UMBC

DISSERTATION APPROVAL SHEET

Title of Dissertation: Hidden Markov Models for High Dimensional Data with Geostatistical Applications

Name of Candidate: Reetam Majumder

Doctor of Philosophy, 2021

Graduate Program: Statistics

Dissertation and Abstract Approved:

Nagaraj Neerchal

Nagaraj Neerchal

Professor

Mathematics and Statistics

10/25/2021 | 6:03:28 PM EDT

NOTE: *The Approval Sheet with the original signature must accompany the thesis or dissertation. No terminal punctuation is to be used.

ABSTRACT

Title of dissertation: HIDDEN MARKOV MODELS FOR HIGH DIMENSIONAL DATA WITH GEOSTATISTICAL APPLICATIONS

Reetam Majumder
Doctor of Philosophy, 2021

Dissertation directed by: Nagaraj K. Neerchal
Professor of Statistics
Department of Mathematics and Statistics
University of Maryland, Baltimore County

Stochastic precipitation generators (SPGs) are a class of statistical models which generate synthetic data that can simulate dry and wet rainfall stretches for long durations. Generated precipitation time series data are used in climate projections, impact assessment of extreme weather events, and water resource and agricultural management. In this thesis, we construct SPGs for daily precipitation data that is specified as a semi-continuous distribution with a point mass at zero for no precipitation and a mixture of Exponential or Gamma distributions for positive precipitation. Our generators are obtained as hidden Markov models (HMMs) where the underlying climate conditions form the states. Maximum likelihood estimation of an HMM's parameters has historically relied on the Baum-Welch algorithm, which is a modification of the Expectation Maximization algorithm. We implement variational Bayes (VB) as an alternative estimation procedure for HMMs with semi-continuous emissions. Stochastic optimization in the form of stochastic variational Bayes (SVB) has been employed for computational speedup in practical cases.

A univariate state process is often unable to adequately capture the underlying weather conditions over large watersheds, since different areas can have local weather

regimes. We extend the HMM to a linked HMM (LHMM) where locations are divided into clusters. Each cluster's emissions are assumed to arise from a cluster-specific state process; the state processes are correlated and together form a multivariate Markov chain (MMC). The MMC provides more flexibility to accommodate heterogeneity that might be present in larger geographical areas. A Gaussian copula is constructed to capture the correlation structure of the MMC. Finally, we also construct a Gaussian copula for the emissions of the HMM to explicitly parameterize the pairwise correlations of observed positive precipitation. Daily precipitation data over the Chesapeake Bay watershed in the Eastern coast of the USA is used as a demonstrative case study. Remote sensing precipitation data is sourced from the GPM-IMERG dataset for the wet season between July to September from 2000-2019. Synthetic data generated from the clustered LHMM can reproduce the monthly precipitation statistics as well as the spatial correlations present in the historical GPM-IMERG data.

HIDDEN MARKOV MODELS FOR HIGH DIMENSIONAL DATA WITH GEOSTATISTICAL APPLICATIONS

by

Reetam Majumder

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, Baltimore County, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2021

Advisory Committee:
Dr. Nagaraj K. Neerchal, Advisor
Dr. Anindya Roy
Dr. Matthias K. Gobbert
Dr. Amita Mehta
Dr. Andrew Raim

DEDICATION

To my parents, Madhumita Majumder and Pradip Kumar Majumder.

ACKNOWLEDGMENTS

I wish to express my gratitude to the members of my dissertation committee for being instrumental in my development as a researcher. Dr. Nagaraj K. Neerchal, my thesis advisor, has allowed me the freedom to explore problems until I found the one I was passionate about. His patience with me and his approach to solving problems makes every day working with him a learning experience. Dr. Anindya Roy taught me the topics which then became research interests, and fundamentally changed my perception of the role of statistics in the modern world. Dr. Matthias K. Gobbert gave me the opportunity to be a Research Assistant at the UMBC High Performance Computing Facility, and has been a supportive mentor who has championed me at every opportunity. Dr. Amita Mehta has been my guide in my exploration of the Earth Sciences, and has also given me opportunities to work with her at the Joint Center for Earth Systems Technology on projects with tangible policy impacts. Finally, Dr. Andrew Raim not only provided me sage advice that has influenced my trajectory as a graduate student, but has also been a constant role model.

I would like to acknowledge the faculty at the Department of Mathematics and Statistics at UMBC for their guidance over the past 4 years. I am also appreciative of the staff at the department for their patience while helping me navigate graduate school. On a similar note, I am grateful to the International Student and Scholar Services at UMBC for being incredibly responsive towards, transparent with, and supportive of the international student community through a period that has at times been quite trying.

All my colleagues in the department, past and present, have formed the tapestry of my graduate school experience, and spending time with them has been enriching. In particular, I would like to give a shout out to Iris Guaran for mentoring me early on, and to my cohort mates Yewon Kim and Mark Ramos for being generous with their friendship,

their knowledge, and their time.

I would also like to thank my friends Puja Bhattacharya and Riddhi Pratim Ghatak for pushing me to go to graduate school, for being the sounding board for my plans, and for offering honest feedback and advice that has always served me well.

Finally, and most importantly, my deepest appreciation and gratitude go towards my family. Without their unwavering support and constant encouragement, I would not be here.

My Research Assistantship with JCET was supported by a grant titled 'A Sustainable Forest Management and Information System (SFMIS) Tool', awarded to UMBC from the Jet Propulsion Laboratory, funded by the NASA Applied Sciences Biodiversity and Ecological Forecasting program. Part of my thesis research was supported by the U.S. National Science Foundation under the CyberTraining (OAC–1730250) and MRI (OAC–1726023) programs. The hardware used in the computational studies is part of the UMBC High Performance Computing Facility (HPCF). The facility is supported by the U.S. National Science Foundation through the MRI program (grant nos. CNS–0821258, CNS–1228778, and OAC–1726023) and the SCREMS program (grant no. DMS–0821311), with additional substantial support from the University of Maryland, Baltimore County (UMBC). See hpcf.umbc.edu for more information on HPCF and the projects using its resources.

My sincerest apology to anyone I have inadvertently left out. Thank you all.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Weather stations have historically been one of the primary sources of meteorological data globally, and remote sensing data from satellites have recently given researchers access to data that is observed over dense spatial grids at frequent intervals. Meteorological data is distributed in the form of a multivariate time series where each univariate component corresponds to a location. However, modeling it directly using time series requires the estimation of a large number of parameters and high-dimensional autocoviariance matrices. The statistical analysis of such datasets at scale calls for specialized methods that are computationally efficient while being able to represent the dynamics of the underlying processes to a satisfactory degree. Hidden Markov models (HMM) are an attractive class of models used to model geostatistical data. HMMs assume that the observed data, known as the emission process, is generated by a finite-valued latent variable. The latent variable is assumed to follow a first order Markov process and is referred to as the state process. The Markov property of the state process serves to capture the temporal dependency in the data, and the emission process at each time point describes the spatial patterns in the data. HMMs were initially introduced and have been studied since the late 1960s, and have found extensive use in speech recognition [Rabiner, 1989], finance [Rydén et al., 1998], genomics [Boys and Henderson, 2004], as well as in meteorology [Hughes and Guttorp, 1994]. In this thesis we will be working with daily precipitation data over a watershed. Our interest lies in modeling the temporal patterns in the data as well as capturing the spatial dependency in multi-site precipitation.

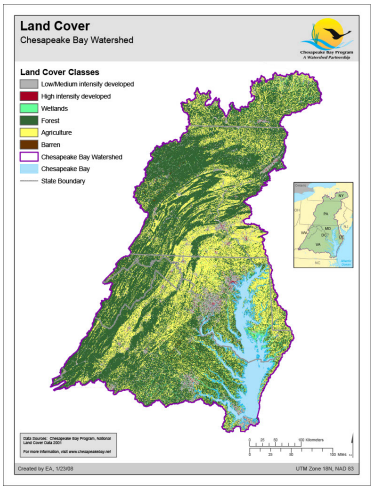

Figure 1.1: Land cover classes within the Chesapeake Bay watershed in the East coast of the USA.

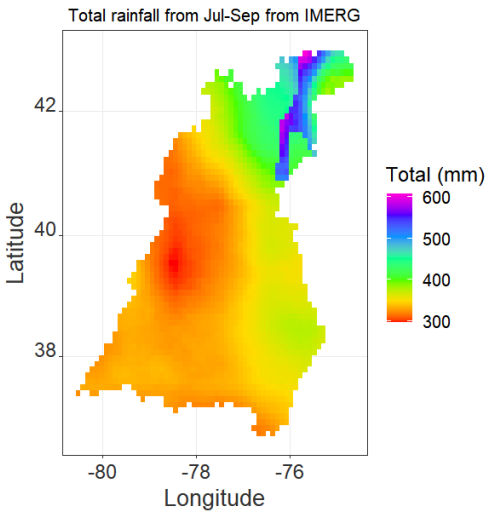

Figure 1.2: Total precipitation for Jul–Sep over the Chesapeake Bay watershed from GPM-IMERG data for 2000–2019.

1.1 Application: Modeling Multi-Site Daily Precipitation Data

Our motivating example pertains to the stochastic modeling of daily precipitation data over large watersheds obtained from remote sensing sources. Precipitation is the major component of the global water cycle and plays an important role in atmospheric and land surface processes in the climate system. While numerical weather models study precipitation over large areas, observed precipitation data are used to develop statistical models for precipitation over smaller areas at higher temporal frequencies and higher spatial resolution. The measurement and modeling of precipitation has historically relied on rain gauges whose presence is often sparse and spatially irregular, with remote sensing data becoming common relatively recently. A common class of statistical models which are of interest for analyzing remote sensing data are known as stochastic weather generators (SWGs). SWGs can be used to generate long time series of synthetic data to simulate weather patterns and are indispensable in weather and climate research. The particular type of SWG we focus on in this thesis is called a stochastic precipitation generator (SPG). The modeling and forecasting of seasonal and inter-annual variations in precipitation is used to

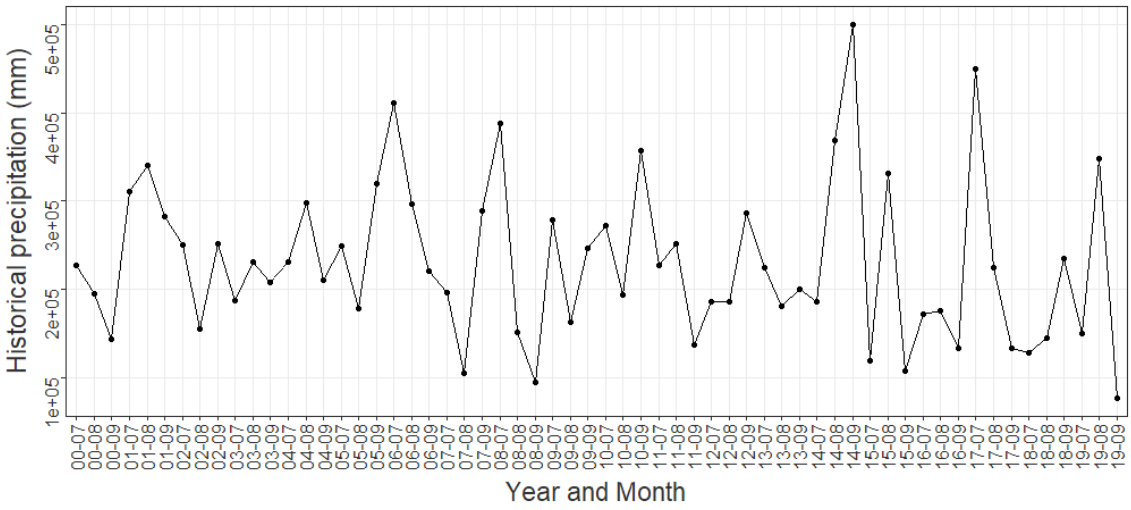

Figure 1.3: Total Precipitation over the Chesapeake Bay watershed for the months of Jul (07)–Sep (09) from 2000–2019

determine water allocation and resource management for regions dependent on precipitation as a primary water source. To this end, SPGs are constructed to produce time series of synthetic data representative of the general rainfall patterns within a region. In particular, they aim to replicate key statistical properties of the historical data like dry and wet stretches, spatial correlations, and extreme weather events. Our interest in SPGs arose from a hydrologic modeling problem [Majumder et al., 2019] where we were trying to assess the seasonal water budget for the Potomac river basin using the VIC model [Hamman et al., 2018]. Calibrating hydrologic models requires precipitation ensembles, which can be provided through SPGs. SPGs are also used in downscaling numerical weather models, and simulations from them are used for climate projections, impact assessments of extreme weather events, water resources and agricultural management, and for public and veterinary health. In general, SWGs complement numerical models which tend to be extremely sensitive to starting values. While the output provided from these models are statistical estimates and therefore have uncertainty built in, ensemble datasets generated from these models can improve other climate and weather models. Breinl et al. [2017] provides a review of current SPG approaches and applications. Our region of interest is the Chesapeake Bay watershed which includes parts of six states and nine major river systems on the East Coast of the USA. Figure 1.1 shows the Chesapeake Bay watershed and the different land cover classes within it. The watershed has a diverse, interconnected ecosystem which is affected by extreme weather potentially related

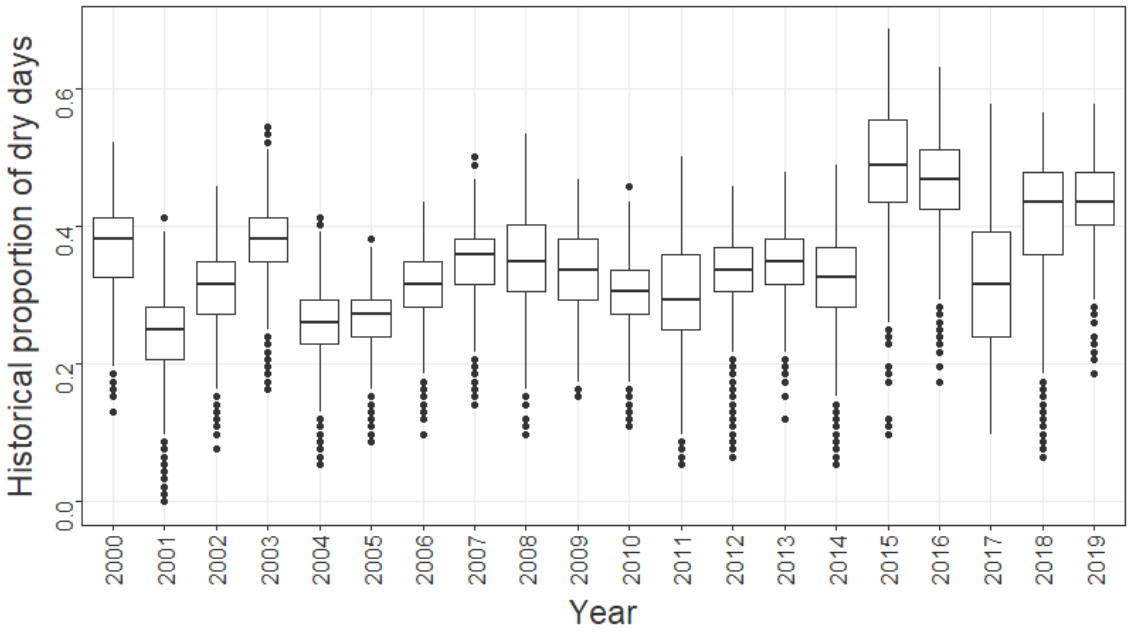

Figure 1.4: Distribution of proportion of dry days for each year's wet season from 2000–2019 across grid points over the Chesapeake Bay watershed.

to climate change [Chesapeake Bay Program, 2012], and has been targeted for restoration as an integrated watershed and ecosystem. Rainfall within the watershed and resulting runoff into the rivers and the bay bring substantial amounts of sediments and nutrients to the bay and impact the water quality of the bay. Therefore, understanding and forecasting rainfall patterns and temporal variability, particularly extreme rainfall events in the Chesapeake Bay watershed, are crucial for monitoring and managing water quality in the bay.

