heart sound signals (Saraçoğlu, 2012), mobile-health service platform for analyzing and classifying heart sound is developed by using HMM (Thiyagaraja et al., 2018). HMM is used to classify and forecast future clinical situations probabilistically using observed vital sign values like heart rate and blood pressure (Forkan and Khalil, 2017).

HMM model is used to identify distinct symptom-functional states and multivariate Cox regression model is used to predict the mortality of cancer patients (Wen et al., 2018).

1.1.1 Comorbidity or Competing Risks

Despite the fact that the term "comorbidity" and "competing risks" have a different meaning in the context of medical research, in this literature review, we research studies including any of these definitions, as both represent a collection of correlated events from a statistical modeling perspective.

Competing risks are events that change the likelihood of, or completely prevent the occurrence of an event of interest. This phenomenon is investigated in 1760 by Daniel Bernoulli (David and Moeschberger, 1978). Assuming existence of comorbidity problem one encounter with the problem of competing causes of death, that is there is higher possibility of morbidity due to complications caused by comorbid conditions rather than due to the main illness (Koczwara, 2016). Comorbidity is a medical term that refers to a patient who is afflicted with two or more diseases concurrently. It can refer to a variety of chronic diseases that are highly interconnected and share risk factors (Feinstein, 1970, Satariano, 2000).

(a) Competing risks

Competing risk modeling is applied to any failure process that involves more than one separate cause or kind of failure; hence, competing risks are extensively investigated in survival analysis. While Kaplan-Meier survival analysis and Cox proportional hazards regression are frequently used to describe disease risk in order to represent the probability of failure, when the underlying data contain competing-risk events (Satagopan et al., 2004, Lunn and McNeil, 1995), these methods can overestimate disease risk by failing to account for competing risk of mortality (Berry et al., 2010, Gooley et al., 1999, Southern et al., 2006, Kim, 2007, Satagopan et al., 2004). Due to unrealistic assumption underlying the KM estimate, that competing risks are statistically independent, Varadhan et al., 2010 use cause-specific hazard, event-free survival, and cumulative incidence functions to address the analytical challenges posed by competing risks.

Multi-state models are useful for studies involving patients who experience competing risks (Jepsen et al., 2015). A multi-state disease model provides numerous disease states, and the transitions between them are characterized by the occurrence of outcome events, each of which has the potential to alter the time required to generate another outcome event. In particular, models with underlying unobserved multi-states may be beneficial for obtaining additional information from limited observed data. Moreover, disease dynamics are relatively easy to explain in such models because they can be completely represented by a single matrix of probabilities that describes the transition rates between disease stages. Competing risks are studied using multi-state Markov models, including HMMs (Putter et al., 2007, Andersen et al., 2002, Aalen, 1978). The competing risks structure for disease progression is taken into account in HMM assumption (Lange et al., 2018).

(b) Comorbidity

The importance of comorbidities for clinical practice has been widely recognized, and yet statistical frameworks for modeling their co-evolution over time have received little attention. However, understanding the underlying disease dynamics could be of significant value for treatment planning. The literature is lack of sophisticated models to understand the relation between diseases having influence on each other.

Statistical models have been used to identify patterns of co-occurrence of diseases (Guisado-Clavero et al., 2018, C.-F. Huang et al., 2017, Violán et al., 2020). These studies, however, address the question of which comorbidities frequently co-occur but do not model their progression.

Comorbidities are represented as networks using dynamic structural equation models (Bringmann et al., 2013, Groen et al., 2020), or deep diffusion processes (Qian et al., 2020). Nodes corresponded to symptoms and edges to potentially causal relationships. The studies provide information on which diseases co-occur but not on the dynamics of diseases. Bayesian networks have been used to examine comorbidities and their temporal relationships (Faruqui et al., 2018, Lappenschaar, Hommersom, Lucas, Lagro, and Visscher, 2013, Lappenschaar, Hommersom, Lucas, Lagro, Visscher, et al., 2013). These works, which are based on a variety of different clinical and demographic data about patients, provide only a brief understanding of the underlying disease states representing the illness development.

The interaction among diabetes and chronic liver disease under Metformin treatment is modeled by a CHMM with a personalized, non-homogeneous transition mechanism (Maag et al., 2021). The model has a fixed number of hidden states but a greater number of states may be more informative about the progression of comorbidity. Also, the model is applied to clinical and demographic data, however the application of the model to restricted administrative data may be limited.

Comorbidity is used in studies employing HMM and its extensions (Leiva-Murillo et al., 2011, Z. Huang et al., 2015,Powell et al., 2019). However, Markov models with memoryless property imply that a patient's current state is distinct from their future trajectory. As a result, HMM-based models are unable of adequately explaining the variability in individuals' progression trajectories, which is frequently caused by their varied clinical histories or chronologies (order and timing) of clinical events. This is a critical shortcoming in models of survival analysis for complex chronic diseases with numerous morbidities (Lee et al., 2019).

1.2 The Model Proposed in the Thesis

The proposed model takes into account time influence of the disease improvement on the patient in both univariate and bivariate forms; exposes the factors having impact on the diseases both in univariate and bivariate forms; controls the dependence structure in bivariate dimensions.

Interaction among diseases is implemented on limited administrative time series data not clinical data. It is critical to note that the majority of studies on competing risks rely on data from detailed clinical observations. However, a lack of data, particularly clinical data on individuals, may make classical statistical models of competing risks difficult to use. This could be due to restricted access to hospital data or a dearth of information on a particular cause or geographic location. As a result, it is necessary to develop alternative statistical models for sparse data.

Comorbidity of diseases have various etiological models, where risk factors have an influential role (Valderas et al., 2009). For example, in the associated risk factors model, risk factors for one disease are correlated with risk factors for another, increasing the likelihood of the diseases occurring concurrently. On the other hand, in the heterogeneity model, disease risk factors are not correlated, but each is capable of causing diseases associated with the other risk factor among diseases. The proposed model can be applicable to study various pathways to comorbidity, assuming that these interactions happen in unobserved process.

1.2.1 Extensions of Hidden Markov Model

The strong side of the HMM is a possibility of extension of the model. In the literature, a linear combination of the prior estimate of transition matrix and the empirical transition matrix is established by (Siu et al., 2005), Monte Carlo Markov chain is used to perform Bayesian inference and evaluate the posterior distribution of transition matrix (Pasanisi et al., 2012). The bivariate HMM have been developed to study the dependency between discrete and continuous observations (Oflaz et al., 2019).

There are main established approaches to model interaction of several processes by combination of two or more HMMs. Hierarchical HMM is a model where each hidden state is an HMM as well, where children states depend on parent states (Fine et al., 1998). Factorial HMM splits the hidden state into multiple variables that are merged at output, and each state has its own transition matrix (Ghahramani and Jordan, 1996). Event-specific HMMs developed by Kristjansson et al., 2000, aims to model a class of weakly connected time series in which only the onset of events are coupled in time. The representation capacity of event-coupled HMMs is clearly constrained by the restrictive structure, which is designed for a relatively narrow class of applications. Coupled HMM factor HMM to many chains in which the present state of each chain is dependent on the prior state of all chains (Brand et al., 1997). J. Kwon and Murphy, 2000 model traffic velocities by using CHMM. Clearly, the completely coupled architecture developed by Brand et al., 1997 is the most powerful in terms of representing interactions between many sequences. This framework can be used to naturally model a wide variety of applications. (Zhong and Ghosh, 2001).

In the frame of these, we propose the underlying theory of the developed model based on CHMM. Our model is distinguished by the fact that we combine interacting processes via a copula function, implying that joint hidden states follow a predefined joint probability. Other coupled HMMs with a variety of architectures combine hidden chains through the use of conditional probabilities; that is, hidden states are connected to preceding states following Markov property, and their distribution is defined solely by transition state probabilities. Assuming that hidden states are dependent on one another and have a joint probability distribution, and as we cannot observe hidden states directly, using a copula to represent the dependence structure is the optimal choice. Copulas have evolved into one of the most widely used statistical tools for describing, analyzing, and modeling the dependence of random variables.

There are several studies on integrating copula with a HMM. To construct a dependency between the intensity levels of the various modalities, a Gaussian copula is utilized, that is, marginal distributions of those are linked by copula (Lapuyade-Lahorgue et al., 2017). Another way of merging two theories is importing hidden Markov chain in the copula parameter (X. Hu, 2015). Instead of the assumption of conditional independence between observed variables and hidden states researches suggest studying the dependence of observed values on unknown states via copula (Derrode and Pieczynski, 2016). Also, there is a copula-based HMM of cylindrical time series, where a mixture of copula-based cylindrical densities approximates the distribution of cylindrical data, the parameters of which rely on the development of a latent Markov chain (Lagona, 2019). Copula is used to construct the dependence structure of the Markov process by providing copula representation of the Markov property (Sun and Jiang, 2018).

To our best knowledge, there is no study that combines hidden chains using the copula function in the context of a coupled HMM.

CHAPTER 2

COUPLED HIDDEN MARKOV MODEL

This chapter presents the design of CHMM after an explicit definition of the base model.

2.1 Hidden Markov Model

A HMM is a probabilistic graphical framework that models stochastic dynamic of time series data. The structure of HMM is displayed in the Figure 2.1 at which S and X define hidden states and the observations, respectively. This refers to the idea that each state in the Markov model is influenced by hidden states through the time.

Consider a stochastic process $\{S_t\}$, $\mathbf{S}^{(T)} = (S_1, ..., S_T)$ in discrete time $t = 1, 2, ..., T$, with S_t representing the process's state at time t and S_1 indicating the initial state. For all $t \in \mathbb{N}$ a sequence of random variables $\{S_t : t = 1, 2, ...\}$ follows a Markov prop-

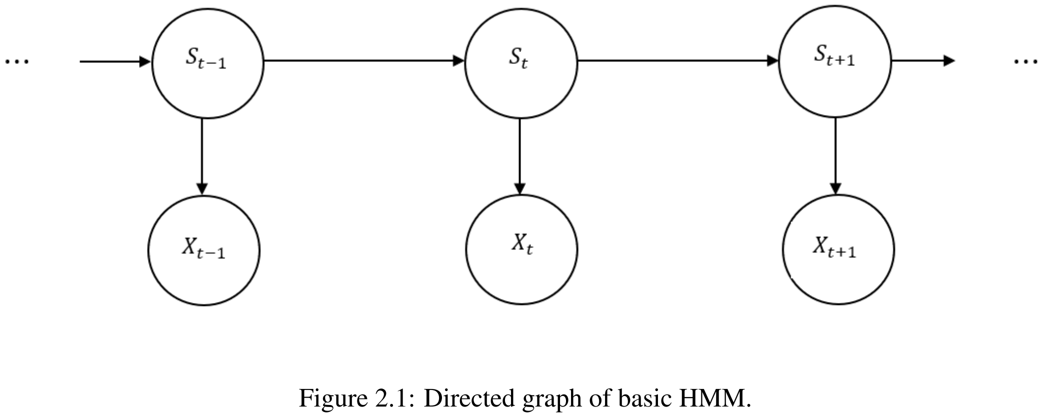


Figure 2.1: Directed graph of basic HMM.

erty defined as

$$\mathbb{P}(S_{t+1}|S_1, S_2, ..., S_t) = \mathbb{P}(S_{t+1}|S_t). \tag{2.1}$$

Thus, the probability distribution of the next state is determined solely by the current state and not by prior states. This condition is also called as no-after-effectiveness (Zhang et al., 2010).

The hidden process is defined by the initial state probability, m_q, and transition probabilities, $\pi_{q,r}$, where hidden states take discrete values, $q, r = \{1, 2, ..., Q\}$, $Q \in \mathbb{N}$. The transition state probability can be expressed as the probability of moving from state q to state r at time t,

$$\pi_{q,r} = \mathbb{P}(S_{t+1} = r|S_t = q)$$

Thus, matrix of transition probabilities, Γ, is defined as a square matrix of probabilities

$$\mathbf{\Gamma} = \begin{pmatrix} \pi_{11} & \cdots & \pi_{1Q} \\ \vdots & \ddots & \vdots \\ \pi_{Q1} & \cdots & \pi_{QQ} \end{pmatrix},$$

with constraint of $\sum_{r=1}^{Q} p_{q,r} = 1$, each row of the matrix sums to one. Initial state probability distribution is defined as follows:

$$m_q = \mathbb{P}(S_1 = q), \quad m_q \geq 0, \quad \text{and} \quad \sum_{q=1}^{Q} m_q = 1.$$

Assume that X_t represents the observation at time t, $t = 1, 2, ..., T$, and consider the vector of observations $\mathbf{X}^{(T)} = (X_1, ..., X_T)$. The model posits that the process that generates X_t is conditional on the Markov property-satisfying hidden state S_t.

Thus, an HMM can be determined by hidden 'parameter process' $\{S_t : t = 1, 2, ...\}$ and the 'state-dependent process' $\{X_t : t = 1, 2, ...\}$, satisfying

$$\mathbb{P}(S_{t+1}|S_t, ...S_1) = \mathbb{P}(S_{t+1}|S_t), \quad t = 2, 3, ...$$

$$\mathbb{P}(X_t|X^{(t-1)}, S^{(t)}) = \mathbb{P}(X_t|S_t), \quad t \in \mathbb{N}.$$

Additionally, we must describe the emission distribution, also known as the state-dependent distribution, $\Phi_q(X_t)$, which establishes the relationship between the observation and hidden states. For discrete-valued observations $\Phi_q(X_t)$ is defined as follows:

$$\Phi_q(X_t) = \mathbb{P}(X_t = x|S_t = q), \quad q = 1, 2, ..., Q, \quad x \in \mathbb{N}.$$

The CDLL of a HMM, i.e. the log-likelihood of observations, $\mathbf{x}^{(T)} = (x_1, ..., x_T)$, and hidden states, $\mathbf{s}^{(T)} = (s_1, ..., s_T)$, is defined as follows:

$$\log\left(\mathbb{P}(\mathbf{x}^{(T)}, \mathbf{s}^{(T)})\right) = \log m_q + \sum_{t=2}^{T} \log \pi_{q,r} + \sum_{t=1}^{T} \log \Phi_q(x_t).$$

Since the sequence of states in a hidden process of an HMM is not observed, those states are considered to be missing data. To represent the sequence of missing data consider the indicator functions as follows:

$$u_r(t) = 1 \quad \text{if and only if} \quad s_t = r, \quad (t = 1, 2, ..., T),$$

and

$$v_{qr} = 1 \quad \text{if and only if} \quad s_{t-1} = q \quad \text{and} \quad s_t = r, \quad (t = 2, 3, ..., T).$$

Treating hidden states as missing data, CDLL of a HMM becomes

$$\begin{aligned} \log\left(\mathbb{P}(\mathbf{x}^{(T)}, \mathbf{s}^{(T)})\right) = \sum_{r=1}^{Q} u_r(1) \log m_q + \sum_{q=1}^{Q} \sum_{r=1}^{Q} \left(\sum_{t=2}^{T} v_{qr}(t) \right) \log \pi_{q,r} \\ + \sum_{r=1}^{Q} \sum_{t=1}^{T} u_r(t) \log \Phi_q(x_t). \end{aligned} \tag{2.2}$$

2.2 Coupled Hidden Markov Model

Numerous applications involve the interaction of multiple sequences, however separate HMM models are incapable of capturing the interactions between them. CHMMs circumvent this constraint by assuming distinct but correlated state sequences to underlie the observed variables.

The standard fully coupled HMM proposed by Brand et al., 1997, determine collection of HMM models in which the state of one model at time t is dependent on the

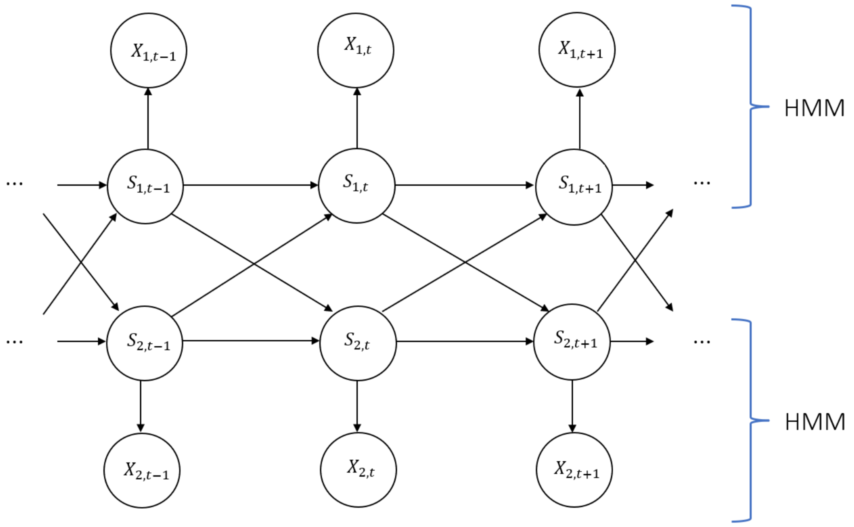


Figure 2.2: Directed graph of CHMM with two underlying Markov chains (Brand et al., 1997).

states of all other models (including itself) at time $t-1$. The main concept of the model is to combine state processes of multiple HMMs. Consider an Q-dimensional observed random variables of length T, denoted by $X_{Q,t}$, $S_{Q,t}$ denotes Q hidden state sequences, where $t = 1, ..., T$. The structure of a CHMM with two underlying state processes is displayed in Figure 2.2. Two HMMs are combined by assuming separate but conditionally dependent state sequences that form a basis for the different variables observed.

CHMM is computationally more expensive than HMM because it allows different number of states to be applied for each individual HMM. Consider observed time series $\mathbf{X}_i = \{X_{i,1}, ..., X_{i,T}\}$, sequence of hidden states $\mathbf{S}_i = \{S_{i,1}, ..., S_{i,T}\}$, $i = 1, ..., I$, where I represents the total number of underlying hidden chains. Similarly to single HMM, coupled model is defined by the following set of parameters.

(i) Initial state distribution:

$$m_{i,q} = \mathbb{P}(S_{i,1} = q_i), \quad m_{i,q} \geq 0, \quad \text{and} \quad \sum_{q=1}^{Q} m_{i,q} = 1,$$

where $1 \leq q_i \leq Q_i$, Q_i is the number of hidden states of ith HMM.

(ii) Transition state distribution, where each hidden state have I parent states from preceding time point. Therefore, transition probabilities are defined as follows:

$$\pi_{q_i|r_1,...,r_I} = \mathbb{P}(S_{i,t+1} = q_i | S_{1,t} = r_1, ..., S_{I,t} = r_I),$$

where $\sum_{q_i=1}^{Q_i} \pi_{q_i,r_1,...,r_I} = 1$.

(iii) Emission distribution can be an arbitrary distribution

$$\Phi_{q_i}(X_{i,t}) = \mathbb{P}(X_{i,t} | S_{i,t} = q_i), \quad q_i = 1, 2, ..., Q_i.$$

The forward probability of the CHMM, or posterior of a state sequence through fully coupled I-chain HMM, is given as follows (Brand et al., 1997):

$$\mathbb{P}(S^I|X) = \prod_i^I \left(m_{i,q} \Phi_{q_i}(X_{i,1}) \prod_{t=2}^T \left(\Phi_{q_i}(X_{i,t}) \prod_j^I \pi_{q_i|q_j} \right) \right) / \mathbb{P}(X), \tag{2.3}$$

where $\Phi_{q_i}(X_{i,t})$ is the probability of the observed value given a state in chain i, $\pi_{q_i|q_j}$ is the probability of a state in chain i given a previous state in chain j. The $\mathbb{P}(X)$ is generally unknown and considered to be constant across sequences, therefore, is ignored in calculations.

2.3 Estimation by the Expectation Maximization Algorithm

To develop a basic HMM model, the transition probabilities, initial probabilities, and emission probability parameters are estimated using the EM algorithm, which is also known as the Baum-Welch algorithm in the context of HMMs. EM algorithm maximizes log-likelihood of the model by treating hidden states as missing data.

2.3.1 Forward and Backward Probabilities

The forward–backward algorithm proposed by Baum et al., 1970 (also see Zucchini and MacDonald, 2009, Scott, 2002, Rabiner, 1989, Chib, 1996) is applicable to any discrete time HMM with a finite state space and offers us with two valuable tools.The forward recursion approach is a computationally efficient way to calculate the likelihood of the observed data, whereas the backward recursion algorithm provides us with the distribution of each hidden state at time t, given the state at the following time point, $t+1$, and all of the observations.

Forward probabilities, $\alpha_t(q)$, and backward probabilities, $\beta_t(q)$ are defined as follows:

$$\alpha_t(q) = \mathbb{P}(X_1, ..., X_{t-1}, S_t = q)\mathbb{P}(X_t|S_t = q), \quad t = 1, 2, ..., T,$$

$$\beta_t(q) = \mathbb{P}(X_{t+1}, ..., X_T|S_t = q), \quad t = 1, 2, ..., T.$$

The likelihood, L_T, is defined as follows:

$$L_T = \sum_q \alpha_t(q)\beta_t(q).$$

Given the structure of the HMM model, the forward-backward recursion necessary to conduct model inference is defined as

$$\alpha_t(q) = \mathbb{P}(X_t|S_t = q)\sum_r \mathbb{P}(S_t = q|S_t = r)\alpha_{t-1}(r),$$

$$\beta_{t-1}(q) = \sum_r \mathbb{P}(X_t|S_t = r)\mathbb{P}(S_t = r|S_{t-1} = q)\beta_t(r).$$

The recursion is initialized with $\alpha_1(q) = \mathbb{P}(X_1|S_1 = q)\mathbb{P}(S_1 = q)$ and $\beta_T(q) = 1$, $q = 1, ..., Q$, $t = 1, ..., T$.

2.3.2 Expectation Maximization Algorithm

The EM algorithm is an efficient iterative procedure for estimating the parameters of an underlying distribution from a given data set in the presence of missing or hidden data using the maximum likelihood estimation. The EM algorithm is divided into two phases. Conditional expectations of unobserved states given observations are computed in the E-step. The CDLL function is maximized in the M-step using the conditional expectation acquired in the E-step. The iteration process is performed until convergence is achieved.

The EM algorithm for a HMM:

(i) **E-part**: $v_{qr}(t)$ and $u_r(t)$ are substituted for the conditional expectations of being in a state r at time t given the observations $x^{(T)}$:

$$\hat{u}_r(t) = \mathbb{P}(S_t = r|X^{(T)} = x^{(T)}) = \alpha_t(r)\beta_t(r)/L_T$$

and

$$\hat{v}_{qr}(t) = \mathbb{P}(S_{t-1} = q, S_t = r | X^{(T)} = x^{(T)}) = \alpha_{t-1}(r)\pi_{qr}\Phi_r(x_t)\beta_t(r)/L_T.$$

(ii) **M-part**: The CDLL maximizes each term with regard to the associated set of parameters, namely the starting distribution m_q, the transition probability matrix $\mathbf{\Gamma}$, and the emission distribution parameters. According to the CDLL of HMM, three distinct maximizations in the M-step are required. Thus:

(a) Setting $u_r(1) = \hat{u}_r(1)/\sum_{r=1}^{Q} \hat{u}_r(1) = \hat{u}_r(1)$, maximize $\sum_{r=1}^{Q} u_r(1) \log m_r$ with respect to initial distribution m_r;

(b) Setting $\gamma_{qr} = \sum_{t=2}^{T} v_{qr}(t)/\sum_{r=1}^{Q} \left(\sum_{t=2}^{T} v_{qr}(t)\right)$, maximize

$$\sum_{q=1}^{Q}\sum_{r=1}^{Q}\Big(\sum_{t=2}^{T} v_{qr}(t)\Big) \log \pi_{qr}$$

with respect to $\mathbf{\Gamma}$;

(c) Depending on the nature of the assumed emission distribution, the third term can be maximized analytically (closed-form solutions are available) or numerically (numerical estimation or approximation is necessary).

