MODELING STATE DURATION AND EMISSION DEPENDENCE IN HIDDEN MARKOV AND HIDDEN SEMI-MARKOV MODELS

A Dissertation presented to
the Faculty of the Graduate School
at the University of Missouri

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

by

SHIRLEY ROJAS SALAZAR

Drs. Erin M. Schliep and Christopher K. Wikle, Dissertation Supervisors

JULY 2022

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled:

MODELING STATE DURATION AND EMISSION DEPENDENCE IN
HIDDEN MARKOV AND HIDDEN SEMI-MARKOV MODELS

presented by Shirley Rojas Salazar,
a candidate for the degree of Doctor of Philosophy and hereby certify that, in their opinion, it is worthy of acceptance.

Dr. Erin M. Schliep

Dr. Christopher K. Wikle

Dr. Scott H. Holan

Dr. Athanasios C. Micheas

Dr. Edgar C. Merkle

ACKNOWLEDGMENTS

I would like to thank my advisors Dr. Erin Schliep and Dr. Chris Wikle for all their support, patience and valuable time, I will be forever grateful. I would also like to thank the professors on my committee for their comments and insights.

I would like to thank my colleagues and friends at the Department of Statistics and at the Office of International Affairs and External Cooperation at the University of Costa Rica for all their support and encouragement. Also, I would like to thank the US Embassy in Costa Rica and the Fulbright-LASPAU program for allowing me to pursue graduate studies.

Lastly, I would like to thank my husband and my family for all their love and support throughout these years.

TABLE OF CONTENTS

CHAPTER

APPENDIX

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

Modeling State Duration and Emission Dependence in Hidden Markov and Hidden Semi-Markov Models

Shirley Rojas Salazar

Drs. Erin M. Schliep and Christopher K. Wikle, Dissertation Supervisors

ABSTRACT

Hidden Markov models (HMM) are composed of a latent state sequence and an observation sequence conditionally independent on the states, which follows an emission distribution. Hidden semi-Markov models (HSMM) extend the HMM by explicitly modeling the duration in the states. This dissertation expands the HSMM by introducing non-homogeneity in the duration model, with duration parameters defined as functions of time-varying covariates, which has not been considered to date. This model is applied to high-frequency environmental data. The variable transition HMM (VTHMM) also expands the HMM by considering the duration in the state transition probabilities. We present a VTHMM for team sports data to obtain inference on the dynamic network of players in a game, and model high temporal resolution player location data. Lastly, the conditional independence assumption in the emission distribution can be violated, in particular with high-frequency data. We propose two novel approaches to address the conditional dependence, by introducing data subsampling in the MCMC sampling algorithm for parameter inference in HMMs and HSMMs, and by considering basis function expansions in the emission distribution.

Chapter 1

Introduction

A hidden Markov model (HMM) is composed of two stochastic processes, a sequence of unobserved discrete states and another set of observable random variables that are assumed conditionally independent given the state at each observed time point (Yu, 2016; Rabiner, 1989). There is a probability defined for the transition from one state to another, which is conditional on the current state. The probability of self-transitioning (i.e., remaining in the same state) is non-zero, which implies that the time spent in each state follows a geometric distribution (Yu, 2010). The duration has to be explicitly modeled in processes for which this distributional assumption may not be realistic. This model extension corresponds to the hidden semi-Markov model (HSMM) (Yu, 2016).

HMMs and HSMMs have been applied in numerous areas. There is an increasing number of high-frequency data sets being collected from new kinds of instruments (e.g., wearable devices, health monitoring, ecological tracking, and behavior sampling (acceleration)). HMMs and HSMMs are an appropriate approach for analyzing these high temporal resolution data when the observations are considered to be related to a latent or unobserved process. Among the first areas to take advantage of these models was speech recognition, but they have since been applied in other areas such as activity recognition (Duong et al., 2005; Natarajan and Nevatia, 2007), event recognition in videos (Motoi et al., 2012; Xie et al., 2004), animal movement (Scharf et al., 2016; Leos-Barajas et al., 2017), and animal behavior (Ruiz-Suarez et al., 2022). For a more comprehensive list of applications of these models see Mor et al. (2021), and Chapter 9 in Yu (2016).

Consider environmental time series, which are increasingly being measured at high temporal resolutions. For example, Woillez et al. (2016) applied an HMM to high-resolution temperature and depth data for geolocation and tracking of pelagic fish. Rousseeuw et al. (2015) analyzed data measured every 20 minutes in a marine station and applied a hybrid HMM to model phytoplankton dynamics. Stoner and Economou (2020) developed an HMM to analyze sub-daily rainfall data, and applied it to a dataset of hourly time series of rainfall in Exeter, UK. An example of the use of an HSMM for high-frequency data corresponds to the study by Sansom and Thompson (2008), which applied an HSMM using a high temporal resolution rainfall dataset collected in New Zealand to study the spatial and temporal variation of rainfall.

Time-series data from wearable devices are another example of high temporal resolution data that have been analyzed with HMMs and HSMMs. Wearable devices are accessories, or smart clothes, worn on or near the body (Motti, 2020) that provide users with information regarding, among other things, their health (Wu and Haick, 2018) and physical activity (Li et al., 2016). These devices have been used in contexts such as health or sports to monitor the performance of athletes during training or competition. The data collected by wearing devices are obtained at an extremely high-frequency for a large number of variables, including acceleration, velocity, and heart rate (Kos and Kramberger, 2017; Motti, 2020). There are several applications of HMMs to these type of data. One such example is the activity recognition analysis by Thomas et al. (2010) who apply both an HMM and a semi-Markov model (SMM) to acceleration data collected at 100 Hz in training sessions of elite swimmers. In another example, Huang et al. (2018) applied an HMM to physical activity data from individuals collected every 5 minutes over several days to obtain inference on activity states.

The chapters of this dissertation present HMMs and HSMMs that deal with high-frequency data in the environmental and sports context. In the sections of this chapter we give a brief overview of the HMM and introduce various ways of modeling the state duration, which includes the use of HSMMs. We also provide insight into the violation of the assumption of conditional independence in the emission distribution when we analyze high-frequency data. Lastly, an overview of the chapters of this dissertation is provided.

1.1 A Brief Overview of Hidden Markov Models

In this section we briefly describe HMMs, variable transition HMMs (VTHMMs) and HSMMs.

1.1.1 HMM

The first of the two stochastic processes in an HMM is the unobserved state sequence, denoted as $\mathbf{s} = (s_1, \ldots, s_T)'$, where $s_t \in \{1, 2, \ldots, S\}$ for time $t = 1, \ldots, T$, and S is the total number of unique states. In addition, we have an observation sequence denoted as $\mathbf{y} = (y_1, \ldots, y_T)'$. In a discrete-time HMM, the probabilities of transitioning from one state to another are denoted as $p_{j,k}$, with $p_{j,k} = P\left(s_{t+1} = k \mid s_t = j\right)$, where $1 \leq j, k \leq S$, and $\sum_{k=1}^{S} p_{j,k} = 1$. These entries make up the transition probability matrix, $\mathbf{P}$. The observations have a distribution $y_t \sim f\left(\theta_{s_t}\right)$ with state-dependent parameters $\theta_{s_t} \in \{\theta_1, \cdots, \theta_S\}$. Figure 1.1A illustrates the HMM, where one observation is emitted by each state, and at each time point there is a transition in the state sequence.

In HMMs, the duration in each state follows a geometric distribution and is not explicitly modeled (Yu, 2010). The duration can be considered in HMMs with the following general approaches: HSMMs, variable transition HMMs (VTHMMs), and expanded state HMMs (Johnson, 2005).

1.1.2 HSMM

Figure 1.1C illustrates the sequences in an HSMM. Here, there is still a sequence of states, but instead of only one observation emitted per state, we have a group of observations emitted in each state. Each of the groups of observations from each state can be viewed as segments, labeled as q, with $q = 1, \ldots, Q$. The number of observations in each of the Q segments defines the duration in the state. Here we consider the state sequence as $\mathbf{s} = (s_1, \ldots, s_Q)'$, where $s_q \in \{1, 2, \ldots, S\}$. The duration in each segment is denoted as τ_q. The distribution of these durations is defined generally as $\tau_q \sim h\left(\phi_{s_q}\right)$ with state-dependent parameters ϕ_{s_q}.

The state-specific parameters ϕ_{s_q} in the duration distribution are assumed to be constant over time. However, with this approach we cannot capture the temporal variation in the duration in each state. There is a need to introduce non-homogeneity explicitly in the state duration modeling. Although non-homogeneity has been defined in the emission distribution, and in the transition probabilities, to date it has not been explicitly presented in HSMMs.

1.1.3 VTHMM

In the VTHMM, as with the HMM, one observation is emitted by each state (Figure 1.1B). However, the transition probabilities $p_{j,k}$ are not constant in time. Rather, they are a function of the duration in the state at time t, and are denoted as $p_{j,k}(d_t) = P\left(s_{t+1} = k \mid s_t = j, d_t\right)$. This model, with duration dependent state transition probabilities, was introduced by Vaseghi (1995) and Ramesh and Wilpon (1992). It is also presented in Azimi et al. (2005) with an improvement in computational efficiency. In a non-homogeneous or inhomogeneous HMM, the transition probabilities are not constant in time. As such, the VTHMM can be regarded as a type of non-homogeneous HMM. In Yu (2016), it is referred to as a special case of a HSMM, where transition to the same state (self-transition) is allowed.

1.2 Conditional independence assumption

An observation in an HMM (or HSMM) is assumed to be independent of previous observations and states conditioned on the current state. Thus, it is conditionally independent given the current state (Pohle et al., 2017; Yu, 2016). In some observation sequences, especially with high-frequency data, it is unlikely that an observation at time t is not related to the observation at $t-1$ or previous observations even after conditioning on the latent state at time t.

A common approach to deal with this conditional independence violation is the use of Markov-switching regression models Hamilton (2010), which incorporate an autoregressive structure into the emission distribution. There are several applications of HMMs and HSMMs that use an AR or a VAR structure (Ruiz-Suarez et al., 2022; Langrock et al., 2017; Xu and Liu, 2021, e.g.). Other approaches, which are computationally more challenging, include the use of neural networks (Dai et al., 2017; Ravuri and Wegmann, 2016).

There are other potential approaches that can be incorporated in HMMs and HSMMs to account for the conditional dependence. These include the use of basis functions, or subsampling the data within the sampling algorithm for parameter inference. Basis function expansions are used to model temporal and spatial dependence Wikle et al. (2019), and can be easily integrated into the emission distribution as nonparametric temporal random effects. Langrock et al. (2015) introduced basis functions (penalized splines) to model the observation sequence more flexibly, but basis functions have not been utilized with the purpose of accounting for residual dependence in the emission distribution. On the other hand, data subsampling has been used to reduce computational time in MCMC sampling algorithms (Maclaurin and Adams, 2015; Quiroz et al., 2019), but it has not been explored as an option to reduce the impact of autocorrelation in the observation sequence.

1.3 Overview of the Chapters

The chapters of this dissertation consider duration modeling and solutions to conditional independence violation in HMMs and HSMMs, with applications to high temporal resolution data. In the second chapter we model the duration explicitly in an HSMM with a flexible approach that defines time-varying state duration parameters. The third chapter considers an implicit modeling of the duration through a VTHMM in the context of networks in team sports. In the fourth chapter, we propose new ways for dealing with conditional dependence in the emission distribution and evaluate their impact through simulation and a real-world data set.

In Chapter 2 we model the duration in HSMMs by defining the duration distribution parameter as a function of time-varying covariates. This is a novel approach since the duration parameters in HSMMs have in the past been assumed to be constant over time. To this end, we provide a flexible method for situations in which this parameter changes over time. We apply the model to environmental high-frequency data that is commonly used as an indicator of cyanobacteria in a lake.

In Chapter 3 we apply a VTHMM in the context of a dynamic network to implicitly model the duration in a state. We define the state transition probabilities as a function of time-varying covariates, and among those covariates we include the duration in the state. We model the locations of players in the context of team sports, specifically soccer. The data come from wearable devices that record the locations at a high temporal resolution (every second). To model these locations we adapt a model originally proposed for animal movement. We assume a two state model where the states define the connection or non-connection between each pair of players. Our model provides a non-deterministic way of defining the adjacency matrix in a team sports dynamic network, which has not been considered in this context. Our approach represents a contribution to the methods available to model team sports data since it provides a framework to model player movement in a game, assuming there is an underlying network and considering the duration of a connection (or non-connection) and other time-varying covariates to explain the probability of having (or not having) a connection throughout the game.

In Chapter 4 we explore how inference of the emission distribution parameters is affected by the presence of conditional dependence in the observation sequence of HMM and HSMMs. We introduce two novel approaches to tackle the conditional independence violation in the emission distribution. The models proposed are applied to the same environmental dataset used in Chapter 2, to illustrate how the analysis of these types of datasets can account for the conditional dependence of high temporal resolution observations. The approaches we propose are not constrained to only tackle dependence with an autoregressive structure, unlike current methods such as Markov-switching autoregressive models, and they also provide a computationally more efficient alternative to current neural network approaches.

The concluding chapter provides a brief overview and discussion of future work to extend the models presented in this dissertation. The future work includes the application of the model developed in Chapter 2 with other duration distributions, such as a geometric mixture, and further exploring the criteria for number of states selection. Also, future work involves the extension of the model in Chapter 3 to consider the graph-coupled hidden Markov model (Dong et al., 2012).

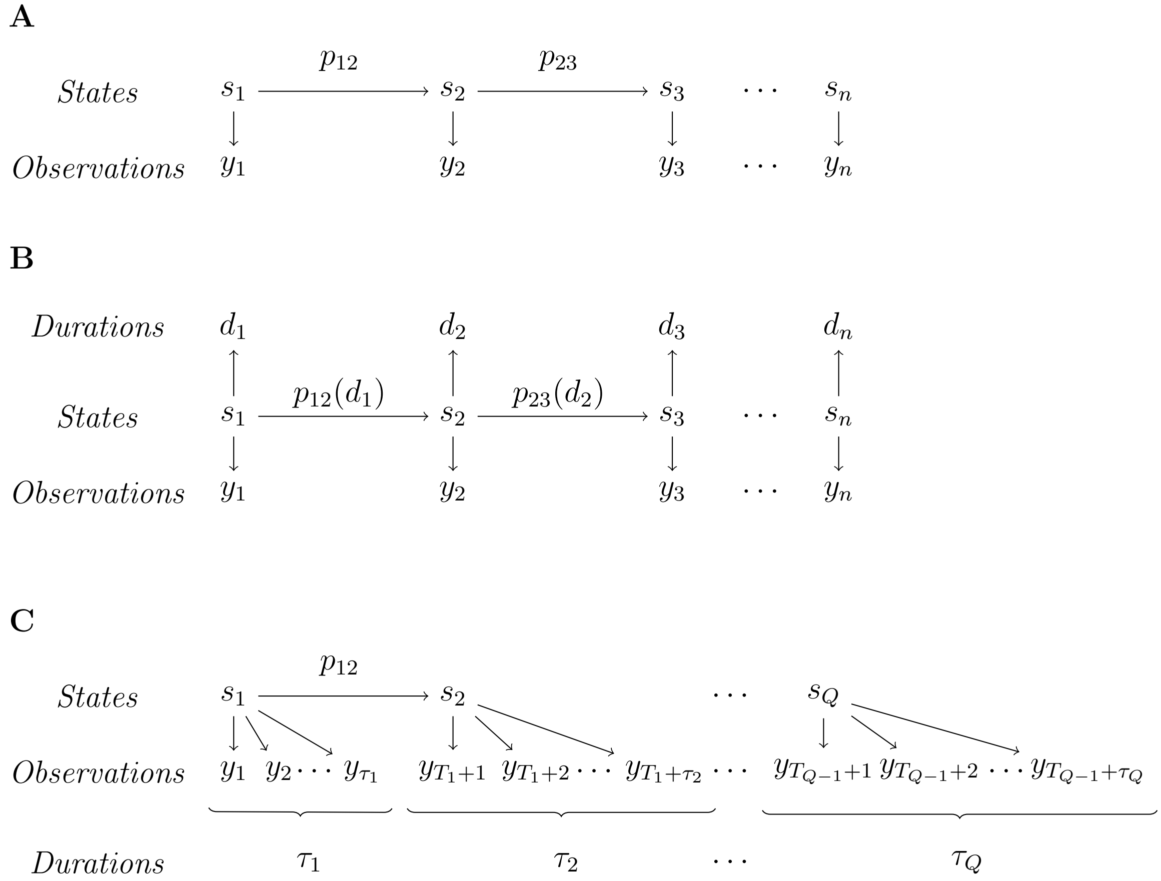

Figure 1.1: State and observation sequences. Panel **A**. HMM: One observation is emitted by each state in the sequence. Panel **B**. VTHMM: One observation is emitted by each state in the sequence and the transition probability is a function of the duration. Panel **C**. HSMM: Several observations are emitted by each state, the number is determined by the duration in the state.

Chapter 2

A Bayesian Hidden Semi-Markov Model with Covariate-Dependent State Duration Parameters for High-Frequency Environmental Data

2.1 Introduction

Environmental time series data are measured at different frequencies, with a general increasing trend towards high temporal resolutions. These high-frequency data can be studied using a wide range of analyses. For example, Li and Sun (2021) presented a stochastic precipitation generator and applied it to high-frequency (30-second) rainfall data. Lin et al. (2020) used high-frequency (1-second) supervisory control and data acquisition (SCADA) wind power data to predict power. They utilized a deep neural network to predict the wind power and incorporated the isolation forest method to identify anomalies in the data points. High-frequency time series data are also collected in lakes and have been analyzed with different models and statistical approaches. Carpenter et al. (2020) studied the dynamics of cyanobacteria in Lake Mendota (Wisconsin, USA) using a drift-diffusion-jump model. The model was applied to phycocyanin concentrations measured every minute for the years 2008 through 2018. They found that for each of the years studied, the concentration of phycocyanin can be summarized with two stable states. Coloso et al. (2011) studied drivers of lake ecosystem metabolism by fitting multiple linear regression models to high-frequency data from two temperate lakes. For each of the dependent variables, gross primary production, respiration, and net ecosystem production, different important drivers were identified, including temperature, wind speed, photosynthetically active radiation, among others.

Hidden Markov and hidden semi-Markov models provide an alternative approach for analyzing high-frequency environmental data. A hidden Markov model (HMM) consists of a sequence of unobserved discrete states and another set of observable random variables that are assumed conditionally independent given the state at each observed time point (Rabiner, 1989). The transition from one state to another depends on a transition probability, which is defined conditionally on the current state, and where the probability of self-transitioning (i.e., remaining in the same state) is non-zero. This non-zero probability implies that the time spent in each state follows a geometric distribution (Yu, 2010). However, this distributional assumption may not be realistic for some processes, making it necessary to additionally model the state duration. This model extension defines the hidden semi-Markov model (HSMM) (Yu, 2016).

Environmental data have been modeled with HMMs and HSMMs. For example, Rousseeuw et al. (2015) applied a hybrid HMM to model phytoplankton dynamics using data measured every 20 minutes in a marine station. They incorporated a spectral clustering method into their HMM modeling in order to build a fully unsupervised HMM. Stoner and Economou (2020) developed an HMM to analyze sub-daily rainfall data, and, through the use of simulations, were able to show that their model can capture characteristics such as long dry periods or seasonal variation. They also applied the model to a real dataset of hourly time series of rainfall in Exeter, UK. Similar types of data have been analyzed with semi-Markov and HSMMs. For example, King and Langrock (2016) present an extension of the Arnason-Schwarz model where they define a semi-Markov model for the state process, and apply the model to capture-recapture data of house finches. Sansom and Thompson (2008) studied the spatial and temporal variation of rainfall with an HSMM using a high temporal resolution rainfall dataset collected in New Zealand.

When the duration in an HSMM is modeled with a Poisson distribution, the duration parameter, which can be different for each hidden state, is assumed to be constant in time. This assumption, however, might not be reasonable in all cases. If we consider, for example, hourly rainfall data observed over the course of a year, and we model it with two different states representing wet and dry episodes, we would expect the length of time of these episodes to be different depending on the time of year due to seasonal rainfall patterns (e.g., monsoon season).

To capture this temporal variation in the duration in each state, we extend the HSMM by modeling the duration parameters as a function of time varying covariates.

This enables the identification of factors associated with the time spent in the different states. For example, when there is a state transition, the duration parameter for the new state could be modeled as a function of covariates observed in the period leading up to the transition, or the value of the covariate at the moment right before the switch. The functional relationship between covariates and the parameters of the duration distribution could be state-specific, and modeling these relationships can provide important inference with regard to their extent and direction. Importantly, inference is not obtained at the high-frequency level at which the data are collected, but rather in terms of the duration intervals.

The goal of this chapter is to develop an HSMM with time-varying duration parameters that are dependent on covariates for studying high-frequency environmental data. Specifically, we model high-frequency concentrations of phycocyanin, an estimate of presence or relative abundance of cyanobacteria (blue-green algae), in Lake Mendota, Wisconsin. We use covariates to explain the variation in the duration in each state and obtain inference on important characteristics. Previous approaches using HMMs or HSMMs have included covariates in the observation model or in the specification of transition probabilities (e.g. Koki et al., 2020; Economou et al., 2014; Titman and Sharples, 2010), but the inclusion of covariates in the model for state durations has not been considered explicitly. By modeling the state transition probabilities with covariates, the durations are modeled but in an implicit way, such as in Stoner and Economou (2020).

Understanding the temporal variation in cyanobacteria concentration as well as drivers of this variation in urban lakes is important to public health. There has been an increase in cyanobacterial blooms (Huisman et al., 2018) and these high concentrations of blue-green algae can produce toxins that are linked to illness in humans and animals (Paerl and Huisman, 2008; Falconer, 1999; Havens, 2008).