We use daily data from the GPM-IMERG dataset [Huffman et al., 2019] for the months of July to September from 2000 – 2019. At a spatial resolution of $0.1^\circ \times 0.1^\circ$, The IMERG dataset covers the 64,000 square mile watershed with 1927 grid points. Figure 1.2 shows the seasonal rainfall at each grid point of the basin. The values are obtained by computing the sum total of daily rainfall between July 1 and September 30 for each year between 2000–2019, and then averaging them over the 20 years. Our analysis focuses on July–September since they are the wettest months of the year for this area. The figure shows high and low precipitation areas over the watershed and is representative of the high degree of spatial correlation present in the data. Figure 1.3 plots monthly precipitation for July–September over the watershed for each year from 2000–2019. The first coordinate of the x-axis in the bottom left corresponds to Jul 2000 (00-07)

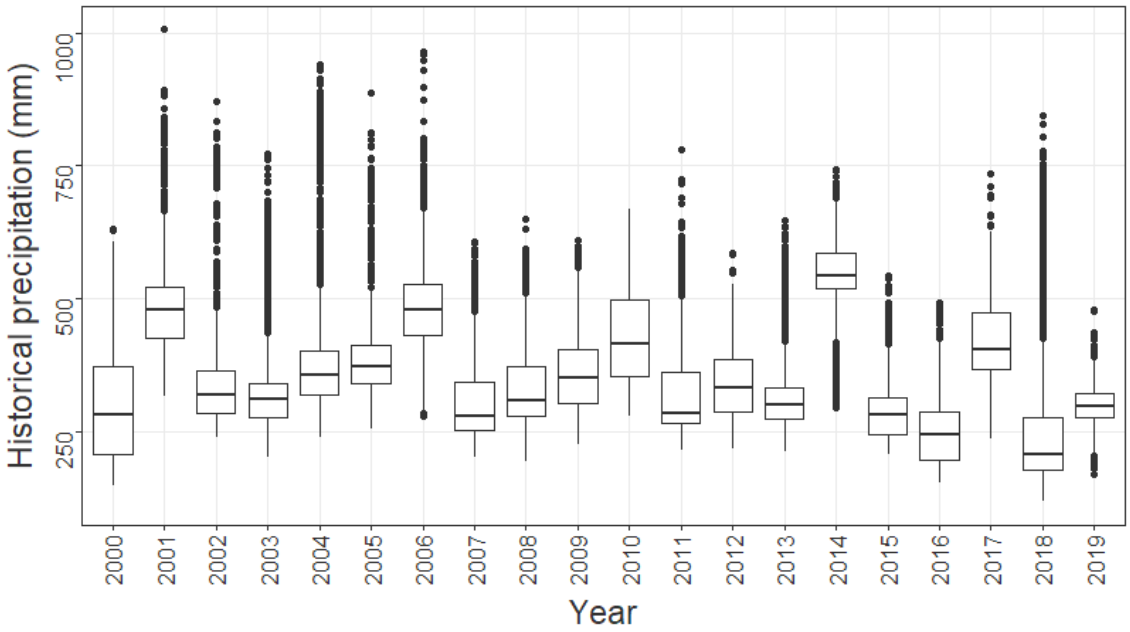

Figure 1.5: Distribution of total precipitation (mm) for each year's wet season from 2000–2019 across grid points over the Chesapeake Bay watershed.

and the final coordinate in the bottom right corresponds to Sep 2019 (19-09). The data shows high inter-annual variability and also a lot of variability from month to month. Figure 1.4 contains boxplots representing the distribution of the proportion of dry days across the watershed over the duration of the wet season months for each year from 2000–2019. Individual points represent the proportion of dry days at a single grid point for a particular year's wet season. The median for most of the boxplots are below 0.4 which indicates precipitation occurs over the basin more than half of all the wet season days. The maximum value in the entire plot is 0.68, which occurs in 2015. Further, most years' boxplots are negatively skewed, indicating predominantly wet days during the wet season over the watershed. Similarly, Figure 1.5 contains boxplots representing the distribution of total precipitation during the wet season for each year at each grid point of the watershed from 2000–2019. We see that the variability in the data changes from year to year, and that precipitation for each year is highly skewed towards the right with long tails corresponding to extreme weather events. These are some of the monthly and annual statistics we want the SPG to be able to reproduce. We also want the SPG to be able to reproduce the Pearson correlations between GPM-IMERG precipitation estimates at the different grid points of the watershed. Since the GPM-IMERG dataset divides the watershed into 1927 grid points, there are 1.85×10^6 pairwise

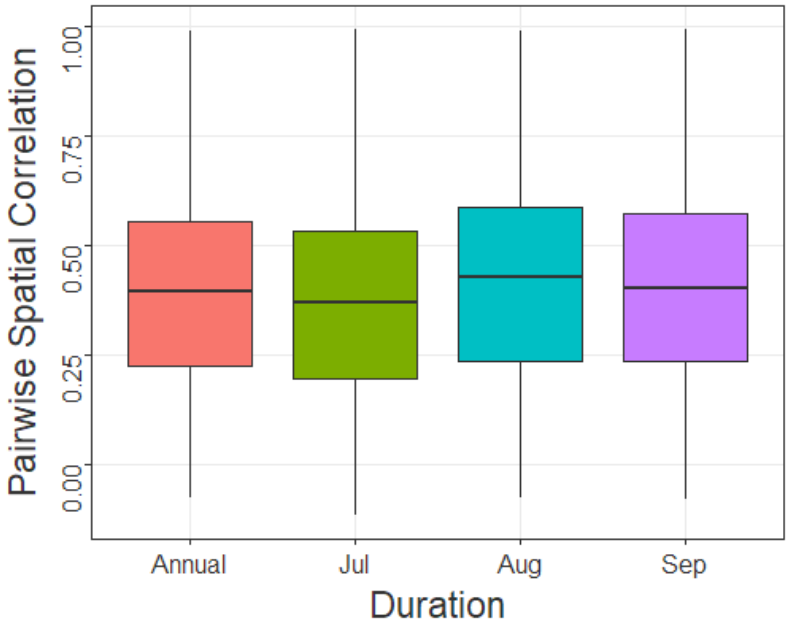

Figure 1.6: Distribution of the pairwise correlation of precipitation between different grid points over the Chesapeake Bay watershed.

correlation values computed for the grid points. Figure 1.6 plots the distribution of these pairwise correlations for each month of the wet season, as well as for the entirety of the wet season. We see very high correlations for all the boxplots, with August having the highest correlations overall. We also note the presence of correlation values less than 0 in the data. Around 2.9% of the 1.85×10^6 Pearson correlation values are less than 0. However, the lowest value is only around -0.07. We do not believe this to be indicative of negative association in the precipitation patterns in these locations. Rather, since the IMERG data is in itself an estimate, we consider this as noise inherent in the data. Finally, we fit a linear model to gauge the relationship between the spatial correlations and distance, in this case measured as the Euclidean distance between two grid points' latitude and longitude coordinates. Unsurprisingly, the model estimated a negative slope that was highly significant, indicating that the correlation in precipitation between two locations decreases as the physical distance between them increases.

1.2 Overview of Related Work

An overwhelming amount of HMM studies use the Baum-Welch algorithm [Baum and Petrie, 1966] for parameter estimation. The algorithm is a variant of the expectation-maximization (EM) algorithm for efficient parameter estimation in HMMs which takes into account the Markov assumptions of the model. Bayesian alternatives which use Gibbs sampling have been outlined in Scott [2002] and in Cappé et al. [2005] but tend to be computationally intensive. Parameter estimation using variational Bayes has been developed more recently [Ji et al., 2006, McGrory and Titterington, 2009] based around initial work by MacKay [1997], but has not been studied as extensively as the Baum-Welch algorithm.

The majority of studies [Wilks, 1998, Hughes and Guttorp, 1994, Robertson et al., 2004] specify the positive part of daily precipitation either as a single Gamma distribution or a mixture of two Exponential distributions depending on the climatology of the area and the season for which the study was conducted. Much of the groundwork for using HMMs for daily precipitation was laid in Hughes and Guttorp [1994], with Bellone [2000], Bellone et al. [2000] proposing different emission distributions for precipitation amounts and precipitation occurrence models. This was extended to non-homogeneous hidden Markov models (NHMM) in [Robertson et al., 2004, 2006, Kirshner, 2005, Kirshner et al., 2004]. NHMMs let the transition probability matrix of the HMM's Markov process depend on time. We restrict ourselves to homogeneous HMMs in this thesis, i.e. the Markov chain's parameters do not change over time. The Baum-Welch algorithm is used for parameter estimation in all these studies. However, being a maximum likelihood based method, it can run into problems when it comes to graphical models like HMMs [Attias, 1999]. In particular, it can lead to model overfitting for graphs with complex structures. This is where variational Bayes (VB) provides an attractive alternative. While MCMC methods use sampling to find the posterior distribution, VB uses optimization to calculate an approximate posterior; the posteriors are obtained by an iterative EM-like algorithm which always converges [Attias, 1999]. Instead of yielding a single point estimate of the model parameters, it learns an ensemble of models, and estimates posterior density functions for the model parameters given a training dataset. The variational posteriors have an analytical form and can be used to perform Bayesian inference. A

recent review of VB methods can be found in Blei et al. [2017], and there is ongoing research into the theoretical properties of the variational approximation [Zhang and Zhou, 2017, Zhang and Gao, 2020, Pati et al., 2018, Wang and Blei, 2019, Yang et al., 2020].

Our own previous work in this area has mostly used the Baum-Welch algorithm for parameter estimation. In Kroiz et al. [2020a], we worked with precipitation data for the Potomac river basin in Eastern USA. We found that the classical HMM severely underestimates spatial correlations, and constructed a Gaussian copula to explicitly model spatial correlations. This was an early motivating example for the research presented in this thesis. We looked at model selection for the Potomac dataset in Kroiz et al. [2020b], comparing BIC scores for a variety of Gamma distribution and Exponential distribution configurations. In general, using mixtures of Gamma distributions instead of mixtures of Exponential distributions tended to have provide Bayesian information criterion (BIC) scores, suggesting better model fit. For either case, using more than 2 mixture components for positive precipitation resulted in at best marginally better fit at added computational cost. Finally, we also looked at fitting a Gaussian copula for the Chesapeake Bay watershed data in Majumder et al. [2020] after using the Baum-Welch algorithm for marginal parameter estimation.

1.3 Contributions

This thesis has two overarching goals:

1. Develop HMMs for modeling semi-continuous daily precipitation data which employ VB for parameter estimation instead of the B-W algorithm and can capture the spatial correlation in the data,

2. Develop a formulation for HMMs with a multivariate Markov chain (MMC) as the underlying state process which is appropriate for large spatial domains.

The novel contributions of this thesis are:

- Variational Bayes parameter estimation in HMMs with semi-continuous emissions; use of mixtures of Gamma and Exponential distributions in a manner similar to existing maximum

likelihood implementations (Sections 3.1 – 3.5),

- Developing the use of modified Gamma shape mixtures (GSM) for modeling positive precipitation and parameter estimation using VB; the modified GSM has fewer mixture components than its original specification (Section 3.6),
- Developing a linked HMM with an MMC as the underlying state process, using a Gaussian copula to capture the correlation structure of the MMC (Section 4.1),
- Constructing a Gaussian copula for the emission distribution of an HMM with semi-continuous emissions to explicitly express the correlation structure in the observed data (Section 4.2),
- Proposing a modified minibatch sampling method for the stochastic implementation of variational Bayes parameter estimation (Section 3.8),
- Deriving empirical priors to use in the variational Bayes estimation process (Section 3.7).

1.4 Outline of the Thesis

The rest of this thesis is organized as follows: Chapter 2 provides concepts and background for the models, estimation methods, and datasets used in this thesis. Chapter 3 covers the estimation of HMM parameters using variational Bayes for single-site and multi-site precipitation. Chapter 4 extends the HMM to have an MMC as the underlying state process using a Gaussian copula and provides a learning algorithm for the copula parameters. It also introduces a copula for the emission distribution to capture the pairwise correlations of observed precipitation at different locations. Chapter 5 brings together all the components for a case study using daily precipitation over the Chesapeake Bay watershed as a demonstrative example. Finally, Chapter 6 summarizes our work and outlines remaining questions and possible future work.

Chapter 2

Background

2.1 The Hidden Markov Model

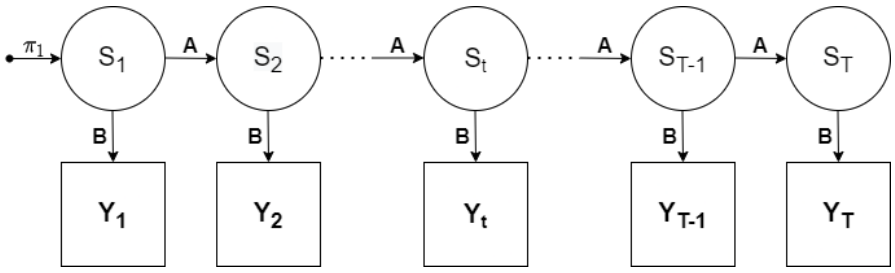

Figure 2.1: A directed acyclic graph (DAG) specifying the conditional independence structure for a hidden Markov model.

A hidden Markov model (HMM) is a pair of stochastic processes $\{S_t, Y_t\}_{t\geq 1}$ where $\{S_t\}$ is a Markov chain, and conditional on it, $\{Y_t\}$ is a sequence of independent random variables such that the distribution of Y_t depends only on S_t. $\{S_t\}$ usually takes values in a finite set; t is often, although not necessarily, an integer index. However, $\{S_t\}$ is unobservable, and instead we observe only $\{Y_t\}_{t\geq 0}$. $\{Y_t\}$ can be univariate or multivariate, and can follow a discrete, continuous, or mixture distribution. $\{S_t\}$ is known as the state process, while $\{Y_t\}$ is called the emission or observation process. A hidden Markov model is characterized by the following:

1. *K*, the **number of states** in the model. While latent in nature, the states often have physical interpretations in the context of the application. In the case of precipitation models, the states correspond to underlying weather regimes dictating spatial and temporal patterns of rainfall over a region. Usually, we assume that any state can be reached from any other state, i.e. the Markov process is ergodic in nature.

2. *R*, the **number of distinct observation symbols** if the observations are discrete. This is also called the discrete alphabet size. If the emissions are continuous and are assumed to arise from a mixture of distributions, *R* denotes the number of mixture components. A HMM for daily precipitation occurrence, whose emissions are dry and wet days, corresponds to a discrete emission process with $R = 2$. Similarly, an HMM for daily precipitation amounts will have a semi-continuous mixture distribution with $R \geq 2$, with a point mass at zero for no rainfall and at least one wet mixture component corresponding to positive rainfall.