It is widely established that EM algorithms can converge to a local optimum rather than a global optimum due to the presence of local modes or saddle-points in the log-likelihood function's lower bound. In a discussion for the Dempster et al., 1977 paper on EM, Murray, 1977 presented an often encountered case in which the EM algorithm converged to a fixed point but not to the log-likelihood global optimum. Murray, 1977 offered a data set that replicated the convergence challenges and proposed a solution that is still widely used today: run the algorithm from numerous initial points and aim for a global optimum. Later, C. J. Wu, 1983 defined the regularity conditions necessary for the EM algorithm to reach a global optimum. However, these conditions of regularity are difficult to verify in practice. Typically, the most practical strategy remains the one advised by Murray, 1977: restart the algorithm with a variety of initial values and keep the parameter estimates that provide the best performance across restarts.

In the case of CHMM, where several hidden sequences are supposed to be dependent, the complexity of E-step becomes numerically inconvenient when number of parameters is too large. When both the number of chains and the number of states are small, that is, when $K = Q^I$ is less than a few tens, the global model can be viewed as a single HMM and the E-step can be performed via forward-backward recursion with complexity $\mathcal{O}(TK^2)$ (X. Wang et al., 2019). As in HMM, the M-step for CHMM is straightforward and tractable.

2.3.3 Variational Expectation Maximization Algorithm

The attractiveness of VEM stems from its computing efficiency and ability to perform deterministic coordinate ascent on the evidence lower bound surface, $J(\mathbf{X}, \theta, \mathbb{P})$ (definition of the function is given in the text following).

There exists a tractable and deterministic approximation to the posterior probability of the hidden states (Ghahramani and Jordan, 1997). The fundamental concept is to estimate the posterior distribution over the hidden variables $\mathbb{P}(S_t|X_t)$ with a tractable distribution $Q(S_T)$. This approximation establishes a lower bound on the log likelihood, which can be utilized to develop an efficient learning algorithm. A lower bound on the log-likelihood can be defined using any distribution over the hidden variables $Q(S_T)$ (L. K. Saul et al., 1996) as follows:

$$\begin{aligned}
\log \mathbb{P}(X_t) &= \log \sum_{S_t} \mathbb{P}(S_t, X_t) \\
&= \log \sum_{S_t} Q(S_t) \left[\frac{\mathbb{P}(S_t, X_t)}{Q(S_t)} \right] \\
&\geq \sum_{S_t} Q(S_t) \log \left[\frac{\mathbb{P}(S_t, X_t)}{Q(S_t)} \right],
\end{aligned}$$

where the final step makes advantage of Jensen's inequality. The Kullback-Leibler divergence is the difference between the left and right sides of this inequality (Thomas and Joy, 2006):

$$KL((\mathbb{P} \parallel Q) = \sum_{S_t} \log \left[\frac{Q(S_t)}{\mathbb{P}(S_t|X_t)} \right].$$

Exact inference difficulty in the approximation supplied by Q is determined by its conditional independence relations, not by its parameters; consequently, Q can be chosen to have a tractable structure. The VEM algorithm attempts to minimize the Kullback-Leibler divergence, and hence to maximize the log-likelihood lower bound with respect to the model parameters, θ,

$$J(\mathbf{X}, \theta, \mathbb{P}) = \sum_{S_t} Q(S_t) \log \left[\frac{\mathbb{P}(S_t, X_t)}{Q(S_t)} \right],$$

where $\mathbf{X}$ represents observed variables.

VEM includes two steps:

(i) VE-step: compute the approximate conditional distribution Q, given the observed data and the current value of the parameter θ, as

$$Q^{h+1} = \arg\max_{Q} J(\mathbf{X}, \theta, \mathbb{P}) = \arg\min_{Q} KL((\mathbb{P} \parallel Q).$$

(ii) M-step: maximize the updated lower bound with respect to the set of parameters as

$$\theta^{h+1} = \arg\max_{\theta} J(\mathbf{X}, \theta, Q^{h+1}).$$

The accuracy of this approximation depends mainly on the class of approximate distributions within which $\tilde{\mathbb{P}}$ is searched for.

There are some statistical and computational constraints to variational inference. Firstly, because the $J(\mathbf{X}, \theta, \mathbb{P})$ is frequently multimodal, depending on the initial parameter estimations, the deterministic method may converge to a local rather than a global optimum. Secondly, because the variational distribution is frequently chosen for computing ease rather than accuracy relative to the posterior distribution, the lower bound may be far from tight.

However, this is not always a disadvantage. Certain models, such as the mixture of Gaussians and the mixed-membership model, exhibit a large number of posterior modes as a result of label switching: shifting the cluster assignment labels results in a large number of symmetric posterior modes. It is sufficient to represent one of these modes for examining latent clusters or predicting new data (Blei et al., 2017).

As in EM algorithm for VEM, when utilizing pure randomization methods, one cannot rely on a single result but must instead run the process numerous times and then choose the best (Bishop and Svensén, 2012) or average the results or average the results (Waterhouse, 1998).

CHAPTER 3

COPULA

3.1 Copula Modeling for Continuous Random Variables

Copulas are utilized to extract the dependency structure of a multivariate distribution. We can construct any multivariate distribution by providing the marginal distributions and its copula separately.

Definition 3.1.1. *A bivariate copula $\mathcal{C}$ is a function from $\mathcal{J} = [0,1]^2$ to $[0,1]$ defined as $\mathcal{C}(u,v) = \mathbb{P}(U \leq u, V \leq v)$, where $U, V \sim \mathcal{U}_{[0,1]}$, the continuous uniform distribution on the unit interval.*

Such copulas naturally arise in statistical modeling as a result of the well-known Sklar's theorem (Sklar, 1959):

Theorem 3.1.1 (Sklar)**.** *Let F_{XY} be the distribution function of a bivariate random vector (X,Y), with marginal distribution functions F_X and F_Y . Then there exists a copula $\mathcal{C}$ such that, for all $(x,y) \in \mathbb{R}^2$,*

$$F_{XY}(x,y) = \mathcal{C}(F_X(x), F_Y(y)). \tag{3.1}$$

If F_X and F_Y are continuous, then $\mathcal{C}$ is unique; otherwise $\mathcal{C}$ is uniquely determined on Ran $F_X\times$ Ran F_Y only, where Ran F_X denotes the range of the cumulative distribution function, F_X. Conversely, for any univariate distribution functions F_X and F_Y and any copula $\mathcal{C}$, the function F_{XY} defined by Equation 3.1 is a valid bivariate distribution function with marginals F_X and F_Y .

For all $u, v \in (0,1]$ the inverses of F and G are left-continuous and defined as

$$F^{-1}(u_{\leftarrow}) = \lim_{s<u, s\to u} F^{-1}(s), \quad G^{-1}(v_{\leftarrow}) = \lim_{t<v, t\to v} G^{-1}(t).$$

Likewise, the left limits of the right-continuous functions F and G are defined as

$$F(x_{\rightarrow}) = \lim_{s>x, s\rightarrow x} F(s), \quad G(y_{\rightarrow}) = \lim_{t>y, t\rightarrow y} G(t).$$

The definition below includes further notation to clarify differences between F and G, whether they are either continuous or discrete.

Definition 3.1.2. *Let (X, Y) be a pair of random variables such that $\mathbb{P}(X \leq x, Y \leq y) = H(x, y)$, $\mathbb{P}(X \leq x) = F(x)$ and $\mathbb{P}(Y \leq y) = G(y)$ for all $x, y \in \mathbb{R}$.*

(i) *$\mathcal{A}$ is the set of "sub-copulas" associated with H, i.e., the collection of functions $A : [0,1]^2 \rightarrow [0,1]$ such that for all $x, y \in \mathbb{R}$,*

$$H(x, y) = A(F(x), G(y)).$$

(ii) *$B : [0,1]^2 \rightarrow [0,1]$ is the function defined for all $u, v \in [0,1]$ by*

$$B(u, v) = H(F^{-1}(u), G^{-1}(v)).$$

(iii) *$C : [0,1]^2 \rightarrow [0,1]$ is the function defined for all $u, v \in [0,1]$ by*

$$C(u, v) = H(F^{-1}(u_{\leftarrow}), G^{-1}(v_{\leftarrow})).$$

(iv) *D is the distribution function of the random pair $(F(X), G(Y))$.*

(v) *E is the distribution function of the random pair $(F(X_{\rightarrow}), G(Y_{\rightarrow}))$.*

When both F and G are continuous, Sklar's Theorem indicates that the various objects presented in Definition 3.1.2 coincide, i.e.,

$$\mathcal{A} = \{B\} = \{C\} = \{D\} = \{E\},$$

and $\mathcal{A}$ consists of the unique copula associated with H (Nelsen, 1999).

When F and G are not continuous, their inverses exhibit plateaus; also, a copula representation for discrete functions exists in $\mathcal{A}$, but it is no longer unique, resulting in an identifiability issue.

3.1.1 Copulas with Discrete Random Variables

Copulas are prominent in the literature due to their ability to isolate the impacts of dependency from the impacts of the marginal distribution. Clearly, if $\mathcal{C}$ is unique, it characterizes how the two marginals F_X and F_Y interact to produce the joint behavior of (X, Y), while remaining ignorant to the marginals' origins. For instance, if X and Y are independent ($X \perp\!\!\!\perp Y$) and if $\mathcal{C}$ is unique, then $\mathcal{C}$ must be the 'independence copula' (or 'product copula') as defined in Equation 3.1

$$\Pi(u, v) = uv \quad \forall (u, v) \in \mathcal{J}$$

and this regardless of F_X and F_Y. It is frequently neglected that this condition is only favorable in the presence of continuous margins. When the joint distribution F_{XY} and the copula $\mathcal{C}$ do not have a one-to-one correspondence, i.e., when X and/or Y are discrete, the preceding reasoning collapses.

Now, in the discrete case of X and/or Y, Ran F_X and/or Ran F_Y are just countable subsets of $[0, 1]$. As a result, the distributions of $F_X(X)$ and/or $F_Y(Y)$ are not $\mathcal{U}_{[0,1]}$, therefore their joint distribution cannot be a copula as defined by Definition 3.1.1. It is actually a subcopula, i.e., a function that satisfies the fundamental structural criteria of copulas but whose support is simply a strict subset of $\mathcal{J}$ comprising 0 and 1 (Definition 2.2.1 in Nelsen, 2006). Any such subcopula may be extended into a copula (Lemma 2.3.5 in Nelsen, 2006): the gaps in $\mathcal{J}$ (Ran $F_X \times$ Ran F_Y) can be filled in a way that preserves the properties of copulas; but, there are an infinite number of methods to do so, thus $\mathcal{C}$ in Equation 3.1 is unidentifiable.

Such unidentifiability leads to major inconsistencies, which Genest and Nešlehová, 2007 explored carefully after preliminary warnings in Marshall, 1996, and calls into doubt the validity of copula modeling for discrete data.

When applied to discrete random vectors, copulas lose their margin-free nature, despite the fact that the whole copula approach is developed to exploit the benefits of margin-freeness (Schweizer, Wolff, et al., 1981).

The absence of uniqueness of Sklar's representation in the discrete situation is the main source of problem. Copulas should be used with caution when dealing with count data since many of its useful qualities do not transfer from the continuous to the discrete situation. (Genest and Nešlehová, 2007).

The discrete situation presents complications because when F and G have jumps, their inverses have plateaus. When this occurs, Sklar's Theorem still guarantees that a copula representation for H exists in $\mathcal{A}$, but the latter is no longer unique, presenting an identifiability problem. The functions defined by (B)–(E) then represent distinct objects; neither is a copula, and some, but not all, of them are members of $\mathcal{A}$ (Genest and Nešlehová, 2007).

Some significant consequences of this unidentifiability issue include the following:

(i) The set of copulas consistent with

$$H(x, y) = \mathcal{C}(F(x), G(y)), \quad x, y \in \mathbb{R}$$

can be quite large. The issue of unidentifiability is expected to be more severe when the variables X and Y have a high concentration of mass in a few values.

(ii) The dependence between X and Y variables is no longer a function of the copula alone; hence, the probabilistic and copula-based definitions of classical measures of concordance no longer correspond. Similar inconsistencies occur with regard to dependency concepts.

3.2 Bivariate Discrete Copula for Count Data

Due to issues with the copula function's unidentifiability for discrete random variables, Geenens, 2020 proposed the bivariate discrete copula for count data.

3.2.1 Theoretical Framework

Let (X, Y) be a bivariate discrete vector where X and Y may only take a finite number of values. Without loss of generality, let $X \in \mathcal{S}_X = \{0, 1, \ldots, R-1\}$ and $Y \in \mathcal{S}_Y = \{0, 1, \ldots, S-1\}$, with $R, S \in \mathbb{N}$, $2 \geq R, S < \infty$. Let $\mathbf{p}$ be its joint probability mass function, defined by $\mathrm{p}_{xy} = \mathbb{P}(X = x, Y = y), (x, y) \in S_X \times S_Y$, and $\mathbf{p}_X = (\mathrm{p}_{0\bullet}, \mathrm{p}_{1\bullet}, \ldots, \mathrm{p}_{R-1\bullet})$ and $\mathbf{p}_Y = (\mathrm{p}_{\bullet 0}, \mathrm{p}_{\bullet 1}, \ldots, \mathrm{p}_{\bullet S-1})$ its marginal distributions: $\mathrm{p}_{x\bullet} = \sum_{y\in\mathcal{S}_Y} \mathrm{p}_{xy} = \mathbb{P}(X = x)$ and $\mathrm{p}_{\bullet y} = \sum_{x\in\mathcal{S}_X} \mathrm{p}_{xy} = \mathbb{P}(Y = y)$. Let $\mathcal{P}_{R\times S}$ be the set of all such bivariate discrete distributions $\mathbf{p}$ with $\mathrm{p}_{x\bullet} > 0\ \forall x \in \mathcal{S}_X$ and $\mathrm{p}_{\bullet y} > 0\ \forall y \in \mathcal{S}_Y$, identified to the $(R \times S)$-matrices

$$\mathbf{p} = \begin{pmatrix} \mathrm{p}_{00} & \mathrm{p}_{01} & \cdots & \mathrm{p}_{0,S-1} \\ \mathrm{p}_{10} & \mathrm{p}_{11} & \cdots & \mathrm{p}_{1,S-1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{p}_{R-1,0} & \mathrm{p}_{R-1,1} & \cdots & \mathrm{p}_{R-1,S-1} \end{pmatrix}.$$

Let

$$\omega_{xy} = \frac{\mathrm{p}_{00}\mathrm{p}_{xy}}{\mathrm{p}_{0y}\mathrm{p}_{x0}}, \quad \forall (x, y) \in \mathcal{S}_X\backslash\{0\} \times S_Y\backslash\{0\}. \tag{3.2}$$

Consider the set $\mathcal{M}^{(+)}_{(R-1)\times(S-1)}$ of all $(R - 1) \times (S - 1)$-matrices with non-negative, possibly infinite, entries. Define the map

$$\Omega : \mathcal{P}_{R\times S} \to \mathcal{M}^{(+)}_{(R-1)\times(S-1)} : \mathbf{p} \to \Omega(\mathbf{p}) = [\omega_{xy}]_{x=1,\ldots,R-1,y=1,\ldots,S-1}, \tag{3.3}$$

where ω_{xy} is given by Equation 3.2. $\Omega(\mathbf{p})$ denotes the odds ratio matrix.

Define $\mathcal{D}^{(1)}_{Q\times Q}$ the set of all diagonal $Q \times Q$ matrices whose entry (1, 1) is equal to 1 and other diagonal entries are positive. For any $\Phi \in \mathcal{D}^{(1)}_{R\times R}$ and $\Psi \in \mathcal{D}^{(1)}_{S\times S}$, let

$$g_{\Phi,\Psi} : \mathcal{P}_{R\times S} \to \mathcal{P}_{R\times S} : g_{\Phi,\Psi}(\mathbf{p}) = \frac{\Phi \cdot \mathbf{p} \cdot \Psi}{\|\Phi \cdot \mathbf{p} \cdot \Psi\|_1}. \tag{3.4}$$

The matrix Φ multiplies the rows of the matrix $\mathbf{p}$, whereas the matrix Ψ multiplies the columns of the matrix $\mathbf{p}$.

Definition 3.2.1 (Geenens, 2020)**.** *A bivariate $(R \times S)$-discrete copula is the bivariate $(R\times S)$-discrete distribution of a vector (U, V) whose both marginal distributions are discrete uniform on $\mathcal{S}_U \doteq \{\frac{1}{R+1}, \frac{2}{R+1}, \ldots, \frac{R}{R+1}\}$ and $\mathcal{S}_V \doteq \{\frac{1}{S+1}, \frac{2}{S+1}, \ldots, \frac{S}{S+1}\}$, respectively. The associated copula probability mass function (p.m.f.) is thus a bivariate discrete copula p.m.f. $\bar{\mathbf{p}}$ on $\mathcal{S}_U \times \mathcal{S}_V$ such that for all $u \in \{0, 1, \ldots, R - 1\}$, $\sum_{v=0}^{S-1} \bar{p}_{uv} = \frac{1}{R}$, and for all $v \in \{0, 1, \ldots, S - 1\}$, $\sum_{u=0}^{R-1} \bar{p}_{uv} = \frac{1}{S}$, where $\bar{p}_{uv} = \mathcal{P}(U = \frac{1}{R+1} + \frac{u}{R+1}, V = \frac{1}{S+1} + \frac{v}{S+1})$.*

The following theorem establishes the existence and uniqueness of the copula p.m.f. Let $\mathcal{C}_{R\times S} = \{\mathbf{p} \in \mathcal{P}_{R\times S} : \mathrm{p}_{x\bullet} = \frac{1}{R}\forall x \in \mathcal{S}_X, \mathrm{p}_{\bullet y} = \frac{1}{S}\forall y \in \mathcal{S}_Y\} \subset \mathcal{P}_{R\times S}$, the set of all $(R \times S)$-discrete copulas as per Definition 3.2.1. For any discrete set A, denote $|A|$ the number of elements in A. Then:

Theorem 3.2.1 (Geenens, 2020)**.** *Let* $\mathbf{p} \in \mathcal{P}_{R\times S}$.

(i) Suppose that, for all $(v_X \times v_Y) \in N(\mathbf{p}), \frac{|v_x|}{R} + \frac{|v_y|}{S} \leq 1$. *Then, there exists a unique* $\bar{\mathbf{p}} \in [\mathbf{p}] \bigcap \mathcal{C}_{R\times S}$*;*

(ii) Suppose that, for all $(v_X \times v_Y) \in N(\mathbf{p}), \frac{|v_x|}{R} + \frac{|v_y|}{S} \leq 1$*, with* $\frac{|\tilde{v}_x|}{R} + \frac{|\tilde{v}_y|}{S} = 1$ *for some* $(\tilde{v}_x \times \tilde{v}_y) \in N(\mathbf{p})$.

(a) If, for all $(\tilde{v}_x \times \tilde{v}_y) \in N(\mathbf{p})$ *such that* $\frac{|\tilde{v}_x|}{R} + \frac{|\tilde{v}_y|}{S} = 1, (\mathcal{S}_X \backslash \tilde{v}_x \times \mathcal{S}_Y \backslash \tilde{v}_y) \in N(\mathbf{p})$*, then there exists a unique* $\bar{\mathbf{p}} \in [\mathbf{p}] \bigcap \mathcal{C}_{R\times S}$*;*

(b) If there exists $(\tilde{v}_x^* \times \tilde{v}_y^*) \in N(\mathbf{p})$ *such that* $\frac{|\tilde{v}_x^*|}{R} + \frac{|\tilde{v}_y^*|}{S} = 1$ *and* $(\mathcal{S}_X \backslash \tilde{v}_x^* \times \mathcal{S}_Y \backslash \tilde{v}_y^*) \notin N(\mathbf{p})$*, then* $[\mathbf{p}] \bigcap \mathcal{C}_{R\times S} = \varnothing$*, but there existst a unique* $\bar{\mathbf{p}} \in CI([\mathbf{p}]) \bigcap \mathcal{C}_{R\times S}$*;*

(iii) Suppose that there exists $(\tilde{\tilde{v}}_x^* \times \tilde{\tilde{v}}_y^*) \in N(\mathbf{p})$ *such that* $\frac{|\tilde{\tilde{v}}_x^*|}{R} + \frac{|\tilde{\tilde{v}}_y^*|}{S} > 1$. *Then,* $CI([\mathbf{p}]) \bigcap \mathcal{C}_{R\times S} = \varnothing$.

A margin-free measure of overall concordance in $\mathbf{p}$ can be defined as Pearson's correlation coefficient on $\bar{\mathbf{p}}$. Yule's coefficient Υ is the discrete analogue to Spearman's ρ. Assume that U is discrete uniform on $\{\frac{1}{R+1}, \frac{2}{R+1}, \ldots, \frac{R}{R+1}\}$, V is discrete uniform on $\{\frac{1}{S+1}, \frac{2}{S+1}, \ldots, \frac{S}{S+1}\}$, and their joint p.m.f. is equal to the copula p.m.f. $\bar{\mathbf{p}}$. Pearson's correlation between U and V can then be determined as (Geenens, 2020)

$$\Upsilon = 3\sqrt{\frac{(R-1)(S-1)}{(R+1)(S+1)}}\Big(\frac{4}{(R-1)(S-1)}\sum_{u=0}^{R-1}\sum_{v=0}^{S-1} uv\bar{p}_{uv} - 1\Big). \tag{3.5}$$

This coefficient is equal to 1 or -1 if and only if $\bar{\mathbf{p}}$ is a diagonal matrix.

3.2.2 Binomial Copula

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent copies of a bivariate Bernoulli random variable with p.m.f. For $x, y \in \{0, 1\}$, $\mathrm{p}_{xy} = \mathbb{P}(X = x, Y = y)$, $\mathrm{p}_{\bullet y} = \mathrm{p}_{0y} + \mathrm{p}_{1y}$, and $\mathrm{p}_{x\bullet} = \mathrm{p}_{x0} + \mathrm{p}_{x1}$, with $\pi_X = \mathrm{p}_{1\bullet}$ and $\pi_Y = \mathrm{p}_{\bullet 1}$,

$$\begin{array}{c|cc|c} X \setminus Y & 0 & 1 & \\ \hline 0 & \mathrm{p}_{00} & \mathrm{p}_{01} & \mathrm{p}_{0\bullet} \\ 1 & \mathrm{p}_{10} & \mathrm{p}_{11} & \mathrm{p}_{1\bullet} \\ \hline & \mathrm{p}_{\bullet 0} & \mathrm{p}_{\bullet 1} & 1 \end{array}.$$

Marshall and Olkin, 1985 defined the bivariate Binomial as the distribution of the vector $(X, Y) = (\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} Y_i)$. Its p.m.f. is, for $(x, y) \in \{0, \ldots, n\} \times \{0, \ldots, n\}$,

$$\mathbb{P}(X = x, Y = y) = \textstyle\sum_{k=max(x+y-n,0)}^{min(x,y)} \binom{n}{k, x-k, y-k, n-x-y+k} \mathrm{p}_{00}^{n-x-y+k} \mathrm{p}_{10}^{x-k} \mathrm{p}_{01}^{y-k} \mathrm{p}_{11}^{k}.$$

Then, it can be checked that the odds ratios (Equation 3.2) are

$$\omega_{xy} = \sum_{k=max(x+y-n,0)}^{min(x,y)} \frac{\binom{n}{k,x-k,y-k,n-x-y+k}}{\binom{n}{x}\binom{n}{y}} \omega^k,$$

where ω is the odds ratio of the initial bivariate Bernoulli defined as

$$\omega = \frac{\mathbb{P}(X = 0, Y = 0)\mathbb{P}(X = 1, Y = 1)}{\mathbb{P}(X = 1, Y = 0)\mathbb{P}(X = 0, Y = 1)}. \tag{3.6}$$

With n constant, the dependency structure of a bivariate Binomial is thus determined solely by the parameter ω, and the associated Binomial(n)-copula is a one-parameter model. It is a $((n + 1) \times (n + 1))$-discrete distribution with uniform margins.