The remainder of this chapter is outlined as follows. In Section 2.2 we present the phycocyanin data used in this analysis. Section 2.3 provides a description of the HSMM, which defines the duration distribution parameter as a function of covariates. The results of the model applied to the phycocyanin dataset are presented in Section 2.4. Lastly, Section 2.5 provides a discussion and conclusion, as well as future extensions.

2.2 High-frequency Lake Mendota data

The high-frequency environmental data in this application correspond to measurements from Lake Mendota, Wisconsin, in 2018. The data can be found in the North Temperate Lakes Long-Term Ecological Research program database (Lead PI et al., 2020). Sensors in an instrumented buoy located in the lake recorded measurements every minute. In 2018, the buoy recorded observations from April 11 to November 15. The dataset consists of several other variables including weather conditions (air temperature, relative humidity, wind speed, and wind direction), and lake characteristics (such as chlorophyll, photosyntetically-active radiation, dissolved oxygen, etc).

Phycocyanin is a pigment of cyanobacteria, and provides an estimate or diagnostic of its presence and concentration (Carpenter et al., 2020). Given that high concentrations of cyanobacteria are a major public health concern, particularly in urban lakes such as Lake Mendota, it is essential to understand the temporal variation in concentration levels as well as possible environmental drivers of this variation. For each year of their study, Carpenter et al. (2020) modeled the phycocyanin concentrations and identified two regimes of low and high phycocyanin, as well as abrupt transitions between the states. The objective of our analysis is to extend on their work by describing the latent states of cyanobacteria as captured by phycocyanin concentration, the duration in each of those states, and the covariates associated with that duration.

Following the methods in Carpenter et al. (2020), we consider the standardized levels of phycocyanin as our observation sequence. We first compute the maximum hourly measurement, which results in a total of 5232 observations (Figure 2.1). These maximum values of phycocyanin are measured in relative fluorescent units (RFU). They are standardized by being log transformed ($\log_{10}$), centered and scaled, and fitted using a dynamic linear model. See the supplementary information in Carpenter et al. (2020) for more details.

In a study done in Lake Mendota, Soranno (1997) found that weather variables can impact the dynamics of algae at finer time scales. Considering this, the variables we examine as possible covariates to capture the time variation in the state durations are hourly average air temperature, wind speed, relative humidity and photosynthetically-active radiation (PAR). During the time period April 11 to November 15, 2018, the air temperature, measured in °C, ranged from -7 to 33, with a mean of 17. The average and maximum wind speed were 3.8 and 15 m/s respectively. Relative humidity ranged from 21% to 100% with an average of 74%. Radiation was measured with a surface sensor in $\mu\mathrm{mol}\,\mathrm{m}^{-2}\,\mathrm{s}^{-1}$ and it ranged from 0 to nearly 2000.

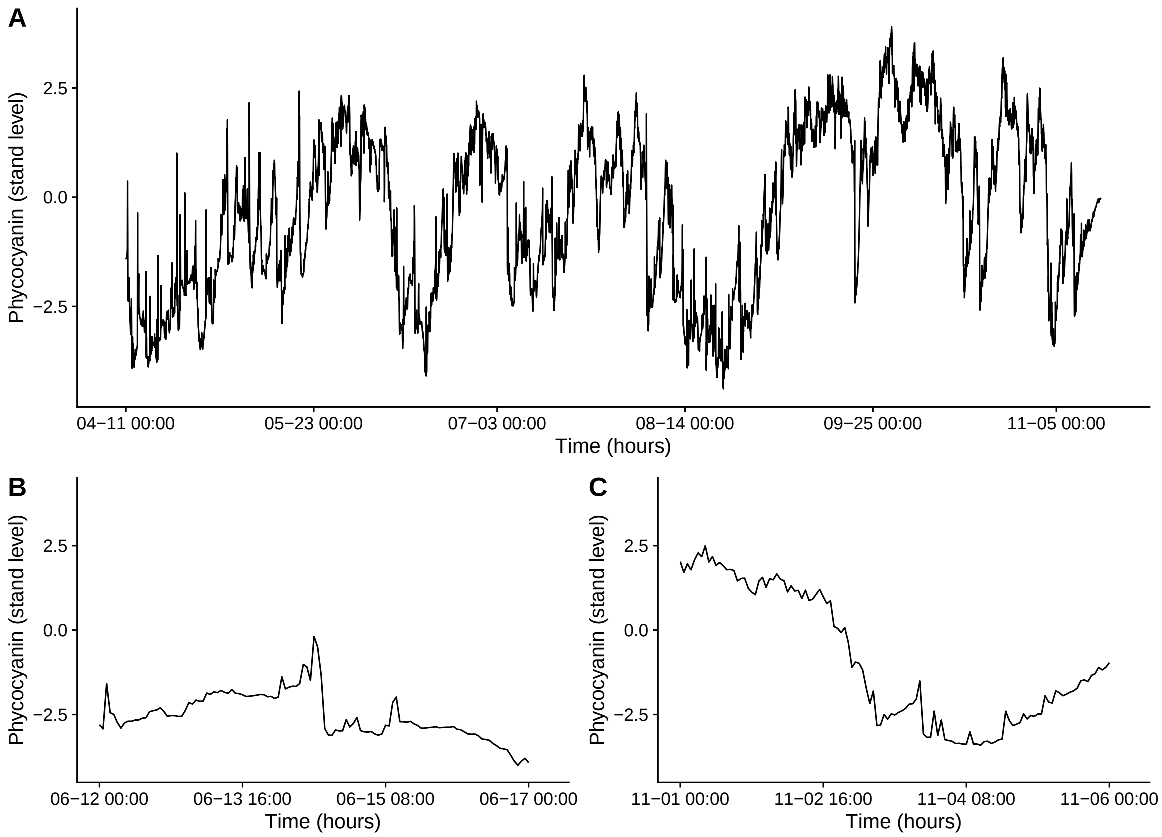

Figure 2.1: Phycocyanin standardized levels in Lake Mendota, 2018. Panel **A**. Full period (mid April to mid November). Panels **B** and **C** show the observations in more detail for two 5-day periods in June and November, respectively

2.3 Hidden semi-Markov model with covariate dependent duration parameters

We begin by specifying the HMM and important notation. Then, we extend the HMM to the HSMM, develop the state-specific duration model as a function of covariates, and describe the methods for Bayesian inference.

2.3.1 Hidden Markov model and notation

An HMM includes two stochastic processes, one represents a Markov chain of states that are hidden, and the other generates a sequence of observations that are influenced by the unobservable states (Rabiner, 1989; Yu, 2010). In a discrete-time HMM, the sequence of observations from time 1 to n can be denoted as $\mathbf{y} = (y_1, \dots, y_n)'$. The corresponding sequence of unobserved states is denoted as $\mathbf{S} = (S_1, \dots, S_n)'$, where $S_i \in \{1, 2, \dots, M\}$, $i = 1, \dots, n$, and M is the total number of unique states. The state at time 1 has a distribution defined by $\rho_j = P(S_1 = j)$, $j = 1, \dots, M$. The transition to the next state, S_2, is conditional on state S_1 according to the Markov property. In general, the transition probability matrix $\mathbf{P}$ provides the probabilities of transitioning from one state to another when the state space is discrete and constant in time. The matrix $\mathbf{P}$, has entries $p_{j,k}$, with $p_{j,k} = P(S_{i+1} = k \mid S_i = j)$, where $1 \leq j, k \leq M$, and $\sum_{k=1}^{M} p_{j,k} = 1$.

Observations are emitted by each of the states in the hidden sequence (Figure 2.2A) following a state-dependent probability distribution $f(\mathbf{y}|\boldsymbol{\theta}, \mathbf{S})$. Assuming the observation distribution is Gaussian, the parameters $\boldsymbol{\theta}$ correspond to the mean and variance for each state: $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_M)'$ and $\boldsymbol{\sigma}^2 = (\sigma_1^2, \cdots, \sigma_M^2)'$. The joint likelihood of the observations can be written as:

$$L(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{S}) = \prod_{i=1}^{n} f(y_i|\mu_{S_i}, \sigma_{S_i}^2),$$

and the likelihood of the Markov chain is:

$$L(\mathbf{S} \mid \boldsymbol{\rho}, \mathbf{P}) = \rho_{S_1} \prod_{i=1}^{n-1} p_{S_i, S_{i+1}}.$$

The complete likelihood of the Markov model is the joint likelihood of observations and states: $L(\mathbf{y}, \mathbf{S} \mid \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\rho}, \mathbf{P}) = L(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{S}) \times L(\mathbf{S} \mid \boldsymbol{\rho}, \mathbf{P})$. In summary, an HMM with M states and n observations has a set of model parameters that includes the emission distribution parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}^2$, the initial distribution probabilities $\boldsymbol{\rho}$, and the transition probability matrix $\mathbf{P}$.

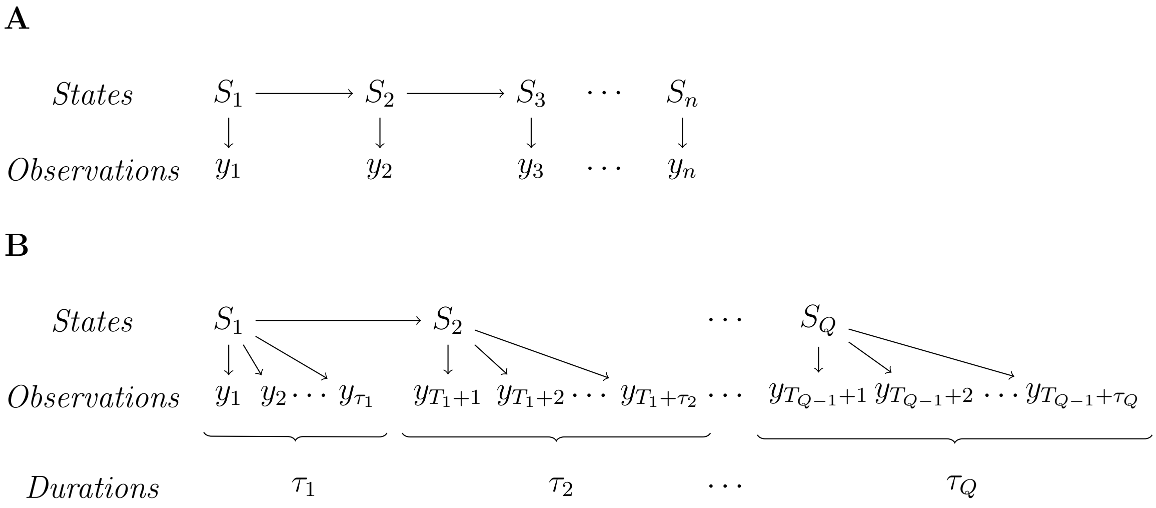

Figure 2.2: State and observation sequences. Panel **A**. HMM: One observation is emitted by each state in the sequence. Panel **B**. HSMM: Several observations are emitted by each state, the number is determined by the duration in the state

2.3.2 Hidden semi-Markov model

Figure 2.2B illustrates the HSMM where instead of assuming there is only one observation per state, a sequence of observations are emitted. The number of observations depends on the amount of time spent in the state. Following the notation in Economou et al. (2014), let τ represent the length of time that the sequence remains in a state before transitioning. These *durations* are labeled in Figure 2.2B as $\tau_1, \ldots, \tau_Q$, where Q is the number of intervals or segments. For $q = 1, \ldots, Q$ we define T_q to be the cumulative duration in segments 1 through q. Lastly, we define $h_j(\tau \mid \phi_j)$ as the duration distribution for each state j, $j = 1, \dots, M$, with parameter ϕ_j.

Similar to the Markov model, the likelihood of the semi-Markov model has two main components consisting of the likelihood of the observations conditional on the states and the likelihood of the semi-Markov chain of states. The joint likelihood of the observations can be specified analogous to the HMM case, but is written incorporating the segment-specific notation

$$L(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \mathbf{S}) = \prod_{i=1}^{n} f(y_i|\mu_{S_i}, \sigma^2_{S_i}) = \prod_{q=1}^{Q} f(\mathbf{y}_{\tau_q}|\mu_{S_q}, \sigma^2_{S_q}), \tag{2.1}$$

where $\mathbf{y}_{\tau_q}$ corresponds to the vector of all the observations in time interval q. The likelihood of the state sequence includes the distribution of the first state, the transition probabilities for the state switches, as well as the information from the duration times

$$L(S_1, \dots, S_Q, \tau_1, \dots, \tau_Q|\boldsymbol{\rho}, \mathbf{P}, \boldsymbol{\phi}) = \rho_{S_1} \prod_{q=1}^{Q-1} h_{S_q}(\tau_q \mid \phi_{S_q}) p_{S_q,S_{q+1}} h_{S_Q}(\tau_Q \mid \phi_{S_Q}). \tag{2.2}$$

Thus, the joint distribution of data, states and durations of the hidden semi-Markov model can be written as

$$\begin{aligned} &L(\mathbf{y}_{\tau_1}, \dots, \mathbf{y}_{\tau_Q}, S_1, \dots, S_Q, \tau_1, \dots, \tau_Q \mid \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\rho}, \mathbf{P}, \boldsymbol{\phi}) \\ &\qquad = \rho_{S_1} \prod_{q=1}^{Q-1} h_{S_q}(\tau_q|\phi_{S_q}) p_{S_q,S_{q+1}} f(\mathbf{y}_{\tau_q}|\mu_{S_q}, \sigma^2_{S_q}) \\ &\qquad \times\ h_{S_Q}(\tau_Q|\phi_{S_Q}) f(\mathbf{y}_{\tau_Q}|\mu_{S_Q}, \sigma^2_{S_Q}). \end{aligned} \tag{2.3}$$

Note we have added the duration distribution parameters of each state to the list of parameters of the HMM. Specifically, the set of model parameters of the HSMM presented includes $\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\rho}, \mathbf{P}$, $\boldsymbol{\phi}$, $\mathbf{S}$, and $\boldsymbol{\tau}$.

2.3.3 Use of covariates to model duration

Previous approaches have specified non-homogeneous HMM and HSMMs by modeling the parameters of the emission distribution or the probabilities of transition using covariates. We propose introducing non-homogeneity in the HSMM duration by letting the parameters of the state duration distribution vary in time as a function of covariates. If we let the duration distribution be a zero-truncated Poisson, we can define the duration parameter $\phi_{S_{q+1}}$ of the interval $q+1$ as a function of the covariate measurements observed prior to the transition at T_q+1. Notice that this specification enables the duration parameter to be both state-specific and vary in time.

Let $\mathbf{X}$ be an $n \times r$ covariate matrix with rows corresponding to times 1 to n, where r is the number of covariates. Let $\boldsymbol{\beta}_{S_{q+1}}$ be an $(r+1)$-dimensional coefficient vector for state S_{q+1} (accounting for an intercept in the model). Then the duration parameter for interval $q+1$, which we denote as $\phi_{S_{q+1}}(\mathbf{X}_{1:T_q}, \boldsymbol{\beta}_{S_{q+1}})$, is a function of the covariate values observed up to time point T_q (the first T_q rows of $\mathbf{X}$), and state specific coefficients $\boldsymbol{\beta}_{S_{q+1}}$. Here, $\phi_{S_{q+1}}(\mathbf{X}_{1:T_q}, \boldsymbol{\beta}_{S_{q+1}})$ can take any functional form of the covariates as long as $\phi_{S_{q+1}} > 0$. For example, we can write the function as

$$\phi_{S_{q+1}}(\mathbf{X}_{1:T_q}, \boldsymbol{\beta}_{S_{q+1}}) = g\left(\beta_{0,S_{q+1}} + \beta_{1,S_{q+1}} f_1\left(\mathbf{x}_{1,1:T_q}\right) + \cdots + \beta_{r,S_{q+1}} f_r\left(\mathbf{x}_{r,1:T_q}\right)\right), \quad (2.4)$$

where $g(\cdot)$ is a specified function that ensures $\phi_{S_{q+1}} > 0$, and $f_1(\cdot), \ldots, f_r(\cdot)$ can be any function of the covariates observed from time 1 to the time previous to the transition, T_q. The joint distribution, which now includes the state-specific duration parameter function, can be written as:

$$\begin{aligned}
&L(y_{\tau_1},\ldots,y_{\tau_Q},S_1,\ldots,S_Q,\tau_1,\ldots,\tau_Q \mid \boldsymbol{\mu},\boldsymbol{\sigma}^2,\boldsymbol{\rho},\mathbf{P},\mathbf{B},\mathbf{X}) \\
&\qquad = \rho_{S_1} h_{S_1}\left(\tau_1 \mid \phi_{S_1}(\mathbf{X}_0,\boldsymbol{\beta}_{S_1})\right) f(\mathbf{y}_{\tau_1}|\mu_{S_1},\sigma^2_{S_1}) \\
&\qquad \times \prod_{q=2}^{Q} h_{S_q}\left(\tau_q \mid \phi_{S_q}(\mathbf{X}_{1:T_{q-1}},\boldsymbol{\beta}_{S_q})\right) p_{S_{q-1},S_q} f(\mathbf{y}_{\tau_q}|\mu_{S_q},\sigma^2_{S_q}),
\end{aligned} \tag{2.5}$$

where $\mathbf{X}_0$ are the initial values for the covariates, and $\mathbf{B}$ is the matrix of β-coefficients with M rows and the number of columns is the number of covariates, r, plus an intercept:

$$\mathbf{B} = \begin{pmatrix} \beta_{0,1} & \beta_{1,1} & \cdots & \beta_{r,1} \\ \beta_{0,2} & \beta_{1,2} & \cdots & \beta_{r,2} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{0,M} & \beta_{1,M} & \cdots & \beta_{r,M} \end{pmatrix}.$$

That is, $\boldsymbol{\beta}'_{S_q}$ is the row of $\mathbf{B}$ that corresponds to state S_q. For example, when the state in interval q is 1, then $\boldsymbol{\beta}_{S_q=1} = (\beta_{0,1},\beta_{1,1},\ldots,\beta_{r,1})'$.

2.3.4 Estimation of model parameters

Model inference can be obtained in a Bayesian framework using Markov chain Monte Carlo (MCMC) and a Metropolis-within-Gibbs sampling algorithm (see Appendix A.1 for the detailed sampling algorithm). To complete the model specification, we assign diffuse priors to the model parameters. The means of the emission distribution are assigned independent Normal priors, $N\,(0,10000)$ and the variances are assigned inverse-Gamma priors, $IG\,(3,3)$. The initial probabilities, as well as each of the rows of the transition matrix, have Dirichlet priors with concentration parameters of 1. The coefficient parameters in the model for the state-specific durations are assumed to be independent and distributed as $N\left(0, 10000\right)$.

The posterior distribution of the states, durations, and rest of the parameters of the HSMM can be summarized as:

$$\begin{aligned} p(\boldsymbol{S}, \boldsymbol{\tau}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\rho}, \mathbf{P}, \mathbf{B} \mid \mathbf{y}, \mathbf{X}) &\propto p(\mathbf{y} \mid \boldsymbol{S}, \boldsymbol{\mu}, \boldsymbol{\sigma}^2) \times p(\boldsymbol{\tau} \mid \boldsymbol{S}, \mathbf{B}, \mathbf{X}) \times p(\boldsymbol{S} \mid \boldsymbol{\rho}, \mathbf{P}) \\ &\times p(\boldsymbol{\mu} \mid \boldsymbol{\theta}_{\mu}, \boldsymbol{\lambda}^2_{\mu}) \times p(\boldsymbol{\sigma}^2 \mid \boldsymbol{\theta}_{\sigma^2}, \boldsymbol{\lambda}_{\sigma^2}) \times p(\mathbf{B} \mid \boldsymbol{\theta}_B, \boldsymbol{\lambda}^2_B) \times p(\boldsymbol{\rho} \mid \boldsymbol{\theta}_{\rho}) \times p(\boldsymbol{P} \mid \boldsymbol{\theta}_P). \end{aligned} \tag{2.6}$$

The state means and variances, initial probabilities, and transition probabilities can be sampled from their full conditionals using a Gibbs update, whereas a Metropolis algorithm is needed for the duration distribution coefficients.

Economou et al. (2014) provide an MCMC implementation of the HSMM using a forward algorithm to estimate the parameters, which alleviates the need to sample the state sequence in the process. However, our model requires sampling the states in order to obtain inference on the parameters of the duration distributions. The state sequence in an HSMM can be sampled with the Gibbs sampler presented in Johnson and Willsky (2013), and we use it to sample the states in each iteration of our MCMC. We verified with a simulation that our sampling algorithm can recover the true parameters.

2.4 Application to Lake Mendota phycocyanin data

The HSMM specified in (2.5) was used to model the hourly maximum standardized levels of phycocyanin in Lake Mendota for the period April 11 to November 15, 2018. In our investigation, we considered different choices of the number of states as well as different functions of the covariates in the duration distribution and chose the model with the lowest deviance information criterion DIC.

Recall that the number of states in HMMs and HSMMs has to be chosen *a priori* and is usually based on information criteria or expert knowledge (Liu and Song, 2020). Carpenter et al. (2020) identified two stable states representing high and low phycocyanin, as well as an unstable equilibrium between the two states. Motivated by their analysis, we investigated both a two state model and three state model.