3. *A*, the **state transition probability** matrix. $A = ((a_{jk}))$, where $a_{jk} = Pr[s_{t+1} = k|s_t = j]$ for $j, k = 1, \ldots, K$. For an ergodic process, we have $a_{jk} > 0$ for all j, k.

4. π_1, **the initial state probability distribution**. It is a *K*-vector with $\pi_{1j} = Pr[\pi_1 = j]$ for $j = 1, \ldots, K$.

5. *B*, the **conditional distribution of emission probabilities** for each state *j*. In the discrete emission case, its components are $B = ((b_{jr}))$, where $b_{jr} \equiv b_{jr}(t) = Pr[y_t = r|s_t = j]$ for $r = 1, \ldots, R$ and $j = 1, \ldots, K$. For semi-continuous and continuous emissions, its components would contain mixture probabilities and probability densities with their own parameters.

The parameters of an HMM are therefore represented by $\Theta = (A, B, \pi_1)$, where the initial probability distribution π_1 and the transition matrix *A* parameterize the Markov process $\{S_t\}$, while *B* consists of the parameters of the emission process $\{Y_t\}$. Figure 2.1 depicts a graphical representation of a hidden Markov model. Given this specification, fitting an HMM to the data involves solving three problems:

1. **Model Selection** - Given a sequence of observations $y_{1:T} = \{y_1, \ldots, y_t, \ldots, y_T\}$ and a model $\Theta = (A, B, \pi_1)$, compute the likelihood of the observations arising from the model, $p(y_{1:T}|\Theta)$.

2. **Optimal State Sequence** - Given a sequence of observations $y_{1:T}$ and a model $\Theta = (A, B, \pi_1)$, choose the most likely sequence of states that explain the data.

3. **Parameter Estimation** - Given a sequence of observations $y_{1:T}$, find values of $\Theta = (A, B, \pi_1)$ that maximize $p(y_{1:T}|\Theta)$.

2.2 Learning Procedures for Discrete HMMs

We now formalize the setup for HMMs with discrete observations and briefly review the learning procedures for them. Let $y_{1:T}$ be a sequence of observations from a discrete random variable taking values in the range $\{1, \ldots, R\}$ at each time point t. The distribution of y_t is generated by a latent state variable s_t which takes values in the range $\{1, \ldots, K\}$. The sequence $s_{1:T} = \{s_1, \ldots, s_t, \ldots, s_T\}$ is generated by a stationary, first order Markov process. Stationary Markov chains are processes where

$$Pr[s_t = j_t, s_{t-1} = j_{t-1}, \ldots, s_1 = j_1] = Pr[s_{t+i} = j_t, s_{t-1+i} = j_{t-1}, \ldots, s_{1+i} = j_1],$$

for all $t \geq 1$ and $i \geq 0$. Additionally, a first order Markov chain is a process where

$$Pr[s_{t+1} = j_{t+1}|s_t = j_t, \ldots, s_1 = j_1] = Pr[s_{t+1} = j_{t+1}|s_t = j_t],$$

which is the source of temporal dependency in the data. The complete data likelihood of a sequence of length T contains both the observed vector $y_{1:T}$ and the unobserved state $s_{1:T}$, and is given by the following expression:

$$p(y_{1:T}, s_{1:T}) = p(s_1)\left[\prod_{t=1}^{T-1} p(s_{t+1}|s_t)\right]\left[\prod_{t=1}^{T} p(y_t|s_t)\right], \tag{2.1}$$

where $p(s_1)$ is the **initial probability** corresponding to the first hidden state, $p(s_{t+1}|s_t)$ denotes the **transition probability** from state s_t to state s_{t+1}, and the conditional probability $p(y_t|s_t)$ denotes the **emission probability** at each time point. The parameters of the HMM are represented

as $\Theta = (A, B, \pi_1)$, where

$$A = ((a_{jk})) : a_{jk} = Pr[s_{t+1} = k | s_t = j], \quad \text{the } K \times K \text{ transition probability matrix,} \tag{2.2}$$

$$B = ((b_{jr}(t))) : b_{jr}(t) = Pr[y_t = r | s_t = j], \quad \text{the } K \times R \text{ emission probability matrix,} \tag{2.3}$$

$$\pi_1 = ((\pi_{1j})) : \pi_{1j} = Pr[\pi_1 = j], \quad \text{the } K \times 1 \text{ initial state vector.} \tag{2.4}$$

For all *j* and *t*, $\sum_{k=1}^{K} a_{jk} = 1$ and $\sum_{r=1}^{R} b_{jr}(t) = 1$. Similarly, $\sum_{j=1}^{K} \pi_{1j} = 1$. The distributions of the states and the emissions can be expressed as:

$$p(s_1 | \pi_1) = \prod_{j=1}^{K} \pi_{1j}^{s_{1,j}}, \tag{2.5}$$

$$p(s_{t+1} | s_t, A) = \prod_{j=1}^{K} \prod_{k=1}^{K} a_{jk}^{s_{t,j} s_{t+1,k}}, \tag{2.6}$$

$$p(y_t | s_t, B) = \prod_{j=1}^{K} \prod_{r=1}^{R} b_{jr}^{s_{t,j} y_{t,r}}, \tag{2.7}$$

and the log-likelihood of the complete data becomes:

$$\log p(y_{1:T}, s_{1:T} | \Theta) = \sum_{j=1}^{K} s_{1,j} \log \pi_{1j} + \sum_{t=1}^{T-1} \sum_{j=1}^{K} \sum_{k=1}^{K} s_{t,j} s_{t+1,k} \log a_{jk} + \sum_{t=1}^{T} \sum_{j=1}^{K} \sum_{r=1}^{R} s_{t,j} y_{t,r} \log b_{jr}, \tag{2.8}$$

where $s_{t,j} = \mathbb{I}(s_t = j)$ and $y_{t,r} = \mathbb{I}(y_t = r)$. The expression on the right hand side of (2.8) cannot be explicitly computed due to the log-likelihood containing latent variables. Since there are *K* states and at each time point *t*, s_t can be in any of these states with non-zero probability, we would be required to sum over over K^T state sequences. This is an intractable problem, and there is no direct way to estimate the parameters. We can, however, use an Expectation-Maximization (EM) approach [Dempster et al., 1977] to choose $\Theta = (A, B, \pi_1)$ such that the log-likelihood is locally maximized. The theory for HMMs, developed by Baum and his colleagues and published in a series of papers [Baum and Petrie, 1966, Baum and Eagon, 1967, Baum and Sell, 1968, Baum et al.,

1970, Baum, 1972], originally dealt primarily with parameter estimation for models where the emission process is discrete. The parameter estimation procedure proposed is commonly known as the Baum-Welch (B-W) algorithm and is a modification of the EM algorithm. Liporace [1982] and Juang et al. [1986] extended it to multivariate, continuous distributions, as well as to their mixtures. Rabiner [1989] contains a comprehensive review of the methods associated with all of the 3 problems of HMMs discussed earlier.

2.2.1 The Baum-Welch Algorithm for parameter estimation

Like the EM algorithm, the B-W algorithm consists of an expectation (E) step and a maximization (M) step. It begins with an initial guess of the parameters, $\Theta^{(0)}$. At the k^{th} iteration where $k \geq 1$,

1. In the E-step, calculate

$$Q(\Theta^{(k)}|\Theta^{(k-1)}) = \mathbb{E}\big(\log p(y_{1:T}, s_{1:T}|\Theta^{(k)}) \mid y_{1:T}, \Theta^{(k-1)}\big),$$

   where $\Theta^{(k)}$ are the parameter values at the k^{th} iteration.

2. In the M-step, maximize $Q(\Theta^{(k)}|\Theta^{(k-1)})$ with respect to $\Theta^{(k)}$.

3. Stop iterations when the supremum norm $\| \Theta^{(k)} - \Theta^{(k-1)} \|_\infty < \epsilon$ for some fixed tolerance $\epsilon > 0$.

To understand how the E-step in the B-W algorithm differs from its EM counterpart, we express the objective function as

$$\begin{aligned} Q(\Theta^{(k)}|\Theta^{(k-1)}) &= \mathbb{E}\big(\log p(y_{1:T}, s_{1:T}|\Theta^{(k)}) \mid y_{1:T}, \Theta^{(k-1)}\big) \\ &= \sum_{s_1=1}^{K} \cdots \sum_{s_T=1}^{K} \log p(y_{1:T}, s_{1:T}|\Theta^{(k)}) p(s_{1:T}|y_{1:T}, \Theta^{(k-1)}), \end{aligned}$$

Using the expression for the log-likelihood of the complete data in (2.8), we have:

$$
\begin{aligned}
Q(\Theta^{(k)}|\Theta^{(k-1)}) = & \sum_{s_1=1}^{K} \cdots \sum_{s_T=1}^{K} \Bigg\{ \sum_{j=1}^{K} s_{1,j} p(s_{1:T}|y_{1:T}, \Theta^{(k-1)}) \log \pi_{1j} \\
& + \sum_{t=1}^{T-1} \sum_{j=1}^{K} \sum_{k=1}^{K} s_{t,j} s_{t+1,k} p(s_{1:T}|y_{1:T}, \Theta^{(k-1)}) \log a_{jk} \\
& + \sum_{t=1}^{T} \sum_{j=1}^{K} \sum_{r=1}^{R} s_{t,j} y_{t,r} p(s_{1:T}|y_{1:T}, \Theta^{(k-1)}) \log b_{jr} \Bigg\} \\
= & \sum_{j=1}^{K} p(s_1 = j|y_{1:T}, \Theta^{(k-1)}) + \sum_{t=1}^{T} \sum_{j=1}^{K} \sum_{k=1}^{K} p(s_{t+1} = k, s_t = j|y_{1:T}, \Theta^{(k-1)}) \\
& + \sum_{t=1}^{T} \sum_{j=1}^{K} \sum_{r=1}^{R} p(s_t = j|y_{1:T}, \Theta^{(k-1)})
\end{aligned}
$$

Note that in the previous expression everything inside the $\log()$ is a function of $\Theta^{(k)}$, while everything outside of it is a function of $\Theta^{(k-1)}$. Following notation in Rabiner [1989], we denote

$$
\begin{aligned}
\gamma_t(j) &= p(s_t = j|y_{1:T}, \Theta^{(k-1)}) \text{ and} \\
\xi_t(j,k) &= p(s_{t+1} = k, s_t = j|y_{1:T}, \Theta^{(k-1)})
\end{aligned}
$$

To maximize Q in the M-step, we need to calculate $\gamma_t(j)$ and $\xi_t(j,k)$ in the E-step. The Forward-Backward recursion [Baum and Eagon, 1967, Baum and Sell, 1968] provides an efficient algorithm to calculate these probabilities and is the main point of difference between the Baum-Welch algorithm and the conventional EM algorithm.

E-step: The Forward-Backward recursion

The Forward Variable is defined as the joint probability of the partial observation sequence up to a time t, and the state s_t at that time point:

$$
F_t(j) = p(y_1, \ldots, y_t, s_t = j|\Theta).
$$

It is calculated for every time point using recursion. To prevent underflow errors, the Forward Variable is scaled at every step, which is equivalent to scaling the entire sequence at the end.

1. **Initialization**: For all $j = 1, \ldots, K$, define

$$F_1(j) = \pi_{1j} \cdot b_{jr}(1) \text{ and}$$
$$\tilde{F}_1(j) = c_1 \cdot F_1(j), \text{ where}$$
$$c_1 = \frac{1}{\sum_{j=1}^{K} F_1(j)}.$$

   Here, π_{1j} is the initial distribution of the state process and b_{jr} is the emission distribution, as defined in (2.4) and (2.3) respectively.

2. **Recursion**: for $t = 2, \ldots, T$ and for each state $k = 1, \ldots, K$, use the recursion

$$F_t(k) = \left[\sum_{j=1}^{K} \tilde{F}_{t-1}(j) \cdot a_{jk}\right] b_{kr}(t)$$
$$\text{and } \tilde{F}_t(j) = c_t \cdot F_t(k), \text{ where}$$
$$c_t = \frac{1}{\sum_{j=1}^{K} F_t(j)}.$$

   Here, a_{jk} are the transition probabilities for the state process, as defined in (2.2).

3. **Termination**: Note that $\tilde{F}_t(j) = (\prod_{\tau=1}^{t} c_\tau) F_t(j)$. Using the definitions provided, this gives us

$$p(y_{1:T}|\tilde{\Theta}) = \sum_{j=1}^{K} F_T(j) = \frac{1}{\prod_{t=1}^{T} c_t},$$

   where $p(y_{1:T}|\tilde{\Theta})$ is the normalizing constant for the posterior.

The Backward Variable is defined as the probability of the last *T-t* observations given that the system is in state *j* at time *t*, i.e.

$$B_{tj} = p(y_{t+1}, \ldots, y_T | s_t = j, \Theta).$$

The Backward Algorithm has similar steps but works its way back from the final time point. We use the same scaling factors that we derived in the Forward Algorithm. We can also compute scaling factors explicitly for the Backward Variable like we did for the Forward Variable.

1. **Initialization**: For each state *j*, set

$$B_T(j) = 1$$
$$\text{and } \tilde{B}_T(j) = c_T \cdot B_T(j).$$

2. **Recursion**: for $t = T-1, \ldots, 1$ and each state *j*, calculate

$$B_t(j) = \sum_{k=1}^{K} a_{jk} \cdot \tilde{B}_{t+1}(k) \cdot b_{kr}(t+1),$$
$$\tilde{B}_t(j) = c_t \cdot B_t(j).$$

The Forward and Backward algorithms can be run in parallel if the Backward Algorithm calculates its own normalizing constants instead of reusing the ones provided by the Forward Algorithm. Once both Forward and Backward variables are calculated, we get

$$\gamma_t(j) = \frac{\tilde{F}_t(j) \cdot \tilde{B}_t(j)}{\sum_{j=1}^{K} \tilde{F}_t(j) \cdot \tilde{B}_t(j)},$$
$$\xi_t(j,k) = \frac{\tilde{F}_t(j) \cdot a_{jk} \cdot b_{kr}(t+1) \cdot \tilde{B}_{t+1}(k)}{\sum_{j=1}^{K} \sum_{k=1}^{K} \tilde{F}_t(j) \cdot a_{jk} \cdot b_{kr}(t+1) \cdot \tilde{B}_{t+1}(k)}.$$

The M-step: Maximizing Q

Note that $\gamma_t(j) = \sum_{k=1}^{K} \xi_t(j,k)$. Furthermore, $\gamma_t(j)$ and $\xi_t(j,k)$ can be interpreted as:

$$\sum_{t=1}^{T-1} \gamma_t(j) = \text{ expected number of state transitions from state } j, \text{ and}$$
$$\sum_{t=1}^{T-1} \xi_t(j,k) = \text{ expected number of transitions from state } j \text{ to state } k.$$

Q is then maximized using the following expressions, described as re-estimation expressions by Rabiner [1989]:

$$\begin{aligned}\bar{\pi}_{1j} &= \gamma_1(j),\\ \bar{a}_{jk} &= \frac{\sum_{t=1}^{T-1}\xi_t(j,k)}{\sum_{t=1}^{T-1}\gamma_t(j)},\\ \bar{b}_{jr}(t) &= \frac{\sum_{t=1}^{T}\gamma_t(j)\cdot y_{t,r}}{\sum_{t=1}^{T}\gamma_t(j)},\end{aligned}$$

where $y_{t,r} = \mathbb{I}(y_t = r)$. The algorithm iterates until $\| \Theta^{(k)} - \Theta^{(k-1)} \|_\infty < \epsilon$ for some predefined tolerance level ϵ.