For example, if $n = 2$, the bivariate Binomial distribution and its odds ratio matrix (Equation 3.3) are, respectively,

$$\mathbf{p} = \begin{pmatrix} \mathrm{p}_{00}^2 & 2\mathrm{p}_{00}\mathrm{p}_{01} & \mathrm{p}_{01}^2 \\ 2\mathrm{p}_{00}\mathrm{p}_{10} & 2(\mathrm{p}_{11}\mathrm{p}_{00} + \mathrm{p}_{10}\mathrm{p}_{01}) & 2\mathrm{p}_{11}\mathrm{p}_{01} \\ \mathrm{p}_{10}^2 & 2\mathrm{p}_{10}\mathrm{p}_{11} & \mathrm{p}_{11}^2 \end{pmatrix},$$

and

$$\Omega(\mathbf{p}) = \begin{pmatrix} \frac{1}{2}(\omega + 1) & \omega \\ \omega & \omega^2 \end{pmatrix}.$$

One may convert the margins of $\mathbf{p}$ to uniforms through Equation 3.4 and obtain

$$\bar{\mathbf{p}} = \frac{1}{3} \begin{pmatrix} \frac{\omega(\omega+1)}{\omega^2+\omega+1+\sqrt{\omega(\omega+2)(2\omega+1)}} & \frac{\sqrt{\omega(\omega+2)(2\omega+1)}-3\omega}{(\omega-1)^2} & \frac{\omega+1}{\omega^2+\omega+1+\sqrt{\omega(\omega+2)(2\omega+1)}} \\ \frac{\sqrt{\omega(\omega+2)(2\omega+1)}-3\omega}{(\omega-1)^2} & \frac{\omega^2+4\omega+1-2\sqrt{\omega(\omega+2)(2\omega+1)}}{(\omega-1)^2} & \frac{\sqrt{\omega(\omega+2)(2\omega+1)}-3\omega}{(\omega-1)^2} \\ \frac{\omega+1}{\omega^2+\omega+1+\sqrt{\omega(\omega+2)(2\omega+1)}} & \frac{\sqrt{\omega(\omega+2)(2\omega+1)}-3\omega}{(\omega-1)^2} & \frac{\omega(\omega+1)}{\omega^2+\omega+1+\sqrt{\omega(\omega+2)(2\omega+1)}} \end{pmatrix} \tag{3.7}$$

for $\omega \neq 1$. For $\omega = 1$, $\bar{\mathbf{p}}$ is the (3 × 3)-independence copula p.m.f. (7.3). One also has

$$\Upsilon = \frac{\omega^2 - 1}{\omega^2 + \omega + 1 + \sqrt{\omega(\omega + 2)(2\omega + 1)}}$$

as Yule's coefficient (Equation 3.5) for this copula p.m.f., which is $\Upsilon = -1$ for $\omega = 0$ and $\Upsilon = 1$ for $\omega = \infty$. Thus, this family of Binomial copulas is exhaustive since it accommodates all values of Yule's coefficients between -1 and 1.

CHAPTER 4

COUPLED HIDDEN MARKOV MODEL WITH COPULA

To be able to explain the joint behavior of two discrete time series variables, we combine two hidden Markov chains by copula function. For three or more time series variables, we suppose that there are identical number of hidden chains and we combine each pair of hidden chains by copula function. The proposed CHMM with discrete bivariate copula function accounts for dependency between diseases, it is a system of multiple interacting processes. The interaction between variables are considered in the hidden space rather than the observation space.

A series of observations $X_i = (X_{i,t})$ is supposed to be the total number of patients with disease i, $(i = 1, ..., I)$, observed at time t, $t = 1, ..., T$. We denote hidden process for disease i as $(\mathbf{S}_i) = (S_{i,1}, S_{i,2}, ..., S_{i,T})$, where $S_{i,t}$ takes Q different values, $Q \in \mathbb{N}$. In this setting, the joint hidden process, denoted as $(S_t)_t$ or $(\mathbf{S}_t)$, with $S_t = (S_{1,t}, ..., S_{I,t})$, consists in Q^I possible values.

Emission distribution or state-dependent conditional distribution $\Phi_{i,q}(X_{i,t})$ might be an arbitrary discrete distribution

$$\Phi_{i,q}(X_{i,t}) = \mathbb{P}(X_{i,t}|S_{i,t} = q).$$

(a) Binomial emission distribution: We assume that the observed variable $X_{i,t}$ conditional on state $S_{i,t}$ follows Binomial distribution with probability of success, $p_{i,q}$, i.e. the probability of having disease i conditional on state q. Therefore, $\Phi_{i,q}(X_{i,t})$ is defined for each disease i as follows:

$$\Phi_{i,q}(X_{i,t}) = \mathbb{P}(X_{i,t} = x_{i,t}|S_{i,t} = q) = \binom{n_i}{x_{i,t}} p_{i,q}^{x_{i,t}} (1 - p_{i,q})^{n_i - x_{i,t}},$$

$$x_{i,t} = 0, 1, 2, ..., n_i,$$

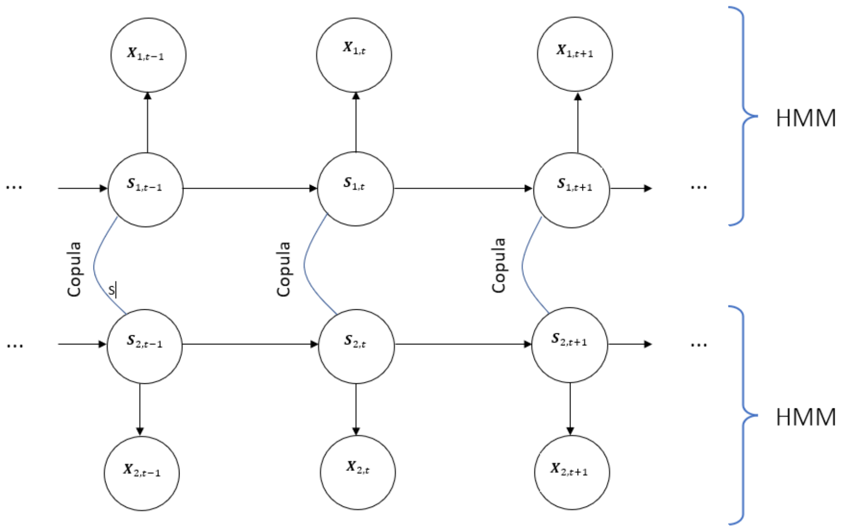


Figure 4.1: Directed graph of CHMM with copula function with two underlying Markov chains.

where n_i is the total number of trials for disease i.

(b) Poisson emission distribution: We assume that the observed variable $X_{i,t}$ conditional on hidden state $S_{i,t}$ follows Poisson distribution with the rate parameter, $\lambda_{i,q}$. Therefore,

$$\Phi_{i,q}(X_{i,t}) = \mathbb{P}(X_{i,t} = x_{i,t}|S_{i,t} = q) = \frac{\lambda_{i,q}^{x_{i,t}} e^{-\lambda_{i,q}}}{x_{i,t}!}, \quad x_{i,t} = 0, 1, 2, ...,$$

where $\lambda_{i,q}$ is the parameter representing average number of patients with disease i conditional on state q.

4.1 Hidden Markov Chain

We assume that the joint hidden process $(\mathbf{S}_t)$ fulfills a Markov property. Hidden dependency structure for two diseases is represented in Figure 4.1. The set of state of all diseases $(S_{i,t})_i$ is a Markov chain and the edges between the state of all diseases at a given time t for all individuals allows to consider disease dependence.

We assume that the transition probabilities of the joint hidden process $(\mathbf{S}_t)$ arise from the product of two terms (both supposed to be constant along time): one accounting for the transitions within each disease and one accounting for the dependency between diseases by the discrete bivariate copula function defined in the following,

$$\mathbb{P}(\mathbf{S}_t = r|\mathbf{S}_{t-1} = q) =: P_{qr} \propto \pi_{q,r} \prod_i \prod_{j \neq i} \max(\bar{\mathbf{p}}(\mathbf{S}_i, \mathbf{S}_j)), \tag{4.1}$$

where

(i) $\boldsymbol{\pi}$ is a $Q \times Q$ transition matrix (each row sums to one) and q (resp. r) indicates the joint hidden state of the comorbidity of diseases;

(ii) the dependency relationship among the diseases is determined by the copula function, see Definition 3.2.1,

$$\bar{\mathbf{p}}(\mathbf{S}_i, \mathbf{S}_j).$$

In particular, $\bar{\mathbf{p}}$ is defined by Equation 3.7 for Binomial copula.

In this model, $\boldsymbol{\pi}$ stands for the transitions within the comorbidity levels of diseases, while copula function introduces the dependency between diseases.

We further assume that the initial joint hidden process $\mathbf{S}_1 = (S_{i,1})_i$ has distribution

$$\mathbb{P}(\mathbf{S}_1 = q) \propto m_q \prod_i \prod_{j \neq i} \max(\bar{\mathbf{p}}(\mathbf{S}_i, \mathbf{S}_j))$$

where m_q is an initial distribution of the states $1 \leq q \leq Q$, whose probability does not change across diseases. However, the model can be extended to include the assumption of varying transition probabilities across diseases; consequently, each hidden chain will have its own transition probability matrix and initial state probabilities.

The proposed method models the dependency of hidden chains by joint distribution of hidden states represented by discrete bivariate copula, whereas CHMM proposed by Brand et al., 1997 couples chains by modeling the causal relationships between their hidden state variables with matrices of conditional probabilities.

We define an indicator function $\chi^r_{i,t} = \mathbf{1}_{\{S_{i,t}=r\}}$.

With this notation, the distribution of the hidden process S is given by

$$\begin{aligned} \mathbb{P}(\mathbf{S}) =& \frac{1}{Z} \prod_{i,q} \prod_{j \neq i} [m_q \max(\bar{\mathbf{p}}(\mathbf{S}_i, \mathbf{S}_j))]^{\chi^q_{i,1}} \times \\ & \times \prod_{t \geq 2, q, r} \prod_i \pi_{q,r}^{\chi^q_{i,t-1} \chi^r_{i,t}} \prod_{j \neq i} \max(\bar{\mathbf{p}}(\mathbf{S}_i, \mathbf{S}_j))^{\chi^q_{i,t-1}, \chi^r_{i,t}}, \end{aligned} \tag{4.2}$$

where Z is a normalizing constant.

Therefore, the joint probability for the sequence of states and observations is defined as follows:

$$\begin{aligned}\mathbb{P}(\mathbf{X},\mathbf{S}) =& \frac{1}{Z}\prod_{i,q}\prod_{j\neq i}[m_q \max(\bar{\mathbf{p}}(\mathbf{S}_i,\mathbf{S}_j))]^{\chi^q_{i,1}}\times \\ &\times \prod_{t\geq 2,q,r}\prod_{i}\pi_{q,r}^{\chi^q_{i,t-1}\chi^r_{i,t}}\prod_{j\neq i}\max(\bar{\mathbf{p}}(\mathbf{S}_i,\mathbf{S}_j))^{\chi^q_{i,t-1},\chi^r_{i,t}}\times \\ &\times \prod_{i,r,t}\Phi_{i,r}(X_{i,t})^{\chi^r_{i,t}}.\end{aligned} \tag{4.3}$$

It follows that the **CDLL** function is

$$\begin{aligned}\log\mathbb{P}(\mathbf{X},\mathbf{S}) =& \sum_{i,q}\chi^q_{i,1}\log m_q + \sum_{i}\sum_{t\geq 2,q,r}\chi^q_{i,t-1}\chi^r_{i,t}\log\pi_{q,r} \\ &+ \sum_{i,r,t}\chi^r_{i,t}\sum_{j\neq i}\log\max(\bar{\mathbf{p}}(\mathbf{S}_i,\mathbf{S}_j)) + \sum_{i,r,t}\chi^r_{i,t}\log\Phi_{i,r}(X_{i,t}) - \log Z.\end{aligned} \tag{4.4}$$

4.2 Variational Expectation Maximization Algorithm

Following the determination of the log-likelihood function, we must find estimators of unknown parameters that maximize this function. When number of hidden states is large, the CDLL becomes computationally intractable. Therefore, we use a variational approximation of the E-step of the EM algorithm. The VEM algorithm maximizes a lower bound of the log-likelihood. In the study, we mainly rely on the approach of (X. Wang et al., 2019) and follow the lines of (Jaakkola, 2001) and (Wainwright and Jordan, 2008) to derive the variational approximation of the log-likelihood.

For any distribution $\tilde{P}$, we have

$$\begin{aligned}\log\mathbb{P}(\mathbf{X}) &\geq \log\mathbb{P}(\mathbf{X}) - KL[\tilde{\mathbb{P}}(\mathbf{S})|\mathbb{P}(\mathbf{S}|\mathbf{X})] \\ &= \tilde{E}\log\mathbb{P}(\mathbf{S},\mathbf{X}) - \tilde{E}\log\tilde{\mathbb{P}}(\mathbf{S}) \\ &=: J(\mathbf{X},\theta,\tilde{\mathbb{P}})\end{aligned} \tag{4.5}$$

where $\tilde{E} = E_{\tilde{\mathbb{P}}}$ and KL denotes the Kullback-Leibler divergence. The maximization of CDLL turns into the maximization of the lower bound $J(\mathbf{X}, \theta, \tilde{\mathbb{P}})$ with respect to the parameter θ. As EM algorithm, VEM includes two steps:

VE-step: compute the approximate conditional distribution $\tilde{\mathbb{P}}$, given the observed data and the current value of the parameter θ^h, as

$$\tilde{\mathbb{P}}^{h+1} = \arg\max_{\tilde{\mathbb{P}}} J(\mathbf{X}, \theta^h, \tilde{\mathbb{P}}) = \arg\min_{\tilde{\mathbb{P}}} KL[\tilde{\mathbb{P}}(\mathbf{S})|\mathbb{P}(\mathbf{S}|\mathbf{X};\theta^h)].$$

M-step: maximize the updated lower bound with respect to the set of parameters as

$$\theta^{h+1} = \arg\max_{\theta} J(\mathbf{X}, \theta, \tilde{\mathbb{P}}^{h+1}).$$

The accuracy of this approximation depends mainly on the class of approximate distributions within which $\tilde{\mathbb{P}}$ is searched for. Here we employ the variational methods for graphical models proposed by L. Saul and Jordan, 1995 and modified for Coupled HMM by Ghahramani and Jordan, 1997. The approximation bases on the setting $\tilde{P}$ to be a product of the independent Markov chains, that is

$$\tilde{\mathbb{P}}(\mathbf{S}) = \prod_i \tilde{\mathbb{P}}(\mathbf{S}_i) \quad where \quad \tilde{\mathbb{P}}(\mathbf{S}_i) = \prod_i \tilde{\mathbb{P}}(S_{i,1}) \prod_{t\geq 2} \tilde{\mathbb{P}}(S_{i,1}|S_{i,t-1}). \tag{4.6}$$

Then we have

$$\tilde{\mathbb{P}}(\mathbf{S}_i) = \frac{1}{\tilde{Z}_i}\Big(\prod_q (m_q h^q_{i,1})^{\chi^q_{i,1}}\Big) \prod_{t\geq 2}\Big(\prod_{q,r} (\pi_{q,r} h^r_{i,t})^{\chi^q_{i,t-1}\chi^r_{i,t}}\Big), \tag{4.7}$$

where $\tilde{Z}_i$ is the normalizing constant that ensures that $\tilde{\mathbb{P}}(\mathbf{S}_i)$ equals one. The variational parameters $h^l_{i,t}$ can be thought of as correction terms for a Markov chain with parameters $(\mathbf{m}, \pi)$.

Let $\tau^r_{i,t} = \tilde{E}(\chi^r_{i,t})$ and $\Lambda^{qr}_{i,t} = \tilde{E}(\chi^q_{i,t-1}\chi^r_{i,t})$ denote the conditional expectations given the observations $\mathbf{X}_{i,t}$. According to these, we define the lower bound as shown in Theorem 4.2.1.

Theorem 4.2.1. *For $i = 1, ..., I$ and $t = 1, ..., T$, the lower bound*

$$\begin{aligned} J(\mathbf{X}, \theta, \tilde{\mathbb{P}}^h) = & \sum_{i,r,t} \tau^r_{i,t}\Big(\log \Phi_{i,r}(X_{i,t}) + \sum_{j\neq i} \log \max(\bar{\mathbf{p}}(\mathbf{S}_i, \mathbf{S}_j)) - \log h^r_{i,t}\Big) \\ & + \sum_i \log \tilde{Z}_i - \log Z \end{aligned} \tag{4.8}$$

Proof. The lower bound of the log-likelihood, for any distribution $\tilde{P}$, is defined as

$$J(\mathbf{X},\theta,\tilde{\mathbb{P}}) = \tilde{E}\log\mathbb{P}(\mathbf{S},\mathbf{X}) - \tilde{E}\log\tilde{\mathbb{P}}(\mathbf{S}) \tag{4.9}$$

Having the log-likelihood of approximating conditional distribution $\mathbb{P}(\widetilde{\mathbf{S}})$ as

$$\begin{aligned}\log\tilde{\mathbb{P}}(\mathbf{S}_i) = &\sum_{i,r}\chi_{i,1}^{r}\log m_r h_{i,1}^{r}\\ &+\sum_{i}\sum_{t\geq 2,q,r}\chi_{i,t-1}^{q}\chi_{i,t}^{r}\log\pi_{q,r}h_{i,t}^{r}\\ &-\sum_{i}\log Z_i,\end{aligned} \tag{4.10}$$

Equation 4.9 becomes

$$\begin{aligned}J(\mathbf{X},\theta,\tilde{\mathbb{P}}) = &\sum_{i,r}\tilde{E}\chi_{i,1}^{r}\tilde{E}\log m_r\\ &+\sum_{i,r,t}\tilde{E}\chi_{i,t}^{r}\sum_{j\neq i}\tilde{E}\log\max(\bar{\mathbf{p}}(\mathbf{S}_i,\mathbf{S}_j))\\ &+\sum_{t\geq 2,q,r}\tilde{E}\chi_{i,t-1}^{q}\chi_{i,t}^{r}\sum_{i}\tilde{E}\log\pi_{q,r}\\ &+\sum_{i,r,t}\tilde{E}\chi_{i,t}^{r}\tilde{E}\log\Phi_{i,r}(X_{i,t})\\ &-\tilde{E}\log Z\\ &-\sum_{i,r}\tilde{E}\chi_{i,1}^{r}\tilde{E}\log m_r h_{i,1}^{r}\\ &-\sum_{i,t\geq 2,q,r}\tilde{E}\chi_{i,t-1}^{q}\chi_{i,t}^{r}\tilde{E}\log\pi_{q,r}h_{i,t}^{r}\\ &+\sum_{i}\tilde{E}\log\tilde{Z}_i\end{aligned} \tag{4.11}$$

With denotation $\tau_{i,t}^{r} = \tilde{E}(\chi_{i,t}^{r})$, $\Lambda_{i,t}^{qr} = \tilde{E}(\chi_{i,t-1}^{q}\chi_{i,t}^{r})$ above equation becomes

$$
\begin{aligned}
J(\mathbf{X},\theta,\tilde{\mathbb{P}}) = &\sum_{i,r} \tau_{i,1}^{r}(\log m_r - \log m_r h_{i,1}^{r}) \\
&+ \sum_{i,r,t} \tau_{i,t}^{r}\Big(\log \Phi_{i,r}(X_{i,t}) + \sum_{j\neq i} \log \max(\bar{\mathbf{p}}(\mathbf{S}_i,\mathbf{S}_j))\Big) \\
&+ \sum_{t\geq 2,q,r} \sum_{i} \Lambda_{i,t}^{q,r}(\log \pi_{q,r} - \log \pi_{q,r} h_{i,t}^{r}) \\
&- \log Z + \sum_{i} \log \tilde{Z}_i
\end{aligned}
\tag{4.12}
$$

Since $\sum_k \Lambda_{i,t}^{q,r} = \tau_{i,t}^{r}$ and according to the multiplication property of the logarithm, the equation becomes

$$
\begin{aligned}
J(\mathbf{X},\theta,\tilde{\mathbb{P}}^{h}) = &\sum_{i,r,t} \tau_{i,t}^{r}\Big(\log \Phi_{i,r}(X_{i,t}) + \sum_{j\neq i} \log \max(\bar{\mathbf{p}}(\mathbf{S}_i,\mathbf{S}_j)) - \log h_{i,t}^{r}\Big) \\
&+ \sum_{i} \log \tilde{Z}_i - \log Z
\end{aligned}
\tag{4.13}
$$

□

Based on these, we define analytically the optimal variation parameter in the E-part with the conditions for forward and backward algorithms. The similar analytical derivations are performed for the Maximization part under the condition of having Binomial and Poisson distribution assumptions.

4.2.1 Expectation Part

Theorem 4.2.2. *For $i = 1, ..., I$ and $t = 1, ..., T$, optimal value for the variation parameter is found as*

$$
h_{i,t}^{r} = \Phi_{i,r}(X_{i,t}) \prod_{j\neq i} \max(\bar{\mathbf{p}}(\mathbf{S}_i,\mathbf{S}_j)).
$$

Proof. Following Ghahramani and Jordan, 1997, Appendix D, we get

$$\begin{aligned}\frac{\partial J(\mathbf{X}, \theta, \tilde{\mathbb{P}}^h)}{\partial \log h_{i,t}^r} =& \Big[\log \Phi_{i,r}(X_{i,t}) + \sum_{j \neq i} \log max(\bar{\mathbf{p}}(\mathbf{S}_i, \mathbf{S}_j)) - \log h_{i,t}^r \Big] \frac{\partial \tau_{i,t}^r}{\partial \log h_{i,t}^r} \\ & - \tau_{i,t}^r + \tau_{i,t}^r = 0\end{aligned} \tag{4.14}$$

because Z does not depend on $h_{i,t}^r$ and $\frac{\partial \log \tilde{Z}_i}{\partial \log h_{i,t}^r} = \tau_{i,t}^r$. This derivative is zero for

$$h_{i,t}^r = \Phi_{i,r}(X_{i,t}) \prod_{j \neq i} \max(\bar{\mathbf{p}}(\mathbf{S}_i, \mathbf{S}_j)).$$

□

4.2.1.1 Forward-Backward Algorithm

The conditional moment, which depend on the normalizing constant $\tilde{Z}_i$, are then computed using an independent forward-backward recursion for each disease and individual.

Forward recursion: set $F_{i,1}^q \propto m_q h_{i,1}^q$ for $t \geq 2$ and compute

$$F_{i,t}^r \propto \sum_q F_{i,t-1}^q \pi_{q,r} h_{i,t}^r.$$

Backward recursion: $\tau_{i,T}^r = F_{i,T}^r$ holds and, for $1 \leq t \leq T-1$, compute

$$G_{i,t+1}^r = \sum_r F_{i,t}^q \pi_{q,r},$$

$$\Delta_{i,t}^{q,r} = \pi_{q,r} \frac{\tau_{i,t+1}^r}{G_{i,t+1}^r} F_{i,t}^q,$$

$$\tau_{i,t}^q = \sum_r \Delta_{i,t}^{q,r}.$$

4.2.2 Maximization Part

The derivation of M-part for transition and initial probabilities is done with respect to the proposed framework. In this respect, we define the M-part derivation under the assumptions of Binomial and Poisson distributions separately.

Theorem 4.2.3. *For $i = 1, ..., I$ and $t = 1, ..., T$, transition probability and initial probability, respectively, are*

$$\hat{\pi}_{q,r} = \frac{\sum_i \tau^r_{i,t+1}}{\sum_r \sum_i \tau^r_{i,t+1}}, \tag{4.15}$$

$$\hat{m}_q = \frac{\sum_i \tau^r_{i,1}}{\sum_r \sum_i \tau^r_{i,1}}. \tag{4.16}$$

Proof. We use Lagrange multiplier since we have a constraint $\sum_r \pi_{q,r} = 1$.

$$f_x = \lambda g_x,$$

$$f_y = \lambda g_y,$$

$$f_z = \lambda g_z,$$

where λ is called the Lagrange multiplier.