The duration (hours) in each state is modeled using a zero-truncated Poisson distribution, where the duration parameter is defined as a function of four covariates. These include temperature, wind speed, relative humidity, and PAR. Including the intercept term, this results in five coefficient parameters for each state. Following the notation in (2.4), the duration parameter function in this application is defined as:

$$\begin{aligned}\phi_{S_{q+1}}(\mathbf{X}_{1:T_q}, \boldsymbol{\beta}_{S_{q+1}}) = \exp\big[&\beta_{0,S_{q+1}} + \beta_{1,S_{q+1}} f(x_{1,T_q-l:T_q}) + \beta_{2,S_{q+1}} f(x_{2,T_q-l:T_q}) \\ &+\beta_{3,S_{q+1}} f(x_{3,T_q-l:T_q}) + \beta_{4,S_{q+1}} f(x_{4,T_q-l:T_q})\big]\,,\end{aligned} \tag{2.7}$$

where $f(x_{r,T_q-l:T_q})$ is a function of covariate r from time $T_q - l$ to T_q. The functions $f(\cdot)$ we considered were the mean, maximum, slope, variance and sum, for time periods of 3, 6, 12 and 24 hours (l=2, 5, 11, 23, respectively).

For each model considered, the MCMC algorithm was run for 60000 iterations. The first 20000 iterations were obtained using an adaptive random walk Metropolis algorithm for the duration distribution coefficients. These iterations were used to select the proposal variances for the random walk and then discarded. The remaining 40000 iterations were obtained based on these fixed proposal variances and these samples were used for parameter inference.

The three-state model with covariates defined as the maximum over the 12 hours ($l = 11$) before a transition resulted in the lowest DIC. All subsequent results and model inference pertain to this model.

The posterior mean and 95% credible intervals for the emission distribution parameters in each of the states is presented in Table 2.1. There is a clear distinction between the mean phycocyanin in each state since none of the credible intervals overlap. The three states represent low, medium, and high cyanobacteria states. The variability in the lower state is notably higher, and the wide range of this state can be seen in Figure 2.3. These low, medium and high states of cyanobacteria can be associated with the regimes found in Carpenter et al. (2020). Recall that they identified two stable states, which are comparable to our low and high states (S1 and S3), while the shifts in between these two regimes correspond to our middle state (S2).

Table 2.1: Posterior mean and 95% CI of the emission distribution parameters

State	Mean	Variance
S1	-2.19 (-2.26, -2.09)	0.74 (0.68, 0.83)
S2	0.08 (-0.05, 0.25)	0.44 (0.36, 0.51)
S3	1.73 (1.65, 1.81)	0.44 (0.39, 0.51)

The posterior probabilities of transitioning are shown in Table 2.2. Overall, it is more likely there is a transition between adjacent states (e.g., 1 to 2) than a jump from 1 to 3. This is also true for transitions from higher to lower cyanobacteria states (e.g., 3 to 2). The transition between adjacent states is expected given that the middle state represents a passing state from regimes of cyanobacteria concentration.

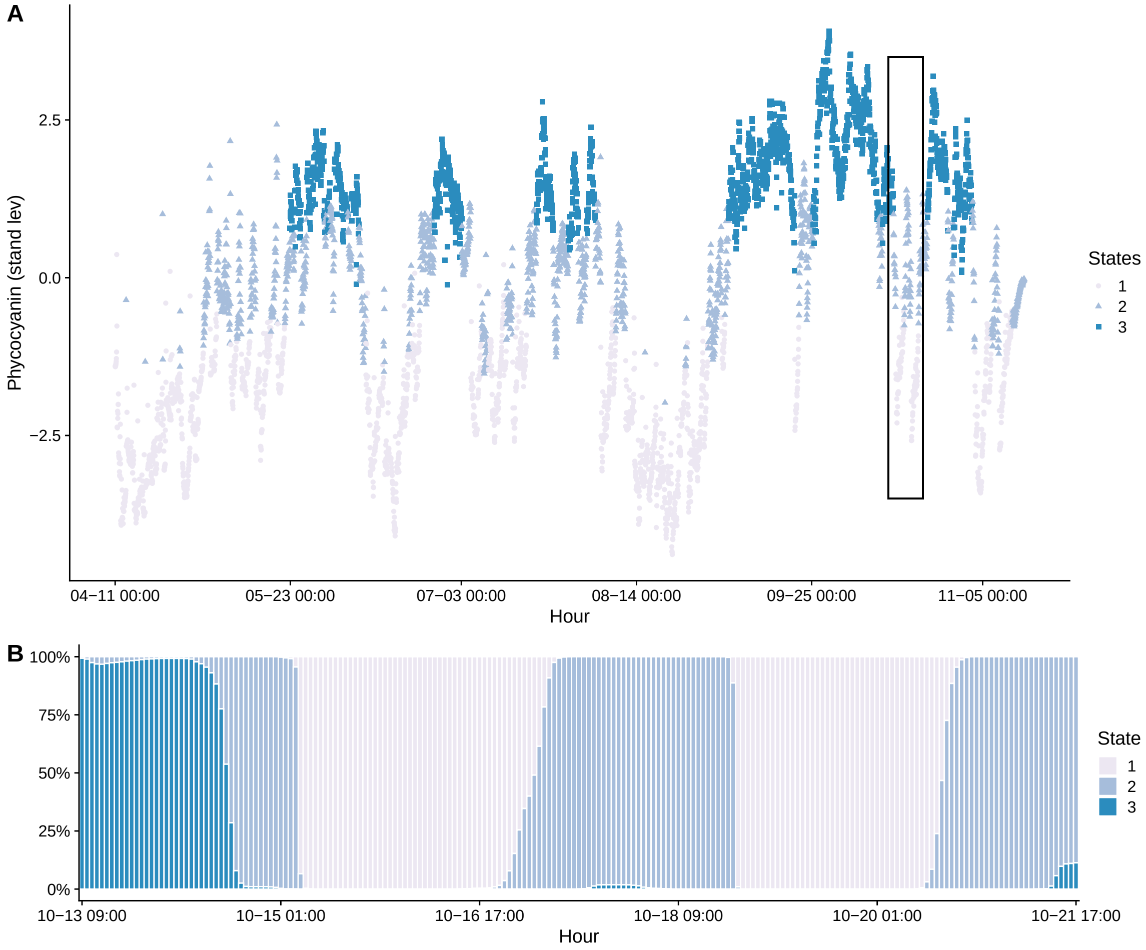

Figure 2.3: Phycocyanin standardized levels classified by latent states of cyanobacteria. Panel **A**. Full state sequence. For each time point, the state is the mode obtained among all iterations. Panel **B** zooms in the period enclosed in the rectangle above. A percent stacked bar for each time point shows the relative distribution of the states sampled in the iterations

Table 2.2: Posterior mean and 95% CI for the state transition probabilities.

Transition	State 1	State 2	State 3
State 1	–	0.96 (0.87, 1)	0.04 (0, 0.13)
State 2	0.59 (0.43, 0.73)	–	0.41 (0.27, 0.57)
State 3	0.09 (0, 0.27)	0.91 (0.73, 1)	–

Table 2.3 provides the number of segments in each state as well as summaries of the duration parameters and durations for each state. The second state has more segments, yet the average duration parameter for this state is much smaller than the other two states. The first and third state have fewer segments, but both the mean and variation in the duration parameter is greater than for the second state. The variation in the duration parameters both within and between states signifies the importance of modeling the durations using covariates and state-specific parameters.

Table 2.3: Segments and duration parameter statistics

State	# segments	Duration parameter			Duration (hours)		
		Mean	Minimum	Maximum	Mean	Minimum	Maximum
S1	27	79.7	55.9	102.3	79.7	20	145
S2	44	31.7	14.1	66.5	31.6	1	94
S3	18	97.5	32.7	220.2	96.4	17	230

The posterior means and credible intervals for the coefficients of the state-specific duration distribution parameters are given in Table 2.4, with the significant coefficients presented in bold font. The coefficients of the duration distribution provide information about the average hourly duration in the states.

Air temperature is significant in capturing the variation in the duration in the high cyanobacteria state and is directly related to the duration. That is, when the maximum temperature in the hours before a transition to this state is warmer, the duration is longer. Wind speed is significant in the low cyanobacteria state. When the maximum of wind speeds in the 12 hours prior to a transition to the low cyanobacteria state is high, we anticipate a shorter duration in that state. Relative humidity is a significant predictor of duration in the second state, with an inverse relation. Lastly, the photosyntetically-active radiation covariate is related to the third state. The negative coefficient for this state indicates that when the maximum value of PAR is at higher levels in the half day before there is a transition to the higher cyanobacteria state, we expect a decrease in the duration.

Table 2.4: Posterior mean and 95% CI of the duration parameter coefficients.

Variable	State 1	State 2	State 3
Intercept	**4.34 (4.25, 4.42)**	**3.40 (3.16, 3.56)**	**4.46 (3.98, 4.87)**
Temperature	0 (-0.08, 0.09)	0.16 (0, 0.32)	**0.40 (0.05, 0.74)**
Wind speed	**-0.12 (-0.23, -0.05)**	0.13 (-0.07, 0.25)	0.01 (-0.34, 0.27)
Relative humidity	-0.02 (-0.15, 0.06)	**-0.27 (-0.41, -0.05)**	-0.24 (-0.61, 0.16)
PAR	0.01 (-0.09, 0.11)	0.04 (-0.13, 0.21)	**-0.55 (-0.86, -0.30)**

2.5 Discussion

In this work we present an extension of the HSMM and apply it to high-frequency environmental data. Using a zero-truncated Poisson distribution for the duration (hours) in each state, we investigated the variation in time spent in each state as a function of time-varying covariates. Although not demonstrated here, this model could be applied to obtain predictions of the next states in the sequence and their expected duration.

In our application, the model enabled the characterization of cyanobacteria concentration in a lake and the detection of important differences in the relationship between the duration in the states and the covariates. The variability of the duration parameters in the different segments supports the introduction of non-homogeneity in the HSMM. While our modeling approach identified similar states of cyanobacteria levels as in Carpenter et al. (2020), we identified unique sets of weather covariates associated with the duration in each of the states.

Higher spring air temperature has been shown to be directly related to cyanobacterial biovolume (Ho and Michalak, 2020). As such, the effect of temperature on the duration in the high cyanobacteria state was expected given that cyanobacteria thrive under warm temperature (Paerl and Huisman, 2008). With respect to wind speed, it is possible that transitions between steady states can be due to wind mixing. For example, Soranno (1997) showed that wind velocity was low during high-pigment conditions, specifically when compared to the wind speeds in the days leading up to these periods. Isles et al. (2015) showed a sharp threshold for physical mixing of the water column at approximately 4.5 m/s, which could break up either stable state (S1 or S3 in our model). Whereas precipitation has been shown to be related to blooms (Reichwaldt and Ghadouani, 2012; Carpenter et al., 2020), these data were not available for this analysis and will be investigated in future analyses. Given the relationship between relative humidity and precipitation, the relationship between humidity and the duration in the middle state detected in our model might be an indirect association. Lastly, Rousso et al. (2021) found that phycocyanin readings may be underestimated when fluorescence measurements are taken under bright light, which may explain the inverse relation we detected in our model between PAR and the duration in the high state.

The inference obtained from the duration in this model can be potentially associated with ecological resilience. Arani et al. (2021) propose to measure resilience, the maximum perturbation that a system endures without transitioning to another state, with life expectancy. They present how to fit a Langevin equation to time series data to capture the different forces that affect a system, and obtain the mean exit time from a state to quantify life expectancy. The information provided by our model in terms of duration in a particular state before transitioning to another can be explored to measure life expectancy as well.

Other approaches exist for introducing non-homogeneity into HSMMs. For example, parameters corresponding to transition probabilities could also be modeled as a function of covariates. However, the focus of the analysis presented here was in capturing the variation in the duration of time in each state. Given that a direct transition from the low to the high state, or vice versa, is unlikely, it was not necessary to model the transition probabilities in terms of covariates for this application.

In both HMMs and HSMMs, observations are assumed to be conditionally independent given the state, meaning they are independent of previous states and observations (Pohle et al., 2017; Yu, 2016). Incorporating more flexible data models in HSMMs, such as those that account for possible dependence between high-frequency observations, is an open research area and will be subject of future work.

Chapter 3

Variable Transition Hidden Markov Model for Network Estimation with Team Sports Data

3.1 Introduction

Network analysis, both static and dynamic, has proven to be a useful tool for analyzing data from team sports. In such applications, a network consists of a set of actors that correspond to the players of a team, and the edges are the relations between the players. Network analysis can help provide inference concerning the performance of players, team strategy, and evolution of relations between players over time, among others. For example, Peña and Touchette (2012) used network analysis in a soccer game to derive a passing network and learn about the strategy of teams. Park and Yilmaz (2010) analyzed video data from a soccer game with a social network based on directed and weighted interactions between the actors. Other examples of social network analysis in team sports can be found in Clemente et al. (2016).

Hidden Markov models (HMM) and hidden semi-Markov models (HSMM) have also been applied to sports and team sports data, mostly in applications related to activity recognition or event detection. These models consider a Markov chain of unobserved states and a sequence of observations dependent on the states (Yu, 2010). Motoi et al. (2012) used a Bayesian HMM to create metadata for sports games through event detection, and their method was evaluated using video data of soccer games to detect events such as kick offs, corner kicks, or goal kicks. Thomas et al. (2010) performed activity recognition using both an HMM and a semi-Markov model (SMM) applied to swimming data collected with a wearable sensor, where the segmentation of the session improved evaluation of the training. Wang et al. (2016) applied two HMMs to classify badminton strokes based on acceleration magnitude. The first HMM was used to recognize whether or not the motions corresponded to a stroke, and the second was used to classify the strokes into different types.

HMMs can be combined with network analysis of sports data. Du et al. (2020) integrated an HMM into the analysis of network data to provide a model for a player's performance in basketball games. The authors analyzed data from 20 games, where the observation data were the points that a player obtained in each game and the states defined their performance. A benefit of the combination of HMMs with networks is that it provides a way of modeling edges as a latent network. Usually, the relations among the players are defined in terms of the passes between them (Peña and Touchette, 2012; Park and Yilmaz, 2010), either as an indicator of whether or not there is a connection present, or as the number of passes. However, the edges in a dynamic network can also be modeled as the hidden states in an HMM, which, to our knowledge, has not been applied in the team sports context.

The transition probabilities between states in an HMM can be nonstationary (time-varying). From a modeling perspective, the transition probabilities can be allowed to change over time as functions of time-varying covariates. This approach has been used in many applications of HMM and HSMMs. For example, in the sports context, Ötting et al. (2021) defined the state transition probability as a function of covariates to account for factors that contribute to a change in the state. In their work, the states represented the level of control of a team and the covariates used for the transitions were score difference, an indicator of home game, the market value of the opposing team and the minute of the game. Another covariate that can be included in the specification of the transition probabilities is the time in the state. Both Vaseghi (1991, 1995) and Ramesh and Wilpon (1992) introduced such duration-dependent state transition probabilities. In this case, the probability of transition to the next state is a function of the time spent in the state before the transition. This type of inhomogeneous HMM is termed a "variable transition hidden Markov model" (VTHMM), and is one of the methods of duration modeling in HMMs (Johnson, 2005) in addition to HSMMs.

In this chapter, we develop a model with a latent network to capture edges between players in team sports with a VTHMM using player location data. The states are the edges between players and the probability of these edges is defined in terms of the time that the players have been connected, as well as other covariates of interest. To analyze location data of players in the context of a network with an HMM, we adapt the model developed by Scharf et al. (2016) for the state-dependent distribution of our HMM (note, they modeled animal telemetry data assuming there is a latent dynamic social network). The contribution of this work is threefold. First, we extend the use of animal movement models to study team sports. Second, we propose modeling the dynamic process using a VTHMM and incorporate the duration in the state transition probability. Lastly, it provides a model-based framework to learn the underlying networks of the players in contrast to previous deterministic approaches.

The chapter is organized as follows. Section 3.2 presents the model, with the adaptation of the model in Scharf et al. (2016) for our observation model and with the transition probability function for the state model. Section 3.3 introduces the player location data from a soccer game that motivated our analysis and the results of applying our methods to the data. Lastly, Section 3.4 includes a discussion and directions for future work.

3.2 Model

An HMM is a mixture model that contains two parts, a finite state Markov chain, which is the mixing distribution, and a sequence of observations that are dependent on the Markov chain of states (Scott, 2002; Zucchini et al., 2017). The state sequence can be denoted as $\mathbf{S} = (S_1, \ldots, S_T)'$, where $S_t \in \{1, 2, \ldots, M\}$, for time $t = 1, \ldots, T$, and with M being the total number of unique states. The transition from state k at time t to state l at time $t+1$ is defined by the probability $p_{k,l} = P\left(S_{t+1} = l \mid S_t = k\right)$. We can model the observation sequence $\mathbf{y} = (y_1, \ldots, y_T)'$ with a state-dependent probability distribution with parameters $\boldsymbol{\theta}$, often referred to as emission distribution, $f(\mathbf{y}|\boldsymbol{\theta}, \mathbf{S})$.

The duration is not explicitly modeled in HMMs. Rather, it implicitly follows a geometric distribution (Yu, 2010). There are various ways to incorporate duration modeling into HMMs. These approaches include hidden semi-Markov models (HSMM), variable transition HMMs (VTHMM), and standard HMMs that have complex state topologies or expanded states (Johnson, 2005).

The VTHMM is an HMM where the transition probabilities depend on the time spent so far in the current state (Vaseghi, 1995). The probabilities are defined as $p_{k,l}(d_t) = P\left(S_{t+1} = l \mid S_t = k, d_t\right)$, with d_t being the duration spent in the current state (i.e., state k) up to time t. The emission distribution is defined in the same way for both types of models. The VTHMM can also be seen as a type of inhomogeneous or nonstationary HMM. In these models the transition probabilities are not constant in time, and will change according to time-dependent covariates. The model we present in this chapter is an expanded VTHMM where other covariates in addition to the duration are considered in the state transition probabilities.

In the following subsections we present the details of a VTHMM applied to player location data in team sports. In Section 3.2.1 we present the function for the transition probabilities, which takes into account time-dependent covariates and duration, and in Section 3.2.2 we describe the model used for the observed player location data.

3.2.1 Duration dependent state transitions

In our model for team sports data we assume that at each time point t there is an underlying network G. This network can be represented as a graph, with an adjacency matrix at t denoted as $\mathbf{W}_t$. The edges in G are unweighted and undirected, meaning that the entries of $\mathbf{W}_t$ only indicate whether or not there is a relation between players i and j. We regard this relation as a latent connection between the players and analyze it as a hidden state sequence in our model. Specifically, we consider a two state model. For each pair of players i and j, there is a state sequence $w_{ij,1}, \ldots, w_{ij,T}$, where the states represent a connection at time t, $w_{ij,t} = 1$, or no connection at time t, $w_{ij,t} = 0$.

The transition probabilities are defined in terms of covariates and the duration in the current state, i.e., the duration of the presence or absence of a connection. The covariates are time-dependent, implying the transition probabilities are also time dependent, and are observed for each pair of players at each time point. We denote the r-th covariate observed for the ij pair at time t as $x_{ij,r,t}$, where $r = 1, \ldots, q$. We denote the time that the ij pair has spent in the current state up to time t as $d_{ij,t}$. If we consider the absence of a connection state at time t, $w_{ij,t} = 0$, the probability of remaining in that same state from time t to $t + 1$ is

$$\begin{aligned} &P(w_{ij,t+1} = 0 \mid w_{ij,t} = 0, \mathbf{x}_{ij,t}, d_{ij,t}) = \\ &\qquad\qquad logit^{-1}(\beta_0^{(0)} + \beta_1^{(0)} x_{ij,1,t} + \cdots + \beta_q^{(0)} x_{ij,q,t} + \beta_{q+1}^{(0)} d_{ij,t}). \end{aligned} \tag{3.1}$$

Conversely, if we consider the presence of a connection state at time t, $w_{ij,t} = 1$, the probability of remaining in that same state from time t to $t + 1$ is

$$\begin{aligned} &P(w_{ij,t+1} = 1 \mid w_{ij,t} = 1, \mathbf{x}_{ij,t}, d_{ij,t}) = \\ &\qquad\qquad logit^{-1}(\beta_0^{(1)} + \beta_1^{(1)} x_{ij,1,t} + \cdots + \beta_q^{(1)} x_{ij,q,t} + \beta_{q+1}^{(1)} d_{ij,t}). \end{aligned} \tag{3.2}$$

Here there are $q + 1$ covariates, including the duration, $d_{ij,t}$. The coefficients $\boldsymbol{\beta}^{(0)} = \left(\beta_0^{(0)}, \beta_1^{(0)}, \ldots, \beta_{q+1}^{(0)}\right)'$ and $\boldsymbol{\beta}^{(1)} = \left(\beta_0^{(1)}, \beta_1^{(1)}, \ldots, \beta_{q+1}^{(1)}\right)'$ represent the vector of coefficients associated with the no connection state and connection state, respectively. Note that the probability of transitioning from state $w_{ij,t} = 0$ to $w_{ij,t+1} = 1$ from time t to time $t + 1$ is $1 - P(w_{ij,t+1} = 0 \mid w_{ij,t} = 0, \mathbf{x}_{ij,t}, d_{ij,t})$. This probability is similarly defined for a transition from the connected to the unconnected state. The state at time $t = 1$ for any ij pair is modeled as $w_{ij,1} \sim Bern(p_1)$, where p_1 is the initial probability of a connection.

3.2.2 Emission model

The emission model in a VTHMM is defined in the same way as in a standard HMM. The emission model we present for player location data in team sports is a hierarchical model. We denote the observed location for an individual i at time t as $\mathbf{s}_{i,t} \in \mathbf{R}^2$, and define it as a function of a latent process $\boldsymbol{\mu}_{i,t} \in \mathbf{R}^2$, which is the true location of the player. The model for the observed location data is $\boldsymbol{s}_{i,t} = \boldsymbol{\mu}_{i,t} + \epsilon\mathbf{1}$, where ϵ is the measurement error and is distributed $N(0, \tau^2)$.