2.2.2 Viterbi Decoding for the most likely sequence of states

Once the model parameters of an HMM have been estimated, the next topic of interest is identifying the most likely sequence of states, given the model, that could have generated the observed data. Being able to label each observation with a state allows for the exploration of the statistical properties of data arising from a certain state, and in many cases aids in interpreting the states. This process is often referred to as decoding. However, this problem does not have a unique solution since it depends on the definition of an *optimal sequence* of states. One possible optimality criteria is to choose the most likely state at each instant and form a sequence out of them. This will maximize the expected number of correct states. However, such an approach does not take into account sequences of states and the Markovian nature of the underlying process. One could instead try to maximize the expected number of correct pairs of consecutive states, or even longer chains. However, the most common optimality criterion considers the most likely sequence, or path, of states. This maximizes $p(s_{1:T}|y_{1:T}, \Theta)$, which is equivalent to maximizing $p(s_{1:T}, y_{1:T}|\Theta)$ since $y_{1:T}$ is known. A dynamic programming method called the Viterbi Algorithm or Viterbi decoding [Viterbi, 1967] is used to find this single best state sequence.

The Viterbi Algorithm

Define the quantity $\delta_t(j)$ as:

$$\delta_t(j) = \max_{s_{1:(t-1)}} p(s_1, s_2, \ldots, s_t = j, y_1, y_2, \ldots, y_t|\Theta)$$

which has the highest probability (likelihood) at time *t* along a single path, that accounts for the first *t* observations and ends in state *j*. By induction,

$$\delta_{t+1}(k) = \big[\max_{1\leq j\leq K} \delta_t(j)a_{jk}\big] \cdot b_{kr}(t+1).$$

We also need to keep track of the arguments which correspond to the highest likelihood for each combination of time point *t* and state *j*. This is stored in the variable $\psi_t(j)$. Thus, decoding the most likely path conditional on the model and data consists of the following steps:

1. **Initialization:**

$$\delta_1(j) = \pi_{1j} \cdot b_{jr}(1), j = 1, \ldots, K,$$
$$\psi_1(j) = 0.$$

2. **Recursion:** For $j, k = 1, \ldots, K$, $t = 2, \ldots, T$,

$$\delta_t(k) = \max_{1\leq j\leq K} \big[\delta_{t-1}(j)a_{jk}\big] \cdot b_{kr}(t),$$
$$\psi_t(k) = \arg\max_{1\leq j\leq K} \big[\delta_{t-1}(j)a_{jk}\big].$$

3. **Termination:** For $j = 1, \ldots, K$,

$$P^* = \max_{1\leq j\leq K} \delta_T(j),$$
$$q_T^* = \arg\max_{1\leq j\leq K} \delta_T(j).$$

4. **Path backtracking:** For $t = T-1, T-2, \ldots, 1$,

$$q_t^* = \psi_{t+1}(q_{t+1}^*).$$

The first three steps of the Viterbi recursion are quite similar to the Forward Algorithm, with the main difference being that we maximize over the previous states instead of summing over them. The sequential maximization results in the path with the highest probability (likelihood), and the last step identifies the states which are on that path.

2.3 The Hidden Markov Model for Precipitation

Our HMM for precipitation follows along the lines of Hughes and Guttorp [1994] and Robertson et al. [2006]. Let $\mathbf{y}_{1:T} = \{\mathbf{y}_1, \ldots, \mathbf{y}_t, \ldots, \mathbf{y}_T\}$, where $\mathbf{y}_t' = (y_{t1}, \ldots, y_{tL})$, be the $L \times T$ matrix of precipitation amounts for a network of L grid points over T days. Let $s_{1:T} = \{s_1, \ldots, s_t, \ldots, s_T\}$ be the set of hidden (unobserved) weather states, where $s_t \in \{1, \ldots, K\}$. At each location,

$$p(y_{tl} = y | s_t = j) = \begin{cases} c_{jl0} & \text{if } y = 0 \\ \sum_{m=1}^{M} c_{jlm} f(y | \omega_{jlm}, \lambda_{jlm}) & \text{if } y > 0 \end{cases} \tag{2.9}$$

with $c_{jlm} \geq 0$ and $\sum_{m=0}^{M} c_{jlm} = 1$ for all $l = 1, \ldots, L$ and $j = 1, \ldots, K$; $f(\cdot|\omega, \lambda)$ is the density function of a Gamma distribution with shape $\omega > 0$ and rate $\lambda > 0$. Studies often simplify this by setting $\omega = 1$ and working with Exponential distributions. The states arise from a stationary, first-order Markov process:

$$p(s_1, \ldots, s_T) = p(s_1) \prod_{t=1}^{T-1} p(s_{t+1} | s_t), \tag{2.10}$$

where $p(s_t|s_{t-1})$ form a $K \times K$ matrix of state transition probabilities $A = ((a_{jk}))$, $1 \leq j \leq K$, $1 \leq k \leq K$, and $p(s_1) = \pi_{1j}$ is the initial distribution. Daily rainfall $\mathbf{y}_t$ depends only on the state

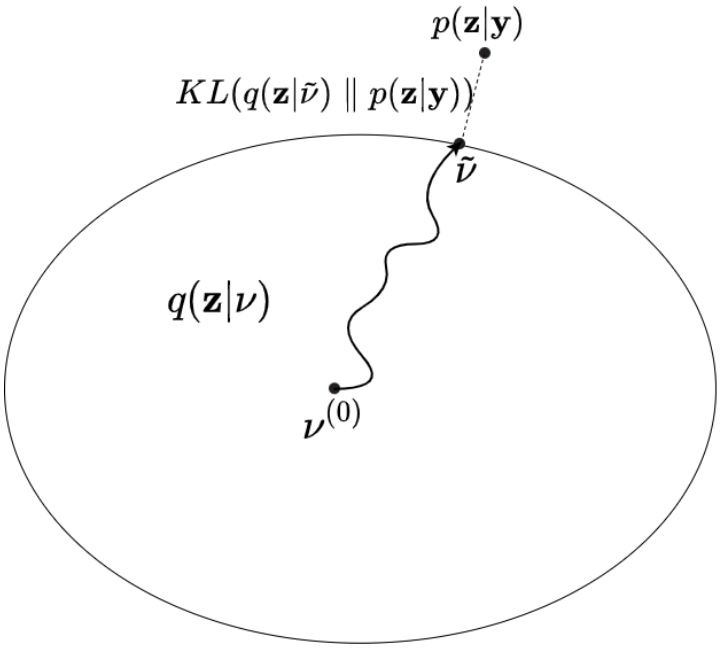

Figure 2.2: Variational inference represented as an optimization problem.

s_t on day t:

$$p(\boldsymbol{y}_{1:T}|s_{1:T}) = \prod_{t=1}^{T} p(\boldsymbol{y}_t|s_t), \text{ and} \tag{2.11}$$

$$p(\boldsymbol{y}_{1:T}, s_{1:T}) = \left\{p(s_1) \prod_{t=1}^{T-1} p(s_{t+1}|s_t)\right\}\left\{\prod_{T=1}^{T} p(\boldsymbol{y}_t|s_t)\right\}. \tag{2.12}$$

Spatial dependence is assumed to be captured implicitly through the dependence of the emissions on the state process $\{s_t\}$. The L location components of $\mathbf{y}_t$ are independent of each other given s_t, i.e.

$$p(\boldsymbol{y}_t|s_t) = \prod_{l=1}^{L} p(y_{tl}|s_t). \tag{2.13}$$

2.4 Variational Bayes for Posterior Approximation

Variational Bayesian methods (alternatively called variational inference, variational approximation, or VB) aim to approximate the posterior distribution using optimization. In many real

life applications where the data is high dimensional or there is a complex hierarchical structure present in the model, Markov chain Monte Carlo (MCMC) methods are impractical or outright infeasible. Variational Bayes first posits a family of approximate distributions $\mathbb{Q}$ over the latent variables and parameters, which have its own variational parameters. It then optimizes within this family to find a member, i.e. the parameter settings, such that the resulting variational distribution is the closest approximation to the true posterior within the chosen variational family.

The computational efficiency and accuracy of the variational approximation depend on three things. The first is the distance metric that assesses the quality of the approximation. The most common approach is to minimize the Kullback-Leibler (KL) divergence [Kullback and Leibler, 1951]. Let KL divergence between two distributions $p(x)$ and $q(x)$ of a discrete random variable x is defined as

$$D_{KL}(p(x) \parallel q(x)) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}.$$

The continuous version of the KL divergence is

$$D_{KL}(p(x) \parallel q(x)) = \int_{-\infty}^{\infty} p(x) \log \frac{p(x)}{q(x)} dx$$

for a continuous variable x. The optimal variational posterior satisfies

$$\tilde{q}(\cdot) = \arg \min_{q(\cdot) \in \mathbb{Q}} KL\big(q(\boldsymbol{z}) \parallel p(\boldsymbol{z}|\boldsymbol{y})\big). \tag{2.14}$$

The optimum values of the hyperparameters are found using a variational EM-like algorithm; Figure 2.2 contains a visual representation of variational inference. Note that the true posterior is typically not in the variational family $\mathbb{Q}$.

The objective function in (2.14) involves the posterior $p(\boldsymbol{z}|\boldsymbol{y})$ which is often difficult to compute in practice. However, the minimization in (2.14) is equivalent to the maximization of a quantity known as the evidence lower bound (ELBO), defined as

$$\text{ELBO}(q) = \mathbb{E}[\log p(\boldsymbol{z}, \boldsymbol{y})] - \mathbb{E}[\log q(\boldsymbol{z})]. \tag{2.15}$$

The equivalence follows from the identity

$$\log p(\boldsymbol{y}) = KL(q(\boldsymbol{z}) \parallel p(\boldsymbol{z}|\boldsymbol{y})) + \text{ELBO}(q), \tag{2.16}$$

and the fact that the evidence log-likelihood $\log p(\boldsymbol{y})$ (also known as the ensemble log-likelihood) is a function of only the data and does not involve the parameters. All expectations in (2.16) are taken with respect to the variational posterior distribution $q(\boldsymbol{z})$. Since the KL divergence is non-negative, it follows that the ELBO is indeed a lower bound; Jordan et al. [1999] obtained the bound directly by applying Jensen's inequality to $\log p(\boldsymbol{y})$. The methodology for using VB optimization to estimate HMM parameters is outlined in MacKay [1997] for a discrete emission process, and in Ghahramani and Beal [2000] for emissions arising from a conjugate exponential family.

While $KL(q \parallel p)$ is the commonest distance metric used since it leads to analytically tractable expectations for conjugate exponential families, it is not the only option. Naesseth et al. [2020] have proposed using $KL(p \parallel q)$. In an extension of the same idea, Dieng et al. [2017] have proposed minimizing the χ-divergence from the posterior to the variational family,

$$D_{\chi^2}(p \parallel q) = \mathbb{E}_{q(\mathbf{z};\lambda)}\left[\left(\frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z};\lambda)}\right)^2 - 1\right].$$

α-divergence metrics have found traction in recent years, which minimize a divergence function between the variational posterior and the joint α-fractional posterior distribution, for $\alpha \in (0, 1)$. The corresponding method is known as α-VB [Li and Turner, 2016, Hernandez-Lobato et al., 2016]. Yang et al. [2020] have shown that when operating in a frequentist setup, point estimates derived from the α-VB procedure converge at an optimal rate to the true parameters in a wide range of problems.

The second factor under consideration is the choice of the variational family. The goal is to choose $\mathbb{Q}$ that is expressive enough to allow for a good approximation to the posterior, and is simple enough to allow tractable optimization. A common choice is to restrict the analysis to a

family of distributions where the latent variables and the parameters are all mutually independent:

$$q(\boldsymbol{z}) = \prod_{j=1}^{m} q_j(z_j), \tag{2.17}$$

and each latent component z_j has its own variational marginal posterior. This is known as the mean-field assumption, and the corresponding family of distributions is known as the mean-field variational family.

An additional assumption that tends to be made is that all variables and their parameters arise from the conjugate exponential family; i.e. the complete data likelihood is in the exponential family and the parameters have conjugate priors. Under this assumption, the posterior distributions are conjugate to the priors. Along with the mean field assumption, this simplifies the variational Bayes EM (VBEM) algorithm significantly. An overview is provided below.

Mean Field Variational Bayes

Consider the case with observed data $\mathbf{y} = (y_1, \ldots, y_n)^T$ and latent variables $\mathbf{x} = (x_1, \ldots, x_n)^T$ parameterized by a vector of parameters β. The parameter vector β follows a distribution with hyperparameters α. The joint distribution of the complete data and the prior thus takes the form

$$p(y, x, \beta|\alpha) = p(\beta|\alpha) \cdot \prod_{i=1}^{n} p(y_i, x_i|\beta).$$

The complete conditionals are assumed to be in the exponential family to take advantage of conjugacy which simplifies calculations. The complete data likelihood is also assumed to have a conditional independence structure,

$$p(y_i, x_i|y_{-i}, x_{-i}, \beta) = p(y_i, x_i|\beta),$$

where y_{-i} denotes all observations except y_i.The conditional independence structure is required for mean field variational Bayes to work. β is considered a **global variable**, i.e. it is the same for all *i*. On the other hand x_i's are **local variables**, since they vary for each *i*. Further, x_i is *J*-dimensional, i.e. $x_i = x_{i,1:J}$. The complete conditionals of the global and local variables are in

the exponential family and have the form:

$$p(\beta|y,x,\alpha) = h(\beta)\exp\big\{\eta_g(y,x,\alpha)^T t(\beta) - a_g\big(\eta_g(y,x,\alpha)\big)\big\}, \tag{2.18}$$

$$p(x_{i,j}|y_i,x_{i,-j},\beta) = h(x_{i,j})\exp\big\{\eta_l(y_i,x_{i,-j},\beta)^T t(x_{i,j}) - a_l\big(\eta_l(y_i,x_{i,-j},\beta)\big)\big\}, \tag{2.19}$$

where $h(\cdot)$ is the base measure, $a(\cdot)$ is the log normalizer, $\eta(\cdot)$ is the natural parameter vector, and $t(\cdot)$ is the vector of sufficient statistics. All throughout this discussion, the subscripts *l* and *g* (for example in $\eta_l(\cdot), \eta_g(\cdot)$) refer to local and global parameters respectively. The complete data likelihood thus has the form:

$$p(y,x|\beta) = \prod_{i=1}^{n} h(y_i,x_i)\cdot\exp\big\{\beta^T t(y_i,x_i) - a_l(\beta)\big\}. \tag{2.20}$$

Similarly, the prior takes the form

$$p(\beta|\alpha) = h(\beta)\cdot\exp\big\{\alpha^T t(\beta) - a_g(\alpha)\big\}, \tag{2.21}$$

where $t(\beta) = \big[\beta, -a_l(\beta)\big]$; α is similarly written as $\alpha = \big[\alpha_1,\alpha_2\big]$. Comparing with (2.18), we get an expression for $\eta_g(y,x,\alpha)$:

$$\eta_g(y,x,\alpha) = \big(\alpha_1 + \sum_{i=1}^{n} t(y_i,x_i), \alpha_2 + n\big). \tag{2.22}$$