Firstly, we calculate derivatives of the lower bound with conditional expectations derived in E-step

$$\begin{aligned} J(\mathbf{X}, \theta, \tilde{\mathbb{P}}^{h+1}) = & \sum_{i,q,t} \sum_r \pi_{q,r} \frac{\tau^r_{i,t+1}}{\sum_q \pi_{q,r} F^q_{i,t}} F^q_{i,t} \times \\ & \times \Big(\log \Phi_{i,q}(X_{i,t}) + \sum_{j \neq i} \log \max(\bar{\mathbf{p}}(\mathbf{S}_i, \mathbf{S}_j)) - \log h^q_{i,t} \Big) \\ & + \sum_i \log \tilde{Z}_i - \log Z \end{aligned} \tag{4.17}$$

The first derivative of the lower bound is as follows:

$$\frac{\partial J(\mathbf{X}, \Theta, \tilde{\mathbb{P}}^{h+1})}{\partial \pi_{q,r}} = \sum_{i,q,t} \sum_r \frac{\tau^r_{i,t+1} F^q_{i,t} A^q_{i,t} \Big(\sum_q F^q_{i,t} \pi_{q,r} - \pi_{q,r} \sum_q F^q_{i,t} \Big)}{\Big(\sum_q F^q_{i,t} \pi_{q,r} \Big)^2}, \tag{4.18}$$

where $A^q_{i,t} = \log \Phi_{i,q}(X_{i,t}) + \sum_{j \neq i} \log \max(\bar{\mathbf{p}}(\mathbf{S}_i, \mathbf{S}_j)) - \log h^q_{i,t}$. Then we have a system of equations

$$\begin{cases} \sum_{i,q,t} \sum_r \frac{\tau^r_{i,t+1} F^q_{i,t} A^q_{i,t} \Big(\sum_q F^q_{i,t} \pi_{q,r} - \pi_{q,r} \sum_q F^q_{i,t} \Big)}{\Big(\sum_q F^q_{i,t} \pi_{q,r} \Big)^2} = \lambda \sum_r 1 \\ \sum_r \pi_{q,r} = 1 \end{cases} . \tag{4.19}$$

The system of equations is true for

$$\hat{\pi}_{q,r} = \frac{\sum_i \tau^r_{i,t+1}}{\sum_r \sum_i \tau^r_{i,t+1}}, \tag{4.20}$$

since

$$\sum_r \hat{\pi}_{q,r} = \sum_r \frac{\sum_i \tau^r_{i,t+1}}{\sum_r \sum_i \tau^r_{i,t+1}} = 1$$

and for $\lambda = 0$

$$\sum_{i,q,t} \sum_r \frac{\tau^r_{i,t+1} F^q_{i,t} A^q_{i,t} \Big(\sum_q F^q_{i,t} \frac{\sum_i \tau^r_{i,t+1}}{\sum_r \sum_i \tau^r_{i,t+1}} - \frac{\sum_i \tau^r_{i,t+1}}{\sum_r \sum_i \tau^r_{i,t+1}} \sum_q F^q_{i,t} \Big)}{\Big(\sum_q F^q_{i,t} \frac{\sum_i \tau^r_{i,t+1}}{\sum_r \sum_i \tau^r_{i,t+1}} \Big)^2} = 0,$$

since

$$\frac{\sum_i \tau^r_{i,t+1}}{\sum_r \sum_i \tau^r_{i,t+1}} \sum_q F^q_{i,t} - \frac{\sum_i \tau^r_{i,t+1}}{\sum_r \sum_i \tau^r_{i,t+1}} \sum_q F^q_{i,t} = 0.$$

Consequently, the estimation for initial probabilities is obtained as follows:

$$\hat{m}_q = \frac{\sum_i \tau^r_{i,1}}{\sum_r \sum_i \tau^r_{i,1}}.$$

□

4.2.2.1 Maximization Part for Binomial Emission Distribution

Theorem 4.2.4. *Given random variable X_i having Binomial $(n; p_{i,q})$ distribution, the EM estimate of emission distribution is*

$$\hat{p}_{i,q} = \frac{\sum_t \sum_r B^{q,r}_{i,t} X_{i,t}}{n \sum_t \sum_r B^{q,r}_{i,t}} \tag{4.21}$$

where

$$B^{q,r}_{i,t} = \pi_{q,r} \frac{\tau^r_{i,t+1}}{G^r_{i,t+1}} F^q_{i,t}.$$

Proof. We have vector of state-dependent parameters for each disease. Let $i = 2$, that is we have two hidden chains representing the occurrence of two diseases (or any two events). Then, the lower bound of CDLL with conditional moments becomes

$$\begin{aligned} J(\mathbf{X}, \theta, \tilde{\mathbb{P}}^{h+1}) &= \sum_{q,t} \sum_{r} \pi_{q,r} \frac{\tau^{r}_{1,t+1}}{G^{r}_{1,t+1}} F^{q}_{1,t} \log \binom{n}{X_{1t}} p_{1,q}^{X_{1t}} (1 - p_{1,q})^{n - X_{1t}} \\ &+ \sum_{q,t} \sum_{r} \pi_{q,r} \frac{\tau^{r}_{2,t+1}}{G^{r}_{2,t+1}} F^{q}_{2,t} \log \binom{n}{X_{2t}} p_{2,q}^{X_{2t}} (1 - p_{2,q})^{n - X_{2t}} \\ &+ \ldots \end{aligned} \tag{4.22}$$

We take the derivative of $J(\mathbf{X}, \theta, \tilde{\mathbb{P}}^{h+1})$ with respect to the probability of having disease i conditional on state q, i.e. $p_{i,q}$. For disease $i = 1$ we have,

$$\begin{aligned} \frac{\partial J(\mathbf{X}, \Theta, \tilde{\mathbb{P}}^{h+1})}{\partial p_{1,q}} &= \frac{\sum_{q,t} \sum_{r} \pi_{q,r} \frac{\tau^{r}_{1,t+1}}{G^{r}_{1,t+1}} F^{q}_{1,t} X_{1,t}}{p_{1,q}} \\ &- \frac{\sum_{q,t} \sum_{r} \pi_{q,r} \frac{\tau^{r}_{1,t+1}}{G^{r}_{1,t+1}} F^{q}_{1,t} (n - X_{1,t})}{1 - p_{1,q}} = 0, \end{aligned} \tag{4.23}$$

then,

$$\sum_{q,t} \sum_{r} B^{q,r}_{1,t} X_{1,t} - \sum_{q,t} n p_{1,q} \sum_{r} B^{q,r}_{1,t} = 0,$$

where

$$B^{q,r}_{1,t} = \pi_{q,r} \frac{\tau^{r}_{1,t+1}}{G^{r}_{1,t+1}} F^{q}_{1,t}.$$

Then

$$\hat{p}_{1,q} = \frac{\sum_{t} \sum_{r} B^{q,r}_{1,t} X_{1,t}}{n \sum_{t} \sum_{r} B^{q,r}_{1,t}} \tag{4.24}$$

Since the second derivative of the function is negative at $\hat{p}_{1,q}$, we conclude that $J(\mathbf{X}, \Theta, \tilde{P}^{h+1})$ has local maximum here,

$$\begin{aligned} \left. \frac{\partial^2 J(\mathbf{X}, \Theta, \tilde{\mathbb{P}}^{h+1})}{\partial p^2_{1,q}} \right|_{p_{1,q} = \hat{p}_{1,q}} &= - \left. \frac{\sum_{q,t} \sum_{r} \pi_{q,r} \frac{\tau^{r}_{1,t+1}}{G^{r}_{1,t+1}} F^{q}_{1,t} X_{1,t}}{p^2_{1,q}} \right|_{p_{1,q} = \hat{p}_{1,q}} \\ &- \left. \frac{\sum_{q,t} \sum_{r} \pi_{q,r} \frac{\tau^{r}_{1,t+1}}{G^{r}_{1,t+1}} F^{q}_{1,t} (n - X_{1,t})}{(1 - p_{1,q})^2} \right|_{p_{1,q} = \hat{p}_{1,q}} < 0. \end{aligned} \tag{4.25}$$

The obtained results can be generalized to any i, $i = 1, \ldots, I$ as follows:

$$\hat{p}_{i,q} = \frac{\sum_{t} \sum_{r} B^{q,r}_{i,t} X_{i,t}}{n \sum_{t} \sum_{r} B^{q,r}_{i,t}}.$$

□

4.2.2.2 Maximization Part for Poisson Emission Distribution

Theorem 4.2.5. *Given random variable X_i having Poisson $(\lambda_{i,q})$ distribution, the EM estimate of emission distribution is*

$$\hat{\lambda}_{i,q} = \frac{\sum_t X_{i,t} \frac{\tau^r_{i,t+1}}{G^r_{i,t+1}} F^q_{i,t}}{\sum_t \frac{\tau^r_{i,t+1}}{G^r_{i,t+1}} F^q_{i,t}} \tag{4.26}$$

Proof. We have vector of state-dependent parameters for each disease. Let $i = 2$, that is we have two hidden chains representing two diseases. Then, the lower bound of CDLL with conditional moments becomes

$$\begin{aligned} J(\mathbf{X}, \theta, \tilde{\mathbb{P}}^{h+1}) = & \sum_{q,t} \sum_{r} \pi_{q,r} \frac{\tau^r_{1,t+1}}{G^r_{1,t+1}} F^q_{1,t} \log \frac{\lambda_{1,q}^{X_{1t}} \exp -\lambda_{1,q}}{X_{1t}!} \\ & + \sum_{q,t} \sum_{r} \pi_{q,r} \frac{\tau^r_{2,t+1}}{G^r_{2,t+1}} F^q_{2,t} \log \frac{\lambda_{2,q}^{X_{2t}} \exp -\lambda_{2,q}}{X_{2t}!} \\ & + ... \end{aligned} \tag{4.27}$$

We take the derivative of $J(\mathbf{X}, \theta, \tilde{\mathbb{P}}^{h+1})$ with respect to the probability of having disease i conditional on state q, i.e. $\lambda_{i,q}$. For disease $i = 1$ we have,

$$\frac{\partial J(\mathbf{X}, \Theta, \tilde{\mathbb{P}}^{h+1})}{\partial \lambda_{1,q}} = \sum_{q,r} \frac{\pi_{q,r}}{\lambda_{1,q}} \sum_t X_{1,t} \frac{\tau^r_{1,t+1}}{G^r_{1,t+1}} F^q_{1,t} - \sum_{q,r} \pi_{q,r} \sum_t \frac{\tau^r_{1,t+1}}{G^r_{1,t+1}} F^q_{1,t} = 0. \tag{4.28}$$

Then

$$\hat{\lambda}_{1,q} = \frac{\sum_t X_{1,t} \frac{\tau^r_{1,t+1}}{G^r_{1,t+1}} F^q_{1,t}}{\sum_t \frac{\tau^r_{1,t+1}}{G^r_{1,t+1}} F^q_{1,t}} \tag{4.29}$$

Since the second derivative of the function is negative at $\hat{\lambda}_{1,q}$, we conclude that the $J(\mathbf{X}, \Theta, \tilde{P}^{h+1})$ has local maximum here,

$$\left.\frac{\partial^2 J(\mathbf{X}, \Theta, \tilde{\mathbb{P}}^{h+1})}{\partial \lambda^2_{1,q}}\right|_{\lambda_{1,q} = \hat{\lambda}_{1,q}} = -\sum_{q,r} \frac{\pi_{q,r}}{\lambda^2_{1,q}} \sum_t X_{1,t} \left.\frac{\tau^r_{1,t+1}}{G^r_{1,t+1}} F^q_{1,t}\right|_{\lambda_{1,q} = \hat{\lambda}_{1,q}} < 0. \tag{4.30}$$

The obtained results can be generalized to any i, $i = 1, ..., I$ as follows:

$$\hat{\lambda}_{i,q} = \frac{\sum_t X_{i,t} \frac{\tau^r_{i,t+1}}{G^r_{i,t+1}} F^q_{i,t}}{\sum_t \frac{\tau^r_{i,t+1}}{G^r_{i,t+1}} F^q_{i,t}}.$$

□

4.2.3 Estimation of the Odds Ratio

According to X. Wang et al., 2019, estimation of dependency function within the M-step result in a poor estimation of the dependency function. They suggest to use grid search of parameters and select the ones that minimize weighted Residual Sum of Squares (RSS).

In our study we suggest to use grid search to select parameters of hidden state marginal distributions and parameters necessary for copula function. Moreover, we use the weighted RSS to select the optimal odds ratio for copula probabilities,

$$RSS = \sum_{i,r,t} \tau^r_{i,t}(x_{i,t} - \mu_{i,r})^2, \tag{4.31}$$

where $\mu_{i,r} = \lambda_{i,r}$ and $\mu_{i,r} = np_{i,r}$ for Poisson and Binomial emission probabilities, respectively.

4.2.4 Model selection criteria

The Akaike Information Criterion (AIC) is a method for choosing the best model among a group of models. The model that minimizes the Kullback-Leibler distance between the model and the truth is presumed to be the superior model. The AIC is defined as

$$AIC = -\log L + 2m,$$

where $\log L$ is the log-likelihood of the fitted model and m is the number of free parameters in the model. The first term rewards the model's goodness of fit and reduces as the number of states increases, whereas the second term, described as the penalty term, increases as the number of states increases. The model with the lowest AIC value is preferred.

The Bayesian Information Criterion (BIC) is the approach to model selection among a set of models. The penalty term distinguishes BIC from AIC:

$$BIC = -2 \log L + m \log K,$$

where $\log L$ is the log-likelihood of the fitted model, m and K denotes the number of parameters and observations in the model, respectively.

4.2.5 Algorithm and Numerical Stability

The proposed model is presented both in algorithm and flowchart to be more elaborated. Figure 4.2 illustrates each steps defined in algorithm and explains the stages from start to its end. In this graph, the stages of algorithm development and a decision nodes are shown in blue and pink colors, whereas, outputs/estimated parameters are in green color. The algorithm includes VE-part in the steps 3-6, and M-part in the step 7.

Probabilities are restricted to exist inside the range $[0, 1]$. Precision is required when manipulating and executing arithmetic on small numbers. When possible, it is recommended to work with logarithms of probabilities. In particular, the "log-sum-exp" trick is useful when dealing with numerical underflow (Nielsen and Sun, 2016, Blei et al., 2017),

$$\log \left[\sum_i \exp(y_i) \right] = \alpha + \log \left[\sum_i \exp(y_i - \alpha) \right].$$

The constant α is typically set to $\max_i y_i$. This ensures that common computations in variational inference processes are numerically stable.

Algorithm 1 VEM algorithm of CHMM with bivariate discrete copula.

```
1: Set initial values for initial state probabilities, $m_r$, transition probabilities, $\pi_{q,r}$, and probabilities of having disease $i$ conditional on state $r$, $p_{i,r}$
2: Set or calculate $\omega$ for each pair of diseases
3: Calculate $\max(\bar{\mathbf{p}}(\mathbf{S}_i, \mathbf{S}_j))$ for each pair of diseases. Calculate emission probabilities, $\Phi_{i,r}(X_{i,t})$, for each disease $i$. Then, for each disease $i$ calculate
       $$h^r_{i,t} = \Phi_{i,r}(X_{i,t}) \prod_{j \neq i} \max(\bar{\mathbf{p}}(\mathbf{S}_i, \mathbf{S}_j))$$
4: Calculate forward probabilities: recursive calculation with initial $F^q_{i,1} \propto m_q h^q_{i,1}$ with normalized $m_q$
5: Calculate backward probabilities using obtained forward probabilities
6: Normalize obtained posterior probabilities and calculate the matrix $\Lambda^{qr}_{i,t}$ using posterior probabilities
7: Calculate/update parameters $m_r$, $\pi_{q,r}$, $p_{i,r}$ based on M-part formulas
8: Calculate AIC, BIC, and weighted RSS criteria for each disease
9: Calculate criteria of convergence.
10: if criteria less than the threshold then
11:     stop the algorithm, return the estimated parameters
12: else
13:     update parameters and do steps 3-9
14: end if
```

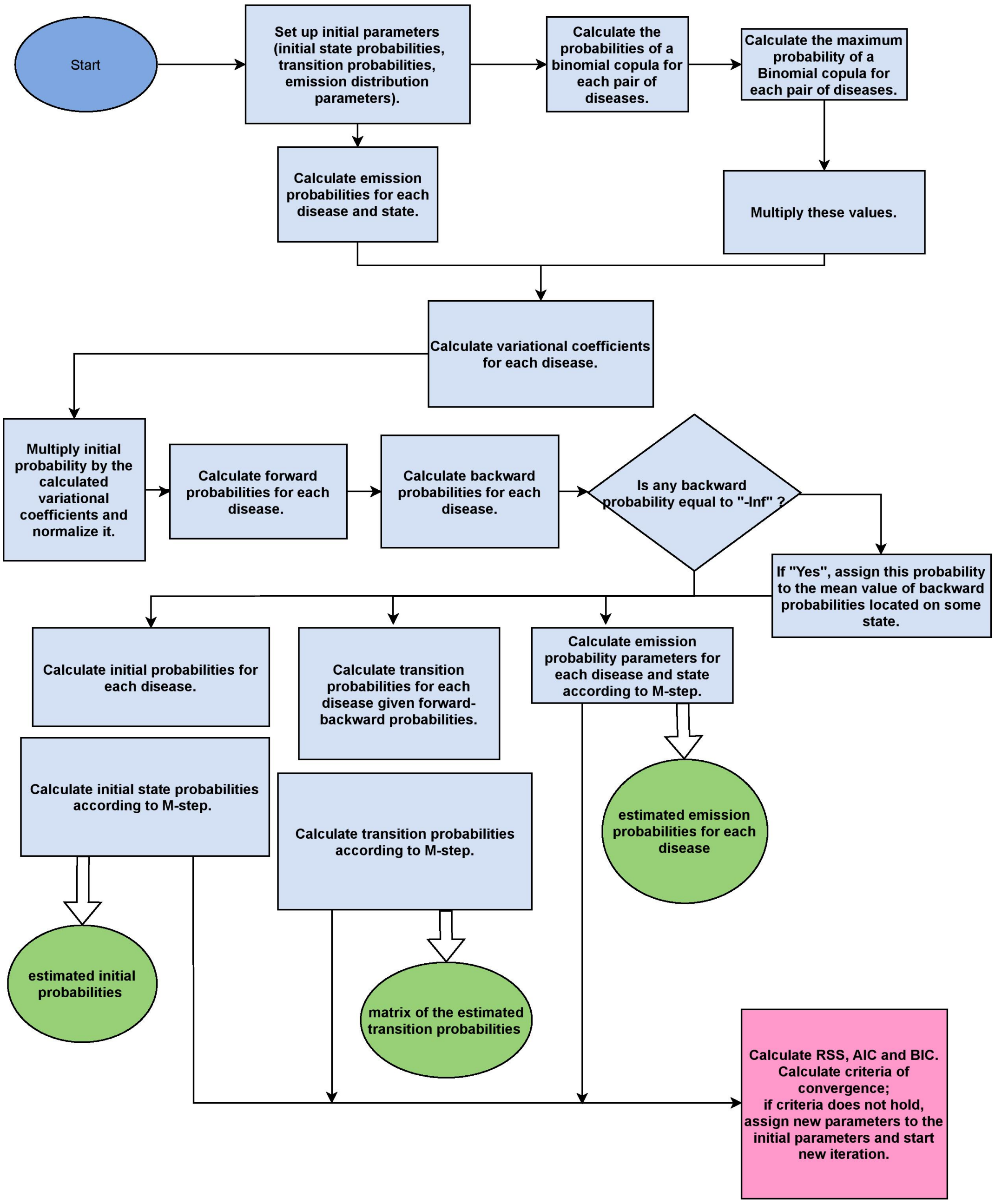

Start
Set up initial parameters (initial state probabilities, transition probabilities, emission distribution parameters).
Calculate the probabilities of a binomial copula for each pair of diseases.
Calculate the maximum probability of a Binomial copula for each pair of diseases.
Calculate emission probabilities for each disease and state.
Multiply these values.
Calculate variational coefficients for each disease.
Multiply initial probability by the calculated variational coefficients and normalize it.
Calculate forward probabilities for each disease.
Calculate backward probabilities for each disease.
Is any backward probability equal to "-Inf" ?
If "Yes", assign this probability to the mean value of backward probabilities located on some state.
Calculate initial probabilities for each disease.
Calculate transition probabilities for each disease given forward-backward probabilities.
Calculate emission probability parameters for each disease and state according to M-step.
Calculate initial state probabilities according to M-step.
Calculate transition probabilities according to M-step.
estimated emission probabilities for each disease
estimated initial probabilities
matrix of the estimated transition probabilities
Calculate RSS, AIC and BIC. Calculate criteria of convergence; if criteria does not hold, assign new parameters to the initial parameters and start new iteration.

Figure 4.2: VEM algorithm of CHMM Copula.

CHAPTER 5

COMORBIDITY BY SIMULATION

To evaluate the proposed method's performance and estimation results, we conduct a simulation study which requires specific modeling to generate the comorbidity to capture the proposed model. Simulation results for data set with an odds ratio of 0.85 and 8 are made to show the sensitivity.

5.1 Simulation Design

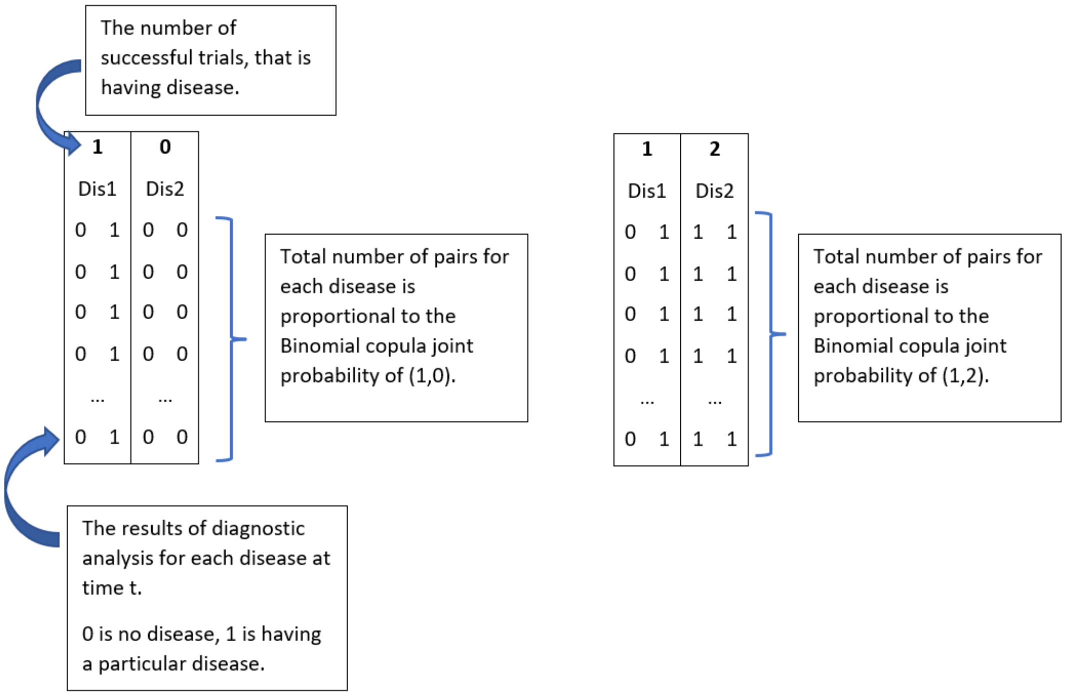
We begin by simulating the hidden states using a correlation structure. Following state derivation, we simulate the realizations from two time series. Once the data set is obtained, the estimation procedure is applied, and the bias and variation of the given versus estimated parameters are compared.

Set up for hidden states. Firstly, consider the Bernoulli distribution with some probability of success. Let 0 represent "no disease" and 1 represent "have disease". Secondly, let n cases be the number of trials, in our case, let it be medical check up/ diagnostic analysis number. So, at time t, having $n = 2$, we might have such result: 0, 1. Meaning that according to the first diagnostic analysis, there is no disease, and according to the second diagnostic analysis, the result is "have disease". So, for two diseases and several time points, we may have such results:

IHD: $\{0, 0\}$ at $t = 1$, $\{0, 1\}$ at $t = 2$, $\{1, 1\}$ at $t = 3$, etc.

Diabetes: $\{0, 1\}$ at $t = 1$, $\{0, 1\}$ at $t = 2$, $\{1, 1\}$ at $t = 3$, etc.

Finally, summing the above values we get these sequences:



The number of successful trials, that is having disease.