The model we consider for the latent location is an extension of the Gaussian Markov random field (GMRF) model presented in Scharf et al. (2016). In modeling animal movement, Scharf et al. (2016) assume there is a latent social network for the animals. They model the positions of the animals at time t with a GMRF conditional on the positions at time $t-1$ and on the latent network at t. Their modeling of movement considers an alignment mechanism to examine the movement of connected animals in the same direction and an attraction mechanism to reflect the movement of an animal towards the others to which it is connected. These collective behaviors of alignment and attraction in animal movement can be used to study the dynamical systems in team sports (Welch et al., 2021). Alignment and attraction between the players are important in order to capture the possible synchrony of the players, as well as possible expansion and contraction of the team on the field. In addition to these two collective behaviors, we also consider the current direction of movement of each player by including a direction vector for each individual in the model.

Here, we consider n players with locations recorded at T time points. Before presenting our model, we first define some of the important terms and notation introduced by Scharf et al. (2016). The number of connections that player i has at time t ranges from 0 to $n-1$. We denote this count as $w^c_{i+,t}$. When a player is not connected to any other, in order to have a nonzero precision, we define $w^c_{i+,t}$ to be a positive constant c . Then, we define the mean location of all players that individual i is connected to, $\bar{\boldsymbol{\mu}}_{i,t}$, as

$$\bar{\boldsymbol{\mu}}_{i,t} = \sum_{j \neq i}^{n} \frac{w_{ij,t}}{w^c_{i+,t}} \boldsymbol{\mu}_{j,t}.$$

The vector that points from the location of player i to the mean of its connections, $\bar{\boldsymbol{\mu}}_{i,t}$, is defined as $\widetilde{\boldsymbol{\mu}}_{i,t}$. When player i has one or more connections, it is defined as

$$\widetilde{\boldsymbol{\mu}}_{i,t} \equiv \frac{\bar{\boldsymbol{\mu}}_{i,t} - \boldsymbol{\mu}_{i,t}}{\|\bar{\boldsymbol{\mu}}_{i,t} - \boldsymbol{\mu}_{i,t}\|_2}.$$

If player i is not connected to any other player at time t, $\widetilde{\boldsymbol{\mu}}_{i,t} = 0$. The unit direction vector for player i at time t is

$$\boldsymbol{\delta}_{i,t} := \frac{\boldsymbol{\mu}_{i,t} - \boldsymbol{\mu}_{i,t-1}}{\|\boldsymbol{\mu}_{i,t} - \boldsymbol{\mu}_{i,t-1}\|}.$$

Next we model $\boldsymbol{\mu}_t$, the latent locations of all the players at time t, conditional on the latent state sequences for all pairs of players as well as the positions at the previous two time points. The multivariate model for $t = 3, \ldots, T$ is defined as

$$\left[\boldsymbol{\mu}_t \mid \boldsymbol{\mu}_{t-1}, \boldsymbol{\mu}_{t-2}, \alpha, \gamma, \eta, \sigma^2, c, \mathbf{W}_t\right] \equiv N\left(\boldsymbol{\mu}_{t-1} + \gamma \widetilde{\boldsymbol{\mu}}_{t-1} + \eta \boldsymbol{\delta}_{t-1}, \mathbf{Q}_t\right), \tag{3.3}$$

where γ is the coefficient of the attraction component $\widetilde{\boldsymbol{\mu}}_t$, and η is the coefficient associated with our unit direction vector term $\boldsymbol{\delta}_t$. The elements of the GMRF precision matrix $\mathbf{Q}_t$ can be specified, following Scharf et al. (2016). That is,

$$\mathbf{Q}_{ij,t} \equiv \begin{cases} -\alpha w_{ij,t}\sigma^{-2}\mathbf{I}_2, & j \neq i \\ w^c_{i+,t}\sigma^{-2}\mathbf{I}_2, & j = i, \end{cases}$$

where α is the coefficient associated with the alignment effect. Recall that $w^c_{i+,t}$ denotes the number of connections of player i, and that $w^c_{i+,t} = c$ when the player i is not connected to any other at time t, in order for the precision of this player to be nonzero. We restrict the value of c to be between 0 and 1 to reflect that the precision of the players latent location will be lower when it has no connections. This general model can be adapted with other terms depending on the specific sport data being analyzed.

3.2.3 Parameter estimation

A Bayesian approach can be used to estimate the parameters in this model. The full conditionals for the model parameters, the location process, and the state sequences are given in Appendix B.2. We assign diffuse priors to the parameters. The coefficients of the probability functions are assigned the prior distribution $N(0, 1000)$, and the initial probability of connection for all player pairs is assigned a Beta(1,1) prior. Regarding the parameters of the emission model, we specify a Half-Cauchy(0.5) prior for the variance, and Beta(1,1) for c. We follow Scharf et al. (2016) and specify a prior with a shifted and scaled Beta distribution to consider the support for the alignment coefficient as (-1,1). Both the attraction coefficient and direction vector coefficient are assigned a $N(0, 1000)$ priors. For the variance of the measurement error, we specify a conjugate inverse-Gamma prior. We verified through simulation that we can capture the true parameters except for σ^2 (see Appendix B.3).

3.3 Application

3.3.1 Soccer data

The data used in this application come from high-resolution tracking data of a Major League Soccer (MLS) game from 2017. The data were recorded with a minimax device from Catapult and recovered through their Sprint software (Catapult-Innovations, 2013). The raw data were collected on each player at a rate of 25Hz for approximately 150 minutes starting at the warm-up and extending through the end of the game. For this analysis, the data were thinned to 1Hz.

Based on the observed locations of the players, we computed two quantities to use as covariates in the state transition probabilities. The first was the difference in turning angle for each pair of players at each time point, which was obtained by first computing the turning angle for each player. The second quantity used as a covariate was the distance between each pair of players at each time point.

We selected four different five minute segments (T=300 time points) of the game to apply the model. Each of these segments were selected based on different events occurring within the game. Segments I & II are from the first half, and III & IV from the second half. A goal was scored in segments II and III, one from each team. The goal in segment II was scored by team B during the middle of the segment. In segment III, team A scored from a penalty kick at the end of the segment.

We ran a Markov chain Monte Carlo algorithm for each of the eight cases (two teams, four segments) for 75000 iterations. The first 30000 were discarded as burn-in. For our observed data model we assumed the variance of the measurement error to be fixed. The Catapult devices are accurate within 10cm, and translating this value to our units, we have $\tau = 8.3e - 07$.

3.3.2 Results

We applied the variable time hidden Markov model to the two teams in the soccer game for each of the segments. The model was applied separately to each team to learn about their attraction, alignment and connection parameters within team. The parameter estimates for Team A are in Tables B.3.1 & 3.2, and those for Team B are in Tables 3.3 & 3.4.

Emission model

From the observation model, we detect that the alignment is high in all segments for both teams, where the posterior mean for $\alpha \approx 0.9$ for all cases. This was expected because of the nature of play in a soccer game. Recall the attraction and direction vectors take values between -1 to 1. When comparing their coefficients, γ and η, we see that the coefficient of the unit direction vector is larger than the coefficient for attraction for all segments for both teams. The variance estimate is also similar in all segments for both teams.

Table 3.1: Posterior mean and 95% CI for model parameters for each segment for Team A, segments I & II.

Component		Parameter		Segment I	Segment II
Emission	Mean	γ	Attraction	0.00077 (0.00044, 0.00112)	0.00037 (0.0007, 0.00068)
		η	Direction	0.011 (0.007, 0.017)	0.01 (0.006, 0.014)
	Precision	α	Alignment	0.88 (0.84, 0.92)	0.87 (0.83, 0.91)
		σ^2	Variance	0.00011 (0.00009, 0.00015)	0.00013 (0.00010, 0.00015)
		c	Non zero count	0.19 (0.14, 0.27)	0.15 (0.1, 0.2)
State transitions	Initial distrib	p_1	Initial prob	0.33 (0.18, 0.5)	0.41 (0.25, 0.58)
	Non connection	$\beta_{0,0}$	Intercept	**3.42 (2.49, 4.57)**	**3.34 (2.09, 4.62)**
		$\beta_{0,1}$	Duration	-0.003 (-0.016, 0.007)	**-0.031 (-0.052, -0.013)**
		$\beta_{0,2}$	Angle	0 (-0.71, 0.6)	0.38 (-0.45, 1.27)
		$\beta_{0,3}$	Distance	0.3 (-2.67, 3.33)	**5.14 (0.45, 10.28)**
		$\beta_{0,4}$	Ang*Dist	0.44 (-1.6, 3)	-2.52 (-5.63, 1.13)
	Connection	$\beta_{1,0}$	Intercept	**3.44 (2.21, 4.79)**	**2.99 (2.18, 3.79)**
		$\beta_{1,1}$	Duration	-0.009 (-0.06, 0.017)	0.001 (-0.019, 0.015)
		$\beta_{1,2}$	Angle	-1.00 (-2.09, 0.05)	-0.27 (-0.84, 0.39)
		$\beta_{1,3}$	Distance	-1.10 (-4.89, 3.08)	-0.81 (-3.47, 2.18)
		$\beta_{1,4}$	Ang*Dist	1.94 (-1.31, 6.18)	0.50 (-1.47, 2.66)

Table 3.2: Posterior mean and 95% CI for model parameters for each segment for Team A, segments III & IV.

Component		Parameter		Segment I	Segment II
Emission	Mean	γ	Attraction	0.00047 (0.00005, 0.00091)	0.00115 (0.00078, 0.00161)
		η	Direction	0.008 (0.002, 0.014)	0.006 (0.003, 0.009)
	Precision	α	Alignment	0.91 (0.85, 0.95)	0.895 (0.856, 0.93)
		σ^2	Variance	0.00010 (0.00008, 0.00014)	0.00009 (0.00008, 0.00010)
		c	Non zero count	0.2 (0.12, 0.36)	0.12 (0.1, 0.15)
State transitions	Initial distrib	p_1	Initial prob	0.26 (0.13, 0.42)	0.57 (0.40, 0.74)
	Non connection	$\beta_{0,0}$	Intercept	**2.34 (1.57, 3.22)**	**2.47 (1.6, 3.49)**
		$\beta_{0,1}$	Duration	0.001 (-0.01, 0.01)	0.007 (-0.001, 0.014)
		$\beta_{0,2}$	Angle	0.66 (-0.08, 1.35)	0.42 (-0.3, 1.11)
		$\beta_{0,3}$	Distance	**3.26 (0.12, 6.42)**	**4.17 (0.41, 7.36)**
		$\beta_{0,4}$	Ang*Dist	-2.05 (-4.05, 0.6)	**-2.83 (-5, -0.27)**
	Connection	$\beta_{1,0}$	Intercept	**2.49 (1.57, 3.48)**	**2.72 (1.88, 3.74)**
		$\beta_{1,1}$	Duration	-0.029 (-0.077, 0.01)	**-0.052 (-0.101, -0.01)**
		$\beta_{1,2}$	Angle	0.18 (-0.67, 1.03)	-0.04 (-0.74, 0.58)
		$\beta_{1,3}$	Distance	-0.82 (-3.04, 1.64)	0.63 (-2.46, 3.47)
		$\beta_{1,4}$	Ang*Dist	1.52 (-1.29, 4.39)	0.12 (-1.96, 2.47)

Table 3.3: Posterior mean and 95% CI for model parameters for each segment for Team B, segments I & II.

Component		Parameter		Segment I	Segment II
Emission	Mean	γ	Attraction	0.00068 (0.00029, 0.00107)	0.00049 (0.00016, 0.00083)
		η	Direction	0.018 (0.015, 0.022)	0.008 (0.005, 0.011)
	Precision	α	Alignment	0.92 (0.91, 0.94)	0.94 (0.92, 0.96)
		σ^2	Variance	0.00009 (0.00007, 0.00010)	0.00011 (0.00009, 0.00013)
		c	Non zero count	0.28 (0.22, 0.35)	0.09 (0.07, 0.11)
State transitions	Initial distrib	p_1	Initial prob	0.47 (0.29, 0.66)	0.4 (0.24, 0.58)
	Non connection	$\beta_{0,0}$	Intercept	-0.21 (-0.63, 0.24)	**2.15 (1.38, 2.85)**
		$\beta_{0,1}$	Duration	0.009 (-0.002, 0.02)	-0.013 (-0.036, 0.005)
		$\beta_{0,2}$	Angle	**2.37 (1.84, 2.95)**	**0.51 (0.13, 1.00)**
		$\beta_{0,3}$	Distance	**8.84 (6.35, 11.34)**	**6.5 (3.61, 9.39)**
		$\beta_{0,4}$	Ang*Dist	**-4.93 (-7.07, -2.57)**	**-2.56 (-3.81, -0.97)**
	Connection	$\beta_{1,0}$	Intercept	**1.96 (0.85, 3.17)**	**2.04 (1.59, 2.51)**
		$\beta_{1,1}$	Duration	-0.088 (-0.356, 0.079)	**0.018 (0.004, 0.03)**
		$\beta_{1,2}$	Angle	**-2.51 (-3.35, -1.72)**	-0.31 (-0.93, 0.32)
		$\beta_{1,3}$	Distance	**6.44 (2.38, 12.28)**	-1.35 (-2.97, 0.18)
		$\beta_{1,4}$	Ang*Dist	1.27 (-2.59, 5.12)	**2.67 (0.09, 5.81)**

Table 3.4: Posterior mean and 95% CI for model parameters for each segment for Team B, segments III & IV.

Component		Parameter		Segment I	Segment II
Emission	Mean	γ	Attraction	8e-05 (-0.00016, 0.00031)	0.00012 (-2e-04, 0.00043)
		η	Direction	0.006 (0.002, 0.01)	0.015 (0.011, 0.02)
	Precision	α	Alignment	0.94 (0.92, 0.96)	0.9 (0.88, 0.93)
		σ^2	Variance	7.1e-05 (6.3e-05, 8.2e-05)	7.2e-05 (5.2e-05, 9.5e-05)
		c	Non zero count	0.13 (0.1, 0.16)	0.22 (0.15, 0.3)
State transitions	Initial distrib	p_1	Initial prob	0.26 (0.13, 0.41)	0.43 (0.25, 0.61)
	Non connection	$\beta_{0,0}$	Intercept	**1.00 (0.09, 1.81)**	**0.99 (0.32, 1.67)**
		$\beta_{0,1}$	Duration	0.003 (-0.009, 0.012)	**0.018 (0.009, 0.028)**
		$\beta_{0,2}$	Angle	**1.1 (0.34, 1.8)**	**1.00 (0.28, 1.85)**
		$\beta_{0,3}$	Distance	**11.55 (7.55, 16.8)**	**4.28 (0.85, 7.5)**
		$\beta_{0,4}$	Ang*Dist	**-5.5 (-8.07, -2.43)**	-1.25 (-4.01, 3.12)
	Connection	$\beta_{1,0}$	Intercept	**2.68 (2.04, 3.46)**	**4.35 (2.68, 6.00)**
		$\beta_{1,1}$	Duration	0.015 (-0.014, 0.033)	**-0.189 (-0.394, -0.045)**
		$\beta_{1,2}$	Angle	**-1.04 (-1.73, -0.52)**	**-2.33 (-3.48, -1.35)**
		$\beta_{1,3}$	Distance	-3.08 (-6.16, 0.28)	-0.03 (-5.36, 4.91)
		$\beta_{1,4}$	Ang*Dist	**5.64 (2.36, 10.02)**	2.4 (-1.31, 7.1)

State transitions

The coefficients associated with the state transition probabilities vary by team and segment. The duration is significant for team A only in two cases. In Table B.3.1, we see that for segment II, the posterior mean of the duration coefficient for the unconnected state probability ($\beta_1^{(0)}$) is negative and significantly different from 0 based on its 95% credible interval. This indicates that the longer a pair of players is unconnected, the lower the probability that they will remain in that state, when all the other covariates are held constant. We find a similar relation between duration and the probability of the connection state in segment IV (Table 3.2). Thus, we can state that when a pair of players is connected, the probability that they remain connected diminishes with time.

For Team B, the duration covariate is significant in three cases. We see the coefficient estimate of β_1 is significantly positive in the unconnected state probability in segment IV (Table 3.4), and in the connection probability in segment II (Table 3.3). These significant positive coefficients indicate that the longer a pair of players has been in the respective connectivity state, the longer they will stay in it, given all the other covariates remain the same. Additionally, in segment IV, we expect the probability of remaining in the connection state to diminish the longer that a pair has been connected.

The angle and distance covariates, as well as their interaction do not significantly capture variation in the probability of remaining connected in any of the segments for team A. However, they do capture variation in the probability of the unconnected state. From Table B.3.1 in segment II and Table 3.2 in segment III, we see the distance coefficient $\beta_3^{(0)}$ is significant, indicating that the further away two players are, the higher the probability that they will remain unconnected. The interaction coefficient $\beta_4^{(0)}$ is significant as shown in the segment IV column, suggesting that the probability of staying unconnected will change depending on how far apart the pair of players are and if they are moving away or toward one another.

For Team B, the interaction between angle and distance is significant for the unconnected state, as seen in segments I, II (Table 3.3) and III (Table 3.4). However, it is not significant in the unconnected probability of segment IV. We find that the main effects of angle and distance each have a positive effect. The bigger the difference in the turning angle between two players, the higher the probability that they remain unconnected. In addition, the further away that two players are from one another on the field, the higher the probability they will remain unconnected in the next time point.

Looking further into the connection state probabilities of Team B, we find for segments II (Table 3.3) and III (Table 3.4) that the interaction between the angle and distance covariates is significant with a positive coefficient $\beta_4^{(1)}$. Some of the main effects in the connection state probability are significant in segments I and IV. For segment I, we find that the bigger the difference in the turning angle between two players, signifying that they are moving away from each other, the lower the probability that they will remain connected. We also see in segment I that the distance coefficient $\beta_3^{(1)}$ is positive, meaning that the further away two players are in the field, the higher the probability that they remain connected. This may seem counter intuitive at first, given that we would expect players that are closer to be connected. However, given the nature of a soccer game, two players can still be in synchrony even if they are far apart in space. Lastly, in segment IV, the difference in angle is significant, and the coefficient $\beta_2^{(1)}$ is negative. This is intuitive since it means that as two players are moving away from each other, it is less likely that they will remain connected.

Estimated dynamic network

As stated previously, the network between players in team sports has traditionally been defined deterministically. For example, in soccer, the edges of the network may be determined by whether or not there is a pass between two players, or by the count of these passes (Peña and Touchette, 2012; Clemente et al., 2016). With our data-driven approach, we obtain probabilistic estimates of connections between each pair of players at each time point. These probabilities are summarized in Figure 3.1 by considering the count of the connected pairs in each of the segments. We can observe the variation in the number of connections throughout the segment, and observe how the connections change after the scored goals in segments II and III.

3.4 Discussion

We presented a VTHMM that considers a network for the latent stochastic component and a GMRF to model the data component. We expanded the VTHMM to accommodate other time varying covariates for the transition probabilities, in addition to the duration in a state. For modeling the team sport observed location data we take advantage of a GMRF model developed in an animal movement ecological context.

In our application we considered data from a professional soccer game. We were

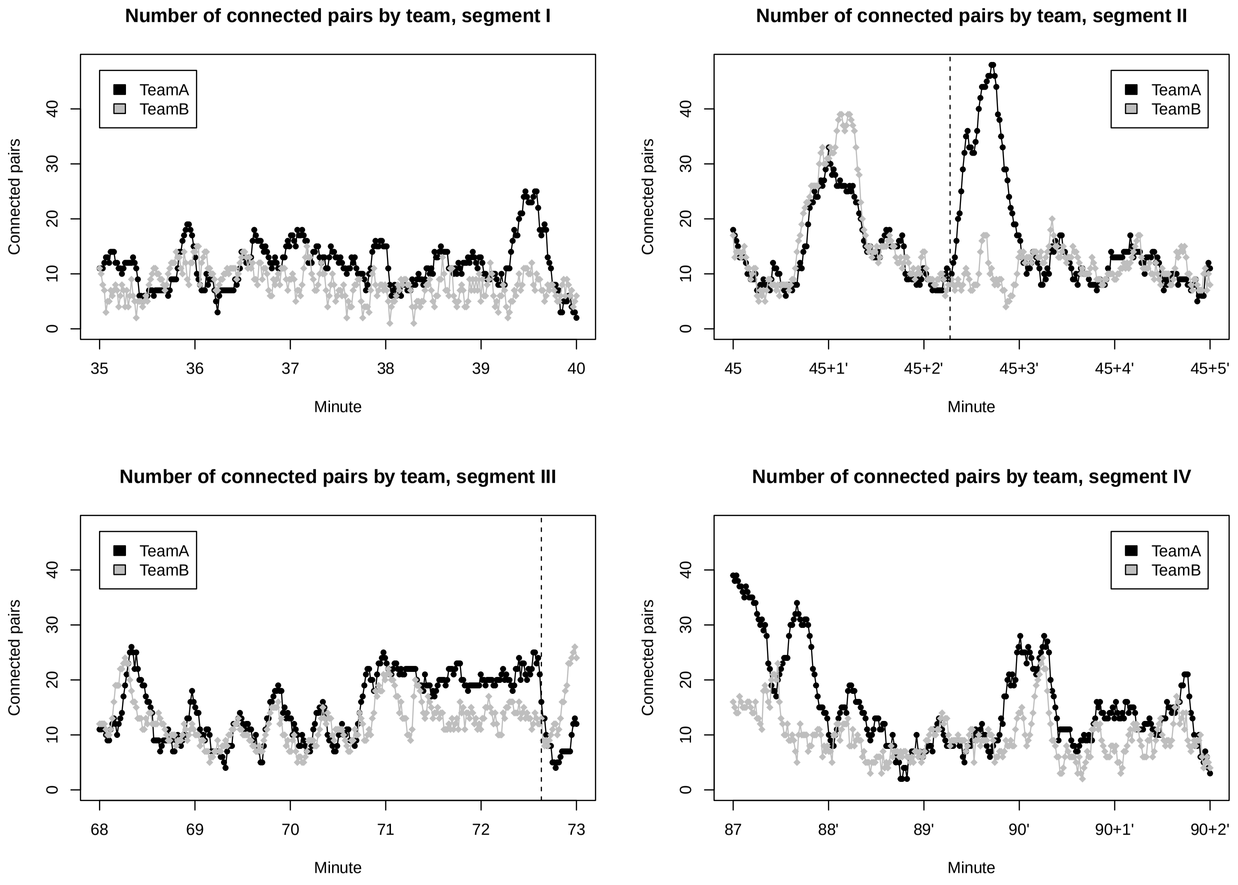

Figure 3.1: Number of connected pairs by team at each time point during each of the four segments of the game. The vertical dashed lines denote times when a goal was scored.

able to learn about the alignment, attraction, and direction of the players in different segments of the game. We also determined which covariates helped explain the variation in the transition probabilities for the two states.