The mean field variational family has posteriors which are separable, that is

$$q(x,\beta) = q(\beta|\lambda)\prod_{i=1}^{n}\prod_{j=1}^{J} q(x_{i,j}|\phi_{i,j}) \tag{2.23}$$

for global variational parameters λ and local variational parameters ϕ. Since the prior and the likelihood are in the exponential family, the variational posteriors are also in the exponential family

with complete conditionals given by:

$$q(\beta|\lambda) = h(\beta) \cdot \exp\big\{\lambda^T t(\beta) - a_g(\lambda)\big\}, \tag{2.24}$$

$$q(x_{i,j}|\phi_{i,j}) = h(x_{i,j}) \cdot \exp\big\{\phi_{i,j}^T t(x_{i,j}) - a_l(\phi_{i,j})\big\}. \tag{2.25}$$

Because of the mean field condition in (2.23), the ELBO $\mathcal{L}$ in (2.15) can be maximized component-wise, and then added up afterwards. The ELBO for the global latent variables takes the form:

$$\mathcal{L}(\lambda) = \mathbb{E}_q \log p(\beta|y,x) - \mathbb{E}_q \log q(\beta) + \text{ constant.} \tag{2.26}$$

Since the complete conditional of β is in the exponential family and of the form (2.18), $\mathbb{E}_q t(\beta) = \nabla_\lambda a_g(\lambda)$. Here $\nabla_\lambda a_g(\lambda)$ is the gradient of the ELBO with respect to λ, i.e. $\nabla_\lambda a_g(\lambda) = \frac{\partial a_g(\lambda)}{\partial \lambda}$. This gives us

$$\mathcal{L}(\lambda) = \mathbb{E}_q[\eta_g(y,x,\alpha)]^T \nabla_\lambda a_g(\lambda) - \lambda^T \nabla_\lambda a_g(\lambda) + a_g(\lambda) + \text{ constant.} \tag{2.27}$$

The first order condition on the ELBO simplifies to

$$\nabla_\lambda \mathcal{L}(\lambda) = \nabla_\lambda^2 a_g(\lambda)\big[\mathbb{E}_q \eta_g(y,x,\alpha) - \lambda\big]. \tag{2.28}$$

The LHS of this equation is 0 when $\mathbb{E}_q \eta_g(y,x,\alpha) = \lambda$. This provides us coordinate updates of the form

$$\lambda^{new} = \lambda^{old} + \mathbb{E}_q \eta_g(y,x,\alpha), \tag{2.29}$$

which optimizes the ELBO for each global variable. Similarly, for the local parameters $\phi_{i,j}$, the first-order condition on the ELBO simplifies to

$$\nabla_{\phi_{i,j}} \mathcal{L} = \nabla_{\phi_{i,j}}^2 a_l(\phi_{i,j})[\mathbb{E}_q \eta_l(y_i, x_{i,-j}, \beta) - \phi_{i,j}], \tag{2.30}$$

which is 0 when $\phi_{i,j} = \mathbb{E}_q \eta_l(y_i, x_{i,-j}, \beta)$. Note that the value of β in the local update depends on the value of λ. Similarly, the global update depends on $\phi_{i,j} = \mathbb{E}_q(x_{i,j})$. The global and local updates therefore form a coupled system which is solved using a variational Bayes EM (VBEM) algorithm [Ghahramani and Beal, 2000, Blei et al., 2017].

There is no unified approach for cases where the mean field assumption does not hold. It is common practice to keep the mean-field assumption for the global variables. Mild relaxations exist where instead of updating hyperparameters of a global variable independently, they are updated sequentially. The assumption often proves too restrictive when the latent variables are not independently distributed. Situations where the mean field assumption is not a plausible assumption are usually referred to by the umbrella term of *structured variational inference*, referring to the structure that needs to be enforced in the variational posterior.

A common example of models which require a structured variational approach are belief networks and Markov graphs. These are models which can be represented as directed and undirected graphs; HMMs are an example of directed acyclic graphs. For such models, Ghahramani and Beal [2000] extended the results for mean field variational Bayes and provided a VBEM algorithm for graphical models. The earliest treatment of HMMs in the variational literature as far as we can tell, was by MacKay [1997]. He laid out the steps of the VBEM algorithm for parameter estimation in HMMs with discrete emissions, similar to the one discussed in Section 2.2. Jordan et al. [1999] also provides a comprehensive discussion on variational methods for graphical models. In general, however, the VBEM algorithm often needs to be adapted on a case-by-case basis depending on the structure that is actually present in the model.

The final factor affecting the quality of the variational approximation is the choice of optimization algorithm. A commonly chosen method for a mean-field variational family is coordinate ascent or coordinate ascent variational inference (CAVI) [Ghahramani and Beal, 2000, Blei et al., 2017], which updates each component of the variational posterior individually. Stochastic gradient optimization [Hoffman et al., 2013] can result in speedups, and is used for large datasets where scalability becomes a requirement. Ranganath et al. [2014] proposed black box variational inference (BBVI) as a scalable, gradient based optimization procedure for the ELBO which is agnostic to the distribution of the prior. BBVI seeks to generalize the optimization to arbitrary variational

families of posterior; a similar approach is taken by Kucukelbir et al. [2017] for automatic differentiation variational inference (ADVI). Most of the black box approaches come with assumptions which are not easy to relax for structured models like HMMs. We conclude this section with an overview of stochastic variational Bayes which we would be using for parameter estimation.

Stochastic Variational Bayes

The primary bottleneck of coordinate ascent algorithms for variational Bayes is that each update requires a pass over all our observed and hidden data for every single variable. One way to reduce this would be to replace the expensive, exact updates with noisy updates that are unbiased and faster to compute. Vanilla stochastic optimization works very well for the mean field variational family, and a brief discussion is provided.

We first make a note of the natural gradient and its use in optimizing probabilistic functions. In gradient ascent optimization problems, the Euclidean gradient points in the direction of the steepest ascent in Euclidean space. For example, if $L(\boldsymbol{w})$ is our objective function under consideration, with $\boldsymbol{w} = (w_1, \ldots, w_n)^T$, the Euclidean gradient is given by the expression

$$\nabla L(\boldsymbol{w}) = \left(\frac{\partial}{\partial w_1} L(\boldsymbol{w}), \ldots, \frac{\partial}{\partial w_n} L(\boldsymbol{w}) \right)^T.$$

However, probabilistic objective functions reside not in Euclidean space but in a curved manifold, known as a Riemannian space. The Euclidean gradient is not ideal in such cases, and the natural gradient is preferable since it points in the direction of the steepest ascent in Riemannian space [Amari, 1998]. For the ELBO, which is a probabilistic objective function, Hoffman et al. [2013] showed that the relationship between the natural gradient $\tilde{\nabla}_\lambda \mathcal{L}(\lambda)$ and the Euclidean gradient $\nabla_\lambda \mathcal{L}(\lambda)$ is

$$\tilde{\nabla}_\lambda \mathcal{L}(\lambda) = \mathcal{G}(\lambda)^{-1} \nabla_\lambda \mathcal{L}(\lambda), \tag{2.31}$$

where $\mathcal{G}(\lambda)$ is the Fisher Information matrix of $q(\lambda)$, i.e.

$$\begin{aligned}\mathcal{G}(\lambda) &= \mathbb{E}_\lambda\big[\big(\nabla_\lambda \log q(\beta|\lambda)\big)\big(\nabla_\lambda \log q(\beta|\lambda)\big)^T\big] \\ &= \nabla^2_\lambda a_g(\lambda)\end{aligned}$$

if $q(\beta|\lambda)$ is in the exponential family. Using the expression for the gradient of the ELBO, we get

$$\tilde{\nabla}_\lambda \mathcal{L}(\lambda) = \mathbb{E}_q \eta_g(y, x, \alpha) - \lambda. \tag{2.32}$$

Thus, the gradient ascent step for optimizing the ELBO for λ takes the form

$$\lambda^{new} = \lambda^{old} + \tau \tilde{\nabla}_\lambda \mathcal{L}(\lambda), \tag{2.33}$$

where τ is the step size. A similar update can be derived for $\phi_{i,j}$, the latent variables. If we use the natural gradient as the direction of ascent, our objective function will have the steepest ascent in the space where the local distance is defined by the KL divergence and not the $\mathbb{L}^2$ norm. Note that the natural gradient update is equivalent to the coordinate ascent update when $\tau = 1$. However, this is still slow as it requires going over the entire data at every iteration of the optimization. Stochastic optimization addresses that by replacing the gradient by a noisy estimate of the gradient that is cheaper to compute.

Consider for the objective function $\mathcal{L}(\lambda)$ a random function $B(\lambda)$ which is an unbiased estimator of the gradient $\nabla_\lambda \mathcal{L}(\lambda)$, i.e. $\mathbb{E}_q B(\lambda) = \nabla_\lambda \mathcal{L}(\lambda)$. For example, $B(\lambda)$ could be the gradient computed based on random samples, or minibatches, taken from the entire data. Then the stochastic gradient ascent step for optimizing the ELBO for λ is

$$\lambda^k = \lambda^{k-1} + \tau_k \cdot b_k(\lambda^{k-1}),$$

where $b_k(\cdot)$ is an independent draw from the noisy gradient B. If τ_k satisfies the Robbins-Monro

conditions [Robbins and Monro, 1951], namely

$$\sum_k \tau_k = \infty$$
$$\text{and } \sum_k \tau_k^2 < \infty,$$

then λ^k converges to a local optimum of $f(\cdot)$. If G_k is any positive definite matrix of appropriate dimensions, a similar property holds [Hoffman et al., 2013]:

$$\lambda^k = \lambda^{k-1} + \tau_k \cdot G_k^{-1} b_k(\lambda^{k-1}). \tag{2.34}$$

In particular, if we choose $G_k = \mathcal{G}_k$, the Fisher information matrix, the natural gradient provides the direction of the steepest ascent for the gradient optimization. In the mean field setup, we choose a noisy gradient by randomly sampling a single data point and doing all computations based on that single data point. In this case,

$$b_k(\cdot) = \tilde{\nabla}\mathcal{L}_i = n \cdot \mathbb{E}_q \eta_g(y_i, x_i, \alpha) - \lambda. \tag{2.35}$$

2.5 Variational Bayes for Hidden Markov Models

The methodology for using VB for parameter estimation in HMMs where the emissions arise from a conjugate exponential family was outlined mainly in three papers. MacKay [1997] did early work for HMMs with discrete emission distributions, and Ji et al. [2006] derived the VB algorithm for HMMs where the emissions are continuous mixtures. Finally, McGrory and Titterington [2009] have discussed model selection in variational HMMs using the Deviance Information Criterion (DIC) [Spiegelhalter et al., 2002] when the size of the model is unknown. Current work on HMMs with continuous emissions and its mixtures often assume the distributions of the emissions to be Gaussian or Gaussian mixtures. Maximum likelihood parameter estimation in HMMs also tend to deal with Gaussian mixtures, e.g, in Rabiner [1989]. In theory, Gaussian mixtures can be replaced by any conjugate exponential family distribution or their mixtures and

the same steps will apply. The primary challenge in modeling HMMs comes from the dependence structure in the state process $\{S_t\}$. While the VBM step can employ the mean field assumption, the VBE step is modified to accommodate the Forward-Backward algorithm, which calculates the variational equivalent of $\gamma_t(j)$ and $\xi_t(j,k)$ defined in (2.2.1).

2.5.1 The variational Forward-Backward Algorithm

As before, the Forward Variable is defined as the joint probability of the partial observation sequence up to a time *t*, and the state s_t at that time point

$$F_{tj} = p(y_1, \ldots, y_t, s_t = j).$$

1. **Initialization**: For all $j = 1, \ldots, K$, define

$$\begin{aligned} F_{1j} &= \pi_1 \cdot p(y_1 | s_1 = j), \\ c_1 &= \frac{1}{\sum_{j=1}^{K} F_{1j}} \text{ and normalize} \\ \tilde{F}_{1j} &= c_1 \cdot F_{1j}. \end{aligned}$$

2. **Recursion**: for $t = 2, \ldots, T$ and for each state $k = 1, \ldots, K$, use the recursion

$$\begin{aligned} F_{tk} &= \left[\sum_{j=1}^{K} \tilde{F}_{t-1,j} \cdot p(s_t = k | s_{t-1} = j) \right] p(y_t | s_t = k) \text{ and normalize} \\ \tilde{F}_{tj} &= c_t \cdot F_{tk} \text{ where} \\ c_t &= \frac{1}{\sum_{j=1}^{K} F_{tj}}. \end{aligned}$$

Note that $\tilde{F}_{tj} = (\prod_{\tau=1}^{t} c_\tau) F_{tj}$. Using the definitions provided, this gives us

$$q(y|\Theta) = \sum_{j=1}^{K} F_{Tj} = \frac{1}{\prod_{t=1}^{T} c_t}, \tag{2.36}$$

where $q(y|\Theta)$ is the normalizing constant for the variational posterior of the latent variables. Re-

call that the Forward Algorithm is used as part of the E-step of the optimization process, with the values of the parameters in Θ set to their means, i.e., $\Theta \equiv \tilde{\Theta}$. Thus $q(y|\Theta)$ can equivalently also be expressed as $p(y|\tilde{\Theta})$.

The Backward Variable is defined as the probability of generating the last $T-t$ observations given that the system is in state *j* at time *t*

$$B_{tj} = p(y_{t+1}, \ldots, y_T | s_t = j).$$

The Backward Algorithm has similar steps but works its way back from the final time point.

1. **Initialization**: For each state *j*, set

$$B_{Tj} = 1 \text{ , and}$$
$$\tilde{B}_{Tj} = c_T \cdot B_{Tj}.$$

2. **Recursion**: for $t = T-1, \ldots, 1$ and each state *j*, calculate

$$B_{tj} = \sum_{k=1}^{K} p(s_{t+1} = k | s_t = j) \cdot \tilde{B}_{t+1,k} \cdot p(y_{t+1} | s_{t+1} = k),$$
$$\tilde{B}_{tj} = c_t \cdot B_{tj}.$$

The two algorithms can be run in parallel. Once both variables are calculated, we get

$$q_s(s_t = j | y_1, \ldots, y_T) \propto \tilde{F}_{tj} \cdot \tilde{B}_{tj}, \text{ and}$$
$$q_s(s_t = j, s_{t+1} = k) \propto \tilde{F}_{tj} \cdot p(s_{t+1} = k | s_t = j) \cdot p(y_{t+1} | s_{t+1} = k) \cdot \tilde{B}_{t+1,k}.$$

2.6 A Conjugate Prior for the Two-parameter Gamma Distribution

The majority of the variational Bayes literature has relied on the conjugacy properties of the exponential family of distributions to simplify computations. In the context of the HMM for precipitation as described in Section 2.3, we noted that Gamma and Exponential distributions are commonly used to model positive precipitation. Exponential distributions require fewer parameters than the Gamma distribution, while Gamma distributions require fewer mixture components and often outperform their Exponential distribution counterparts in terms of model fit. We have also noted this in our previous research when estimating parameters using maximum likelihood methods [Kroiz et al., 2020a,b]. In this section, we go over a conjugate prior for the two-parameter Gamma distribution that can be used for variational Bayes parameter estimation in an HMM for precipitation where positive precipitation is specified as a mixture of Gamma distributions.