1		0	
Dis1		Dis2	
0	1	0	0
0	1	0	0
0	1	0	0
0	1	0	0
...		...	
0	1	0	0

Total number of pairs for each disease is proportional to the Binomial copula joint probability of (1,0).

The results of diagnostic analysis for each disease at time t.

0 is no disease, 1 is having a particular disease.

1		2	
Dis1		Dis2	
0	1	1	1
0	1	1	1
0	1	1	1
0	1	1	1
...		...	
0	1	1	1

Total number of pairs for each disease is proportional to the Binomial copula joint probability of (1,2).

Figure 5.1: Simulation set up of hidden states for two diseases: generation of Bernoulli trials.

IHD: $0, 1, 2, ...$

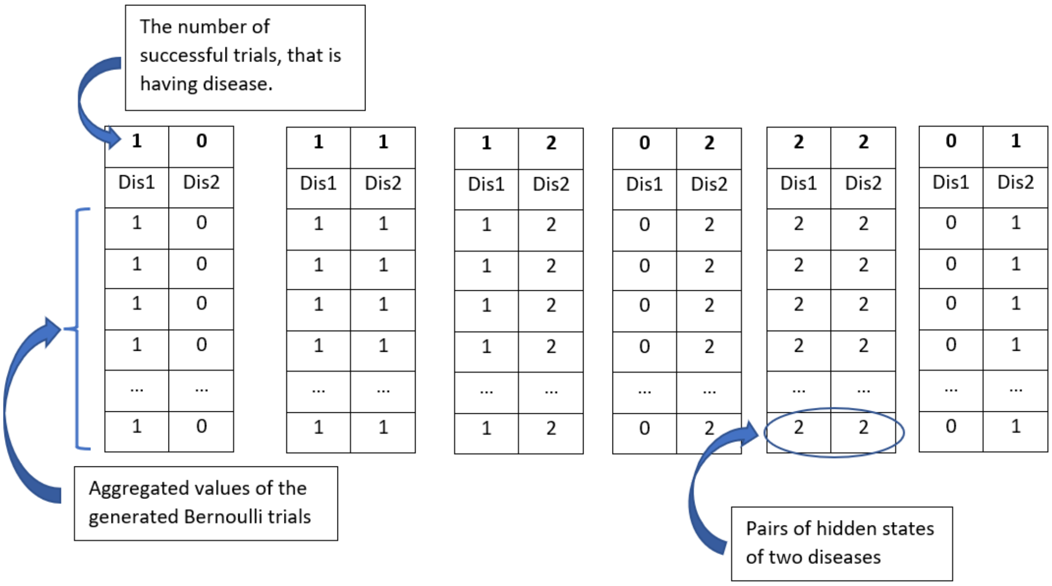
Diabetes: $1, 1, 2, ...$

These values represent hidden states of two hidden chains and follow Binomial distribution with $n = 2$ and probability of success, p. To remind, one of the assumptions of the CHMM Copula model is that hidden chains follow Binomial discrete copula. So, the above hidden chains are dependent.

Simulation design steps for three state model.

S1: For particular odds ratio (Equation 3.6) calculate Binomial copula probabilities. For $n = 2$, Binomial copula distribution is a 3×3-matrix, see Equation 3.7. Rows represent $0, 1, 2$ number of successful trials for disease 1, columns represent $0, 1, 2$ number of successful trials for disease 2.

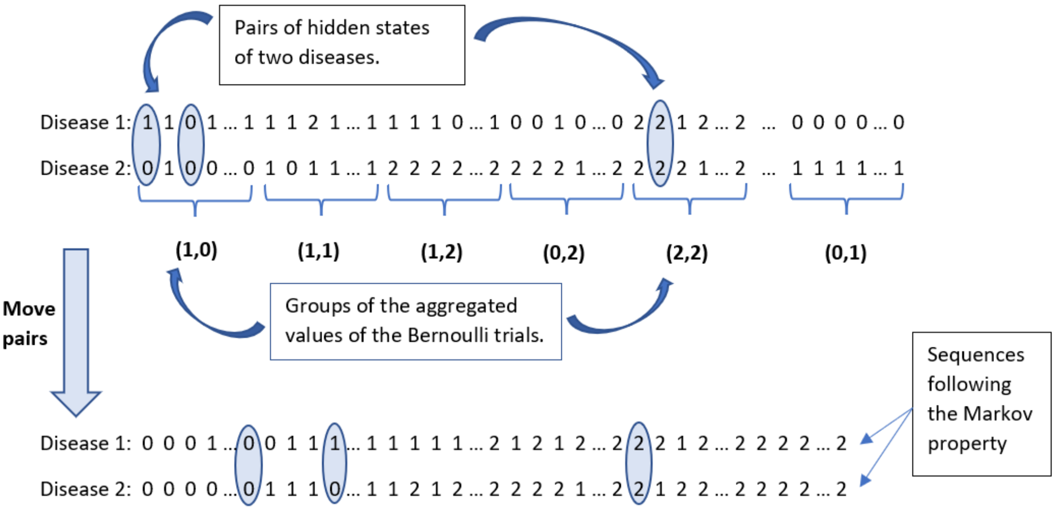
S2: Generate values from Bernoulli distribution for each trial and disease at time t with predefined probability of success. If the generated sequences of states do not satisfy Markov property or do not follow Binomial copula distribution, the


The number of successful trials, that is having disease.

Aggregated values of the generated Bernoulli trials

Pairs of hidden states of two diseases

Figure 5.2: Simulation set up of hidden states for two diseases: aggregated Bernoulli trials at time t for each disease. For illustration purpose, only 6 out of all possible scenarios are displayed.


Pairs of hidden states of two diseases.

Groups of the aggregated values of the Bernoulli trials.

Sequences following the Markov property

Disease 1: 1 1 0 1 ... 1 1 1 2 1 ... 1 1 1 1 0 ... 1 0 0 1 0 ... 0 2 2 1 2 ... 2 ... 0 0 0 0 ... 0
Disease 2: 0 1 0 0 ... 0 1 0 1 1 ... 1 2 2 2 2 ... 2 2 2 2 1 ... 2 2 2 2 1 ... 2 ... 1 1 1 1 ... 1
(1,0)
(1,1)
(1,2)
(0,2)
(2,2)
(0,1)
Move pairs
Disease 1: 0 0 0 1 ... 0 0 1 1 1 ... 1 1 1 1 1 ... 2 1 2 1 2 ... 2 2 2 1 2 ... 2 2 2 2 ... 2
Disease 2: 0 0 0 0 ... 0 1 1 1 0 ... 1 1 2 1 2 ... 2 2 2 2 1 ... 2 2 1 2 2 ... 2 2 2 2 ... 2

Figure 5.3: Simulation set up of hidden states for two diseases: Markov sequences.

probability of success is redefined.

S3: Total number of the generated pairs with particular outcomes, as (0,1), (1,1), (1,0), (1,2), etc., should be proportional to the Binomial copula joint probability of (0,1), (1,1), (1,0), (1,2), etc. (see Figure 5.1).

S4: Sum the generated Bernoulli values for each disease at time t. Example of aggregated values are shown in Figure 5.2.

S5: Finally, two sequences of the generated Binomial values represent sequences of hidden states for two diseases. Pairs of hidden states of two diseases are relocated that each sequence of hidden states follow Markov property (Figure 5.3).

S6: To check whether the generated sequences follow the Binomial copula distribution, repeat 2-4 steps n times. Therefore, $n \times 2$ sequences for two diseases are generated. The joint probabilities of hidden state pairs of two diseases are calculated. The joint probabilities of the generated pairs and Binomial copula probabilities are compared based on bias and Mean Squared Error.

S7: To check whether the generated sequences follow the Markov property, chi-square based test proposed by Anderson and Goodman, 1957 is conducted. *verifyMarkovProperty* function from *markovchain* package (Spedicato et al., 2016), is used. If the high proportion of sequences satisfy the Markov property, we do not drop non-Markov sequences and suppose that established simulation is optimal.

S8: Calculate transition matrices for each hidden chain, calculate comorbidity transition matrix of two diseases.

S9: Calculate initial probability of hidden states.

S10: Finally, generate observations from Poisson distribution given λ parameter dependent of hidden state.

So, to test whether the CHMM Copula model optimally fits the simulated data, estimated transition and initial probabilities, λ parameters are compared with the parameters of the simulated data.

5.1.1 Model Performance on the Simulated Data

The three state hidden Markov chains representing two diseases are simulated. According to the simulation design, firstly, two Bernoulli trials representing two diagnostics of the disease i, $i = \{1, 2\}$ at each time t, $\{1, ..., T\}$ were generated. A successful diagnostic outcome indicates the presence of disease i, while an adverse result indicates the absence of disease i. These pairs are aggregated to obtain values representing hidden states of disease i. The order of the generated Bernoulli trials for each disease is such that the joint hidden states follow the Binomial copula and the sequence of hidden states for each disease follows the Markov property. For each disease, observed values were generated using a Poisson distribution with hidden state-dependent rate parameters. For odds ratios of 0.85 and 8, two data sets were simulated 500 times.

The observed values are generated from the Poisson distribution with the state-dependent rate parameters, for disease 1,

$$\begin{pmatrix} 8 & 9 & 9.5 \end{pmatrix}$$

and for disease 2,

$$\begin{pmatrix} 7 & 8 & 8.5 \end{pmatrix}.$$

To distinguish estimated values of the generated data set from the model's estimation, we call them true values of the generated data.

5.1.1.1 Simulation Results for Data Set with an Odds Ratio of 0.85

The binomial copula p.m.f. for an odds ratio of 0.85 is as follows, as derived using Equation 3.7:

$$\bar{\mathrm{p}} = \begin{pmatrix} 0.10 & 0.11 & 0.12 \\ 0.11 & 0.11 & 0.11 \\ 0.12 & 0.11 & 0.10 \end{pmatrix}.$$

The error metrics were obtained using 500 randomly generated data sets. The following matrix summarizes the bias values between measurements of the joint probability of generated hidden states and their true values

$$Bias = \begin{pmatrix} -0.05 & 0.01 & -0.04 \\ 0.01 & 0.10 & 0.01 \\ -0.03 & 0.02 & -0.04 \end{pmatrix},$$

and MSE values are obtained as follows:

$$MSE = \begin{pmatrix} 0.0029 & 0.0009 & 0.0020 \\ 0.0011 & 0.0122 & 0.0010 \\ 0.0013 & 0.0012 & 0.0021 \end{pmatrix}.$$

According to the calculated bias and MSE values, the generated hidden states follow Binomial copula distribution with an odds ratio of 0.85.

The true value of the hidden state transition probabilities is taken as the average of the transition probabilities for 500 generated data sets,

$$\boldsymbol{\pi} = \begin{pmatrix} 0.354 & 0.464 & 0.182 \\ 0.257 & 0.469 & 0.273 \\ 0.171 & 0.455 & 0.374 \end{pmatrix}.$$

True values of the initial state probabilities are $(0.5 \quad 0.4 \quad 0.1)$.

Bias and MSE values of the transition probabilities are obtained as follows:

$$Bias_{\boldsymbol{\pi}} = \begin{pmatrix} 1.73 \cdot 10^{-17} & -1.11 \cdot 10^{-19} & 1.24 \cdot 10^{-17} \\ 2.39 \cdot 10^{-17} & -2.01 \cdot 10^{-17} & -4.44 \cdot 10^{-18} \\ 6.11 \cdot 10^{-18} & -2.20 \cdot 10^{-17} & 1.78 \cdot 10^{-17} \end{pmatrix}$$

and, respectively,

$$MSE_{\boldsymbol{\pi}} = \begin{pmatrix} 0.004 & 0.005 & 0.002 \\ 0.002 & 0.002 & 0.002 \\ 0.002 & 0.004 & 0.003 \end{pmatrix}.$$

According to the chi-square based Markovian test, 98.6% of the generated hidden states for disease 1 and 99.4% for disease 2 satisfy the Markov property.

The results of statistical analyses (Table 5.1) indicate that a significant number of the 500 generated observations for both diseases are stationary.

Table 5.1: Proportion of stationary sequences. Tests are conducted with significance level of 0.05.

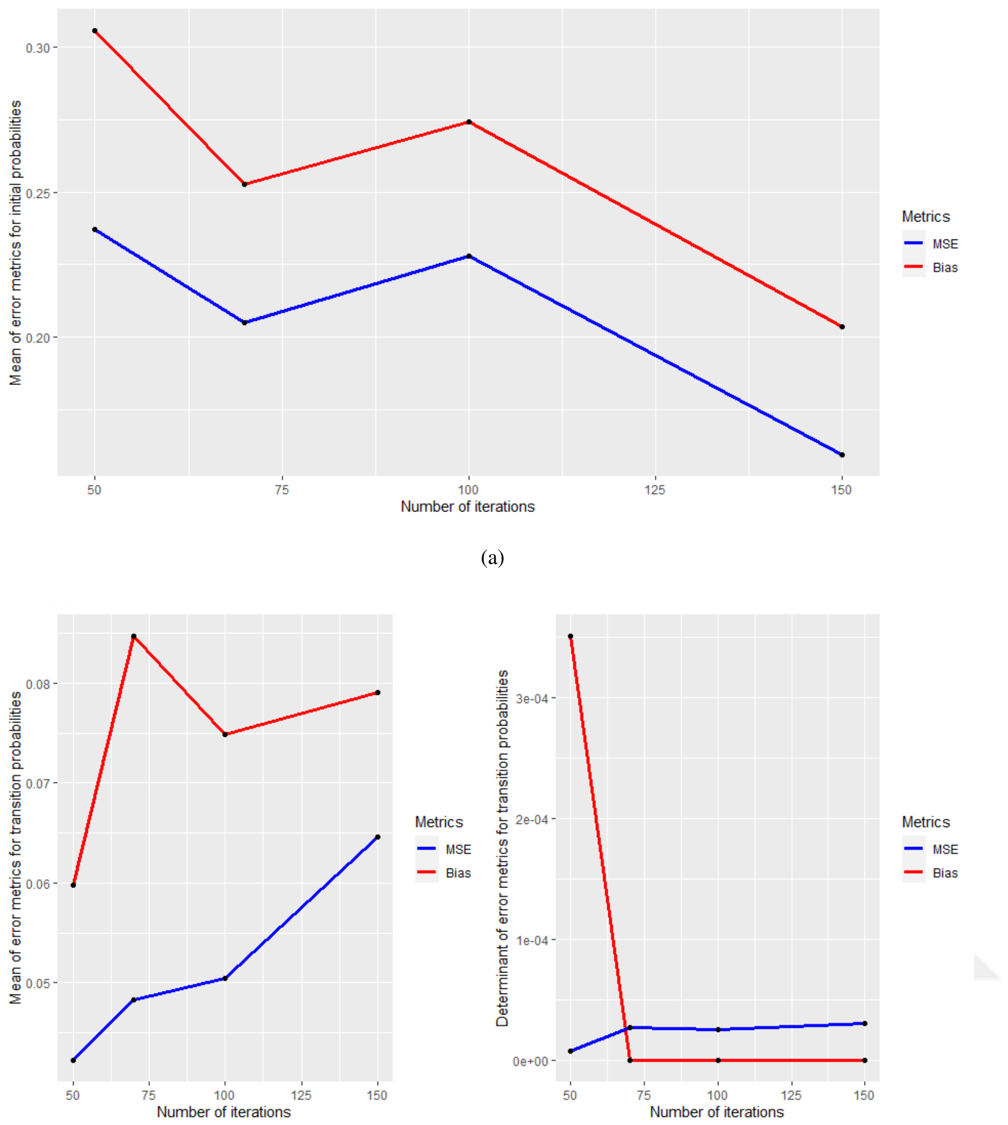
	ADF	KPSS	PP
Disease 1	99.2%	90.8%	100%
Disease 2	98.8%	94.4%	100%

To evaluate the proposed model's performance, it was applied to 50 simulated data sets with the identical true parameters. Moreover, the models with identical initial parameter settings but a varied inner iteration number (50, 70, 100, 150) were applied to the simulated data sets.

Using bias and MSE metrics, the true value of the simulated data parameters are compared to the model's estimated parameters. The mean of the bias and the mean of MSE of the initial state probabilities are calculated and compared for various inner iteration numbers, see Figure 5.4a. In compared to models with less iterations, the model with 150 inner iterations has the lowest mean of both metrics. Mean and determinant of the bias and MSE of the transition state probabilities are calculated. According to Figure 5.4b (left), the mean of bias values varies across iteration numbers, whereas the MSE values increase as the number of iterations increases. Figure 5.4b (right) indicates that the error metrics' determinants are close to zero and that when the iteration number exceeds 70, the determinant values remain constant for both metrics.

We obtain the mean of bias and the mean of MSE of rate parameters for each disease. For iterations greater than 70, the mean of bias and MSE for the first disease are inversely proportional; the model with 100 iterations has the lowest MSE but the highest bias value, as shown in Figure 5.5 (left). On the other hand, the model with 100 iterations produces the highest MSE and the second lowest bias for the second disease (Figure 5.5 (right).

Additionally, the Akaike and Bayesian criteria, as well as weighted RSS, are used to evaluate the performance of models with varying iteration numbers (Figure 5.6). AIC and BIC both exhibit similar dynamics, with the lowest values occurring after 50 iterations and the highest occurring after 150 iterations, whereas weighted RSS exhibits



(a)

(b)

Figure 5.4: Metric values of the models with an odds ratio of 0.85: a) mean of bias and MSE of initial probabilities b) mean of bias and MSE of transition probabilities (left); determinant of bias and MSE of transition probabilities (right).

the lowest values after 150 iterations and the highest following 70 iterations. Because the proposed model has a large number of parameters to estimate, it increases the AIC and BIC criteria; however, weighted RSS is unaffected by the model's complexity and may be a useful tool for evaluating the model's performance.

Generally, for any number of iterations, the calculated bias and MSE between the

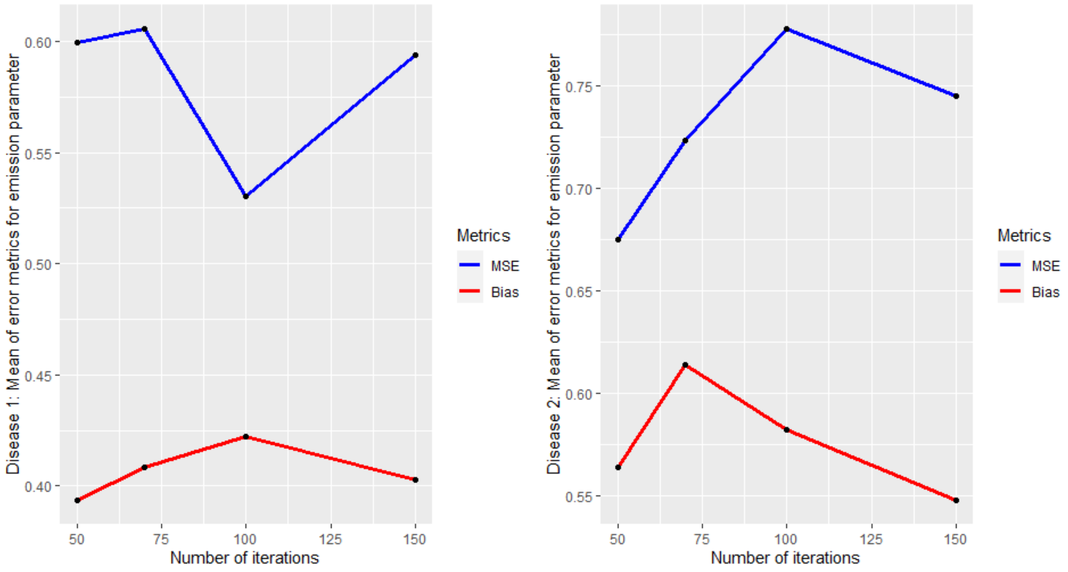


Figure 5.5: Metric values of the models with an odds ratio of 0.85: mean of bias and MSE of emission parameters for disease 1 (left); mean of bias and MSE of emission parameters for disease 2 (right).

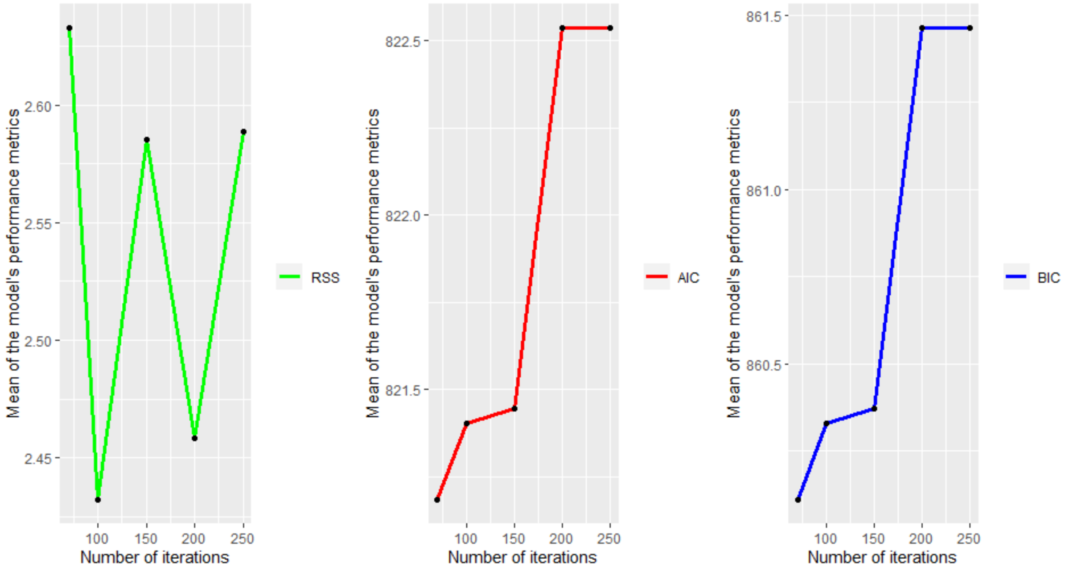


Figure 5.6: Performance metric values of the models with an odds ratio of 0.85.

estimated and true parameters of the simulated data set with an odds ratio of 0.85 are close to zero or slightly greater than zero. The behavior of error metrics varies depending on the initial and transition probabilities, as well as the emission distribution rate parameters. However, iteration numbers of 100 or 150 may be optimal on average for obtaining parameters that are close to their true values.

5.1.1.2 Simulation Results for Data Set with an Odds Ratio of 8

The binomial copula p.m.f. for an odds ratio of 8 is as follows, as derived using Equation 3.7:

$$\bar{\mathrm{p}} = \begin{pmatrix} 0.218 & 0.088 & 0.027 \\ 0.088 & 0.158 & 0.088 \\ 0.027 & 0.088 & 0.218 \end{pmatrix}.$$

500 randomly generated data sets were used to derive the error metrics. The following matrix summarizes the bias values between measurements of the generated hidden states' joint probability and their true values

$$Bias = \begin{pmatrix} -0.019 & 0.009 & 0.005 \\ 0.008 & -0.004 & 0.008 \\ 0.005 & 0.007 & -0.019 \end{pmatrix}$$

and MSE values are derived as follows:

$$MSE = \begin{pmatrix} 0.0006 & 0.0003 & 0.0001 \\ 0.0003 & 0.0003 & 0.0002 \\ 0.0001 & 0.0002 & 0.0005 \end{pmatrix}.$$

Bias and MSE values indicate that the generated hidden states follow a Binomial copula distribution with an odds ratio of 8.

The true value of the hidden state transition probabilities is determined by averaging the transition probabilities for 500 randomly generated data sets,

$$\boldsymbol{\pi} = \begin{pmatrix} 0.858 & 0.124 & 0.018 \\ 0.090 & 0.794 & 0.116 \\ 0.018 & 0.097 & 0.885 \end{pmatrix}.$$

True values of the initial state probabilities are $(0.95 \quad 0.05 \quad 0)$

Bias and MSE values of the transition probabilities are obtained as follows:

$$Bias_{\boldsymbol{\pi}} = \begin{pmatrix} 3.11\cdot^{-17} & -4.27\cdot^{-18} & -8.33\cdot^{-20} \\ 6.76\cdot^{-18} & -2.18\cdot^{-17} & 2.43\cdot^{-18} \\ 4.86\cdot^{-19} & 6.45\cdot^{-18} & -5.08\cdot^{-17} \end{pmatrix}$$

Table 5.2: Proportion of stationary sequences. Tests are conducted with significance level of 0.05. KPSS test also includes the results for significance level of 0.01.