In addition to parameter estimation, we obtain estimates of the underlying dynamic network. Importantly, we estimate the adjacency matrix rather than define it in a deterministic way through information such as passes between players. With this estimated network we have the information necessary to compute different metrics associated with networks in team sports, at both the player and team level. For example, Clemente et al. (2016) present, among other measures, the degree centrality index, which is a count of the connections of a player, to identify players that are more important for the general structure of the team network.

Our approach is flexible in that covariates that are included in either the emission model or the state transition model can be modified or replaced by other covariates that may provide more information specific to the team sport data being analyzed. For example, sport specific positional covariates, ball possession, occurrence of particular events, or proximity to specific locations on the playing field (e.g., the three point line in basketball) could all be considered.

A disadvantage of the model presented here is that computational time increases as more individuals are considered. This is because the analysis is based on information between all pairs of players. In future work we will consider alternative estimation approaches to reduce the computational time in order to enable modeling both teams simultaneously. For an application like the one presented here, this will allow us to compare and contrast the information obtained for each team with the information from all individuals in the game. The computational cost of the current model also prohibits the analysis of an entire game when data are at high temporal frequency.

Chapter 4

Accounting for dependence in the emission distribution of Hidden Semi-Markov Models

4.1 Introduction

In both HMMs and HSMMs, observations are modeled with an emission distribution conditioned on a latent state process. Typically, these observations are assumed to be conditionally independent given the state, meaning they are independent of previous states and observations (Pohle et al., 2017; Yu, 2016). However, high-frequency data are more likely to be highly correlated such that this conditional independence assumption may be inappropriate. The violation of the conditional independence assumption can have an impact on statistical inference. For example, Pohle et al. (2017) presented a simulation study to determine the effects of assumption violations in the selection of the number of states in HMMs, and determined that when the Akaike and Bayesian information criteria are used to select between models, the number of states will be overestimated when the assumption is violated.

Markov-switching regression models (MSR) and neural networks (NN) are two approaches that have been used when the conditional independence assumption is not met. In MSR, the observations are modeled as a function of covariates or as an autoregressive model (Langrock et al., 2017). In the context of NN, Ravuri and Wegmann (2016) apply a deep neural network HMM (DNN-HMM) and show that deeper NNs compensate for the conditional independence assumption violation more than shallow NNs. In addition, Dai et al. (2017) consider a recurrent hidden semi-Markov model (R-HSMM) that incorporates a recurrent neural network (RNN) in the observation model of an HSMM. This model formulation can accommodate more complex dependencies in the observation sequence.

The disadvantages of approaches such as NNs are that they are computationally expensive to implement and difficult to interpret and obtain uncertainty quantification. Thus, it is useful to consider a more computationally tractable approach for mitigating the conditional dependence. One approach that has not been considered in this context is data subsampling. Data subsampling is used for reducing computational cost or in determining the sampling distribution of a statistic. For instance, when dealing with large datasets, it has been used to increase efficiency in Firefly Monte Carlo (FlyMC) (Maclaurin and Adams, 2015) and in subsampling Markov chain Monte Carlo (MCMC) (Quiroz et al., 2019). Both of these approaches present an MCMC sampling algorithm that considers only a subset of the data at each iteration. Experiments conducted to evaluate the performance of FlyMC found it to be more efficient than standard MCMC sampling in many instances. Similarly, Quiroz et al. (2019) showed that subsampling MCMC is often more efficient than standard MCMC and other competing subsampling algorithms. Bradley (2021) provides an approach to incorporate subsampling into the Bayesian hierarchical framework. The author adds a subset model to the general notation of this hierarchical framework, and the data model is then considered a data subset model. Lastly, the subsampling approach in MCMC is also explored by Li and Wong (2018) in their mini-batch tempered MCMC approach and by Wu et al. (2019) with their mini-batch Metropolis-Hastings MCMC. These studies suggest that random data subsampling might be introduced as part of the MCMC algorithm for HSMMs as an attempt to reduce or eliminate the conditional dependence in the data. We investigate this approach in this chapter, along with other, model-based, approaches that directly account for dependence in the observation sequence.

In this chapter we present and compare methods that either mitigate or accommodate conditional dependence in the observation sequence of HMMs and HSMMs. First, we propose a novel approach to mitigate the dependence by introducing data subsampling. As an alternative approach, we consider the use of basis function expansions as random effects (Ruppert et al., 2003) in the emission distribution to capture dependencies in the data. Whereas basis functions have been used as an alternative to parametric models in HMMs (Langrock et al., 2015), they have not been considered with the goal of dealing with the conditional dependence. Lastly, we compare these two approaches with traditional autoregressive models in HSMMs.

In our investigation, we conduct two simulation studies. The first simulation study investigates the implementation and benefits of using subsampling when fitting HSMMs. This includes comparing various subsampling rates as well as levels of autocorrelation in the observation sequence. Then, we compare the three approaches (subsampling, basis functions, and autoregressive models) under a set of generative models with varying dependence.

We apply the models to high-frequency environmental data where there is dependence in the observation sequence. Specifically, we model high-frequency phycocyanin measurements, an indicator of lake cyanobacteria concentration, in Lake Mendota, Wisconsin. We investigate the similarities and differences in posterior inference obtained for these data under the different approaches.

The remaining part of this chapter is outlined as follows. Section 4.2 presents three different approaches to deal with conditional dependence when fitting HMMs. A simulation study for the subsampling approach is included in Section 4.3.1. A more general simulation study comparing the subsampling approach, basis function model, and autoregressive model is included in Section 4.3.2. Section 4.4 shows the application of these modeling approaches to high-frequency lake phycocyanin data. Lastly, in Section 4.5 we discuss future work related to conditional dependence modeling.

4.2 Dependence modeling

HMM and HSMM models have two components, each of which are stochastic processes. One is a Markov chain of hidden states and the other is a time series of observed data, which are conditional on the hidden states (Rabiner, 1989; Yu, 2010). The difference between them is that in the HSMM case the duration in a state is modeled explicitly, since, unlike the HMM where we have one observation per state in the sequence, in the HSMM we have several observations being emitted by each state in the sequence. In the discrete-time case, the number of observations emitted corresponds to the duration in the state.

In a discrete-time HMM, the sequence of observations from time 1 to T can be denoted as $\mathbf{y} = (y_1, \dots, y_T)'$. The corresponding sequence of unobserved states is denoted as $\mathbf{s} = (s_1, \dots, s_T)'$, where $s_t \in \{1, 2, \dots, S\}$, $t = 1, \dots, T$, and S is the total number of unique states. Observations are emitted by each of the states in the hidden sequence following a state-dependent probability distribution, which is usually a Gaussian distribution with the respective mean and variance for each state, that is $y_t \sim N\left(\mu_{s_t}, \tau^2_{s_t}\right)$, with $\mu_{s_t} \in \{\mu_1, \cdots, \mu_S\}$ and $\tau^2_{s_t} \in \{\tau^2_1, \cdots, \tau^2_S\}$.

Both the HMM and HSMM typically rely on a conditional independence assumption in the emission distribution. That is, conditioned on the true (latent) state, the observations are independent of each other. This is a simplifying assumption that is based on historical use of such models for observation sequences that were fairly low frequency. However, given the influx of high-frequency data from modern sensors (e.g., wearable devices), this conditional independence assumption may not be tenable, and one must address the potential for dependence in the emission distribution conditioned on the true state process.

The next subsections outline options for mitigating and accommodating conditional dependence in the observation sequence. These include a model with an autoregressive error structure (AR(1)), a model with random effects represented by a basis function expansion, and a new approach based on data subsampling. The AR case explicitly models the dependence through an autocorrelation parameter. The basis function model approximates the dependence in the data functionally through a combination of basis functions with random coefficients. Lastly, the subsampling approach considers a standard HMM (or HSMM) but mitigates the dependence by thinning the observation sequence at each iteration of the sampling algorithm (in a Bayesian setting).

4.2.1 AR structure model

Several applications of HMM and HSMM have already considered an autoregressive structure in the emission model. Ruiz-Suarez et al. (2022) summarizes how the state-dependent distributions in HMM and HSMM can include autoregressive processes of order p (AR(p)-HMM and AR(p)-HSMM). They use these models, along with standard HMM and HSMM, to compare the error classification in supervised learning of sheep behavioral states given that their observed accelerometer data is collected at a high frequency, and thus has high temporal dependence. In addition, Xu and Liu (2021) incorporate a vector autoregressive structure in an analysis of multivariate financial time-series data, given that financial returns data appear to have temporal dependence.

For our purposes of comparison, we consider here in (4.1) a univariate state-dependent distribution with an AR(1) structure, where the autocorrelation parameter is denoted as ψ. More generally, we can allow the autocorrelation parameter to be state-specific (i.e., ψ_{s_t}).

$$y_t = \mu_{s_t} + \psi_{s_t} \left(y_{s_{t-1}} - \mu_{s_{t-1}} \right) + \theta_t, \qquad \theta_t \sim N(0, \tau^2_{s_t}). \tag{4.1}$$

4.2.2 Basis function model

Here, we account for the conditional dependence in the observation sequence through the use of basis function expansions with random expansion coefficients. Basis functions are applied in several contexts to model dependence in time and space (e.g. Wikle et al., 2019; Hefley et al., 2017). However, this nonparametric approach has not been applied in the context of HMM and HSMM with the goal of mitigating the violation of conditional independence. By using a combination of basis functions, we do not aim to explicitly model the residual autocorrelation in the data, but rather to approximate the dependence in a flexible functional manner.

We consider a general model with this nonparametric structure as presented in (4.2). Here we have K temporal basis functions, which we denote as $\phi_{k,t}$ where $k = 1, \ldots, K$. The random coefficients associated with these functions are denoted c_k. In our simulations, we model these coefficients using independent Normal distributions, but other choices could be considered. Similar to the AR model, these coefficients can be shared across all observations or defined to be state-specific (i.e., c_{k,s_t}).

$$y_t = \mu_{s_t} + \sum_{k=1}^{K} c_k \phi_{k,t} + \theta_t, \qquad \theta_t \sim N(0, \tau_{s_t}^2). \tag{4.2}$$

4.2.3 Subsampling

We propose an alternative method to these model-based approaches for handling dependence in the observation data for situations where one does not expect the residual dependence to affect the latent state. Specifically, we model the observations as in a standard HMM or HSMM, but mitigate the dependence by thinning the observation sequence, which we denote as $\tilde{\mathbf{y}}$ (4.3). Here, thinning refers to randomly subsampling the data to diminish the dependence. In a Bayesian setting, where we have an MCMC algorithm to sample from the conditional distributions of our emission distribution parameters, we perform this random subsampling at each iteration of the MCMC algorithm, such as in approaches like Firefly Monte Carlo (Maclaurin and Adams, 2015).

$$\tilde{y}_t = \mu_{s_t} + \theta_t, \qquad \theta_t \sim N(0, \tau^2_{s_t}), \tag{4.3}$$

where $\tilde{y}_t \in \tilde{\mathbf{y}}_t$ is a subset of size $\tilde{n}$ from $1, \ldots, T$.

4.3 Simulation study

We consider two simulation studies here. The first looks at the proposed subsampling approach since this is new to the HMM and HSMM literature. This is then followed by a more general robustness study that compares subsampling as well as the AR(1) and basis function approaches described in Section 4.2.

4.3.1 Subsampling simulation study

As an illustration of the subsampling approach, we consider a simulation study to investigate the effect of the conditional independence assumption violation in the emission distribution parameter estimation, as well as to determine how the subsampling approach helps in the estimation when this violation is present. We simulate dependent data using an AR(1) model under several scenarios. Then, we fit an HSMM model to each generated dataset using a data subsampling approach, where at each iteration of the MCMC algorithm, an independent and random subset of the data is selected according to a specified sampling rate, and we will have a new observation sequence denoted by $\tilde{\mathbf{y}}$ of size $\tilde{n}$. Once the subsampled observation sequence is defined, we sample the state means and variances from their full conditional distribution ((4.4) and (4.5), respectively). Using a Normal prior distribution with parameters θ_μ and λ_μ^2, the full conditional for each mean is:

$$\mu_j|\cdot \sim N\left(\frac{\frac{\sum_{t=1}^{T} y_t I(S_t=j, y_t \in \tilde{\mathbf{y}})}{\tau_j^2} + \frac{\theta_\mu}{\lambda_\mu^2}}{\frac{\tilde{n}_j}{\tau_j^2} + \frac{1}{\lambda_\mu^2}}, \frac{1}{\frac{\tilde{n}_j}{\tau_j^2} + \frac{1}{\lambda_\mu^2}}\right), \tag{4.4}$$

where $I(\cdot)$ is the indicator function and $\tilde{n}_j$ is the total number of observations of vector $\tilde{\mathbf{y}}$ emitted by state j. Using an inverse-Gamma (IG) prior with parameters θ_{τ^2} and λ_{τ^2}, the full conditional for each variance is:

$$\tau_j^2|\cdot \sim IG\left(\theta_{\tau^2} + \frac{\tilde{n}_j}{2}, \lambda_{\tau^2} + \frac{1}{2}\sum_{t=1}^{T}\left(y_t I(S_t = j, y_i \in \tilde{\mathbf{y}}) - \mu_j\right)^2\right). \tag{4.5}$$

The different scenarios included in this simulation study are defined according to sample size, autocorrelation parameter, and number of states. Three different number of states are considered (S = 2, 3 and 4), while the possible values for the autocorrelation parameter are 0.25, 0.50, and 0.75, as well as a case with no autocorrelation. Although the number of simulated observations varies across all realizations, we consider two cases consisting of approximately 2500 and 5000 observations. Overall, 24 scenarios were considered, each with 100 realizations.

The procedure for simulating the data is as follows. First, the initial state, s_1, is sampled from its distribution. Next, the first duration, d_1, is generated from the corresponding zero-truncated Poisson distribution. The observations, $y_1, y_2, \ldots, y_{d_1}$, are generated from a Normal distribution with a specified autocorrelation. Then, the second state, s_2, is sampled conditional on s_1 according to a transition probability matrix $\mathbf{P}$. After that, the duration is sampled as well as the observations emitted by that second state. The process continues until the specified sample size is reached.

Each of the simulated realizations is modeled with an HSMM assuming independence in the observations. For ease of computation, the states and other parameters are fixed and only the emission distribution parameters are estimated.

In this approach, each iteration of the MCMC algorithm uses a different random subset of data according to a pre-specified percentage. The sampling rates considered are 100, 90, 80, ..., 10%. The 90% credible intervals (CI) for the mean and variance parameters for all 100 datasets across all scenarios and sampling rates are calculated. We assess the subsampling approach by comparing the empirical coverage, and determine the preferred data sampling rate as the one that results in the nominal coverage.

The results of the simulation study are provided in Tables 4.1 to 4.3. Overall, the correlation in the data affects the estimation of the emission distribution parameters, but the effect can be reduced by subsampling during the model fitting procedure. The empirical coverage is calculated as the percentage of 90% CIs that captured the true parameter values. Tables 4.1 to 4.3 present the empirical coverage for the different autocorrelation parameters, sample size and sampling rates utilized. This coverage corresponds to the average coverage of the different state means. The empirical coverage is similar for the two sample size and number of states scenarios.

There are two important results provided in the tables. First, as indicated by the first row of the tables where no subsampling was used, the more correlated the data are, the worse the model does in recovering the true parameter values. Second, the empirical coverage increases as the sampling rate decreases. As we reduce the percent of data used, we are able to reduce the dependence in the data and thus improve our estimates of uncertainty. For a case where the autocorrelation is approximately 0.75 and the entire dataset is utilized (i.e., sampling rate = 100%), the coverage is low, indicating the need to use a smaller sampling rate. Lastly, for data having autocorrelation close to 0.25, we obtain nominal coverage of the emission distribution means when using approximately 80% of the dataset in the MCMC sampling algorithm.

Table 4.1: Coverage percentage of the emission distribution means, 2 states case.

Sampling rate %	ψ (T≈2500)			ψ (T≈5000)		
	0.25	0.50	0.75	0.25	0.50	0.75
100	78	68	44	76	68	55
90	83	73	50	81	74	60
80	87	80	54	86	76	66
70	92	84	62	90	81	72
60	96	85	67	93	84	77
50	98	88	70	94	88	81
40	98	93	76	97	94	86
30	100	96	88	100	98	91
20	100	99	92	100	100	96
10	100	100	100	100	100	100

Table 4.2: Coverage percentage of the emission distribution means, 3 states case.

Sampling rate %	ψ (T≈2500)			ψ (T≈5000)		
	0.25	0.50	0.75	0.25	0.50	0.75
100	77	68	50	82	67	50
90	82	75	56	86	73	55
80	89	80	60	90	76	60
70	92	84	65	92	81	64
60	96	87	71	94	86	72
50	99	92	79	96	92	79
40	99	96	84	99	96	84
30	100	99	90	100	98	90
20	100	100	95	100	100	95
10	100	100	100	100	100	99

Table 4.3: Coverage percentage of the emission distribution means, 4 states case.

Sampling rate %	ψ (T≈2500)			ψ (T≈5000)		
	0.25	0.50	0.75	0.25	0.50	0.75
100	80	66	52	80	67	53
90	86	70	57	86	72	56
80	89	75	61	89	78	58
70	94	81	66	93	83	62
60	96	85	69	95	86	69
50	98	90	75	97	90	76
40	100	96	82	98	95	83
30	100	99	87	99	98	90
20	100	100	94	100	100	96
10	100	100	99	100	100	99

4.3.2 Robustness study

In this section we consider a simulation study to evaluate the effectiveness of the three approaches outlined in Section 4.2 for accommodating/mitigating dependence in HMM/HSMM emission distributions.

Data generation

We consider data generated under three different scenarios. In the scenario 1, after conditioning on the states, there is no dependence in the observation sequence. For the other two scenarios we considered dependence in the observation sequence. Specifically, in scenario 2 there is AR(1) dependence in the observations, with the same autocorrelation parameter ψ, whereas in scenario 3 we have AR(1) dependence but the autocorrelation parameter is different for each state, ψ_{s_t}. The other parameters of the model are the same for all scenarios; i.e., we considered the same transition probabilities and the same duration parameters. Once we generated the state sequence, we generated the emission data according to the three scenarios. The datasets have 2011, 2012 and 2021 observations, respectively. We also take into account measurement error and define an observation model the same way for the three scenarios:

$$y_t = \eta_{s_t} + \epsilon_t, \qquad \epsilon_t \sim N(0, \sigma^2), \tag{4.6}$$

where y_t is an observation at time t, η_t is the true emitted value at that time, and σ^2 is the measurement error variance.

The generative model in scenario 1 defines the true observation at time t, using a Normal distribution with a state-specific mean, μ_{s_t} and state-specific variance, $\tau^2_{s_t}$. The model is given by:

$$\eta_t = \mu_{s_t} + \theta_t, \qquad \theta_t \sim N(0, \tau^2_{s_t}). \tag{4.7}$$

In the second scenario, we have general dependence. We assume a Normal distribution where the mean corresponds to the mean of the state at time t plus the autocorrelation component. The variance is also state specific. The model is written:

$$\eta_t = \mu_{s_t} + \psi \left(\eta_{t-1} - \mu_{s_{t-1}}\right) + \theta_t, \qquad \theta_t \sim N(0, \tau^2_{s_t}). \tag{4.8}$$

Lastly, scenario 3 expands on the scenario 2 by considering a state-dependent autocorrelation parameter:

$$\eta_t = \mu_{s_t} + \psi_{s_t} \left(\eta_{t-1} - \mu_{s_{t-1}}\right) + \theta_t, \qquad \theta_t \sim N(0, \tau^2_{s_t}). \tag{4.9}$$

Again, we retain state specific mean and variance parameters

Models applied to generated data

We fitted six different models to the datasets generated under the three scenarios. The first two models assume conditional independence (4.10).

$$\eta_t = \mu_{s_t} + \theta_t, \qquad \theta_t \sim N(0, \tau^2_{s_t}). \tag{4.10}$$

In the first case, we obtain parameter estimates as in a standard HSMM. In the second, we introduce subsampling in the sampling algorithm when obtaining posterior draws of the emission distribution parameters.