Miller [1980] provides a conjugate prior for the two-parameter Gamma distribution to be used for Bayesian analysis, based on the work of Damsleth [1975]. Let the population density have the Gamma form with shape α and rate θ,

$$g(y_i|\alpha,\theta) = \frac{\theta^\alpha}{\Gamma(\alpha)} y_i^{\alpha-1} \exp\{-y\theta\}, \tag{2.37}$$

with $y_i > 0, \alpha > 0, \theta > 0$. Then for a sample of size n, the likelihood function is

$$p(y|\alpha,\theta) = \frac{\theta^{n\alpha}}{(\Gamma(\alpha))^n} (\prod_{i=1}^{n} y_i)^{\alpha-1} \exp\{-\theta \sum_{i=1}^{n} y_i\}. \tag{2.38}$$

The conjugate prior when both parameters are unknown takes the form

$$\pi(\alpha,\theta|p,q,r,s) = \frac{1}{C} \frac{\theta^{\alpha s-1}}{(\Gamma(\alpha))^r} p^{\alpha-1} \exp\{-q\theta\}, \tag{2.39}$$

$$\text{where } p > 0, q > 0, r > 0, s > 0 \ni \frac{\sqrt[r]{p}}{q/r} < 1,$$

$$\text{and } C = \int_0^\infty \frac{p^{\alpha-1}\Gamma(\alpha s)}{(\Gamma(\alpha))^r q^{\alpha s}} d\alpha.$$

The posterior joint density is then proportional to the product of Eqns. (2.37) and (2.39)

$$p(\alpha, \theta | y, p, q, r, s) \propto \frac{\theta^{\alpha\tilde{s}-1}}{(\Gamma(\alpha))^{\tilde{r}}} \exp\{-\tilde{q}\theta\}(\tilde{p})^{\alpha-1}, \tag{2.40}$$

$$\begin{aligned} \text{where } \tilde{s} &= s + n, & \tilde{r} &= r + n, \\ \tilde{q} &= q + \sum_{i=1}^{n} y_i, & \tilde{p} &= p \prod_{i=1}^{n} y_i. \end{aligned}$$

Thus the posterior conditional density of the rate follows a Gamma distribution

$$p(\theta | \alpha, s, q, y) = \text{Gamma}(\theta | \alpha\tilde{s}, \tilde{q}), \tag{2.41}$$

and the marginal posterior density of the shape has the kernel

$$p(\alpha) \propto \frac{\Gamma(\alpha\tilde{s})}{(\Gamma(\alpha))^{\tilde{r}}} \left[\sqrt[\tilde{s}]{\tilde{p}}/\tilde{q}\right]^{\alpha\tilde{s}}, \tag{2.42}$$

whose moments and functionals need to be computed through numerical integration.

Estimating the posterior of the shape parameter

Computing the density function and the moments of the posterior of α would require us to evaluate integrals of the form $\int_0^\infty a^i p(\alpha) d\alpha$ for $i = 0, 1$ etc. The right hand side of Equation 2.42 could be difficult to evaluate directly since all three terms tend to have large magnitudes due to their dependence on the sample and the sample size. As with the original paper, we calculate the log of the integrand as

$$\log(\alpha^i p(\alpha)) = i \log \alpha + \log \Gamma(\alpha\tilde{s}) + \alpha\tilde{s} \log(\sqrt[\tilde{s}]{\tilde{p}}/\tilde{q}) - \tilde{r} \log \Gamma(\alpha), \tag{2.43}$$

and exponentiate afterwards.

Miller provides an approximation to calculate the first four moments of the posterior distribution of θ as functions of the cumulants of the posterior of α, and fitting a Pearson family curve

to the data. However, if we are just interested in calculating the posterior mean of θ, this is even simpler. Since we know the posterior conditional density of θ follows a Gamma distribution, we can write the unconditional mean in terms of the moments of α

$$\begin{aligned} \mathbb{E}_\theta(\theta) &= \mathbb{E}_\alpha \mathbb{E}_{\theta|\alpha}(\theta|\alpha) \\ &= \frac{\tilde{s}}{\tilde{q}} \mathbb{E}_\alpha \alpha \end{aligned} \tag{2.44}$$

2.6.1 The Deviance Information Criterion

Model comparison techniques are usually based on a trade-off between model fit and model complexity. The Bayesian information criterion (BIC) is a common metric, but it does not always perform well for model selection in HMMs [Bellone, 2000]. It fares especially poorly when sample size is low or model complexity is high. The deviance information criterion (DIC) was proposed as an alternative by Spiegelhalter et al. [2002], and is still based on the core premise of trading-off Bayesian measures of model complexity and fit. The DIC initially focused on exponential family models, but has been extended using variational approximations to mixture models [McGrory and Titterington, 2007] and HMMs [McGrory and Titterington, 2009]. For HMMs, McGrory and Titterington [2009] define the DIC as

$$\text{DIC} = \overline{D(\theta)} + p_D,$$

where $D(\theta) = -2 \log p(y|\theta)$, and $\overline{D(\theta)}$ is its expectatation with respect to $p(\theta|y)$. $D(\theta)$ measures model fit and p_D measures model complexity. The latter is defined as

$$\begin{aligned} p_D &= \overline{D(\theta)} - D(\tilde{\theta}) \\ &= \mathbb{E}_{\theta|y}\{-2 \log p(y|\theta)\} + 2 \log p(y|\tilde{\theta}), \end{aligned}$$

where $\tilde{\theta}$ is the posterior mean or mode of the parameters of interest. Based on the expressions above, the DIC is expressed as

$$\text{DIC} = -2\log p(y|\tilde{\theta}) + 2p_D, \tag{2.45}$$

$$\text{and } p_D \approx -2\int q_\theta(\theta)\log\left\{\frac{q_\theta(\theta)}{p(\theta)}\right\}d\theta + 2\log\left\{\frac{q_\theta(\tilde{\theta})}{p(\tilde{\theta})}\right\} \tag{2.46}$$

where p_D is approximated with the aid of the variational distribution $q(\cdot)$. $p(y|\tilde{\theta})$ can be obtained as part of the Forward Algorithm for HMMs. The DIC can be used to compare different models, and is interpreted in a manner similar to the BIC.

2.7 Copula Distributions as a Measure of Dependence

A copula (Latin: link) is the joint distribution of random variables $U_1, \ldots, U_d$, where each U_i is marginally distributed as Uniform(0,1). More formally, a d-copula $\mathcal{C} : [0,1]^d \to [0,1]$ is the joint cumulative distribution function (CDF) of a d-dimensional random vector with Uniform marginals. Sklar's Theorem [Sklar, 1959, Durante et al., 2013] allows us to model the joint dependency of random variables by using its univariate marginals and a copula which captures all information about the dependence structure of the variable of interest.

For any CDF $F(\cdot)$, the generalized inverse of F_X is defined as:

$$F^-(x) = \inf\{u : F(u) \geq x\}.$$

Then, if $U \sim U[0,1]$ is a uniform random variable, and F_X is the CDF of a random variable X,

$$Pr[F^-(U) \leq x] = F_X(x).$$

In the opposite direction, if X has a continuous CDF F_X, then

$$F_X(X) \sim U[0,1].$$

With these fundamentals in place, we now formally state Sklar's theorem.

Theorem 2.1 (Sklar, 1959)**.** *Let $X_1, \dots X_d$ be random variables with a joint CDF*

$$F(x_1, \dots, x_d) = Pr(X_1 \leq x_1, \dots, X_d \leq x_d)$$

and marginal CDFs $F_{X_i}(x) = Pr(X_i \leq x) = u_i, i = 1, \dots, d$. Then there exists a copula $\mathcal{C} : [0,1]^d \to [0,1]$ such that:

$$\begin{aligned} &\mathcal{C}(u_1, \dots, u_d) = F(x_1, \dots, x_d) \\ \Leftrightarrow\ &F(x_1, \dots x_d) = \mathcal{C}[F_{X_1}(x_1), \dots, F_{X_d}(x_d)], \end{aligned} \tag{2.47}$$

for all x_i satisfying $u_i = F_{X_i}(x_i)$. If F_{X_i} is continuous for all $i = 1, \dots, d$, then $\mathcal{C}$ is unique; otherwise it is uniquely determined only on $Range(F_{X_1}) \times \dots \times Range(F_{X_d})$, where $Range(F_{X_i})$ denotes the range of the CDF of F_{X_i}. In the opposite direction, consider a copula $\mathcal{C}$ and univariate CDFs $F_{X_1}, \dots, F_{X_d}$. Then F as defined in (2.47) *is a multivariate CDF with marginals given by $F_{X_1}, \dots, F_{X_d}$.*

Sklar's theorem allows us to separate the modeling of the marginals from the modeling of their dependence structure. For an absolutely continuous F with strictly increasing marginals $F_{X_1}, \dots, F_{X_d}$, we can further use the chain rule to get

$$\begin{aligned} f(x_1, \dots, x_d) &= \Big[\prod_{i=1}^{d} f_{X_i}(x_i)\Big] c\big(F_{X_1}(x_1), \dots, F_{X_d}(x_d)\big), \\ \text{that is, } \frac{f(x_1, \dots, x_d)}{\prod_{i=1}^{d} f_{X_i}(x_i)} &= c\big(F_{X_1}(x_1), \dots, F_{X_d}(x_d)\big), \end{aligned} \tag{2.48}$$

where $c(\cdot)$ is the probability density function (pdf) of the copula distribution:

$$c(u_1, \dots, u_d) = \frac{\partial^d}{\partial u_1 \dots \partial u_d} \mathcal{C}(u_1, \dots, u_d).$$

The copula pdf is thus the ratio of the joint pdf of the random variables to what it would have been under independence, and can be considered as an adjustment to convert a pdf under an

independence assumption into a joint pdf.

A common class of copula distributions are elliptical copulas, formed using elliptical distributions. The Gaussian copula is an elliptical copula, as is the *t*-copula. Every d-dimensional copula is parameterized by their marginal parameters $(\boldsymbol{\alpha_1}, \ldots, \boldsymbol{\alpha_d})$, and a copula parameter $\boldsymbol{\theta}$ which is often d-dimensional as well and captures the dependence among the margins. With that in mind, we can rewrite (2.48) as:

$$f(x_1, \ldots, x_d; \boldsymbol{\alpha_1}, \ldots, \boldsymbol{\alpha_d}, \boldsymbol{\theta}) = \Big[\prod_{i=1}^{d} f_{X_i}(x_i; \boldsymbol{\alpha_i})\Big] c\big(F_{X_1}(x_1; \boldsymbol{\alpha_1}), \ldots, F_{X_d}(x_d; \boldsymbol{\alpha_d}); \boldsymbol{\theta}\big). \tag{2.49}$$

The log-likelihood can be written as

$$l(\boldsymbol{\alpha_1}, \ldots, \boldsymbol{\alpha_d}, \boldsymbol{\theta}) = \sum_{i=1}^{d} l_i^M(\boldsymbol{\alpha_i}) + l^C(\boldsymbol{\theta}, \boldsymbol{\alpha_1}, \ldots, \boldsymbol{\alpha_d}), \tag{2.50}$$

$$\text{where } l_i^M(\boldsymbol{\alpha_i}) = \log f_i(x_i; \boldsymbol{\alpha_i}), \text{ and} \tag{2.51}$$

$$l^C(\boldsymbol{\theta}, \boldsymbol{\alpha_1}, \ldots, \boldsymbol{\alpha_d}) = \log c\big(F_{X_1}(x_1; \boldsymbol{\alpha_1}), \ldots, F_{X_d}(x_d; \boldsymbol{\alpha_d}); \boldsymbol{\theta}\big). \tag{2.52}$$

Here, $l_i^M(\cdot)$ is the marginal log-likelihood for $\boldsymbol{\alpha_i}$ and $l^C(\cdot)$ is the copula log-likelihood for the parameters $(\boldsymbol{\theta}, \boldsymbol{\alpha_1}, \ldots, \boldsymbol{\alpha_d})$. The maximum likelihood estimate (MLE) $(\hat{\boldsymbol{\alpha}}_1, \ldots, \hat{\boldsymbol{\alpha}}_d, \hat{\boldsymbol{\theta}})$ is found by solving

$$(\partial l/\partial \boldsymbol{\alpha_1}, \ldots, \partial l/\partial \boldsymbol{\alpha_d}, \partial l/\partial \boldsymbol{\theta}) = \mathbf{0}. \tag{2.53}$$

This can be difficult to solve, especially when d is even moderately large. A workaround is to use the inference functions for margins (IFM) approach described by Joe and Xu [1996] which is a two-step approach to estimating the parameters:

1. Maximize the log likelihoods of the d univariate marginals $l_i^M(\alpha), i = 1, \ldots, d$ to get the estimates $\tilde{\boldsymbol{\alpha}}_1, \ldots, \tilde{\boldsymbol{\alpha}}_d$.

2. Maximize $l(\boldsymbol{\theta}, \tilde{\boldsymbol{\alpha}}_1, \ldots, \tilde{\boldsymbol{\alpha}}_d)$ over $\boldsymbol{\theta}$ to get the estimate $\tilde{\boldsymbol{\theta}}$.

The IFM estimate $(\tilde{\boldsymbol{\alpha}}_1, \ldots, \tilde{\boldsymbol{\alpha}}_d, \tilde{\boldsymbol{\theta}})$ is therefore the solution to

$$(\partial l_1^M / \partial \boldsymbol{\alpha}_1, \ldots, \partial l_d^M / \partial \boldsymbol{\alpha}_d, \partial l / \partial \boldsymbol{\theta}) = \mathbf{0}. \tag{2.54}$$

There exist other parametric, semi-parametric, and non-parametric alternatives to MLE and IFM. There are also Bayesian alternatives to IFM; we refer to Grazian and Liseo [2017] for recent Bayesian advances. All methods have their merits and demerits depending on the form of the copula and the properties of the marginals. Assumptions are often made to simplify the likelihood resulting in trade offs between computational simplicity and efficiency of the estimates. Pair copula constructions which simplify multivariate copulas by presenting them as products of pairwise copulas have also become common recently for dealing with high dimensional copulas [Aas et al., 2009, Brechmann et al., 2012]. In this thesis, we use the pair copula approximation to specify correlation structures in our HMM. This is done both for the state distribution and the conditional emission distribution.