	ADF	KPSS (0.05)	KPSS (0.01)	PP
Disease 1	94%	61.4%	88%	100%
Disease 2	93.8%	68.2%	88.6%	100%

and, respectively,

$$MSE_{\boldsymbol{\pi}} = \begin{pmatrix} 0.0009 & 0.0008 & 0.0001 \\ 0.0007 & 0.0011 & 0.0006 \\ 0.0001 & 0.0009 & 0.0009 \end{pmatrix}.$$

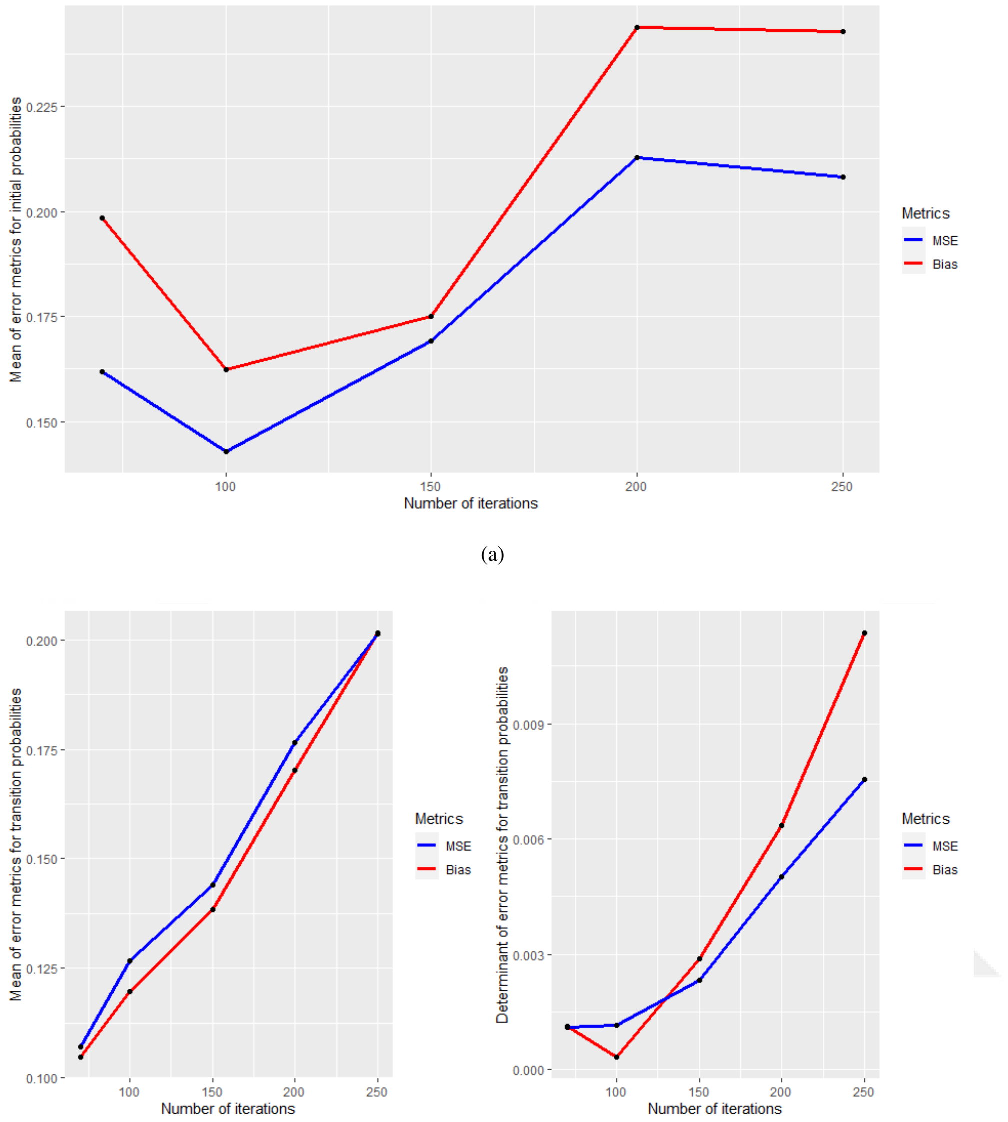
According to the chi-square based Markovian test, 87% of the generated hidden states for disease 1 and 93% for disease 2 satisfy the Markov property.

According to the results of statistical tests (Table 5.2), high proportion of the 500 generated observations for both diseases are stationary.

To assess the proposed model's performance, 50 simulated data sets with equal true parameters were used. Additionally, models with identical starting parameter settings but a varied number of inner iterations (70, 100, 150, 200, 250) were applied to the simulated data sets.

The true values of the simulated data parameters are compared to the model's estimated parameters using bias and MSE metrics. For various inner iteration numbers, the mean of the bias and the mean of the MSE of the initial state probabilities are calculated and compared, see Figure 5.7a. The model with 100 inner iterations had the smallest mean across both measures. We determine the mean and determinant of the bias, as well as the mean square error of the transition state probabilities. As illustrated in Figure 5.7b (left), as the number of iterations grows, the mean of bias and mean of MSE increase. Figure 5.7b (right) demonstrates that as the iteration number increases over 100, the determinant values for both metrics increase, but the increase is negligible because all determinant values for different iterations are near to zero.

We obtain the mean of bias and the mean of MSE of rate parameters for each disease.



(a)

(b)

Figure 5.7: Metric values of the models with an odds ratio of 8: a) mean of bias and MSE of initial probabilities b) mean of bias and MSE of transition probabilities (left); determinant of bias and MSE of transition probabilities (right).

The mean of bias and MSE for the first disease are the lowest for 200 and 250 iterations, as illustrated in Figure 5.5 (left). On the other hand, for the second disease, the model with 100 and 150 iterations delivers the lowest mean of MSE and bias (Figure 5.5 (right). According to both plots, the model with 100 iterations estimates state-dependent rate parameters close to their true values for both diseases.

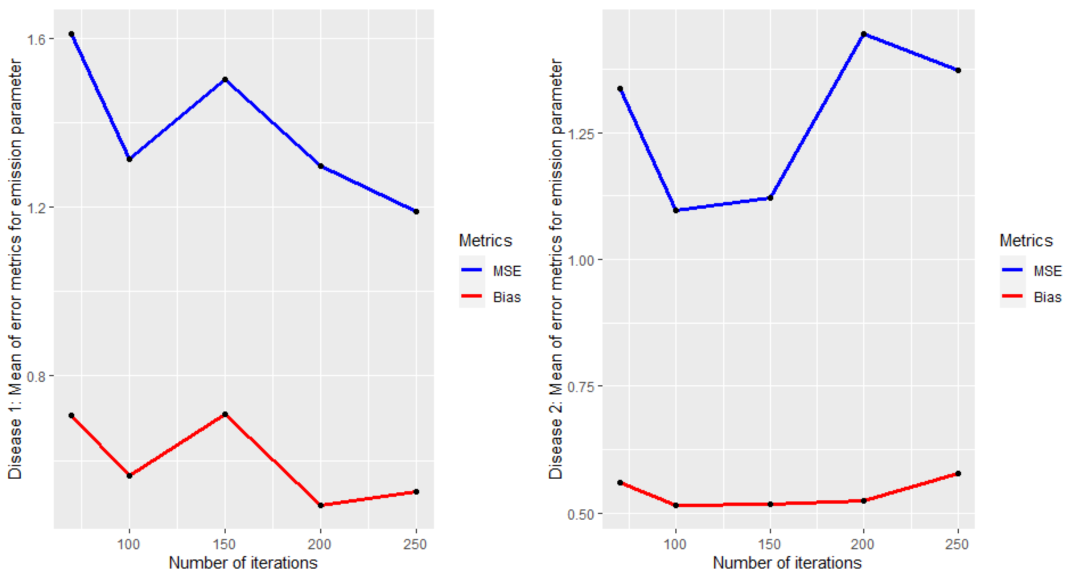


Figure 5.8: Metric values of the models with an odds ratio of 8: mean of bias and MSE of emission parameters for disease 1 (left); mean of bias and MSE of emission parameters for disease 2 (right).

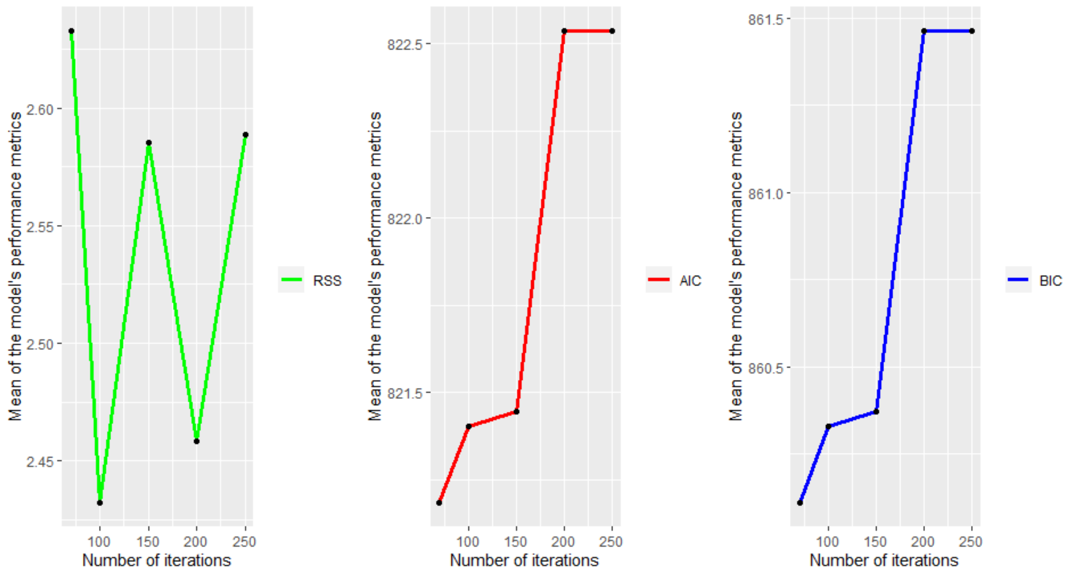


Figure 5.9: Performance metric values of the models with an odds ratio of 8.

AIC and BIC both exhibit similar dynamics, with the lowest values occurring after 70 iterations and the highest values occurring after 200 and 250 iterations, respectively. The weighted RSS, on the other hand, exhibits a fluctuating dynamic, with the lowest values occurring after 100 iterations and the second lowest following 200 iterations (Figure 5.9).

Generally, the calculated bias and MSE between the estimated and true parameters of the simulated data set with an odds ratio of 8 are close to or slightly greater than zero for any number of iterations. Only the mean MSE of rate parameters for both diseases occasionally exceeds one. The behavior of error metrics varies depending on the initial and transition probabilities, as well as the emission distribution rate parameters. According to a broad review of parameter measurements, weighted RSS, and information criteria, iteration number 100 may be best for acquiring parameters close to their true values on average.

For simulated data with an odds ratio of 0.85 and a negative and close to zero Yule's coefficient of association of -0.054, as well as for data with an odds ratio of 8 and a positive Yule's coefficient of association of 0.57, CHMM with discrete copula exhibits a satisfactory goodness of fit.

CHAPTER 6

REAL LIFE APPLICATION

In this section of the thesis, we would like to present the applicability of the proposed method onto the real data set.

6.1 Data and its Analysis

The data set includes daily information relating to the heart disease and the hypertension collected from 49,713 patients from January 2015 to December 2020. Data set is authorized by a private hospital in Turkey, whose name has to be kept confidential due to privacy of the data content. Each entry has the following attributes:

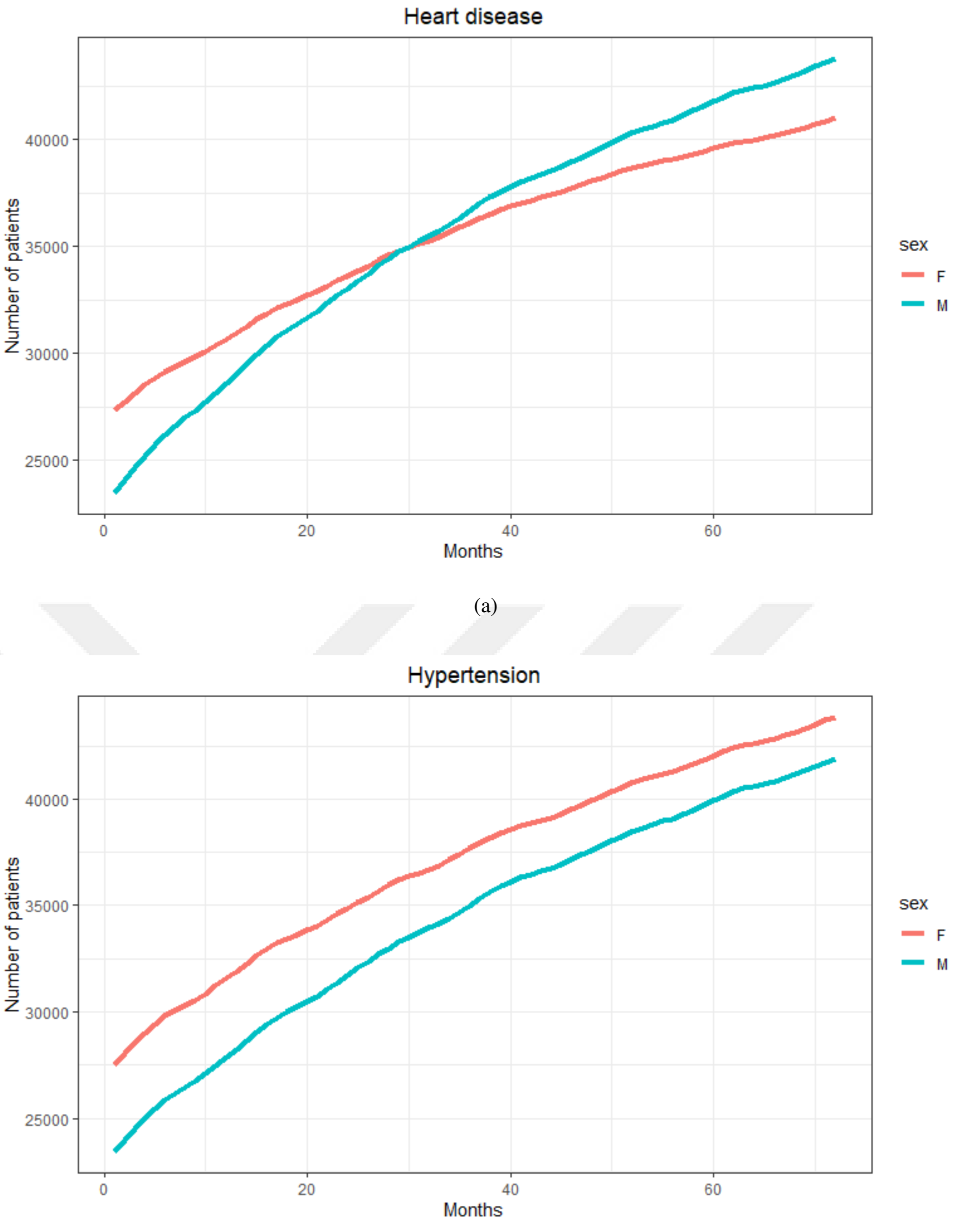
- ID of a patient.
- Age of a patient. Does not appear to change over time, even with patients who span multiple years.
- Gender of a patient.
- Code of disease that the patient is diagnosed with.
- Medical name of the disease that the patient is diagnosed with.
- Date in which the entry was made.

Each patient's data is altered to cover 72 months, beginning with the first month of 2015 and ending with the last month of 2020. Thus, each patient is associated with 72 rows. The raw data set's illness codes were classified into two categories: heart disease and hypertension. The repeated rows have been removed. After the raw data is transformed, the main data set has the following columns:

- ID: describes unique patient ID.
- Age (discrete variable): describes patient age in years. Age is determined by assuming that the patient's initial entry in the raw data contains their true age and incrementing or decrementing it for preceding and previous years, respectively.
- Gender (nominal categorical variable): it describes patient gender as male or female.
- Heart disease and hypertension (nominal categorical variables): each column includes two different possible values, 0 or 1, where 1 is representing the presence of the indicated disease. It is important to notice that once a patient has been diagnosed with a disease, the value for that disease remains 1 for the remainder of the column.
- Month: it ranges from 0 to 72, where 1 represents the first month of 2015, and 72 is the last month of 2020.
- Frequency (discrete variable): it represents the frequency with which a patient was admitted to the hospital during a given month, regardless of the reason.
- Frequency-heart disease and frequency-hypertension (discrete variables): each column indicates the frequency with which a patient presented to the hospital with the indicated ailment.

After cleaning and transforming panel data, it was aggregated to create time series data by counting patients with heart disease or hypertension at a specific time point. Thus, the new data set comprises two columns, heart disease and hypertension, which show the total number of patients diagnosed with a certain ailment during a given month. Additionally, the new data set includes information on gender and the number of hospital admissions associated with the identified disease.

The total number of patients with heart disease or/and hypertension was analyzed using time series analysis and exploratory data analysis.



(a)

(b)

Figure 6.1: The number of patients by gender from January 2015 to December 2020 with: (a) heart disease, (b) hypertension.

Figures 6.1a and 6.1b depict time series plots of the total number of patients with a specific disease divided by gender. Both diseases have an increasing trend due to assumptions that if a patient is reported to have a particular disease, he will be assumed to have that disease until the study is completed. While both genders experience an almost proportional increase in hypertension, until the 30th month, females have a greater prevalence of heart disease than males, and the situation reverses after the 30th month.

Figures 6.2a and 6.2b show the total number of hospital appointments for both diseases by gender. Hospital admissions for heart disease or/and hypertension follow a seasonal pattern. Females visit the hospital less frequently than males (check descriptive statistics), while hospital visits for patients with hypertension are comparable for both genders.

As illustrated in Figures 6.3a and 6.3b, the number of patients with heart disease and hypertension, as well as patient hospital admissions, are highly linearly correlated. For both diseases, the total number of patients and their hospital visits follow a left-skewed distribution.

We calculated the average time interval between disease occurrences. When hypertension occurs first, it takes an average of 12.96 months to be diagnosed with heart disease later; however, it takes an average of 18.88 months to be diagnosed with hypertension following heart disease. There are 4236 patients who develop heart disease after hypertension and 3108 patients who develop hypertension following heart disease.

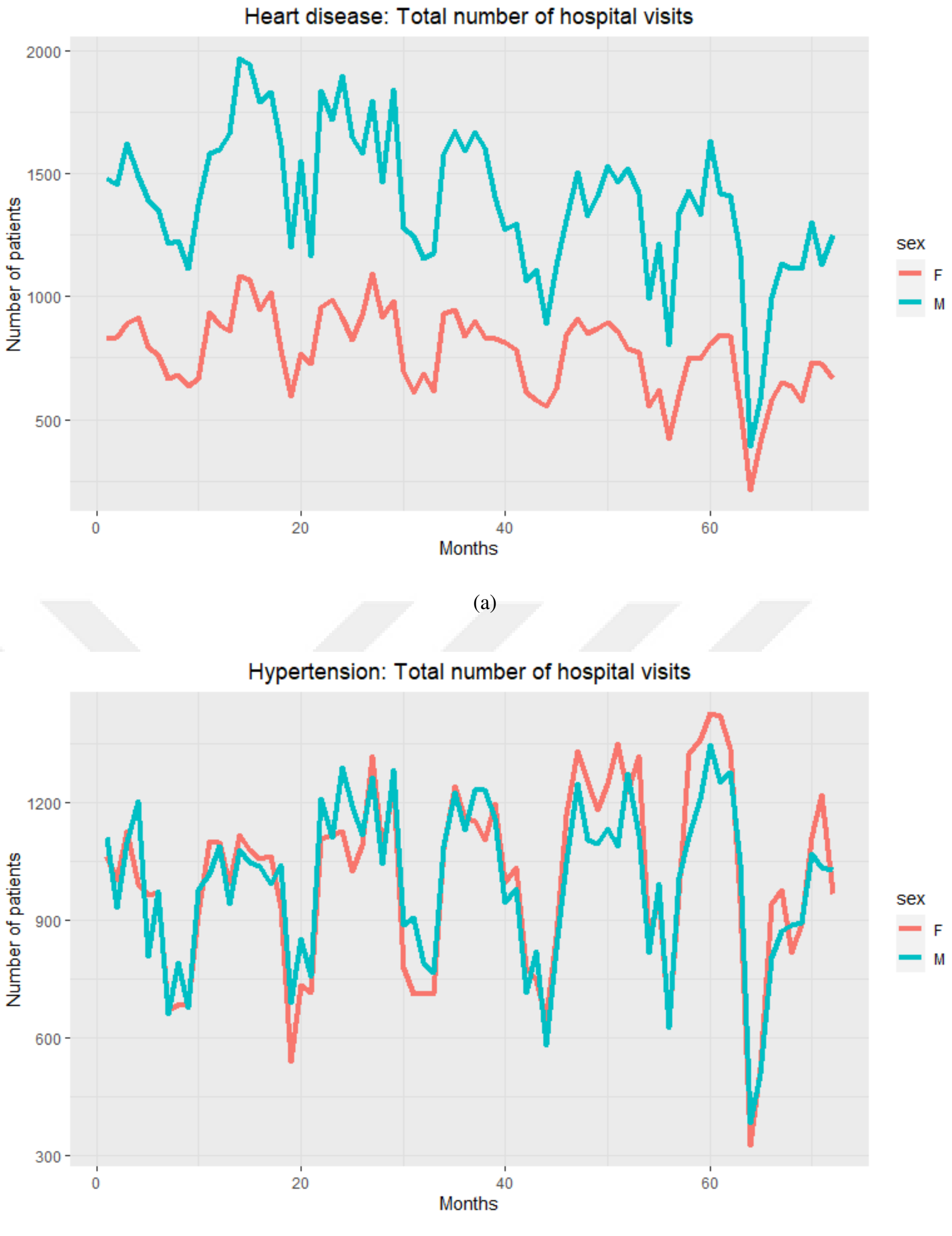
Table 6.1: The number of patients having two diseases. The rows denote the first disease that occurred, while the columns denote the subsequent disease.

	Heart disease	Hypertension
Heart disease	-	3108
Hypertension	4236	-

Table 6.2: Time interval between two diseases on average (months). The rows denote the first disease that occurred, while the columns denote the subsequent disease.

	Heart disease	Hypertension
Heart disease	-	18.88
Hypertension	12.96	-

Figure 6.4 show an increasing trend of monthly number of patients with heart disease

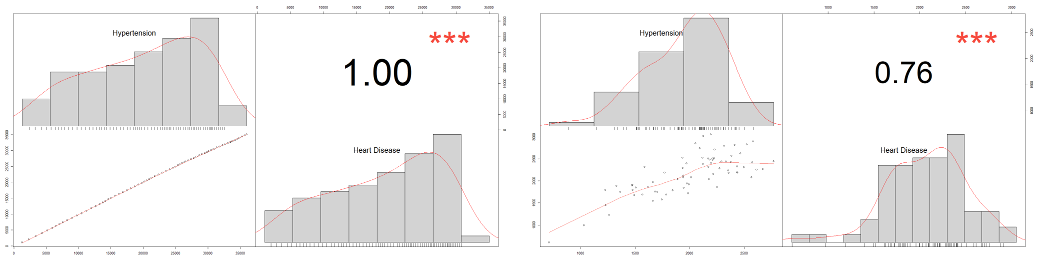


(a)

(b)

Figure 6.2: The number of hospital visits by gender from January 2015 to December 2020: (a) heart disease, (b) hypertension.