We then fitted models that consider dependence in the observations. Two models assume the data have an AR(1) structure and another two models employ basis function expansions. The AR(1) models include a model where the autocorrelation parameter is independent of the state (4.11) and another where the autocorrelation parameter is dependent on the state (refeqmod4).

$$\eta_t = \mu_{s_t} + \psi \left(\eta_{t-1} - \mu_{s_{t-1}}\right) + \theta_t, \qquad \theta_t \sim N(0, \tau^2_{s_t}). \tag{4.11}$$

$$\eta_t = \mu_{s_t} + \psi_{s_t}\left(\eta_{t-1} - \mu_{s_{t-1}}\right) + \theta_t, \qquad \theta_t \sim N(0, \tau^2_{s_t}). \tag{4.12}$$

The basis function models use Fourier basis functions (Wikle et al., 2019). In one case, the coefficients are independent of the state (4.13) and in the other we have state-dependent coefficients (4.14). In both models we considered 6 basis functions ($K = 6$). In both models, we assume independent Normal distributions with mean 0 and variance $\rho^2 = 100$ for each basis function coefficient.

$$\eta_t = \mu_{s_t} + \sum_{k=1}^{K} c_k \phi_{k,t} + \theta_t, \qquad \theta_t \sim N(0, \tau^2_{s_t}). \tag{4.13}$$

$$\eta_t = \mu_{s_t} + \sum_{k=1}^{K} c_{k,s_t} \phi_{k,t} + \theta_t, \qquad \theta_t \sim N(0, \tau^2_{s_t}). \tag{4.14}$$

Results of models applied to the generated data

After fitting the 6 models under each of the 3 data scenarios, we obtained the posterior mean and credible intervals (CIs) of the emission distribution parameters. Figure 4.1 shows the estimates for the state means, where the true means used to generate the data in each scenario are marked with a horizontal line. Figure labels M1 through M6 correspond to the models introduced in Section 4.3.2 in their respective order.

For scenario 1, with no dependence in the observation sequence we observe that, as expected, the first model (M1) captures the true state values in the CIs, and so do all the other models, except model 6 (M6). We note that the subsampling approach (M2) produces wider credible intervals. Again, this is expected because we have less (independent) data available for the estimation.

In the second data scenario, in which the generative model matches model M3,

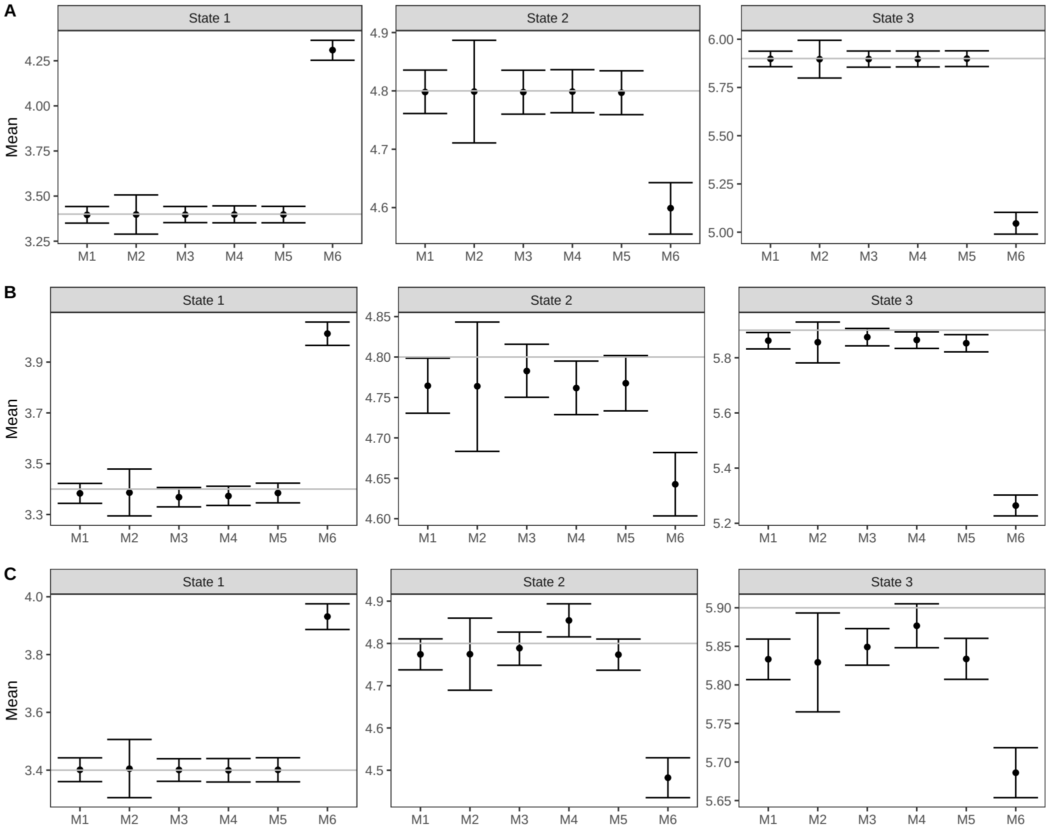

Figure 4.1: Credible intervals (bars) and posterior means (points) of emission distribution mean by state and scenario. Panels **A, B & C** correspond to the three data scenarios. The horizontal lines represent the true means.

we see that this third model captures the true values as expected. It is worth noting that model M2 performs equally as well as M3. The model that does not account for the dependence (M1) underestimates the true mean of states 2 and 3, similar to M4. The basis function model M5 captures the true means of states 2 and 3, but underestimates the mean of state 3. Again, M6 does not recover the true means.

All models, except M6, capture the true means of the first state in the third data scenario. The second state mean is captured by all models except M4 and M6, but model M4 does capture the third state mean which is missed by all the other models. Model M4 is the same model used to generate the data in this last scenario, and we would have expected it to perform better than the rest of models. There appears to be a trade-off in this model in terms of parameter estimation for states 2 and 3, given that the autoregressive coefficients are over/underestimated for the second and third states, respectively (see Table C.1.1 in the Appendix).

In Figure 4.2, the true variance for each state is indicated with a horizontal line. In the first scenario, all models can recover the true variance for all states, except M6 which only recovers the second state variance. In the other two scenarios, several models fail to capture some of the true values. The credible intervals obtained in models M1, M2, and M5 capture most of the true variances, but the autoregressive models (M3 and M4) fail to capture all three variances in scenarios 2 and 3. We expected these AR models to perform equally or better as the other models that do not directly model the autocorrelation in the emission distribution.

4.4 Real data example

This section presents an analysis of a real-world dataset using the subsampling and the flexible basis function approach. Recall, subsampling would typically be applied when one does not believe *a priori* that conditional dependence will affect the estimation of the state process, while the basis function approach would be used if one thought this dependence could affect the state process. Here, our goal is to illustrate the methods on real-world data and to examine the differences in the associated posterior inference.

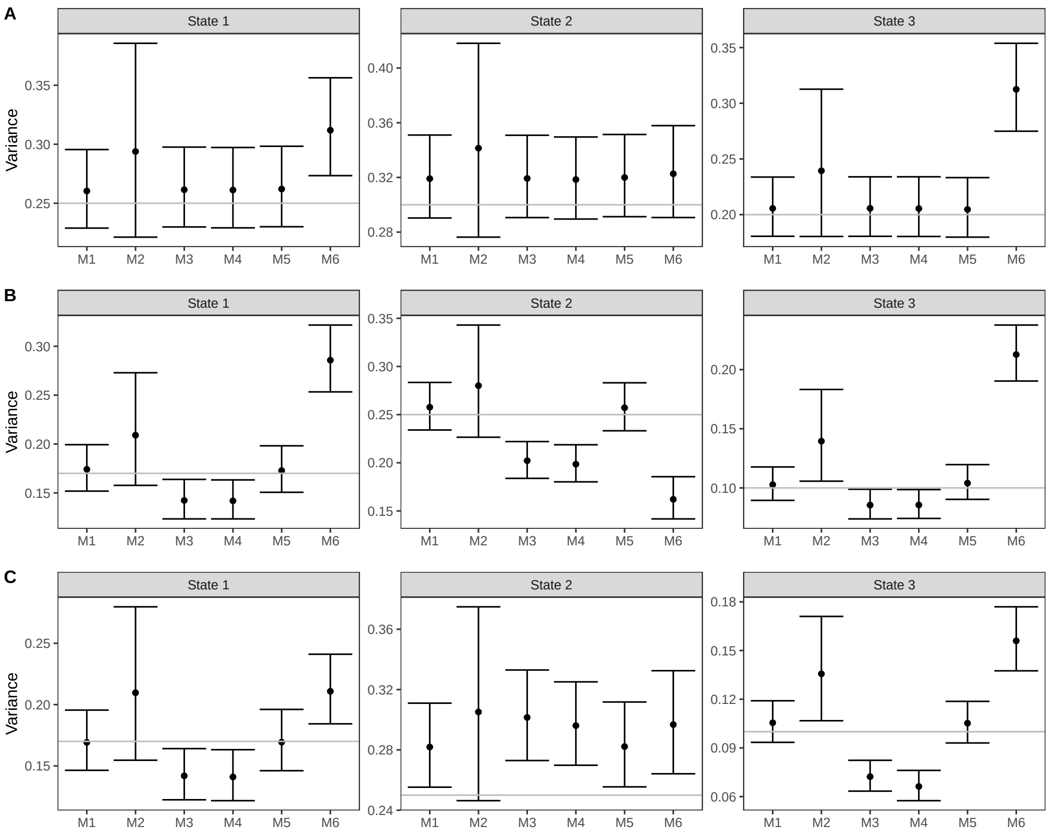

Figure 4.2: Credible intervals (bars) and posterior means (points) of emission distribution variance by state and scenario. Panels **A, B & C** correspond to the three data scenarios. The horizontal lines represent the true variances.

4.4.1 Lake Mendota phycocyanin data

We illustrate the use of two of the models presented in Section 4.2 with high-frequency phycocyanin measurements from Lake Mendota, Wisconsin, in 2018. These data were analyzed in Chapter 2, where we fitted a HSMM to explore the temporal variation in concentration levels of cyanobacteria as well as identify possible environmental drivers of this variation. Recall that phycocyanin is an indicator of cyanobacteria concentration in the lake. High levels of cyanobacteria can lead to adverse health effects in both humans and the environment. The data can be found in the North Temperate Lakes Long-Term Ecological Research program database (Lead PI et al., 2020). In 2018, sensors on an instrumented buoy located in the lake recorded measurements from April 11 to November 15. In our analysis, we model hourly measurements of phycocyanin in Lake Mendota during the data collection period.

4.4.2 HSMM with subsampling applied to phycocyanin data

To determine the optimal subsampling rate for our analysis, we first fitted the HSMM model without subsampling and determined an approximate value for the size of the segments. The durations within each state varied by state and through time. The average duration ranged from 21 to 55. We divided the data into segments to resemble the groups of emitted data by a state and calculated the mean autocorrelation across all groups. By considering group sizes ranging between 21 to 55, we obtained autocorrelation estimates ranging between 0.65 to 0.80. We compared these estimates to those shown in Table 4.2 in order to identify an optimal subsampling rate. Using $\psi = 0.75$ suggests a subsampling rate of approximately 30% for estimating the emission distribution parameters. The remaining results presented in this subsection assume $\psi = 0.75$ and a subsampling rate of 30%.

Table 4.4 presents the posterior means and CIs for the estimation of the emission distribution parameters under subsampling. We modeled the duration in each state with a zero-truncated Poisson (ZTP) distribution, where the intensity parameter was state-specific. The posterior mean and CIs for these estimates are also given in Table 4.4. The states of cyanobacteria concentration identified with this model can be classified as low, medium and high. These levels can be associated with the regimes found in Carpenter et al. (2020).

Table 4.4: Posterior mean and credible intervals for emission distribution and duration parameters of HSMM with subsampling for phycocyanin data

State	Mean	Variance	ZTP parameter
S1	-2.26 (-2.35,-2.16)	0.69 (0.60,0.80)	79.2 (73.9,85.3)
S2	-0.13 (-0.24,-0.01)	0.50 (0.40,0.60)	34.0 (29.8,37.3)
S3	1.60 (1.52,1.68)	0.53 (0.45,0.61)	135.9 (126.3,144.2)

Figure 4.3 shows the estimated cyanobacteria state sequence. There is a clear distinction between the states, and we can see there is more variability among the observations of the low cyanobacteria state, while the medium and high states have observations with similar variances, which are smaller than in the low state. The middle state, has less duration on average than the other two states, and we can identify that the lake is at high levels of cyanobacteria for longer periods of time.

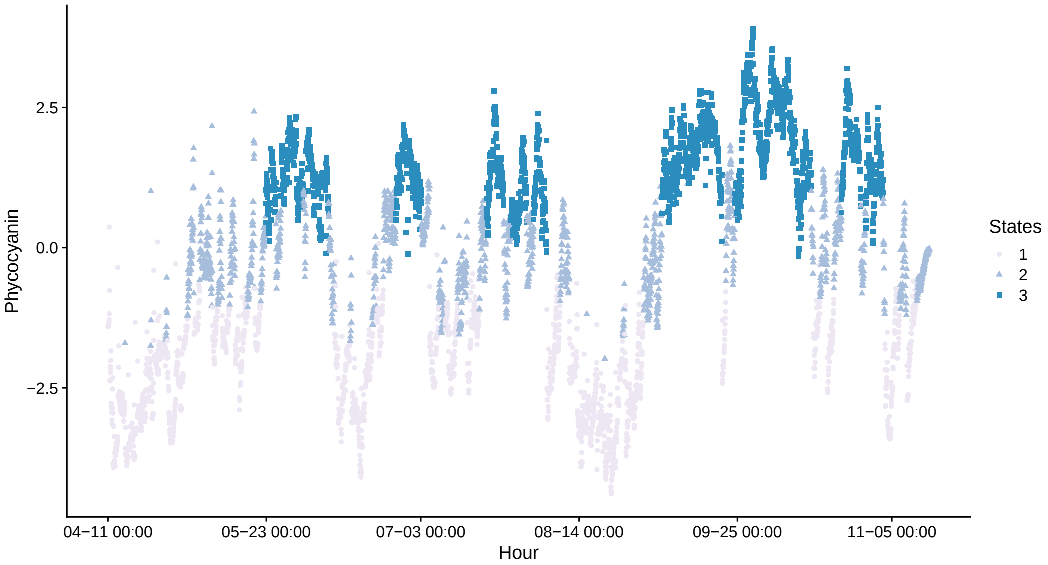

Figure 4.3: Estimated cyanobacteria state sequence with HSMM with subsampling

We fitted a HSMM that does not account for conditional dependence, and show the results in Table 4.5 and Figure 4.4 for comparison. The state sequence is very similar, and although the parameter estimates are not the same, they are close. In general, with this standard approach we see narrower CIs than with the HSMM with subsampling, as expected since the effective degrees of freedom are less than the full sample size given conditional dependence.

Table 4.5: Posterior mean and credible intervals for emission distribution and duration parameters of HSMM for phycocyanin data

State	Mean	Variance	ZTP parameter
S1	-2.19 (-2.23,-2.15)	0.55 (0.49,0.62)	80.53 (76.6,85.39)
S2	0.03 (-0.01,0.08)	0.23 (0.18,0.29)	30.47 (28.63,32.47)
S3	1.69 (1.66,1.72)	0.28 (0.22,0.33)	97.9 (93.2,102.75)

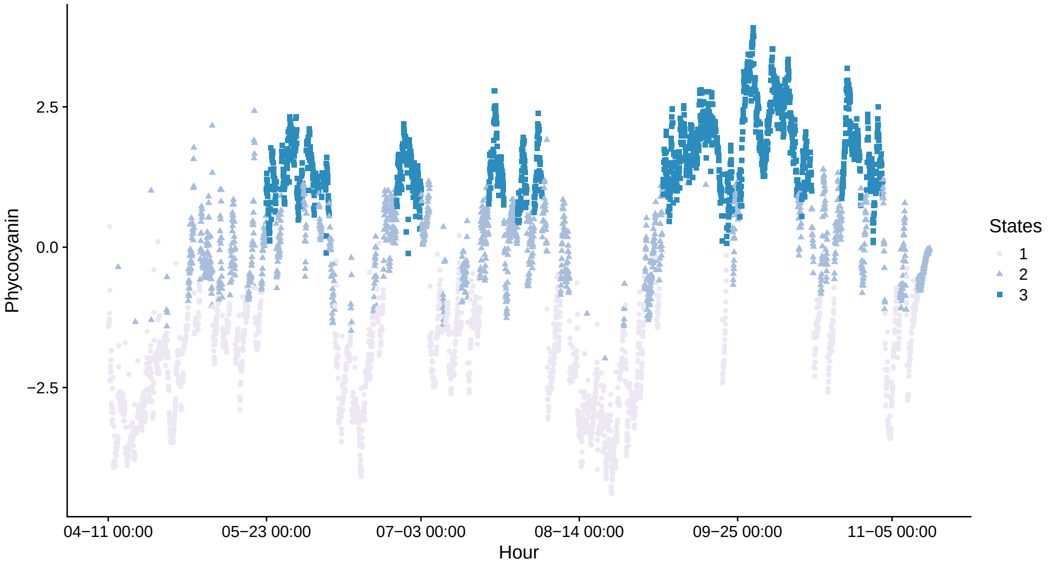

Figure 4.4: Estimated cyanobacteria state sequence with HSMM

4.4.3 HSMM with basis functions applied to phycocyanin data

Next, we applied the basis function approach for HSMM in (4.13) to the phycocyanin observations. In this model, we used the same number of states as in Section 4.4.2 as well as the same state-specific ZTP distribution for the durations. We used 10 Fourier basis functions and assumed the random coefficients were the same across the three states. Table 4.6 presents the posterior means and CIs for the estimation of emission distribution parameters, as well as the duration parameters. The basis function coefficient estimates are given in Table 4.7. Note that the basis coefficients are essentially nuisance parameters in this analysis. However, if it was of interest, one could evaluate the sample cloud of implied time series from the basis expansion, or consider the time series implied by posterior mean coefficients.

Table 4.6: Posterior mean and credible intervals for emission distribution and duration parameters of HSMM with basis functions for phycocyanin data

State	Mean	Variance	ZTP parameter
S1	-1.87 (-1.91,-1.82)	0.37 (0.33,0.41)	66.66 (62.95,70.85)
S2	-0.27 (-0.32,-0.23)	0.17 (0.14,0.2)	27.7 (26.03,29.42)
S3	0.95 (0.93,0.98)	0.14 (0.12,0.17)	67.4 (64.09,71.57)

The parameter estimation is different than in the HSMM with subsampling. As expected given our simulation study in Section 4.3.2, the credible intervals are narrower than in the subsampling HSMM. The state-variance estimates are smaller with the basis function model and in Figure 4.5 we see more overlap in the estimated state sequence. This model differs in all parameter estimates from the subsampling case. In both models we observe that the middle state, which is a transition state between the two regimes identified in Carpenter et al. (2020), has smaller duration than the

Table 4.7: Posterior mean and credible intervals for basis functions coefficients

Coefficient	Estimation
1	-21.59 (-22.89,-20.28)
2	-9.46 (-10.73,-8.18)
3	-42.67 (-44.21,-41.14)
4	-42.2 (-43.67,-40.73)
5	7.29 (5.74,8.91)
6	-40.42 (-41.65,-39.19)
7	-27.92 (-29.27,-26.6)
8	11.65 (10.16,13.21)
9	-18.97 (-20.6,-17.48)
10	-0.28 (-1.56,1.03)

other two states.

Both approaches are helpful in identifying the low, medium and high levels of cyanobacteria in the lake, their mean and variance, and the length of time the lake stays in each of these states. If the residual dependence is not thought to be tied to what the true biological states are, we can use the inference from the subsampling approach.

4.5 Discussion

We explored options for mitigating and accommodating conditional dependence in HSMMs, given that this assumption violation affects the estimation of the emission distribution parameters. Markov-switching models have been more commonly used for this purpose but we have proposed new ways to deal with this dependence.

In general, incorporating basis function expansion random effects has proven successful in dealing with temporal dependence in data. It was then a reasonable step to consider the use of basis functions in the context of HMM and HSMMs for the same

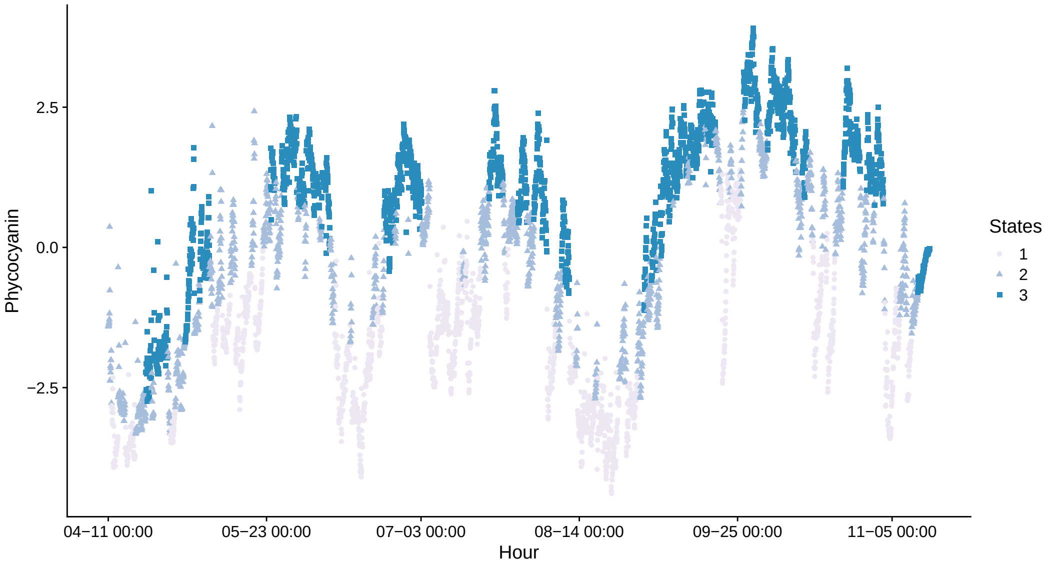

Figure 4.5: Estimated cyanobacteria state sequence with HSMM with basis functions

purpose. The basis function model offers flexibility to address the dependence in this context, without directly having to model the autocorrelation in the data.