2.8 Remote Sensing Data from GPM-IMERG

The Global Precipitation Measurement (GPM) mission[1] is an international satellite mission co-led by the National Aeronautics and Space Administration (NASA) and the Japanese Aerospace and Exploration Agency (JAXA). It aims to unify precipitation measurements from multiple research and operational microwave sensors for delivering next-generation global precipitation products. The GPM 'core' satellite has been deployed by NASA and JAXA carrying an advanced active/passive sensor package. The core satellite works alongside a constellation of satellites provided by a consortium of international partners to compute consistent precipitation estimates. Each constellation member has their own scientific and operational objectives, but also provide the GPM mission with microwave measurements which allows for precipitation products with global coverage and high precipitation frequency. GPM precipitation products also include precipitation gauge information where possible to adjust satellite estimates to reduce biases in

[1] https://gpm.nasa.gov/missions/GPM/constellation

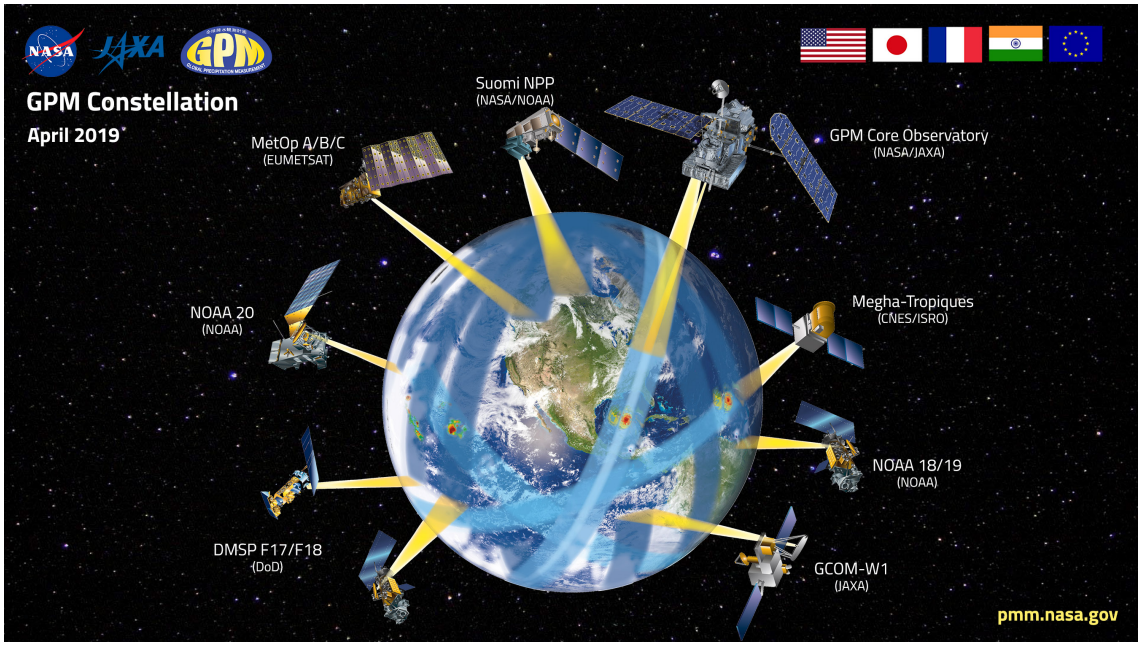

Figure 2.3: Diagram of the GPM satellite constellation as of early 2019. Credit: NASA GSFC.

monthly averages. Figure 2.3 shows the consortium partners and the satellites that are part of the GPM mission as of 2019. The Integrated Multi-satellitE Retrievals for GPM (IMERG) algorithm[2] [Huffman et al., 2019] combines the information that is made available through the GPM satellites to provide global precipitation estimates. The GPM-IMERG product is currently in Version 6 where it fuses precipitation estimates collected during the operation of the Tropical Rainfall Measuring Mission (TRMM) research satellite[3] (2000–2015) with more recent precipitation estimates from GPM (2014–present). This provides a longer record of global precipitation that can be used by researchers to formulate better climate and weather models and understand long-term global mean and extreme precipitation patterns. IMERG provides multiple data products from Level 1 (unprocessed instrument data at full resolution) to Level 3 (research-quality gridded estimates with time interpolation, gauge data, and climatological adjustment). Higher level products are better calibrated and easier to use in research. However, the calibration and postprocessing leads to higher latency, i.e., there is a longer time gap between data collection and the dataset being made available. This thesis relies on daily IMERG Final Run data (V06B), which is a Level 3 product available from June 2000 – present. IMERG V06B is calibrated using monthly gauge data, and is

[2] https://gpm.nasa.gov/data/imerg
[3] https://gpm.nasa.gov/missions/trmm

Table 2.1: Details of the GPM-IMERG dataset used for this study

Version	V06B
Availability	2000-present
Latency	$>$3.5 months
Spatial Resolution	$0.1^\circ \times 0.1^\circ$
Temporal Resolution	Daily
Spatial Coverage	Full coverage for $60^\circ N$ – $60^\circ S$ Partial coverage beyond 60° up to 90°

available at different temporal frequencies, from half-hourly to daily. The GPM-IMERG dataset covers the Chesapeake Bay watershed using 1927 grid points, as shown in Figure 1.2. Table 2.1 provides further details of the data that we use in this thesis.

2.9 Hardware and Software Used

The hardware used in the computational studies is part of the UMBC High Performance Computing Facility (`hpcf.umbc.edu`). The study used CPU nodes with two 18-core Intel Xeon Gold 6140 Skylake CPUs (2.3 GHz clock speed, 24.75 MB L3 cache, 6 memory channels) and 384 GB memory. The nodes are connected by a network of four 36-port EDR (Enhanced Data Rate) InfiniBand switches with 100 Gb/s bandwidth and 90 ns latency. All statistical computations were done in `R 3.6.x` and `R 4.0.x`. Plots were generated using base `R` packages as well as `ggplot2` [Wickham, 2016].

`R` uses double precision arithmetic which supports absolute maximum and minimum values of the order of 10^{308} and 10^{-308} respectively. However, these values can and do get exceeded when computing likelihoods for even modest datasets. For case studies involving real data, the `R Multiple Precision Floating-Point Reliable (Rmpfr)` package [Maechler, 2021], which was built to work with data which require arbitrary floating point precision, was used. In practice, we found that converting a regular floating point variable into an `rmpfr` object or vice versa were the only computational bottlenecks, and operations of `rmpfr` objects did not take perceptibly longer than operations of floating point variables. Simulation studies were run using the `RMPI` package [Yu, 2002], which is an `R` wrapper to the `Message`

`Passing Interface` protocol for parallel computation. Several of our simulation studies require running the same algorithm over multiple datasets generated from the same underlying model. With `RMPI`, we were able to spread out 1000 simulation studies among 3 CPU nodes of the HPCF which correspond to over 100 processes, each of which would only need to run 10 simulations. The processes work independently and communicate only once they have all run their respective simulation studies to provide combined estimates and an array of all the data that has been used for each of their studies. This significantly cuts down on computation time. The code that is used for this thesis is being made available on GitHub[4].

[4] https://github.com/reetamm

Chapter 3

Variational Bayes Parameter Estimation in HMMs for Semi-Continuous Daily Precipitation Data

In this chapter, we develop variational Bayes (VB) estimation for HMMs with semi-continuous emission distributions. We will refer to these HMMs as VB-HMMs and will use it to model daily precipitation data. We first cover univariate HMMs with a single observation sequence, where the data is available for consecutive time steps with no breaks. The emission process which corresponds to the observed precipitation has a semi-continuous distribution, with a point mass at 0 for no rainfall, and a mixture of Exponential distributions for positive rainfall. Next, we discuss parameter estimation when there are multiple observation sequences, e.g., seasonal data for multiple years where there is a break in between the end of the season for a year and the beginning of the season for the next year. These cases all have natural extensions to multivariate emissions corresponding to multi-site precipitation data. Mixtures of Exponential distributions are used to specify the distribution of positive precipitation as previous studies have shown it to be a good choice for daily precipitation amounts [Wilks, 1998, Robertson et al., 2006]. We also derive expressions for the Evidence Lower Bound (ELBO) and the Deviance Information Criterion (DIC) which are used to assess model convergence. A univariate state process corresponding to Figure 2.1 is assumed for all models in this chapter, an assumption which will be relaxed in the following chapters.

Once we have looked at formulations involving a mixture of exponential distributions for specifying positive precipitation, we turn our attention to some alternate emission distribution specifications. The Gamma distribution is a common alternative for precipitation [Robertson et al., 2006, Mhanna and Bauwens, 2012, Popuri, 2019]. We have previously used a mixture of Gamma distributions to estimate HMM parameters using maximum likelihood for daily precipitation over

the Potomac river basin [Kroiz et al., 2020a] and for the Chesapeake Bay watershed [Majumder et al., 2020]. We derive expressions for a VB-HMM which uses Gamma distributions for its emissions. We also derive expressions for a VB-HMM with Gamma shape mixtures (GSM) [Venturini et al., 2008] for positive precipitation. The GSM distribution has fewer parameters than Exponential or Gamma mixtures and has not been applied to precipitation models in the past. Empirical Bayes priors are derived for the GSM setup, which allows a relatively straightforward approach to specifying informative priors. All these estimation approaches fall under the umbrella of coordinate ascent variational Bayes (CAVI), where the entire data is used for the VBEM algorithm.

Finally, we discuss stochastic variational Bayes (SVB) as a practical parameter estimation approach for real life applications, where the dimensionality of the emission process makes CAVI infeasible. We derive expressions for parameter updates when positive precipitation is distributed as a mixture of exponential distributions. We propose a modified minibatch sampling method for stochastic optimization which adds more variability in the samples and improves the parameter estimates of the emission distribution parameters. Simulation study results for both CAVI and SVB are compared to gauge the trade-offs of using different prior specifications and for using stochastic optimization.

3.1 VB-HMM with Univariate Emissions

Let $y_{1:T} = \{y_1, \ldots, y_T\}$ be the precipitation time series of length *T*, with $y_t \geq 0$. The data is generated by a set of underlying hidden states $s_{1:T} = \{s_1, \ldots, s_t, \ldots, s_T\}$, where each state $s_t \in \{1, \ldots, K\}$. Let s_{tj} denote an indicator variable for being in state j at time t, i.e. $s_{tj} = \mathbb{I}(s_t = j)$. For each state *j*, we define an indicator variable r_{tjm} to connect the underlying state to the emission distribution, such that:

$$r_{tjm}s_{tj} = \mathbb{I}\{y_t \text{ comes from the } m^{th} \text{ mixture component and } s_t = j\},\ m = 0, 1, \ldots, M,$$

where $r_{tj} = (r_{tj0}, r_{tj1}, \ldots, r_{tjM})$ is encoded as a standard unit vector, also known as a *one-hot* vector; r_{tj0} indicates no-rainfall events. We assume that the number of states (*K*) and mixture

components (*M*+1) in the HMM are known. Note that for $j = 1, \ldots, K$ and $m = 0, \ldots, M$, we have:

$$\begin{aligned}\mathbb{E}[r_{tjm}s_{tj}] &= Pr[y_t \text{ comes from the } m^{th} \text{ mixture component and } s_t = j] \\ &= Pr[y_t \text{ comes from the } m^{th} \text{ mixture component}|s_t = j] \cdot Pr[s_t = j] \\ &= \mathbb{E}[r_{tjm}|s_{tj}] \cdot \mathbb{E}[s_{tj}].\end{aligned}$$

For each state *j*, r_{tj} follows a categorical distribution which corresponds to a single draw from a multinomial distribution, given by

$$p(r_{tj}|s_t = j) = \prod_{m=0}^{M} c_{jm}^{r_{tjm}},\ m = 0, 1, \ldots, M, \tag{3.1}$$

where $c_j = (c_{j0}, \ldots, c_{jM})$ is the vector of mixture probabilities parameterizing r_{tj}, with $c_{jm} \geq 0$ for all *m*, and $\sum_{m=0}^{M} c_{jm} = 1$. If we assume that positive rainfall for the m^{th} mixture component (where $m > 0$) from state *j* follows an Exponential distribution with rate λ_{jm}, the distribution of an observation from state *j* is given by

$$\begin{aligned}p(y_t, r_{tj}|\lambda_j, c_j, s_t = j) &= p(r_{tj}|c_j, s_t = j) \cdot p(y_t|\lambda_j, r_{tj}, s_t = j) \\ &= c_{j0}^{r_{tj0}} \prod_{m=1}^{M} \left[c_{jm}\lambda_{jm} \exp\{-\lambda_{jm}y_t\}\right]^{r_{tjm}}.\end{aligned} \tag{3.2}$$

The complete data likelihood is given by

$$\begin{aligned}p(y, s, r|\Theta) &= p(y, r|s, \Theta) \cdot p(s|\Theta) \\ &= \prod_{t=1}^{T}\prod_{j=1}^{K}\{p(y_t, r_{tj}|s_t = j, \Theta)p(s_t|\Theta)\} \\ &= p(s_1)\prod_{t=1}^{T}\prod_{j=1}^{K}\{p(y_t, r_{tj}|s_t = j, \Theta)\}\prod_{t=1}^{T}\prod_{j=1}^{K}\{p(s_{t+1}|s_t = j, \Theta)\},\end{aligned}$$

where $p(s|\Theta)$ is the distribution of the states which factorizes into the distribution of the initial state $\pi_1 = p(s_1)$ and the distribution of the state transitions $p(s_{t+1}|s_t)$. For $j, k = 1, \ldots, K$, $\pi_{1j} = Pr[s_1 = j]$ are the initial state probabilities and $a_{jk} = P[s_{t+1} = k|s_t = j]$ are the transi-

tion probabilities, $A = ((a_{jk}))$ is the $K \times K$ transition probability matrix, and $C = ((c_{jm}))$ is the $K \times (M+1)$ matrix of mixture probabilities. Similarly, $\Lambda = ((\lambda_{jm}))$ is a $K \times M$ matrix whose elements are the rate parameters as described in (3.2). Taken together, $\Theta = (A, C, \Lambda, \pi_1)$ parameterizes the HMM. We assign a prior on Θ which factorizes into a product over its components, i.e.