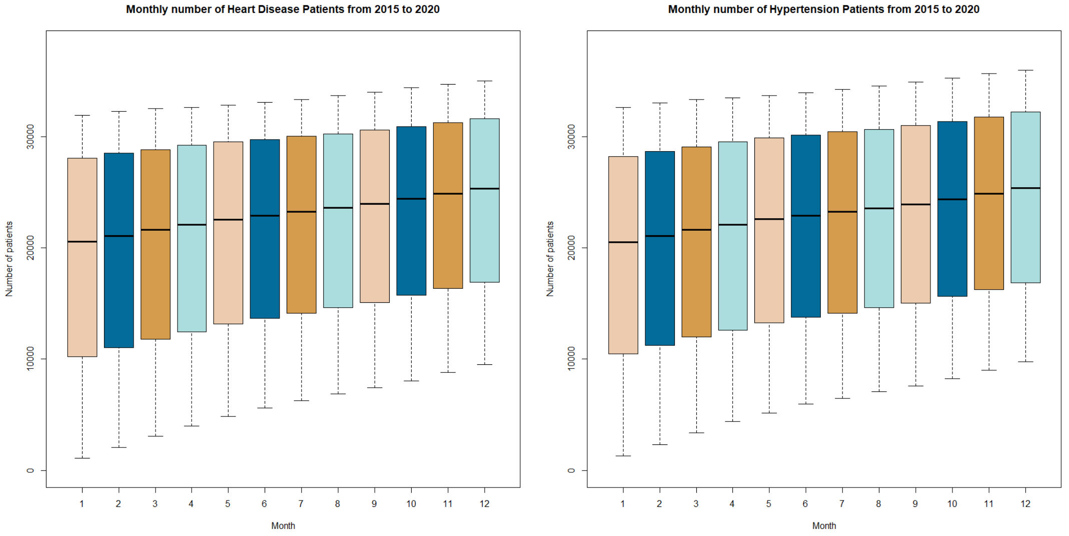
and hypertension over 5 years. Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF) plots of total number of patients over 72 months and those with the second-order difference for both diseases are represented in Figure 6.5. Time series of patients number for both diseases are nonstationary, have an increasing trend.



(a)

(b)

Figure 6.3: Heart disease and hypertension correlation plots and histograms: (a) patients number (b) number of hospital visits.

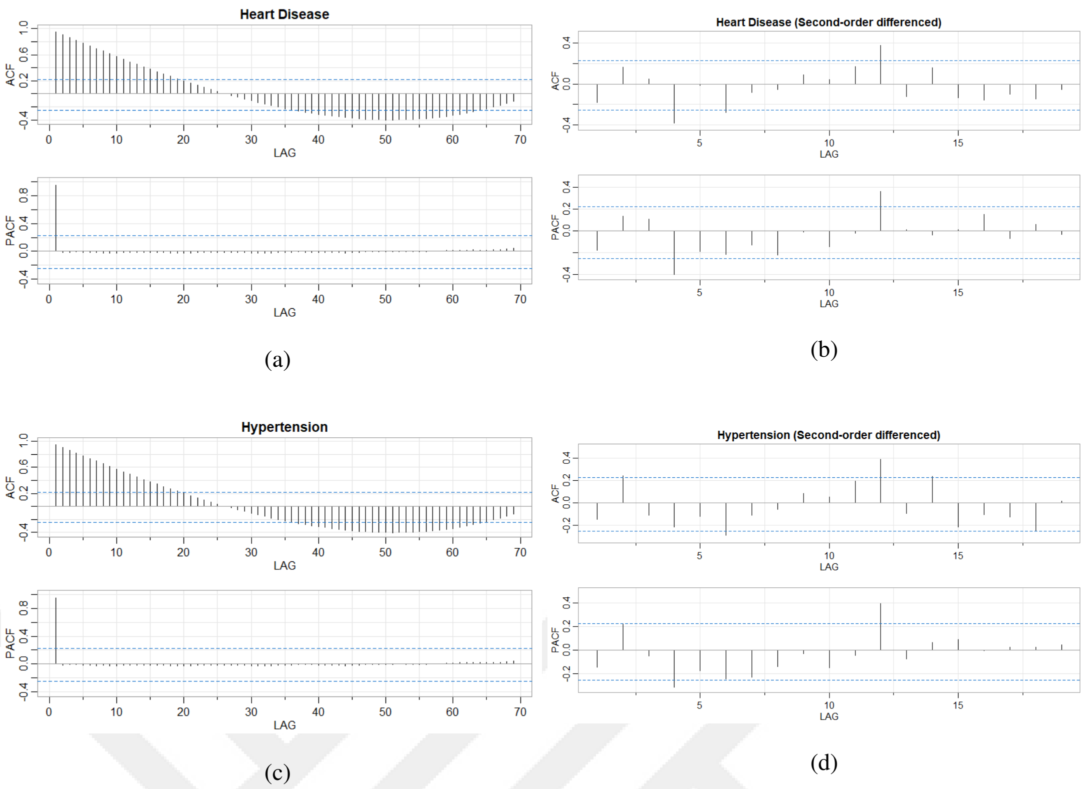

Monthly number of Heart Disease Patients from 2015 to 2020

Monthly number of Hypertension Patients from 2015 to 2020

Figure 6.4: Graph of the estimated transition probabilities.

While HMMs are applicable to both stationary and nonstationary time series (Zucchini et al., 2017), CHMMs, the proposed model's underlying model, do not include any theoretical information indicating whether this type of model can be used for nonstationary time series (Brand et al., 1997). Additionally, application studies used stationary series or transformed nonstationary ones to work with CHMM or models based on CHMM (Kubanek et al., 2012, Ghosh et al., 2017, Darmanjian et al., 2006, Abdelaziz et al., 2014).

A Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test is used to determine the number of differences required to make time series stationary at the significance level of 0.05.



(a)

(b)

(c)

(d)

Figure 6.5: ACF and PACF plots of the number of patients over 72 months: a) heart disease b) heart disease (second-order difference) c) hypertension d) hypertension (second-order difference).

Second-order differencing is performed on the time series of both diseases based on the test results. According to Figure 6.5, heart disease and hypertension series with second-order lag are stationary. The augmented Dickey-Fuller (ADF) test with p-values less than 0.01 for both diseases, the KPSS test with p-values greater than 0.01 for both diseases, and the Phillips-Perron (PP) test with p-value of 0.01 for heart disease and hypertension indicate that the series are stationary.

6.2 Application of the Proposed Model

The proposed model was applied to the number of patients with heart disease and hypertension resulting from the second-order differencing. Because the transformation produces negative values, the observations are increased by the minimum absolute value of the resulting observations, 270. Following that, the observed values are scaled by ten to obtain a stable VEM algorithm fit. Initial setting for the model parameters are arranged as follows:

(i) transition probability matrix with initial state probability, $(0.9 \quad 0.1 \quad 0)$, is defined as

$$\begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{pmatrix};$$

(ii) rate parameters for emission distribution are arranged according to the mean and median values of the observations, the initial values set to $(22 \quad 26 \quad 30)$ for heart disease and $(20 \quad 24 \quad 28)$ for hypertension.

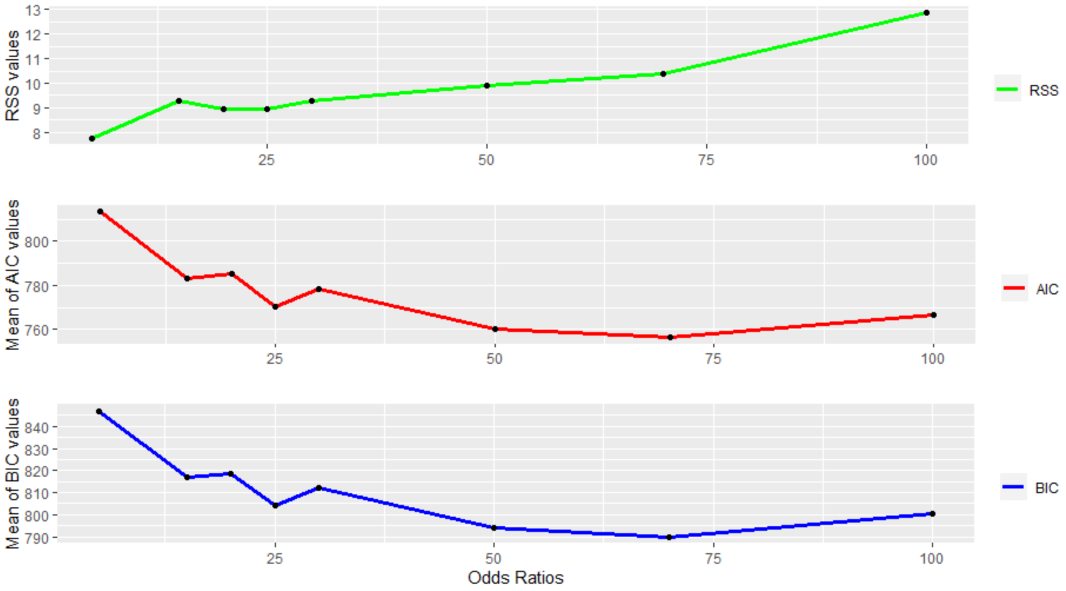


Figure 6.6: Performance metrics for models with various odds ratios.

To obtain the model with the optimum fit, the odds ratios between two hidden states of the diseases, $(5, 15, 20, 25, 30, 50, 70, 100)$, were utilized. Due to the fact that we cannot observe the hidden states, in order to determine the odds ratio between the occurrences of the diseases' hidden states, we examined a range of odds ratio values and fitted the model for each odds ratio. The optimal odds ratio value is chosen based on the one with the lowest weighted RSS. To remind, the weighted RSS of the model is a sum of the weighted RSSs, calculated for each disease.

Various inner iteration numbers were used, $(150, 200, 250, 300, 350, 400)$; the optimal iteration number for each odds ratio is determined by the lowest weighted RSS value. Figure 6.6 summarizes the performance of the selected models with varying odds ratios. Despite the odds ratio of 5 has the lowest weighted RSS, AIC and BIC values are the highest ones. Therefore, all metrics should be considered when determining the optimal odds ratio. According to the plots, the optimal model has an odds ratio of 70 based on AIC, BIC, and RSS values. To ensure that the decision made on the basis of the visual interpretation of the plots is correct, the metrics values displayed in the plots are aggregated; the odds ratio of 70 has the lowest aggregated value.

The estimated parameters of the model after 350 iterations are as follows:

- transition probability matrix,

$$\begin{array}{c} \begin{array}{ccc} 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0.000 & 0.000 & 1.000 \\ 0.023 & 0.972 & 0.005 \\ 0.000 & 0.074 & 0.926 \end{pmatrix} \end{array} \begin{array}{c} \\ 1 \\ 2 \\ 3 \end{array}$$

- initial state probabilities,

$$(0.001 \quad 0.999 \quad 0);$$

- emission distribution parameters of the state-dependent observations for heart disease and hypertension,

$$(25.23 \quad 26.08 \quad 26.46),$$

  and, respectively,

$$(24.38 \quad 27.68 \quad 28.70).$$

The weighted RSS value for the model is 10.38. AIC values for heart disease are 716.85 and hypertension are 795.69, respectively; thus, the average AIC value is 756.27. BIC values for heart disease are 750.58 and hypertension are 829.42; the average BIC value is 790.

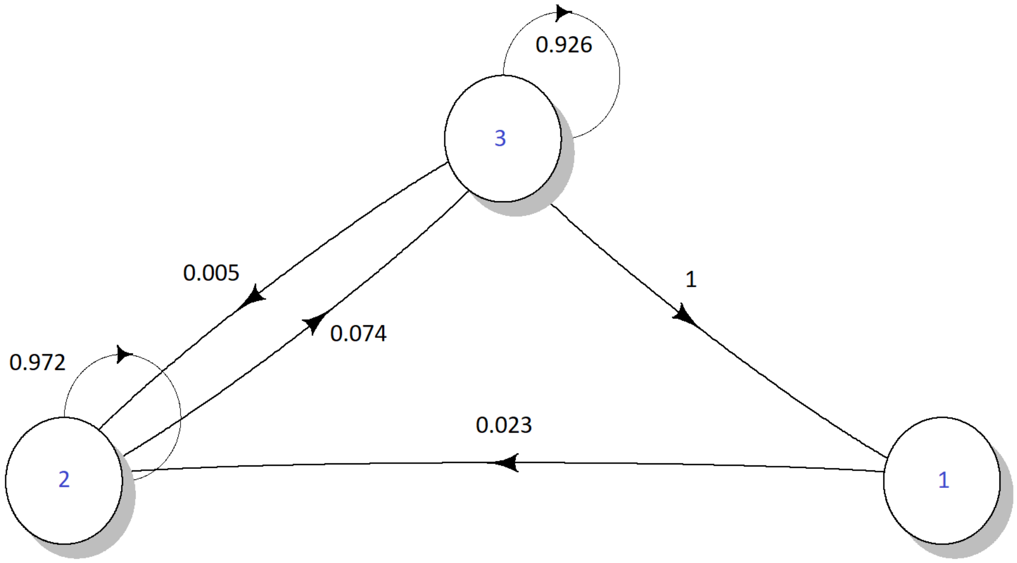


Figure 6.7: Graph of the estimated transition probabilities.

To evaluate the developed model's performance, separate 3-state HMMs were applied to the second-order differenced number of patients with heart disease and hypertension, respectively. The observed values are also scaled by 10. The initial values of the parameters are identical to the initial settings used in the proposed model's application.

For heart disease AIC and BIC values are 494.33 and 514.57, for hypertension are 516.86 and 537.10, respectively.

Some studies developed a method for aggregating obtained information criteria and using them to evaluate model performance and optimize parameters. According to the study of Ngatchou-Wandji and Bulla, 2011, AIC or BIC values are computed for each cluster and then combined; the model with the lowest summed AIC or BIC values is selected. Inouye et al., 1995 computed the AIC or BIC for separated sequences of autoregressive model and then aggregated the information criteria.

Second, the separate HMM models are not able to capture the interactions among different models. Zhong and Ghosh, 2001 In many applications, multiple sequences are interacting with one another The aggregated AIC and BIC values of HMMs for heart disease and hypertension, respectively, are 1011.19 and 1051.67. In comparison to the proposed model's averaged AIC and BIC values, independent HMMs underperform than the proposed model with discrete copula accounting for the dependency of the hidden states underlying two diseases.

6.2.1 Interpretation of the Findings

The Binomial copula distribution for estimated odds ratio of 70 is given as follows:

$$\begin{array}{cc} \begin{array}{ccc} 1 & 2 & 3 \end{array} & \\ \begin{pmatrix} 0.2849 & 0.0443 & 0.0041 \\ 0.0443 & 0.2448 & 0.0443 \\ 0.0041 & 0.0443 & 0.2849 \end{pmatrix} & \begin{array}{c} 1 \\ 2 \\ 3. \end{array} \end{array}$$

The Yule's coefficient of association between hidden states representing heart disease and hypertension is 0.84; correlation analysis conducted on the data supports this result. High association of hidden chains representing heart disease and hypertension is supported by medical studies (J. Wu et al., 2017, Schellevis et al., 1993). For example, pulmonary hypertension complicates the course of many adults with congenital heart diseases (Beghetti and Tissot, 2009), severity of pulmonary hypertension in patients with left heart diseases is studied (Vachiéry et al., 2013, Jin et al., 2017).

According to the transition dynamic of hidden states and copula probabilities of joint states, we designate states as follows:

(i) State 2 represents a single disease (no comorbidity reason);

(ii) State 1 corresponds to the presence of light comorbidity (one comorbidity reason);

(iii) State 3 corresponds to the presence of moderate comorbidity (two comorbidity reasons).

According to the Binomial copula matrix, joint hidden states with different state numbers, such as $(1, 2)$, $(2, 1)$, $(2, 3)$, $(3, 2)$, $(1, 3)$, and $(3, 1)$, have a low probability of occurrence, implying that there is a low probability that two diseases occur on distinct states. For instance, there is a low possibility that heart disease is on a single disease state, while hypertension is on a moderate comorbidity. On the other hand, the probability of both diseases occurring in a light comorbidity state, as well as a moderate comorbidity state, is highest at 0.2849. The probability of both diseases coexisting in a single disease state is 0.2448.

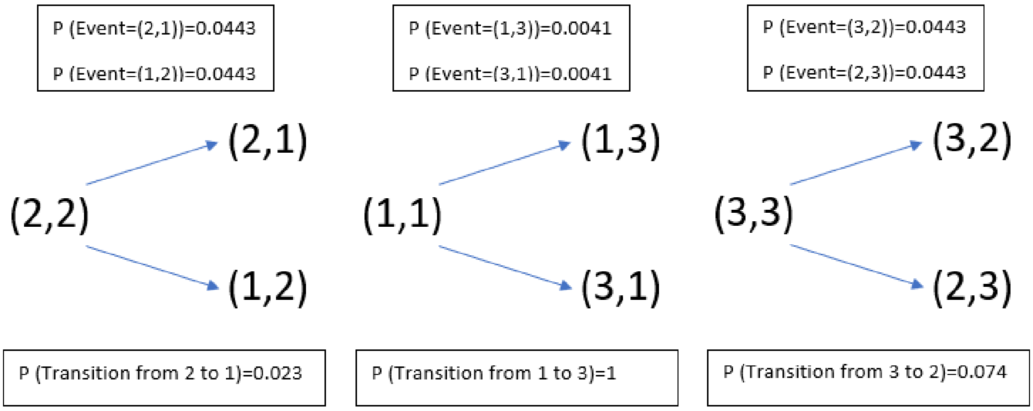


Figure 6.8: Example structure of joint hidden states with Binomial copula and transition probabilities.

Figure 6.8 displays example schemes for transitioning between joint hidden states. By multiplying the joint state and transition probabilities, we can calculate the probability of one of the diseases moving from a single disease state to a light comorbidity state, that is $(2,2) \rightarrow (2,1)$ or $(2,2) \rightarrow (1,2)$. Thus,

$$P((2,2) \rightarrow (2,1)) = 0.023 * 0.0443 = 0.00102.$$

The probability of both diseases remaining in a moderate comorbidity state is 0.2638, the probability of both diseases remaining in a single disease state is 0.2379, the probability of both diseases remaining in a light comorbidity state is 0, and the probability of one of diseases transitioning from light comorbidity to moderate comorbidity is 0.0041.

The proposed model enables investigation of the joint behavior of hidden chains and hidden states, as well as their transition dynamics. Thus, we gain a better understanding of the dependency structure of unobserved knowledge regarding diseases with limited patient data based on hospital visit data across time.

CHAPTER 7

CONCLUDING COMMENTS

The main focus of the proposition in this thesis is to use a combination of hidden Markov theory and copula function to model parallel interacting processes describing two or more chronic diseases.

We develop a novel coupled hidden Markov model in this study that incorporates a bivariate discrete copula function into the hidden process. We employ a novel type of discrete copula, the Binomial copula. We compute a CDLL and develop the necessary inference to implement the model. Due to the large number of parameters required to estimate parameters using the EM algorithm, even for two hidden chains, estimation becomes computationally intractable. As a result, we estimate the model's parameter using a VEM algorithm. Because the variational expectation part of the algorithm requires computing the CDLL's lower bound approximation, we use Kullback-Leibler divergence to derive the lower bound model's log-likelihood. The VEM algorithm is carried out by computing conditional expectations based on forward-backward probabilities and estimators of parameters that maximize the CDLL.

The most frequently encountered problem in computing forward-backward probabilities, numerical underflow of the calculated probabilities, is resolved by transforming the functions used in the algorithm; additionally, scaling the observed values helps overcome numerical underflow.

Due to the unidentifiability of copula functions defined on discrete space, Geenens, 2020 developed a new bivariate discrete copula. Despite extensive theoretical development, statistical inference for parameter estimation of the joint copula function has not been developed yet. Thus, we designed a structure for joint states in such a way that they follow a predefined Binomial copula probability math function with a given odds ratio while also satisfying the Markov property of hidden states in each chain. To verify and confirm the correctness of the designed structure, we calculated bias and MSE metrics and performed a Markovian test on 500 simulated data. The simulation study was conducted for two different odds ratio, and the developed model was applied. The simulation study's findings are satisfactory.

The proposed model is applied to real data from a private hospital from January 2015 to December 2020, including information on hospital appointments. The observed variable is the total number of patients diagnosed with a particular disease in a given month. The purpose of this application is to define the dependency structure of unobserved disease data. The application results demonstrate that the CHMM with discrete copula is useful for investigating disease comorbidity when only population dynamics over time are available and no clinical data are available.

During the application, one of the difficulties encountered is that the proposed model, like all HMM-based models, is sensitive to initial parameter settings. Because the developed model has a large number of parameters to estimate, the optimal fit is defined using combined information from weighted RSS and information criteria. Additionally, the algorithm have been run multiple times with varying inner iteration counts in order to overcome a local optimum.

The model is applicable to the study of more than two diseases; more than two hidden chains can be included in the model. Additionally, the proposed model is general enough to account for any dependent events. The model can be extended to include covariates in both observed and hidden spaces. The proposed approach models interacting time series observations; however, by combining CDLL and GLM functions, it is possible to extend this model to apply it to longitudinal data.

REFERENCES

Aalen, O. (1978). Nonparametric estimation of partial transition probabilities in multiple decrement models. *The Annals of Statistics*, 534–545.

Abdelaziz, A. H., Charaf, L. A., Zeiler, S., & Kolossa, D. (2014). On dynamic stream weight learning for coupled-hmm-based audio-visual speech recognition.

Akat, F., Selcuk-Kestel, A. S., & Tank, F. (2019). The estimation of adopted mortality and morbidity rates using model and the phase type law: The turkish case. *Communications in Statistics-Simulation and Computation*, *48*(9), 2552–2565.

Andersen, P. K., Abildstrom, S. Z., & Rosthøj, S. (2002). Competing risks as a multi-state model. *Statistical methods in medical research*, *11*(2), 203–215.

Anderson, T. W., & Goodman, L. A. (1957). Statistical inference about markov chains. *The annals of mathematical statistics*, 89–110.

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, *41*(1), 164–171.

Beghetti, M., & Tissot, C. (2009). Pulmonary arterial hypertension in congenital heart diseases. *Seminars in respiratory and critical care medicine*, *30*(04), 421–428.

Begun, A., Icks, A., Waldeyer, R., Landwehr, S., Koch, M., & Giani, G. (2013). Identification of a multistate continuous-time nonhomogeneous markov chain model for patients with decreased renal function. *Medical Decision Making*, *33*(2), 298–306.

Berry, S. D., Ngo, L., Samelson, E. J., & Kiel, D. P. (2010). Competing risk of death: An important consideration in studies of older adults. *Journal of the American Geriatrics Society*, *58*(4), 783–787.

Bishop, C. M., & Svensén, M. (2012). Bayesian hierarchical mixtures of experts. *arXiv preprint arXiv:1212.2447*.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, *112*(518), 859–877.

Brand, M., Oliver, N., & Pentland, A. (1997). Coupled hidden markov models for complex action recognition. *Proceedings of IEEE computer society conference on computer vision and pattern recognition*, 994–999.

Bringmann, L. F., Vissers, N., Wichers, M., Geschwind, N., Kuppens, P., Peeters, F., Borsboom, D., & Tuerlinckx, F. (2013). A network approach to psychopathology: New insights into clinical longitudinal data. *PloS one*, *8*(4), e60188.

Chen, Y.-H. (2010). Semiparametric marginal regression analysis for dependent competing risks under an assumed copula. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *72*(2), 235–251.

Chib, S. (1996). Calculating posterior distributions and modal estimates in markov mixture models. *Journal of Econometrics*, *75*(1), 79–97.

Cooper, G. F., Aliferis, C. F., Ambrosino, R., Aronis, J., Buchanan, B. G., Caruana, R., Fine, M. J., Glymour, C., Gordon, G., Hanusa, B. H., et al. (1997). An evaluation of machine-learning methods for predicting pneumonia mortality. *Artificial intelligence in medicine*, *9*(2), 107–138.

Darmanjian, S., Kim, S.-P., Nechyba, M. C., Principe, J., Wessberg, J., & Nicolelis, M. A. (2006). Independently coupled hmm switching classifier for a bimodel brain-machine interface. *2006 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*, 379–384.

David, H. A., & Moeschberger, M. L. (1978). *The theory of competing risks: Ha david, ml moeschberger*. C. Griffin.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, *39*(1), 1–22.

Derrode, S., & Pieczynski, W. (2016). Unsupervised classification using hidden markov chain with unknown noise copulas and margins. *Signal Processing*, *128*, 8–17.