From our subsampling simulation study, we confirmed that the autocorrelation in the data affects parameter estimation. In fact, as the autocorrelation increases, the credible intervals are less likely to capture the true parameter values. Subsampling with decreased sampling rates can be employed to compensate for these effects. In the robustness study used to compare the different models we see that, in terms of mean estimation, the model with subsampling and the basis function model without state-specific coefficients perform better than the autoregressive models and the model that ignores the residual dependence. In the variance estimation, we observed that the autoregressive models do not capture the true parameters, even when the data actually have an autoregressive error structure.

Ruiz-Suarez et al. (2022) show that for classification purposes, models with an AR structure generally perform similarly to models that ignore dependence. Even though classification was not the goal of our analysis, we saw similar results in all our models, except M6. The classification error was 2% or less in models M1 through M5, except for the autoregressive model with state-specific autocorrelation parameters (M4) in data scenario 3, where it was 5%. Even though we usually consider unsupervised applications and concentrate only on inference of the emission distribution parameters as in this chapter, it is interesting to note that the different approaches have comparable state sequence estimations.

In the real-world data application we compared the subsampling (M2) and the basis function (M5) approaches. Parameter inference is not the same with these two models, and the basis function model shows more overlap in the estimated states and the categorization of states as low, medium and high concentration of cyanobacteria is less evident. The subsampling approach produces wider credible intervals for the estimated parameters and the states are more clearly separated in terms of their means, but this also implies higher state variances. The basis function model, on the other hand, shows more overlap in the estimated state sequence, but less state variance. The subsampling approach is recommended in this case given that the residual dependence prevents identification of scientific meaningful states.

Although our simulation studies presented cases with autoregressive dependence, real-world data are likely to have more complex dependence. As future work, we will explore how our proposed subsampling method and the basis function model perform in more complex settings. We expect this will also require the use of more complex basis functions such as multi-scale bases (Wikle et al., 2019).

Currently, we estimate the state sequence at each iteration of our sampling algorithm. There are other Bayesian approaches (as well as frequentist) for HMMs and HSMMs that do not require the state sequence estimation. We will examine possibilities for incorporating subsampling into these other estimation approaches.

Chapter 5

Summary and Future Work

The two main subjects addressed in this dissertation were state duration modeling and solutions to conditional dependence in the emission distribution of HMMs and HSMMs. We developed a model that allows for state duration modeling in cases where the duration distribution parameters cannot be considered constant over time by defining these parameters as functions of time-varying covariates. We also presented a framework for analyzing team sports location data. The focus of this work was in estimating the underlying network of players and exploring the covariates associated with the probability of connections between pairs of players. Lastly, we provide alternative approaches to modeling and mitigating conditional dependence in the emission distribution. However, these methods and models can be further extended. In the following paragraphs, we briefly point out some of the future work related to these efforts.

In the model in Chapter 2, we assumed a zero-truncated Poisson distribution for the state durations. There are other distributions that can be considered and this choice is application specific. The parameters of the duration distribution can also be defined as functions of time-varying covariates. One possible choice might be another discrete distribution, such as a negative binomial, or a more complex distribution like a geometric mixture.

An important consideration in HMMs and HSMMs is the number of states, which is usually determined through information criteria or expert knowledge (Liu and Song, 2020). In our application in Chapter 2, we relied on expert knowledge and the DIC to inform our decision. However, information criteria can lead to selecting a higher number of states given that they under-penalize model complexity, as mentioned in Celeux and Durand (2008) and in the discussion of Spiegelhalter et al. (2002). We are interested in exploring an *ad hoc* approach that seemed to perform well in preliminary analyses. Our proposed approach uses an MCMC convergence diagnostic as an alternative to choosing the appropriate number of states. This convergence measure is the multivariate potential scale reduction factor (MPSRF) discussed in Brooks and Gelman (1998). Our experience has shown that multiple chains do not converge when the number of states is too large, and thus, favor models with fewer states. As future research We would like to explore this approach to model selection more formally.

We can extend the methods developed in Chapter 3 to coupled hidden Markov models (CHMMs), which considers several HMMs jointly (Brand, 1997). In these models, the probability of transition in a one model depends on not only its current state but also the current state of the other models. Following the notation in Touloupou et al. (2019), for a chain $c \in \{1, 2, \ldots, C\}$ at time $t \in \{1, 2, \ldots, T\}$, $s_t^{[c]}$ is the hidden state and $y_t^{[c]}$ is an observation emitted by that state. The transition probability is defined as $P(s_t^{[c]} = j | s_{t-1}^{[c]} = i, \mathbf{s}_{t-1}^{[-c]})$, where i, j belong to the state space with S states, and $\mathbf{s}_{t-1}^{[-c]}$ is the vector of the states at time $t-1$ of all the chains except chain c. In modeling team sport data, we can explore the inclusion of states of other pairs of players as covariates in the function of the probability of connection. This approach is similar to the approach taken in Dong et al. (2012) who extend the CHMM to include dependencies from a network in the graph-coupled hidden Markov model (GCHMM). Other extensions of the model in Chapter 3 would be to consider the case where the probability of connection between two players depends on group membership. This is similar to stochastic block models Snijders and Nowicki (1997). Lastly, one could focus on the structure of the network as a whole when defining the states of the HMM rather than on pairs of players.

Appendix A

Supplement for Chapter 2

A.1 Sampling algorithm

Algorithm 1 MCMC sampling algorithm

```
Initial values
Define initial values for parameters $\boldsymbol{\rho}^{(0)}, \mathbf{P}^{(0)}, \mathbf{B}^{(0)}$.
Generate the state sequence $\mathbf{S}^{(0)}$ as follows:
    a: Sample the first state $S_{q=1}^{(0)}$ using $\boldsymbol{\rho}^{(0)}$.
    b: Calculate $\phi_{S_{q=1}}^{(0)}$ using $\mathbf{X}$ and $\boldsymbol{\beta}_{S_{q=1}}^{(0)}$.
    c: Sample $\tau_1^{(0)}$ from a zero-truncated Poisson with parameter $\phi_{S_{q=1}}^{(0)}$.
    d: Define $\mathbf{S}_{1:T_1} = S_{q=1}$.
    e: Sample $S_{q=2}^{(0)}$ conditional on $S_{q=1}^{(0)}$ using $\mathbf{P}^{(0)}$.
    f: Calculate $\phi_{S_{q=2}}^{(0)}$ using $\mathbf{X}$ and $\boldsymbol{\beta}_{S_{q=2}}^{(0)}$.
    g: Sample $\tau_2^{(0)}$ from a zero-truncated Poisson with parameter $\phi_{S_{q=2}}^{(0)}$.
    h: Define $\mathbf{S}_{T_1+1:T_2} = S_{q=2}$.
    i: Continue until $T_q = n$.
Calculate $\boldsymbol{\mu}^{(0)}$ and $\boldsymbol{\sigma}^{2(0)}$ based on $\mathbf{S}^{(0)}$.

Iterations
for iteration $l = 1, 2, \ldots$ do
    Update $\boldsymbol{\rho}^{(l-1)}$ using Gibbs sampling:
        $\boldsymbol{\rho}^{(l)} \sim Dir\left(I\left(S_1^{(l-1)} = 1\right) + \theta_{\rho_1}, \ldots, I\left(S_1^{(l-1)} = M\right) + \theta_{\rho_M}\right),$
    where $I(\cdot)$ is the indicator function.
    Update the $j$-th row of $\mathbf{P}^{(l-1)}$ for $j = 1, 2, \ldots, M$, using Gibbs sampling:
        $\mathbf{P}_j^{(l)} \sim Dir\left(n_{j1} + \theta_{P_{j1}}, \ldots, n_{jM} + \theta_{P_{jM}}\right),$
    where $n_{jk} = \sum_{q=1}^{Q-1} I\left(S_q^{(l-1)} = j, S_{q+1}^{(l-1)} = k\right)$ is the total number of transitions from state $j$ to state $k$.
```

Algorithm 1 MCMC sampling algorithm (continued)

```
    Update $\boldsymbol{\beta}_j^{(l-1)}$ for $j = 1, 2, \ldots, M$ using random-walk Metropolis. Sample $\mathbf{z} \sim N(\mathbf{0}, \kappa_{\beta_j}^2 \mathbf{I}_{r+1})$, where $I_{r+1}$ is an identity matrix of order $r+1$, and define the proposal vector $\boldsymbol{\beta}_j^{(*)} = \boldsymbol{\beta}_j^{(l-1)} + \mathbf{z}$. Then calculate the Metropolis ratio as:
    $$m_{\boldsymbol{\beta}_j} = \left( \frac{\prod_{q=1}^{Q} ZTP\left(\tau_q^{(l-1)} \mid \phi_q^{(*)}\right)}{\prod_{q=1}^{Q} ZTP\left(\tau_q^{(l-1)} \mid \phi_q^{(l-1)}\right)} \right) \times \left( \frac{N\left(\boldsymbol{\beta}_j^{(*)} \mid \boldsymbol{\theta}_{\beta_j}, \lambda_{\beta_j}^2 \mathbf{I}_{r+1}\right)}{N\left(\boldsymbol{\beta}_j^{(l-1)} \mid \boldsymbol{\theta}_{\beta_j}, \lambda_{\beta_j}^2 \mathbf{I}_{r+1}\right)} \right),$$
    and if $u < m_{\boldsymbol{\beta}_j}$, with $u \sim Unif(0,1)$, let $\boldsymbol{\beta}_j^{(l)} = \boldsymbol{\beta}_j^{(*)}$, and update $\boldsymbol{\phi}^{(l-1)}$, $\boldsymbol{\phi}^{(l)} = \boldsymbol{\phi}^{(*)}$.
    Update $\mathbf{S}^{(l-1)}$ and $\boldsymbol{\tau}^{(l-1)}$ with the sampler in Johnson and Willsky (2013) for the finite HSMM.
    Update $\mu_j^{(l-1)}$ for $j = 1, 2, \ldots, M$ using Gibbs sampling:
    $$\mu_j \sim N\left( \frac{\frac{\sum_{i=1}^{n} y_i I(S_i = j)}{\sigma_j^{2(l-1)}} + \frac{\theta_\mu}{\lambda_\mu^2}}{\frac{n_j}{\sigma_j^{2(l-1)}} + \frac{1}{\lambda_\mu^2}}, \frac{1}{\frac{n_j}{\sigma_j^{2(l-1)}} + \frac{1}{\lambda_\mu^2}} \right),$$
    where $I(\cdot)$ is the indicator function and $n_j$ is the total number of observations emitted by state $j$.
    Update $\sigma_j^{2(l-1)}$ for $j = 1, 2, \ldots, M$ using Gibbs sampling:
    $$\sigma_j^2 \sim IG\left( \theta_{\sigma^2} + \frac{n_j}{2}, \lambda_{\sigma^2} + \frac{1}{2} \sum_{i=1}^{n} \left( y_i I(S_i = j) - \mu_j^{(l-1)} \right)^2 \right).$$
    Save $\boldsymbol{\rho}^{(l)}, \mathbf{P}^{(l)}, \mathbf{B}^{(l)}, \boldsymbol{\phi}^{(l)}, \mathbf{S}^{(l)}, \boldsymbol{\tau}^{(l)}, \boldsymbol{\mu}^{(l)}$ and $\boldsymbol{\sigma}^{2(l)}$.
end for
```

Appendix B

Supplement for Chapter 3

B.1 Posterior distribution

$$
\begin{aligned}
\big[&\boldsymbol{\mu}, \mathbf{W}, \tau^2, \gamma, \eta, c, \alpha, \sigma^2, p_1, \boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)} \mid \mathbf{s}\big] \\
&\propto [\mathbf{s} \mid \boldsymbol{\mu}, \tau^2]\ [\boldsymbol{\mu} \mid \mathbf{W}, \gamma, \eta, c, \alpha, \sigma^2]\ [\mathbf{W} \mid p_1, \boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)}] \\
&\times [\tau^2|\alpha_\tau, \beta_\tau]\ [\gamma|\mu_\gamma, \sigma_\gamma^2]\ [\eta|\mu_\eta, \sigma_\eta^2]\ [c|a_c, b_c]\ [\alpha|a_\alpha, b_\alpha]\ [\sigma^2|\alpha_\sigma, \beta_\sigma] \\
&\times [p_1|a_{p_1}, b_{p_1}]\ [\boldsymbol{\beta}^{(0)}|\mu_{\boldsymbol{\beta}^{(0)}}, \sigma^2_{\boldsymbol{\beta}^{(0)}}]\ [\boldsymbol{\beta}^{(1)}|\mu_{\boldsymbol{\beta}^{(1)}}, \sigma^2_{\boldsymbol{\beta}^{(1)}}]
\end{aligned}
$$

B.2 Full conditional distributions

τ^2

$$\tau^2|\cdot \sim IG(\alpha_\tau^*, \beta\tau^*)$$

$$\alpha_\tau^* \equiv \alpha_\tau + nT$$

$$\beta_\tau^* \equiv \beta_\tau + \frac{1}{2}\sum_{t=1}^{T} \left(\mathbf{s}(t) - \boldsymbol{\mu}(t)\right)' \left(\mathbf{s}(t) - \boldsymbol{\mu}(t)\right)$$

γ

$$\gamma|\cdot \sim N(\mu_\gamma^*, \sigma_\gamma^{2*})$$

$$\sigma_\gamma^{2*} \equiv \left(\frac{1}{\sigma_\gamma^2} + \sum_{t=2}^{T} \widetilde{\boldsymbol{\mu}}(t-1)'\mathbf{Q}(t)\widetilde{\boldsymbol{\mu}}(t-1)\right)^{-1}$$

$$\mu_\gamma^* \equiv \sigma_\gamma^{2*}\left(\mu_\gamma + \sum_{t=2}^{T} \widetilde{\boldsymbol{\mu}}(t-1)'\mathbf{Q}(t)(\boldsymbol{\mu}(t) - \boldsymbol{\mu}(t-1) - \eta\mathbf{d}(t-1))\right)$$

Here $\mathbf{Q}(t) = \sigma^{-2}\mathbf{K}(t)$, with $\mathbf{K}(t) \equiv (\mathbf{W}_+^c(t) - \alpha\mathbf{W}(t)) \otimes \mathbf{I}_2$.

η

$$\eta|\cdot \sim N(\mu_\eta^*, \sigma_\eta^{2*})$$

$$\sigma_\eta^{2*} \equiv \left(\frac{1}{\sigma_\eta^2} + \sum_{t=2}^{T} \mathbf{d}(t-1)'\mathbf{Q}(t)\mathbf{d}(t-1)\right)^{-1}$$

$$\mu_\eta^* \equiv \sigma_\eta^{2*}\left(\mu_\eta + \sum_{t=2}^{T} \mathbf{d}(t-1)'\mathbf{Q}(t)(\boldsymbol{\mu}(t) - \boldsymbol{\mu}(t-1) - \gamma\widetilde{\boldsymbol{\mu}}(t-1))\right)$$

c

MH step with kernel:

$$[c|\cdot] \propto c^{\sum\limits_{i=1}^{n}\sum\limits_{t=2}^{T} I_{\{w_{i+}(t)=0\}}} \exp\left(\frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{t=2}^{T} cI_{\{w_{i+}(t)=0\}}\mathbf{h}_i(t)'\mathbf{h}_i(t)\right)$$
$$\times\, c^{a_c-1}(1-c)^{b_c-1}$$

with $\mathbf{h}_i(t) \equiv \boldsymbol{\mu}_i(t) - \boldsymbol{\mu}_i(t-1) - \gamma\widetilde{\boldsymbol{\mu}}_i(t-1) - \eta\mathbf{d}_i(t-1)$, and proposal distribution $N(c^{(l-1)}, \sigma^2_{c-tune})$ (the ratio includes the correction for the transformed proposal).

α

M-H step, where the kernel of the full conditional density is:

$$[\alpha \mid \cdot] \propto \prod_{t=2}^{T} |\mathbf{Q}(t)|^{1/2} \exp\left\{\frac{-1}{2}\mathbf{h}(t)'\mathbf{Q}(t)\mathbf{h}(t)\right\}(1+\alpha)^{a_\alpha - 1}(1-\alpha)^{b_\alpha - 1}\mathbf{1}_{(-1,1)}(\alpha),$$

keeping in mind that we find α in $\mathbf{Q}(t)$, since $\mathbf{Q}(t) = \sigma^{-2}(\mathbf{W}_+^c(t) - \alpha\mathbf{W}(t)) \otimes \mathbf{I}_2$. Also, we have $\mathbf{h}(t) = \boldsymbol{\mu}(t) - [\boldsymbol{\mu}(t-1) + \gamma\widetilde{\boldsymbol{\mu}}(t-1) + \eta\mathbf{d}(t-1)]$, and the proposal distribution is $N(\alpha^{(l-1)}, \sigma^2_{\alpha-tune})$.

σ^2

$$\sigma^2 \mid \cdot \sim IG(\alpha^*_\sigma, \beta^*_\sigma)$$

$$\alpha^*_\sigma \equiv \alpha_\sigma + nT$$

$$\beta^*_\sigma \equiv \beta_\sigma + \frac{1}{2}\sum_{t=2}^{T} A'\mathbf{K}(t)A,$$

with $A = (\boldsymbol{\mu}_{i,t} - \boldsymbol{\mu}_i(t-1) - \beta\widetilde{\boldsymbol{\mu}}_i(t-1) - \eta\mathbf{d}_i(t-1))$

p_1

$$p_1 \mid \cdot \sim Beta(a^*_{p_1}, b^*_{p_1})$$

$$a^*_{p_1} \equiv a_{p_1} + \sum_{i<j} w_{ij}(1)$$

$$b^*_{p_1} \equiv b_{p_1} + \sum_{i<j}(1 - w_{ij}(1))$$

$\boldsymbol{\beta}^{(0)}, \boldsymbol{\beta}^{(1)}$

M-H step for each vector. The kernel of the full conditional density is:

$$[\boldsymbol{\beta}^{(0)} \mid \cdot] \propto \prod_{t=2}^{T} \left[logit^{-1}(\boldsymbol{\beta}^{(0)\prime}\mathbf{x})\right]^{(1-w(t-1))(1-w(t))} \left[1 - logit^{-1}(\boldsymbol{\beta}^{(0)\prime}\mathbf{x})\right]^{(1-w(t-1))w(t)}$$

$$\times \exp\left(\frac{-1}{2}\boldsymbol{\beta}^{(0)\prime}\sigma^{-2}_{\boldsymbol{\beta}^{(0)}}\mathbf{I}\boldsymbol{\beta}^{(0)}\right)$$

$$[\boldsymbol{\beta}^{(1)} \mid \cdot] \propto \prod_{t=2}^{T} \left[logit^{-1}(\boldsymbol{\beta}^{(1)\prime}\mathbf{x})\right]^{w(t-1)w(t)} \left[1 - logit^{-1}(\boldsymbol{\beta}^{(1)\prime}\mathbf{x})\right]^{w(t-1)(1-w(t))}$$

$$\times \exp\left(\frac{-1}{2}\boldsymbol{\beta}^{(1)\prime}\sigma^{-2}_{\boldsymbol{\beta}^{(1)}}\mathbf{I}\boldsymbol{\beta}^{(1)}\right)$$

The proposal distribution is: $N({\boldsymbol{\beta}^{(k)}}^{(l-1)}, \sigma^2_{\boldsymbol{\beta}^{(k)}-tune}\mathbf{I}_{r+1})$, where r corresponds to the number of covariates, and $k = 0, 1$.

$\boldsymbol{\mu}$

M-H step at each time point, with the following as the kernel of the each full conditional density:

$$[\boldsymbol{\mu}_i(1)|\cdot] \propto [\mathbf{s}_i(1)|\boldsymbol{\mu}_i(1), \tau^2\mathbf{I}_2] \times [\boldsymbol{\mu}_i(2)|\boldsymbol{\mu}_i(1), \ldots] \times [\boldsymbol{\mu}_i(3)|\boldsymbol{\mu}_i(2), \boldsymbol{\mu}_i(1), \ldots]$$

$$[\boldsymbol{\mu}_i(t)|\cdot] \propto [\mathbf{s}_i(t)|\boldsymbol{\mu}_i(t), \tau^2\mathbf{I}_2] \times [\boldsymbol{\mu}_i(t)|\boldsymbol{\mu}_i(t-1), \boldsymbol{\mu}_i(t-2), \ldots]$$

$$\times [\boldsymbol{\mu}_i(t+1)|\boldsymbol{\mu}_i(t), \boldsymbol{\mu}_i(t-1), \ldots] \times [\boldsymbol{\mu}_i(t+2)|\boldsymbol{\mu}_i(t+1), \boldsymbol{\mu}_i(t), \ldots]$$

$$[\boldsymbol{\mu}_i(T-1)|\cdot] \propto [\mathbf{s}_i(T-1)|\boldsymbol{\mu}_i(T-1), \tau^2\mathbf{I}_2] \times [\boldsymbol{\mu}_i(T-1)|\boldsymbol{\mu}_i(T-2), \boldsymbol{\mu}_i(T-3), \ldots]$$

$$\times [\boldsymbol{\mu}_i(T)|\boldsymbol{\mu}_i(T-1), \boldsymbol{\mu}_i(T-2), \ldots]$$

$$[\boldsymbol{\mu}_i(T)|\cdot] \propto [\mathbf{s}_i(T)|\boldsymbol{\mu}_i(T), \tau^2\mathbf{I}_2] \times [\boldsymbol{\mu}_i(T)|\boldsymbol{\mu}_i(T-1), \boldsymbol{\mu}_i(T-2), \ldots]$$

At each time point, the proposal distribution is $N(\boldsymbol{\mu}_i(t)^{(l-1)}, \sigma^2_{\boldsymbol{\mu}-tune}\mathbf{I}_2)$, where l represents the iteration.

W

M-H step for the state at each time point. For the state at time t, we consider the information from the segment it belongs to, given that the states in that segment will be dependent on the state at t. Let there be Q segments, with $q = 1, ..., Q$, and T timepoints with $t = 1, ...T$. The state for pair i, j at time t is denoted as $w(t)$. The state $w(t)$ is in segment q, and that goes from $t_{q,ini}$ to $t_{q,fin}$, where $t_{q,ini} \leq t \leq t_{q,fin}$. The full conditional distribution of $w(t)$ is proportional to:

$$\begin{aligned}&[w(t) \mid \cdot] \propto p_{w(t-1),w(t)} \times p_{w(t),w(t+1)} \times \cdots \times p_{w(t_{q,fin}),w(t_{q,fin}+1)} \\ &\times [\boldsymbol{\mu}(t) \mid w(t), w(t-1), \dots] \times [\boldsymbol{\mu}(t+1) \mid w(t+1), w(t), \dots].\end{aligned} \tag{B.1}$$

The probability of transition from the state at time t to the state at $t+1$ is defined as:

$$p_{w(t),w(t+1)} = h[\beta_0^{(w(t))} + \beta_1^{(w(t))} f(w(t), w(t-1), \dots) + \dots],$$

with $h(\cdot)$ being either $logit^{-1}(\cdot)$ if the transition is to the same state or $1 - logit^{-1}(\cdot)$ if the transition is to a different state. The proposed state is $1 - w(t)^{(l-1)}$, that is, the opposite of the state from the last iteration $l - 1$.

B.3 Simulation results

Table B.3.1: Posterior mean and 95% CI for model parameters

Component		Parameter		True	Mean & CI
Emission	Mean	γ	Attraction	0.00037	0.00023 (-0.00018, 0.00063)
		η	Direction	0.01035	0.013 (0.01, 0.016)
	Precision	α	Alignment	0.875	0.862 (0.841, 0.881)
		σ^2	Variance	0.000125	0.000144 (0.000128, 0.000161)
		c	Non zero count	0.147	0.16 (0.13, 0.19)
State transitions	Initial distrib	p_1	Initial prob	0.409	0.43 (0.19, 0.7)
	Non connection	$\beta_{0,0}$	Intercept	3.22	**4.82 (3.1, 7.37)**
		$\beta_{0,1}$	Duration	-0.0255	**-0.04 (-0.086, -0.01)**
		$\beta_{0,2}$	Angle	0.351	-0.41 (-1.4, 0.47)
		$\beta_{0,3}$	Distance	4.890	3.72 (-1.41, 7.27)
		$\beta_{0,4}$	Ang*Dist	-2.365	-1.51 (-2.93, 0.88)
	Connection	$\beta_{1,0}$	Intercept	2.97	**3.38 (1.95, 4.66)**
		$\beta_{1,1}$	Duration	0.0022	-0.002 (-0.024, 0.018)
		$\beta_{1,2}$	Angle	-0.328	-1.15 (-2.12, 0.29)
		$\beta_{1,3}$	Distance	-0.519	-1.63 (-3.4, 0.94)
		$\beta_{1,4}$	Ang*Dist	0.414	2.97 (-0.28, 6.27)

Appendix C

Supplement for Chapter 4

C.1 Posterior estimates of autoregressive model (M4) in scenario 3

Table C.1.1: Posterior mean estimates and CI, M4 data scenario 3

State	AR(1) coefficients	Estimated coefficients
S1	0.5	0.418 (0.319,0.514)
S2	0.05	0.202 (0.129,0.273)
S3	0.8	0.565 (0.485,0.644)

C.2 Full conditional distributions

Latent process η

Metropolis-hastings update, where the proportional to the log-posterior for $\eta_{S_t,t}$ is:

$$\propto -\frac{1}{2\sigma^2}\{y - (\eta_{S_t,t})\}^2 - \frac{1}{2\tau^2}\left\{\eta_{S_t,t} - \left(\mu_{S_t} + \sum_{k=1}^{K} c_k\phi_{k,t}\right)\right\}^2$$

Basis coefficients

The prior of the vector of basis coefficients $\mathbf{c}$ is $N(\mathbf{m}, \mathbf{S})$, with $\mathbf{m}$ a vector of K zeros, and $\mathbf{S} = \rho^2\mathbf{I}_K$. The full conditional distribution for $\mathbf{c}$ is $N(\mu^*, \mathbf{\Sigma}^*)$, with

$$\mu^* = \mathbf{\Sigma}\mathbf{\Phi}^\top\boldsymbol{\mathcal{T}}^{-1}\left(\boldsymbol{\eta} - \boldsymbol{\mu}\right),$$

and

$$\mathbf{\Sigma}^* = \left(\mathbf{\Phi}^\top\boldsymbol{\mathcal{T}}^{-1}\mathbf{\Phi} + \mathbf{S}^{-1}\right)^{-1}$$

where $\mathbf{\Phi}$ is the matrix of basis functions with dimension $T \times K$, $\boldsymbol{\eta}$ is the vector of true observations of length T and $\boldsymbol{\mu}$ is the vector of state-specific means associated to each observation, that is $\boldsymbol{\mu} = (\mu_{s_1}, \mu_{s_2}, \ldots, \mu_{s_T})'$. The matrix $\boldsymbol{\mathcal{T}}$ is a $T \times T$ diagonal matrix, with diagonal elements $\tau^2_{s_1}, \tau^2_{s_2}, \ldots, \tau^2_{s_T}$.

Bibliography

Arani, B. M. S., Carpenter, S. R., Lahti, L., van Nes, E. H., and Scheffer, M. (2021). Exit time as a measure of ecological resilience. *Science*, 372(6547):eaay4895.

Azimi, M., Nasiopoulos, P., and Ward, R. (2005). Offline and online identification of hidden semi-Markov models. *IEEE Transactions on Signal Processing*, 53(8):2658–2663. Conference Name: IEEE Transactions on Signal Processing.

Bradley, J. R. (2021). An Approach to Incorporate Subsampling Into a Generic Bayesian Hierarchical Model. *Journal of Computational and Graphical Statistics*, 30(4):889–905.

Brand, M. (1997). Coupled hidden Markov models for modeling interacting processes. Technical report.

Brooks, S. P. and Gelman, A. (1998). General Methods for Monitoring Convergence of Iterative Simulations. *Journal of Computational and Graphical Statistics*, 7(4):434–455. Publisher: [American Statistical Association, Taylor & Francis, Ltd., Institute of Mathematical Statistics, Interface Foundation of America].

Carpenter, S. R., Arani, B. M. S., Hanson, P. C., Scheffer, M., Stanley, E. H., and

Van Nes, E. (2020). Stochastic dynamics of Cyanobacteria in long-term high-frequency observations of a eutrophic lake. *Limnology and Oceanography Letters*, 5(5):331–336.

Catapult-Innovations (2013). Sprint help for 5.1 and subsequent releases. Melbourne: Catapult Sports Ltd.

Celeux, G. and Durand, J.-B. (2008). Selecting hidden Markov model state number with cross-validated likelihood. *Computational Statistics*, 23(4):541–564.

Clemente, F. M., Martins, F. M. L., and Mendes, R. S. (2016). *Social Network Analysis Applied to Team Sports Analysis*. SpringerBriefs in Applied Sciences and Technology. Springer International Publishing, Cham.

Coloso, J. J., Cole, J. J., and Pace, M. L. (2011). Difficulty in Discerning Drivers of Lake Ecosystem Metabolism with High-Frequency Data. *Ecosystems*, 14(6):935.

Dai, H., Dai, B., Zhang, Y.-M., Li, S., and Song, L. (2017). Recurrent Hidden Semi-Markov Model. In *International Conference on Learning Representations*.

Dong, W., Pentland, A. S., and Heller, K. A. (2012). Graph-coupled HMMs for modeling the spread of infection. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, UAI'12, pages 227–236, Arlington, Virginia, USA. AUAI Press.

Du, X., Cai, W., Liu, J., Yu, D., Xu, K., and Li, W. (2020). Basketball Player's Value Evaluation by a Networks-based Variant Parameter Hidden Markov Model. *arXiv:2012.15734 [cs]*. arXiv: 2012.15734.

Duong, T., Bui, H., Phung, D., and Venkatesh, S. (2005). Activity Recognition and Abnormality Detection with the Switching Hidden Semi-Markov Model. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 838–845, San Diego, CA, USA. IEEE.

Economou, T., Bailey, T. C., and Kapelan, Z. (2014). MCMC implementation for Bayesian hidden semi-Markov models with illustrative applications. *Statistics and Computing*, 24(5):739–752.

Falconer, I. R. (1999). An Overview of problems caused by toxic blue-green algae (cyanobacteria) in drinking and recreational water. *Environmental Toxicology*, 14(1):5–12.

Hamilton, J. D. (2010). Regime switching models. In Durlauf, S. N. and Blume, L. E., editors, *Macroeconometrics and Time Series Analysis*, The New Palgrave Economics Collection, pages 202–209. Palgrave Macmillan UK, London.

Havens, K. E. (2008). Cyanobacteria blooms: effects on aquatic ecosystems. In Hudnell, H. K., editor, *Cyanobacterial Harmful Algal Blooms: State of the Science and Research Needs*, Advances in Experimental Medicine and Biology, pages 733–747. Springer, New York, NY.

Hefley, T. J., Broms, K. M., Brost, B. M., Buderman, F. E., Kay, S. L., Scharf, H. R., Tipton, J. R., Williams, P. J., and Hooten, M. B. (2017). The basis function approach for modeling autocorrelation in ecological data. *Ecology*, 98(3):632–646. Publisher: [Wiley, Ecological Society of America].

Ho, J. C. and Michalak, A. M. (2020). Exploring temperature and precipitation impacts on harmful algal blooms across continental U.S. lakes. *Limnology and Oceanography*, 65(5):992–1009. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/lno.11365.

Huang, Q., Cohen, D., Komarzynski, S., Li, X.-M., Innominato, P., Lévi, F., and Finkenstädt, B. (2018). Hidden Markov models for monitoring circadian rhythmicity in telemetric activity data. *Journal of the Royal Society Interface*, 15(139).

Huisman, J., Codd, G. A., Paerl, H. W., Ibelings, B. W., Verspagen, J. M. H., and Visser, P. M. (2018). Cyanobacterial blooms. *Nature Reviews Microbiology*, 16(8):471–483. Number: 8 Publisher: Nature Publishing Group.

Isles, P. D., Giles, C. D., Gearhart, T. A., Xu, Y., Druschel, G. K., and Schroth, A. W. (2015). Dynamic internal drivers of a historically severe cyanobacteria bloom in Lake Champlain revealed through comprehensive monitoring. *Journal of Great Lakes Research*, 41(3):818–829.

Johnson, M. (2005). Capacity and complexity of HMM duration modeling techniques. *IEEE Signal Processing Letters*, 12(5):407–410. Conference Name: IEEE Signal Processing Letters.

Johnson, M. and Willsky, A. (2013). Bayesian Nonparametric Hidden Semi-Markov Models. *Journal of Machine Learning Research*, 14:673–701.

King, R. and Langrock, R. (2016). Semi-Markov Arnason–Schwarz models. *Biometrics*, 72(2):619–628. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12446.

Koki, C., Meligkotsidou, L., and Vrontos, I. (2020). Forecasting under model uncertainty: Non-homogeneous hidden Markov models with Pòlya-Gamma data augmentation. *Journal of Forecasting*, page for.2645.

Kos, M. and Kramberger, I. (2017). A Wearable Device and System for Movement and Biometric Data Acquisition for Sports Applications. *IEEE Access*, 5:6411–6420. Conference Name: IEEE Access.

Langrock, R., Kneib, T., Glennie, R., and Michelot, T. (2017). Markov-switching generalized additive models. *Statistics and Computing*, 27(1):259–270.

Langrock, R., Kneib, T., Sohn, A., and DeRuiter, S. L. (2015). Nonparametric inference in hidden Markov models using P-splines. *Biometrics*, 71(2):520–528. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/biom.12282.

Lead PI, N., Magnuson, J., Carpenter, S., and Stanley, E. (2020). North Temperate Lakes LTER: High Frequency Data: Meteorological, Dissolved Oxygen, Chlorophyll, Phycocyanin - Lake Mendota Buoy 2006 - current. Type: dataset.

Leos-Barajas, V., Gangloff, E. J., Adam, T., Langrock, R., van Beest, F. M., Nabe-Nielsen, J., and Morales, J. M. (2017). Multi-scale Modeling of Animal Movement and General Behavior Data Using Hidden Markov Models with Hierarchical Structures. *Journal of Agricultural, Biological and Environmental Statistics*, 22(3):232–248.

Li, D. and Wong, W. H. (2018). Mini-batch Tempered MCMC. arXiv:1707.09705 [stat].

Li, R. T., Kling, S. R., Salata, M. J., Cupp, S. A., Sheehan, J., and Voos, J. E. (2016). Wearable Performance Devices in Sports Medicine. *Sports Health*, 8(1):74–78. Publisher: SAGE Publications.

Li, Y. and Sun, Y. (2021). A multi-site stochastic weather generator for high-frequency precipitation using censored skew-symmetric distribution. *Spatial Statistics*, 41:100474.

Lin, Z., Liu, X., and Collu, M. (2020). Wind power prediction based on high-frequency SCADA data along with isolation forest and deep learning neural networks. *International Journal of Electrical Power & Energy Systems*, 118:105835.

Liu, H. and Song, X. (2020). Bayesian analysis of hidden Markov structural equation models with an unknown number of hidden states. *Econometrics and Statistics*.

Maclaurin, D. and Adams, R. P. (2015). Firefly Monte Carlo: Exact MCMC with Subsets of Data. *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015)*, page 7.

Mor, B., Garhwal, S., and Kumar, A. (2021). A Systematic Review of Hidden Markov Models and Their Applications. *Archives of Computational Methods in Engineering*, 28(3):1429–1448.

Motoi, S., Misu, T., Nakada, Y., Yazaki, T., Kobayashi, G., Matsumoto, T., and Yagi, N. (2012). Bayesian event detection for sport games with hidden Markov model. *Pattern Analysis and Applications*, 15(1):59–72.

Motti, V. G. (2020). Introduction to Wearable Computers. In Motti, V. G., editor, *Wearable Interaction*, Human–Computer Interaction Series, pages 1–39. Springer International Publishing, Cham.

Natarajan, P. and Nevatia, R. (2007). Coupled Hidden Semi Markov Models for Activity Recognition. In *2007 IEEE Workshop on Motion and Video Computing (WMVC'07)*, pages 10–10, Austin, TX, USA. IEEE.

Paerl, H. W. and Huisman, J. (2008). Blooms Like It Hot. *Science*, 320(5872):57–58. Publisher: American Association for the Advancement of Science Section: Perspective.

Park, K.-J. and Yilmaz, A. (2010). Social Network Approach to Analysis of Soccer Game. In *2010 20th International Conference on Pattern Recognition*, pages 3935–3938. ISSN: 1051-4651.

Peña, J. L. and Touchette, H. (2012). A network theory analysis of football strategies. *arXiv:1206.6904 [physics, stat]*. arXiv: 1206.6904.

Pohle, J., Langrock, R., van Beest, F. M., and Schmidt, N. M. (2017). Selecting the Number of States in Hidden Markov Models: Pragmatic Solutions Illustrated Using Animal Movement. *Journal of Agricultural, Biological and Environmental Statistics*, 22(3):270–293.

Quiroz, M., Kohn, R., Villani, M., and Tran, M.-N. (2019). Speeding Up MCMC by Efficient Data Subsampling. *Journal of the American Statistical Association*, 114(526):831–843. Publisher: Taylor & Francis.

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

Ramesh, P. and Wilpon, J. (1992). Modeling state durations in hidden Markov models for automatic speech recognition. In *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 381–384 vol.1. ISSN: 1520-6149.

Ravuri, S. and Wegmann, S. (2016). How Neural Network Depth Compensates for HMM Conditional Independence Assumptions in DNN-HMM Acoustic Models. pages 2736–2740.

Reichwaldt, E. S. and Ghadouani, A. (2012). Effects of rainfall patterns on toxic cyanobacterial blooms in a changing climate: Between simplistic scenarios and complex dynamics. *Water Research*, 46(5):1372–1393.

Rousseeuw, K., Caillault, P., Lefebvre, A., and Hamad, D. (2015). Hybrid Hidden Markov Model for Marine Environment Monitoring. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8(1):204–213.

Rousso, B. Z., Bertone, E., Stewart, R. A., Rinke, K., and Hamilton, D. P. (2021). Light-induced fluorescence quenching leads to errors in sensor measurements of phytoplankton chlorophyll and phycocyanin. *Water Research*, 198:117133.

Ruiz-Suarez, S., Leos-Barajas, V., and Morales, J. M. (2022). Hidden Markov and Semi-Markov Models When and Why are These Models Useful for Classifying States in Time Series Data? *Journal of Agricultural, Biological and Environmental Statistics*, 27(2):339–363.

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric Regression.*

Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Sansom, J. and Thompson, C. S. (2008). Spatial and temporal variation of rainfall over New Zealand. *Journal of Geophysical Research*, 113(D6):D06109.

Scharf, H. R., Hooten, M. B., Fosdick, B. K., Johnson, D. S., London, J. M., and Durban, J. W. (2016). Dynamic social networks based on movement. *The Annals of Applied Statistics*, 10(4):2182–2202.

Scott, S. L. (2002). Bayesian Methods for Hidden Markov Models. *Journal of the American Statistical Association*, 97(457):337–351.

Snijders, T. A. and Nowicki, K. (1997). Estimation and Prediction for Stochastic Blockmodels for Graphs with Latent Block Structure. *Journal of Classification*, 14(1):75–100.

Soranno, P. A. (1997). Factors affecting the timing of surface scums and epilimnetic blooms of blue-green algae in a eutrophic lake. *Canadian Journal of Fisheries and Aquatic Sciences*, 54(9):1965–1975. Publisher: NRC Research Press.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Linde, A. V. D. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4):583–639.

Stoner, O. and Economou, T. (2020). An advanced hidden Markov model for hourly rainfall time series. *Computational Statistics & Data Analysis*, 152:107045.

Thomas, O., Sunehag, P., Dror, G., Yun, S., Kim, S., Robards, M., Smola, A., Green, D., and Saunders, P. (2010). Wearable sensor activity analysis using semi-Markov models with a grammar. *Pervasive and Mobile Computing*, 6(3):342–350.

Titman, A. C. and Sharples, L. D. (2010). Semi-Markov Models with Phase-Type Sojourn Distributions. *Biometrics*, 66(3):742–752.

Touloupou, P., Finkenstädt, B., and Spencer, S. E. F. (2019). Scalable Bayesian Inference for Coupled Hidden Markov and Semi-Markov Models. *Journal of Computational and Graphical Statistics*, pages 1–12.

Vaseghi, S. V. (1991). Hidden Markov models with duration-dependent state transition probabilities. *Electronics Letters*, 27(8):625–626. Publisher: IET Digital Library.

Vaseghi, S. V. (1995). State duration modelling in hidden Markov models. *Signal Processing*, 41(1):31–41.

Wang, Z., Guo, M., and Zhao, C. (2016). Badminton Stroke Recognition Based on Body Sensor Networks. *IEEE Transactions on Human-Machine Systems*, 46(5):769–775.

Welch, M., Schaerf, T. M., and Murphy, A. (2021). Collective states and their transitions in football. *PLOS ONE*, 16(5):e0251970. Publisher: Public Library of Science.

Wikle, C. K., Zammit-Mangion, A., and Cressie, N. A. C. (2019). *Spatio-temporal statistics with R*. Chapman & Hall/CRC the R series. CRC Press, Taylor & Francis Group, Boca Raton.

Woillez, M., Fablet, R., Ngo, T.-T., Lalire, M., Lazure, P., and de Pontual, H. (2016). A HMM-based model to geolocate pelagic fish from high-resolution individual temperature and depth histories: European sea bass as a case study. *Ecological Modelling*, 321:10–22.

Wu, T.-Y., Wang, Y. X. R., and Wong, W. H. (2019). Mini-batch Metropolis-Hastings MCMC with Reversible SGLD Proposal. arXiv:1908.02910 [cs, stat].

Wu, W. and Haick, H. (2018). Materials and Wearable Devices for Autonomous Monitoring of Physiological Markers. *Advanced Materials*, 30(41):1705024. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/adma.201705024.

Xie, L., Xu, P., Chang, S.-F., Divakaran, A., and Sun, H. (2004). Structure analysis of soccer video with domain knowledge and hidden Markov models. *Pattern Recognition Letters*, 25(7):767–775.

Xu, Z. and Liu, Y. (2021). A Regularized Vector Autoregressive Hidden Semi-Markov model, with Application to Multivariate Financial Data. *The International FLAIRS Conference Proceedings*, 34.

Yu, S.-Z. (2010). Hidden semi-Markov models. *Artificial Intelligence*, 174(2):215–243.

Yu, S.-Z. (2016). *Hidden semi-Markov models: theory, algorithms and applications.* Computer science reviews and trends. Elsevier, Amsterdam ; Boston. OCLC: ocn932126653.

Zucchini, W., MacDonald, I. L., and Langrock, R. (2017). *Hidden Markov Models for Time Series: An Introduction Using R, Second Edition.* CRC Press. Google-Books-ID: KlWzDAAAQBAJ.

Ötting, M., Langrock, R., and Maruotti, A. (2021). A copula-based multivariate hidden Markov model for modelling momentum in football. *AStA Advances in Statistical Analysis*.

VITA

Shirley Rojas Salazar was born on August 15, 1989 in Cartago, Costa Rica. She attended the University of Costa Rica and earned a Bachelor degree in Statistics in 2012. She moved to the United States of America in 2015 to pursue graduate studies at the University of Missouri-Columbia. She completed a M.A. in Statistics in the Spring of 2017 and started the Ph.D program in Statistics in the Fall of 2017. She has accepted a position at the Department of Statistics at the University of Costa Rica.