Escarela, G., & Carriere, J. F. (2003). Fitting competing risks with an assumed copula. *Statistical Methods in Medical Research*, *12*(4), 333–349.

Esteban, C., Rodrıguez, P., Escudero, J. R., Clarà, A., Fernández, A., Fernández, S., & Agúndez, I. (2019). Anaemia in patients who underwent vascular surgery: A significant predictor of amputation and death. *Medicina Clınica (English Edition)*, *152*(1), 6–12.

Faruqui, S. H. A., Alaeddini, A., Jaramillo, C. A., Potter, J. S., & Pugh, M. J. (2018). Mining patterns of comorbidity evolution in patients with multiple chronic conditions using unsupervised multi-level temporal bayesian network. *PloS one*, *13*(7), e0199768.

Feinstein, A. R. (1970). The pre-therapeutic classification of co-morbidity in chronic disease. *Journal of chronic diseases*, *23*(7), 455–468.

Fine, S., Singer, Y., & Tishby, N. (1998). The hierarchical hidden markov model: Analysis and applications. *Machine learning*, *32*(1), 41–62.

Forkan, A. R. M., & Khalil, I. (2017). Peace-home: Probabilistic estimation of abnormal clinical events using vital sign correlations for reliable home-based monitoring. *Pervasive and Mobile Computing*, *38*, 296–311.

Geenens, G. (2020). Copula modeling for discrete random vectors. *Dependence Modeling*, *8*(1), 417–440.

Genest, C., & Nešlehová, J. (2007). A primer on copulas for count data. *ASTIN Bulletin: The Journal of the IAA*, *37*(2), 475–515.

Ghahramani, Z., & Jordan, M. I. (1996). Factorial hidden markov models. *Advances in Neural Information Processing Systems*, 472–478.

Ghahramani, Z., & Jordan, M. I. (1997). Factorial hidden markov models. *Machine learning*, *29*(2), 245–273.

Ghosh, S., Li, J., Cao, L., & Ramamohanarao, K. (2017). Septic shock prediction for icu patients via coupled hmm walking on sequential contrast patterns. *Journal of biomedical informatics*, *66*, 19–31.

Gonçalves, J. G. F., Silva, V. J. D., Borges, M. C. C., Prata, A., & Correia, D. (2010). Mortality indicators among chronic chagas patients living in an endemic area. *International journal of cardiology*, *143*(3), 235–242.

Goodman, L. A., & Kruskal, W. H. (1979). Measures of association for cross classifications. *Measures of association for cross classifications* (pp. 2–34). Springer.

Gooley, T. A., Leisenring, W., Crowley, J., & Storer, B. E. (1999). Estimation of failure probabilities in the presence of competing risks: New representations of old estimators. *Statistics in medicine*, *18*(6), 695–706.

Groen, R. N., Ryan, O., Wigman, J. T., Riese, H., Penninx, B. W., Giltay, E. J., Wichers, M., & Hartman, C. A. (2020). Comorbidity between depression and anxiety: Assessing the role of bridge mental states in dynamic psychological networks. *BMC medicine*, *18*(1), 1–17.

Guisado-Clavero, M., Roso-Llorach, A., López-Jimenez, T., Pons-Vigués, M., Foguet-Boreu, Q., Muñoz, M. A., & Violán, C. (2018). Multimorbidity patterns in the elderly: A prospective cohort study with cluster analysis. *BMC geriatrics*, *18*(1), 1–11.

Hsieh, H.-J., Chen, T. H.-H., & Chang, S.-H. (2002). Assessing chronic disease progression using non-homogeneous exponential regression markov models: An illustration using a selective breast cancer screening in taiwan. *Statistics in medicine*, *21*(22), 3369–3382.

Hu, W.-S., & Lin, C.-L. (2017). Cha2ds2-vasc score in the prediction of ischemic bowel disease among patients with atrial fibrillation: Insights from a nationwide cohort. *International journal of cardiology*, *235*, 56–60.

Hu, X. (2015). *A copula-based quantile risk measure approach to hedging under regime switching* (Master's thesis). University of Waterloo.

Huang, C.-F., Liu, J.-C., Huang, H.-C., Chuang, S.-Y., Chen, C.-I., & Lin, K.-C. (2017). Longitudinal transition trajectory of gouty arthritis and its comorbidities: A population-based study. *Rheumatology international*, *37*(2), 313–322.

Huang, Z., Dong, W., Wang, F., & Duan, H. (2015). Medical inpatient journey modeling and clustering: A bayesian hidden markov model based approach. *AMIA Annual Symposium Proceedings*, *2015*, 649.

Huang, Z.-y., Lin, S., Long, L.-l., Cao, J.-y., Luo, F., Qin, W.-c., Sun, D.-m., & Gregersen, H. (2020). Predicting the morbidity of chronic obstructive pulmonary disease based on multiple locally weighted linear regression model with k-means clustering. *International Journal of Medical Informatics*, *139*, 104141.

Inouye, T., Toi, S., & Matsumoto, Y. (1995). A new segmentation method of electroencephalograms by use of akaike's information criterion. *Cognitive brain research*, *3*(1), 33–40.

Jaakkola, T. (2001). 10 tutorial on variational approximation methods. *Advanced mean field methods: theory and practice*, 129.

Jaakkola, T. (2000). Tutorial on variational approximation methods. advanced mean field methods: Theory and practice.

Jepsen, P., Vilstrup, H., & Andersen, P. K. (2015). The clinical course of cirrhosis: The importance of multistate models and competing risks analysis. *Hepatology*, *62*(1), 292–302.

Jin, P., Gu, W., Lai, Y., Zheng, W., Zhou, Q., & Wu, X. (2017). The circulating microrna-206 level predicts the severity of pulmonary hypertension in patients with left heart diseases. *Cellular Physiology and Biochemistry*, *41*(6), 2150–2160.

Kaishev, V. K., Dimitrova, D. S., & Haberman, S. (2007). Modelling the joint distribution of competing risks survival times using copula functions. *Insurance: Mathematics and Economics*, *41*(3), 339–361.

Kim, H. T. (2007). Cumulative incidence in competing risks data and competing risks regression analysis. *Clinical cancer research*, *13*(2), 559–565.

Koczwara, B. (2016). *Cancer and chronic conditions*. Springer.

Kristjansson, T. T., Frey, B. J., & Huang, T. S. (2000). Event-coupled hidden markov models. *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)*, *1*, 385–388.

Kubanek, M., Bobulski, J., & Adrjanowicz, L. (2012). Characteristics of the use of coupled hidden markov models for audio-visual polish speech recognition. *Bulletin of the Polish Academy of Sciences. Technical Sciences*, *60*(2), 307–316.

Kwon, B. C., Anand, V., Severson, K. A., Ghosh, S., Sun, Z., Frohnert, B. I., Lundgren, M., & Ng, K. (2021). Dpvis: Visual analytics with hidden markov models for disease progression pathways. *IEEE transactions on visualization and computer graphics*.

Kwon, J., & Murphy, K. (2000). *Modeling freeway traffic with coupled hmms* (tech. rep.). Technical report, Univ. California, Berkeley.

Lagona, F. (2019). Copula-based segmentation of cylindrical time series. *Statistics & Probability Letters*, *144*, 16–22.

Lange, J. M., Gulati, R., Leonardson, A. S., Lin, D. W., Newcomb, L. F., Trock, B. J., Carter, H. B., Cooperberg, M. R., Cowan, J. E., Klotz, L. H., et al. (2018). Estimating and comparing cancer progression risks under varying surveillance protocols. *The annals of applied statistics*, *12*(3), 1773.

Lappenschaar, M., Hommersom, A., Lucas, P. J., Lagro, J., & Visscher, S. (2013). Multilevel bayesian networks for the analysis of hierarchical health care data. *Artificial intelligence in medicine*, *57*(3), 171–183.

Lappenschaar, M., Hommersom, A., Lucas, P. J., Lagro, J., Visscher, S., Korevaar, J. C., & Schellevis, F. G. (2013). Multilevel temporal bayesian networks can model longitudinal change in multimorbidity. *Journal of clinical epidemiology*, *66*(12), 1405–1416.

Lapuyade-Lahorgue, J., Xue, J.-H., & Ruan, S. (2017). Segmenting multi-source images using hidden markov fields with copula-based multivariate statistical distributions. *IEEE Transactions on Image Processing*, *26*(7), 3187–3195.

Lee, C., Yoon, J., & Van Der Schaar, M. (2019). Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data. *IEEE Transactions on Biomedical Engineering*, *67*(1), 122–133.

Leiva-Murillo, J., Rodrguez, A., & Baca-Garca, E. (2011). Visualization and prediction of disease interactions with continuous-time hidden markov models. *NIPS 2011 Workshop on Personalized Medicine*.

Llobet, R., Pérez-Cortés, J. C., Toselli, A. H., & Juan, A. (2007). Computer-aided detection of prostate cancer. *International Journal of Medical Informatics*, *76*(7), 547–556.

Lo, S. M., & Wilke, R. A. (2010). A copula model for dependent competing risks. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *59*(2), 359–376.

Lunn, M., & McNeil, D. (1995). Applying cox regression to competing risks. *Biometrics*, 524–532.

Luo, Y., Stephens, D. A., Verma, A., & Buckeridge, D. L. (2021). Bayesian latent multi-state modeling for nonequidistant longitudinal electronic health records. *Biometrics*, *77*(1), 78–90.

Maag, B., Feuerriegel, S., Kraus, M., Saar-Tsechansky, M., & Züger, T. (2021). Modeling longitudinal dynamics of comorbidities. *Proceedings of the Conference on Health, Inference, and Learning*, 222–235.

Marshall, A. W. (1996). Copulas, marginals, and joint distributions. *Lecture Notes-Monograph Series*, 213–222.

Marshall, A. W., & Olkin, I. (1985). A family of bivariate distributions generated by the bivariate bernoulli distribution. *Journal of the American Statistical Association*, *80*(390), 332–338.

Maté, T., Guaita, R., Pichiule, M., Linares, C., & Dıaz, J. (2010). Short-term effect of fine particulate matter (pm2. 5) on daily mortality due to diseases of the circulatory system in madrid (spain). *Science of the Total Environment*, *408*(23), 5750–5757.

Meenaxi, D. S., & Singh, N. (2018). A reliability model for the progression of chronic heart failure. *International Journal of Applied Engineering Research*, *13*(21), 15351–15355.

Mosteller, F. (1968). Association and estimation in contingency tables. *Journal of the American Statistical Association*, *63*(321), 1–28.

Murray, G. D. (1977). Contribution to discussion of paper by ap dempster, nm laird and db rubin. *J. Roy. Statist. Soc. Ser. B*, *39*, 27–28.

Najjar, A., Reinharz, D., Girouard, C., & Gagné, C. (2018). A two-step approach for mining patient treatment pathways in administrative healthcare databases. *Artificial intelligence in medicine*, *87*, 34–48.

Nelsen, R. B. (1999). An introduction to copulas, volume 139 of. *Lecture Notes in Statistics*.

Nelsen, R. B. (2006). Archimedean copulas. *An Introduction to Copulas*, 109–155.

Ngatchou-Wandji, J., & Bulla, J. (2011). On choosing a mixture model for clustering.

Nielsen, F., & Sun, K. (2016). Guaranteed bounds on the kullback-leibler divergence of univariate mixtures using piecewise log-sum-exp inequalities. *arXiv preprint arXiv:1606.05850*.

Oflaz, Z. N., Yozgatligil, C., & Selcuk-Kestel, A. S. (2019). Aggregate claim estimation using bivariate hidden markov model. *ASTIN Bulletin: The Journal of the IAA*, *49*(1), 189–215.

Ormerod, J. T., & Wand, M. P. (2010). Explaining variational approximations. *The American Statistician*, *64*(2), 140–153.

Parente, J. D., Möller, K., Shaw, G. M., & Chase, J. G. (2018). Hidden markov models for sepsis classification. *IFAC-PapersOnLine*, *51*(27), 110–115.

Pasanisi, A., Fu, S., & Bousquet, N. (2012). Estimating discrete markov models from various incomplete data schemes. *Computational Statistics & Data Analysis*, *56*(9), 2609–2625.

Powell, G., Verma, A., Luo, Y., Stephens, D., & Buckeridge, D. (2019). Modeling chronic obstructive pulmonary disease progression using continuous-time hidden markov models. *Studies in health technology and informatics*, *264*, 920–924.

Putter, H., Fiocco, M., & Geskus, R. B. (2007). Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in medicine*, *26*(11), 2389–2430.

Qian, Z., Alaa, A., Bellot, A., Schaar, M., & Rashbass, J. (2020). Learning dynamic and personalized comorbidity networks from event data using deep diffusion processes. *International Conference on Artificial Intelligence and Statistics*, 3295–3305.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Saif, I., Hassou, N., Allali, K., & Ennaji, M. M. (2018). Effect of hypermethylation in ovarian cancer: Computational approach. *Meta Gene*, *18*, 157–162.

Salles, G. F., Xavier, S. S., Sousa, A. S., Hasslocher-Moreno, A., & Cardoso, C. R. (2004). T-wave axis deviation as an independent predictor of mortality in chronic chagas' disease. *The American journal of cardiology*, *93*(9), 1136–1140.

Saraçoğlu, R. (2012). Hidden markov model-based classification of heart valve disease with pca for dimension reduction. *Engineering Applications of Artificial Intelligence*, *25*(7), 1523–1528.

Satagopan, J., Ben-Porat, L., Berwick, M., Robson, M., Kutler, D., & Auerbach, A. (2004). A note on competing risks in survival data analysis. *British journal of cancer*, *91*(7), 1229–1235.

Satariano, W. A. (2000). Comorbidities and cancer. *Cancer in the elderly* (pp. 486–508). CRC Press.

Saul, L., & Jordan, M. (1995). Exploiting tractable substructures in intractable networks. *Advances in neural information processing systems*, *8*, 486–492.

Saul, L. K., Jaakkola, T., & Jordan, M. I. (1996). Mean field theory for sigmoid belief networks. *Journal of artificial intelligence research*, *4*, 61–76.

Schellevis, F. G., van der Velden, J., van de Lisdonk, E., Van Eijk, J. T. M., & van Weel, C. v. (1993). Comorbidity of chronic diseases in general practice. *Journal of clinical epidemiology*, *46*(5), 469–473.

Schmidt, S., Norman, M., Misselwitz, B., Piedvache, A., Huusom, L. D., Varendi, H., Barros, H., Cammu, H., Blondel, B., Dudenhausen, J., et al. (2019). Mode of delivery and mortality and morbidity for very preterm singleton infants in a breech position: A european cohort study. *European Journal of Obstetrics & Gynecology and Reproductive Biology*, *234*, 96–102.

Schweizer, B., Wolff, E. F. et al. (1981). On nonparametric measures of dependence for random variables. *Annals of Statistics*, *9*(4), 879–885.

Scott, S. L. (2002). Bayesian methods for hidden markov models: Recursive computing in the 21st century. *Journal of the American statistical Association*, *97*(457), 337–351.

Shih, H.-C., Chou, P., Liu, C.-M., & Tung, T.-H. (2007). Estimation of progression of multi-state chronic disease using the markov model and prevalence pool concept. *BMC Medical Informatics and Decision Making*, *7*(1), 1–12.

Siu, T.-K., Ching, W.-K., Fung, S. E., & Ng¶, M. K. (2005). On a multivariate markov chain model for credit risk measurement. *Quantitative Finance*, *5*(6), 543–556.

Sklar, M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, *8*, 229–231.

Southern, D. A., Faris, P. D., Brant, R., Galbraith, P. D., Norris, C. M., Knudtson, M. L., Ghali, W. A., Investigators, A., et al. (2006). Kaplan–meier methods yielded misleading results in competing risk scenarios. *Journal of clinical epidemiology*, *59*(10), 1110–1114.

Souza, A. C. J., Salles, G., Hasslocher-Moreno, A. M., Sousa, A. S., Brasil, P. E. A. A., Saraiva, R. M., & Xavier, S. S. (2015). Development of a risk score to predict sudden death in patients with chaga's heart disease. *International journal of cardiology*, *187*, 700–704.

Spedicato, G. A., Kang, T. S., Yalamanchi, S. B., Yadav, D., & Cordón, I. (2016). The markovchain package: A package for easily handling discrete markov chains in r. *Accessed Dec*.

Stöber, J., Hong, H. G., Czado, C., & Ghosh, P. (2015). Comorbidity of chronic diseases in the elderly: Patterns identified by a copula design for mixed responses. *Computational Statistics & Data Analysis*, *88*, 28–39.

Suciu, L., Cristescu, C., Suciu, M., Voicu, M., Buda, V., & Tomescu, M. (2019). Evaluation of morbidity and life expectancy based on the markov prediction model, in a group of patients with essential hypertension in romania. *Atherosclerosis*, *287*, e158–e159.

Sun, F., & Jiang, Y. (2018). A hidden resource in wireless channel capacity: Dependence control in action. *arXiv preprint arXiv:1805.00812*.

Tengnah, M. A. J., Sooklall, R., & Nagowah, S. D. (2019). A predictive model for hypertension diagnosis using machine learning techniques. *Telemedicine technologies* (pp. 139–152). Elsevier.

Thiyagaraja, S. R., Dantu, R., Shrestha, P. L., Chitnis, A., Thompson, M. A., Anumandla, P. T., Sarma, T., & Dantu, S. (2018). A novel heart-mobile interface for detection and classification of heart sounds. *Biomedical Signal Processing and Control*, *45*, 313–324.

Thomas, M., & Joy, A. T. (2006). *Elements of information theory*. Wiley-Interscience.

Uğuz, H., Arslan, A., Saraçoğlu, R., & Türkoğlu, İ. (2008). Detection of heart valve diseases by using fuzzy discrete hidden markov model. *Expert Systems with Applications*, *34*(4), 2799–2811.

Vachiéry, J.-L., Adir, Y., Barberà, J. A., Champion, H., Coghlan, J. G., Cottin, V., De Marco, T., Galiè, N., Ghio, S., Gibbs, J. S. R., et al. (2013). Pulmonary hypertension due to left heart diseases. *Journal of the American College of Cardiology*, *62*(25S), D100–D108.

Valderas, J. M., Starfield, B., Sibbald, B., Salisbury, C., & Roland, M. (2009). Defining comorbidity: Implications for understanding health and health services. *The Annals of Family Medicine*, *7*(4), 357–363.

Valdez-Ortiz, R., Navarro-Reynoso, F., Olvera-Soto, M. G., Martin-Alemañy, G., Rodrıguez-Matıas, A., Hernández-Arciniega, C. R., Cortes-Pérez, M., Chávez-López, E., Garcıa-Villalobos, G., Hinojosa-Heredia, H., et al. (2018). Mortality in patients with chronic renal disease without health insurance in mexico: Opportunities for a national renal health policy. *Kidney international reports*, *3*(5), 1171–1182.

Varadhan, R., Weiss, C. O., Segal, J. B., Wu, A. W., Scharfstein, D., & Boyd, C. (2010). Evaluating health outcomes in the presence of competing risks: A review of statistical methods and clinical applications. *Medical care*, S96–S105.

Violán, C., Fernández-Bertolın, S., Guisado-Clavero, M., Foguet-Boreu, Q., Valderas, J. M., Vidal Manzano, J., Roso-Llorach, A., & Cabrera-Bean, M. (2020). Five-year trajectories of multimorbidity patterns in an elderly mediterranean population using hidden markov models. *Scientific reports*, *10*(1), 1–11.

Wainwright, M. J., & Jordan, M. I. (2008). *Graphical models, exponential families, and variational inference*. Now Publishers Inc.

Wang, X., Lebarbier, E., Aubert, J., & Robin, S. (2019). Variational inference for coupled hidden markov models applied to the joint detection of copy number variations. *The international journal of biostatistics*, *15*(1).

Wang, Y., & Pham, H. (2011). Modeling the dependent competing risks with multiple degradation processes and random shock using time-varying copulas. *IEEE Transactions on Reliability*, *61*(1), 13–22.

Wang, Y., Zhou, X., Mascolo, C., Noulas, A., Xie, X., & Liu, Q. (2018). Predicting the spatio-temporal evolution of chronic diseases in population with human mobility data.

Waterhouse, S. R. (1998). *Classification and regression using mixtures of experts* (Doctoral dissertation). Citeseer.

Wen, F.-H., Chen, J.-S., Chou, W.-C., Hsieh, C.-H., Chang, W.-C., Hou, M.-M., & Tang, S. T. (2018). Distinct patterns of conjoint symptom distress and functional impairment in the last year of life predict terminally ill cancer patients' survival. *Journal of pain and symptom management*, *55*(6), 1443–1451.

Wong, W. K., Furst, D. E., Clements, P. J., & Streisand, J. B. (2007). Assessing disease progression using a composite endpoint. *Statistical methods in medical research*, *16*(1), 31–49.

Wu, C. J. (1983). On the convergence properties of the em algorithm. *The Annals of statistics*, 95–103.

Wu, J., Xun, P., Tang, Q., Cai, W., & He, K. (2017). Circulating magnesium levels and incidence of coronary heart diseases, hypertension, and type 2 diabetes mellitus: A meta-analysis of prospective cohort studies. *Nutrition journal*, *16*(1), 1–13.

Yen, A. M.-F., & Chen, H.-H. (2013). Stochastic models for multiple pathways of temporal natural history on co-morbidity of chronic disease. *Computational Statistics & Data Analysis*, *57*(1), 570–588.

Yule, G. U. (1912). On the methods of measuring association between two attributes. *Journal of the Royal Statistical Society*, *75*(6), 579–652.

Zhang, Y.-F., Zhang, Q.-F., & Yu, R.-H. (2010). Markov property of markov chains and its test. *2010 International Conference on Machine Learning and Cybernetics*, *4*, 1864–1867.

Zhong, S., & Ghosh, J. (2001). A new formulation of coupled hidden markov models. *Dept. Elect. Comput. Eng., Univ. Austin, Austin, TX, USA*.

Zucchini, W., & MacDonald, I. L. (2009). *Hidden markov models for time series: An introduction using r*. Chapman; Hall/CRC.

Zucchini, W., MacDonald, I. L., & Langrock, R. (2017). *Hidden markov models for time series: An introduction using r*. Chapman; Hall/CRC.

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Oflaz, Zarina

EDUCATION

Degree	Institution	Year of Graduation
M.S.	Middle East Technical University Department of Statistics	2016
B.S.	L. N. Gumilyov Eurasian National University Department of Mathematics	2014

PROFESSIONAL EXPERIENCE

Year	Place	Enrollment
2017-Present	KTO Karatay University Department of Insurance and Social Security	Lecturer (Öğretim Görevlisi)
2020-Present	KTO Karatay University	Coordinator of the Artificial Intelligence and Data Science Working Group
July 2018-March 2019	Simsoft Computer Technology	Academic Advisor

Foreign Languages

Native Kazakh, Native Russian, Advanced English, Advanced Turkish

PUBLICATIONS

Research Articles

Oflaz, K., Oflaz, Z., Ozaytekin, I., Dincer, K. and Barstugan, R., 2021. Time and volume-ratio effect on reusable polybenzoxazole nanofiber oil sorption capacity investigated via machine learning. Journal of Applied Polymer Science, 138(30), p.50732, ***(SCI, SCI Exp)***.

Oflaz, Z.N., Yozgatligil, C. and Selcuk-Kestel, A.S., 2019. Aggregate Claim Estimation Using Bivariate Hidden Markov Model. ASTIN Bulletin: The Journal of the IAA, 49(1), pp.189-215, ***(SSCI, SCI Exp)***.

Oflaz, Z., 2017. Structural Break, Nonlinearity and the Hysteresis hypothesis: Evidence from new unit root tests. Econometrics Letters, 4(2), pp.1-16.

International Conferences

Oflaz, Z., Yozgatlıgil, C. and Kestel, S.A., 2019. Ischemic Heart Disease Morbidity Rates Estimation using Hidden Markov Model Regression, 23rd International Congress on Insurance: Mathematics and Economics.

Oflaz, Z., Yozgatlıgil, C. and Kestel, S.A., 2018. Estimation Of Claim Amounts Using Bivariate Hidden Markov Models, 31st International Congress of Actuaries.