$$p(\Theta|\nu^{(0)}) = p(\pi_1) \cdot p(A) \cdot p(C) \cdot p(\Lambda),$$

where $\nu^{(0)}$ are hyperparameters whose values are known. We assign independent Dirichlet priors to the rows of *A*, denoted by $a_j = (a_{j1}, \ldots, a_{jK})$, and to the initial distribution $\pi_1 = (\pi_{11}, \ldots, \pi_{1K})$. Similarly, independent Dirichlet priors are assigned to the rows of *C*, denoted by $c_j = (c_{j0}, \ldots, c_{jM})$. Note that if the elements making up the parameter vector of a Dirichlet distribution are equal, it constitutes a symmetric Dirichlet distribution. The sum of the elements of the parameter vector is known as its concentration. A symmetric Dirichlet distribution indicates no prior knowledge favoring one component over another. Finally, independent Gamma priors are assigned to each element of Λ. Thus,

$$\begin{aligned} p(\pi_1) &= Dirichlet(\pi_1|\xi^{(0)}), \\ p(A) &= \prod_{j=1}^{K} Dirichlet(a_j|\alpha_j^{(0)}), \\ p(C) &= \prod_{j=1}^{K} Dirichlet(c_j|\zeta_j^{(0)}), \\ \text{and } p(\Lambda) &= \prod_{j=1}^{K}\prod_{m=1}^{M} Gamma(\lambda_{jm}|\gamma_{jm}^{(0)}, \delta_{jm}^{(0)}), \end{aligned}$$

where $\zeta_j^{(0)} = (\zeta_{j0}^{(0)}, \ldots, \zeta_{jM}^{(0)})$, $\alpha_j^{(0)} = (\alpha_{j1}^{(0)}, \ldots, \alpha_{jK}^{(0)})$, and $\xi^{(0)} = (\xi_1^{(0)}, \ldots \xi_K^{(0)})$. $\gamma_{jm}^{(0)}$ and $\delta_{jm}^{(0)}$ are the shape and rate parameters of the Gamma distribution respectively. The complete data

likelihood can be expressed as

$$\begin{aligned} p(y,s,r|\Theta) &= \prod_{j=1}^{K}\{\pi_{1j}\}^{s_{1j}} \prod_{t=1}^{T}\prod_{j=1}^{K}\{p_j(y_t,r_{tj}|\Theta)\}^{s_{tj}} \prod_{t=1}^{T-1}\prod_{j=1}^{K}\prod_{k=1}^{K}\{a_{jk}\}^{s_{tj}s_{t+1,k}} \\ &= \exp\Bigg\{\sum_{j=1}^{K} s_{1j}\log\pi_{1j} + \sum_{t=1}^{T}\sum_{j=1}^{K}\Bigg[\sum_{m=1}^{M} s_{tj}r_{tjm}(\log c_{jm} + \log\lambda_{jm} - y_t\lambda_{jm}) \\ &\quad + s_{tj}r_{tj0}\log c_{j0}\Bigg] + \sum_{t=1}^{T-1}\sum_{j=1}^{K}\sum_{k=1}^{K} s_{tj}s_{t+1,k}\log a_{jk}\Bigg\}, \end{aligned} \tag{3.3}$$

where $s_{tj} = \mathbb{I}\{s_t = j\}$ denotes the daily state and the product $s_{tj}s_{t+1,k}$ denotes a typical state transition. For convenience, we have denoted $p(\cdot|s_t = j, \Theta)$ as $p_j(\cdot|\Theta)$. We write the prior as

$$\begin{aligned} p(\Theta|\nu^{(0)}) &= p(\pi_1)\cdot p(\lambda)\cdot p(C)\cdot p(A) \\ &= \exp\Bigg\{\sum_{j=1}^{K}\{(\xi_j^{(0)}-1)\log\pi_{1j} + \sum_{m=1}^{M}\left[-\delta_{jm}^{(0)}\lambda_{jm} + (\gamma_{jm}^{(0)}-1)\log\lambda_{jm}\right] \\ &\quad + (\zeta_{j0}^{(0)}-1)\log c_{j0} + \sum_{m=1}^{M}(\zeta_{jm}^{(0)}-1)\log c_{jm} + \sum_{k=1}^{K}(\alpha_{jk}^{(0)}-1)\log a_{jk}\} - \log h^{(0)}\Bigg\}, \end{aligned} \tag{3.4}$$

where $h^{(0)} = h(\nu^{(0)})$ is the normalizing constant for the prior. Comparing this expression with the canonical form for the conjugate exponential family, we arrive at the following expressions for the natural parameters $\phi(\Theta)$, their sufficient statistics $u(s,y,r)$, and the hyperparameters $\nu^{(0)}$:

$$\phi(\Theta) = \begin{bmatrix} \log\pi_{1j} \\ \log c_{j0} \\ \log c_{jm} \\ \log\lambda_{jm} \\ \lambda_{jm} \\ \log a_{jk} \end{bmatrix}, \qquad u(s,y,r) = \begin{bmatrix} s_{1j} \\ s_{tj}r_{tj0} \\ s_{tj}r_{tjm} \\ s_{tj}r_{tjm} \\ y_t s_{tj}r_{tjm} \\ s_{tj}s_{t+1,k} \end{bmatrix}, \qquad \nu^{(0)} = \begin{bmatrix} \xi_j^{(0)}-1 \\ \zeta_{j0}^{(0)}-1 \\ \zeta_{jm}^{(0)}-1 \\ \gamma_{jm}^{(0)}-1 \\ \delta_{jm}^{(0)} \\ \alpha_{jk}^{(0)}-1 \end{bmatrix}, \tag{3.5}$$

for $m = 1, \ldots, M, j = 1, \ldots, K, k = 1, \ldots, K$. The variational family $\mathbb{Q}$ is constrained to distributions which are separable in the following manner:

$$q_z(z) = q_\Theta(\Theta) \cdot q_{s,r}(s, r), \tag{3.6}$$

$$\text{where } q_\Theta(\Theta) = q(\pi_1) \cdot q(A) \cdot q(C) \cdot q(\Lambda). \tag{3.7}$$

As discussed in Section 2.4, $q_\Theta(\Theta)$ and $q_{s,r}(s, r)$ are coupled in (3.6), and the optimization problem in (2.14) cannot be solved analytically. Instead, it is solved numerically by iteratively optimizing $q_\Theta(\Theta)$ and $q_{s,r}(s, r)$ using the VBEM algorithm, which is the variational Bayes generalization for the EM algorithm. In the variational M-step (VBM step), $q_{s,r}(s, r)$ is fixed and $q_\Theta(\Theta)$ is updated, with the posterior taking the same form as the conjugate prior. Since $q_\Theta(\Theta)$ factors as in (3.7), each of its components can be updated individually while holding the others fixed. In the variational E-step (VBE step), we seek to update $q_{s,r}(s, r)$ while holding $q_\Theta(\Theta)$ fixed. Since the HMM's states are first order Markov, we need to take the temporal dependency into consideration if we want meaningful estimates for the latent variables. We accomplish this by adapting the Forward-Backward algorithm, a central part of the Baum-Welch algorithm, into our VBEM.

VBM step: *With the variational posterior $q_{s,r}(s, r)$ of the latent variables fixed at their expected value, update $q_\Theta(\Theta)$, the variational posterior of the model parameters.*

Since the variational posterior $q_\Theta(\Theta)$ is conjugate to the prior in (3.30), the posterior distribution for each component of $\phi(\Theta)$ in (3.31) is obtained by updating the coordinates of $\nu^{(0)}$ with the expected values of the corresponding sufficient statistics $u(s, y, r)$. To this end, we denote the expectations of the latent variables in (3.3) under $q_{s,r}(s, r)$ as

$$\begin{aligned} q_{1j} &= \mathbb{E}(s_{1j}), \\ q_{tj} &= \mathbb{E}(s_{tj}), \\ q_{tjm} &= \mathbb{E}(r_{tjm}), \\ \text{and } q_{jk} &= \mathbb{E}(s_{tj} s_{t+1,k}), \end{aligned}$$

where $j, k = 1, \ldots, K$ and $m = 0, 1, \ldots, M$. The variational updates at each iteration of the

VBM step are then given by

$$\begin{aligned}
\xi_j &= \xi_j^{(0)} + q_{1j}, \\
\zeta_{j0} &= \zeta_{j0}^{(0)} + \sum_{t=1}^{T} q_{tj} q_{tj0}, \\
\zeta_{jm} &= \zeta_{jm}^{(0)} + \sum_{t=1}^{T} q_{tj} q_{tjm}, \\
\gamma_{jm} &= \gamma_{jm}^{(0)} + \sum_{t=1}^{T} q_{tj} q_{tjm}, \\
\delta_{jm} &= \delta_{jm}^{(0)} + \sum_{t=1}^{T} q_{tj} q_{tjm} y_t, \\
\alpha_{jk} &= \alpha_{jk}^{(0)} + \sum_{t=1}^{T-1} q_{jk},
\end{aligned}$$

where $j, k = 1, \ldots, K$ and $m = 1, \ldots, M$.

VBE step: *With the variational posterior $q_\Theta(\Theta)$ of the model parameters fixed at their expected values, update the variational posterior $q_{s,r}(s, r)$ of the latent variables.*

The variational posterior $q_{s,r}(s, r)$ has a form similar to the complete data likelihood in (3.3) with the natural parameters $\phi(\Theta)$ replaced by their expectations under $q_\Theta(\Theta)$. Thus,

$$q_{s,r}(s, r) \propto \prod_{j=1}^{K} \{a_{1j}^*\}^{s_{1j}} \prod_{t=1}^{T} \prod_{j=1}^{K} \prod_{m=0}^{M} \{b_{tjm}^*\}^{s_{tj} r_{tjm}} \prod_{t=1}^{T-1} \prod_{j=1}^{K} \prod_{k=1}^{K} \{a_{jk}^*\}^{s_{tj} s_{t+1,k}}, \tag{3.8}$$

$$\text{with} \quad a_{1j}^* = \exp\big\{\mathbb{E}_q \log \pi_{1j}\big\} = \exp\big\{\Psi(\xi_j) - \Psi(\xi_\cdot)\big\},$$

$$\text{and} \quad a_{jk}^* = \exp\big\{\mathbb{E}_q \log a_{jk}\big\} = \exp\big\{\Psi(\alpha_{jk}) - \Psi(\alpha_{j\cdot})\big\},$$

where $\Psi(\cdot)$ is the digamma function and $\xi_\cdot = \sum_{j=1}^{K} \xi_j$, $\alpha_{j\cdot} = \sum_{k=1}^{K} \alpha_{jk}$. Similarly,

$$b_{tjm}^* = \begin{cases} \exp\big\{\mathbb{E}_q \log\big[c_{j0}\big]\big\} & \text{if } m = 0, \\ \exp\big\{\mathbb{E}_q \log\big[c_{jm} f(y_t|\lambda_{jm})\big]\big\} & \text{if } m > 0. \end{cases}$$

The expectations of the individual terms in b^*_{tjm} are:

$$c^*_{jm} = \exp\{\mathbb{E}_q \log c_{jm}\} = \exp\{\Psi(\zeta_{jm}) - \Psi(\zeta_{j.})\}, \text{ where } \zeta_{j.} = \sum_{m=0}^{M} \zeta_{jm},$$

$$\lambda^*_{jm} = \exp\{\mathbb{E}_q \log \lambda_{jm}\} = \exp\{\Psi(\gamma_{jm}) - \log \delta_{jm}\},$$

$$\hat{\lambda}_{jm} = \mathbb{E}_q \lambda_{jm} = \gamma_{jm}/\delta_{jm}.$$

Therefore,

$$b^*_{tjm} = \begin{cases} \exp\{\Psi(\zeta_{j0}) - \Psi(\zeta_{j.})\} & \text{if } m = 0, \\ \exp\{\Psi(\zeta_{jm}) - \Psi(\zeta_{j.}) + \Psi(\gamma_{jm}) - \log \delta_{jm} - y_t \frac{\gamma_{jm}}{\delta_{jm}}\} & \text{if } m > 0. \end{cases}$$

Here a^*_{1j} is the expected initial state probability, a^*_{jk} are the expected transition probabilities from state *j* to state *k*, and $b^*_{tj} = \sum_{m=0}^{M} b^*_{tjm}$ is the expected emission probability distribution conditional on the system being in state *j* at time *t*. As mentioned previously, the expectations are computed with respect to the variational posterior distribution of the parameters at the current iteration of the VBEM. These can now be used in the Forward-Backward Algorithm described in Section 2.5.1 to get the desired variational posterior estimates for the state probabilities as well as the cluster assignment probabilities. The updates to the variational posterior on the latent variables are:

$$q_{tj} = \frac{\tilde{F}_{tj} \cdot \tilde{B}_{tj}}{\sum_{k=1}^{K} \tilde{F}_{tk} \cdot \tilde{B}_{tk}},$$

$$q_{jk} = \frac{\tilde{F}_{tj} \cdot a^*_{jk} \cdot b^*_{t+1,k} \cdot \tilde{B}_{t+1,k}}{\sum_{j=1}^{K} \sum_{k=1}^{K} \tilde{F}_{tj} \cdot a^*_{jk} \cdot b^*_{t+1,k} \cdot \tilde{B}_{t+1,k}}.$$

where $\tilde{F}_{tj}$ and $\tilde{B}_{tj}$ are the scaled Forward and Backward variables respectively. The posterior

update of q_{1j} is the first entry of q_{tj}. The posterior for the mixture assignments is given by

$$q_{tjm} \propto \begin{cases} 1 & \text{if } m = 0, y_t = 0 \\ 0 & \text{if } m > 0, y_t = 0 \text{ or } m = 0, y_t > 0 \\ c^*_{jm} f(y_t | \lambda^*_{jm}, \hat{\lambda}_{jm}) & \text{if } m > 0, y_t > 0 \end{cases}$$

where $c^*_{jm} f(y_t|\lambda^*_{jm}, \hat{\lambda}_{jm}) = \exp\big\{\Psi(\zeta_{jm}) - \Psi(\zeta_{j\cdot}) + \Psi(\gamma_{jm}) - \log \delta_{jm} - y_t \frac{\gamma_{jm}}{\delta_{jm}}\big\}$. Note that when there is exactly one mixture component for positive rainfall ($M = 1$), observations are assigned to mixture components in a deterministic manner, fixing r_{tj}.

Assessing convergence using the ELBO

Using Equations (3.3) – (3.8), we can rewrite the ELBO as

$$ELBO(q) = \mathbb{E}_{q(s,r)} \log p(y, s, r) + \mathbb{E}_{q(\Theta)} \log p(\Theta) + H\big(q(s, r)\big) - \mathbb{E}_{q(\Theta)} \log q(\Theta),$$

where $H\big(q(s, r)\big)$ is the entropy of the variational posterior distribution over the latent variables. Beal [2003] and Ji et al. [2006] have shown that this simplifies to

$$\begin{aligned} ELBO(q) = \log q(y|\Theta) - KL\big(q(\pi_1) \parallel p(\pi_1)\big) - KL\big(q(A) \parallel p(A)\big) \\ - KL\big(q(C) \parallel p(C)\big) - KL\big(q(\Lambda) \parallel p(\Lambda)\big), \end{aligned} \tag{3.9}$$

where the first term on the right hand side is calculated as part of the Forward Algorithm in (2.36). This relationship is used to compute the ELBO at each iteration, and we declare convergence once the change in ELBO falls below a desired threshold.

3.2 VB-HMM with Multiple Observation Sequences

Multiple observation sequences can arise, for example, when we are collecting data for only a certain number of days or months, but for multiple years. For precipitation, we are usually interested in looking at a particular season at any point. Thus we end up with 3 months of data

for consecutive years, with a 9 month gap between seasons from different years. Rabiner [1989] modified the maximum likelihood estimation procedure by considering data like this as multiple observation sequences from the underlying process. Let there be N independent sequences, denoted by:

$$y \equiv y_{1:T} = \left[y^{(1)}, y^{(2)}, \ldots, y^{(N)}\right],$$

where $y^{(i)} = (y_1^{(i)}, \ldots, y_{T_i}^{(i)})$ is the i^{th} observation sequence. In our example, $T_i = 92$ days (3 months) for each i. N correspondingly represents the number of years for which we have data. As in Rabiner [1989], we make the assumption that each observation sequence is independent of every other observation sequence, that is:

$$p(y_{1:T}|\Theta) = \prod_{i=1}^{N} p(y^{(i)}|\Theta).$$

We can then modify the VBEM algorithm as follows:

VBE step: Run the Forward-Backward algorithm independently for each observation sequence, giving us the posterior estimates $q_{1j}^{(i)}, q_{tj}^{(i)}, q_{tjm}^{(i)}, q_{jk}^{(i)}$ for $t = 1, \ldots, T_i$, $i = 1, \ldots, N$, $j, k = 1, \ldots, K$, and $m = 0, \ldots, M$.

VBM step: Update each component of $q_\Theta(\Theta)$ by evaluating the expected values of the sufficient statistics $u(s, y, r)$. Since $y = \left[y^{(1)}, y^{(2)}, \ldots, y^{(N)}\right]$, $u(s, y, r) = \sum_{i=1}^{N} u(s^{(i)}, r^{(i)}, r^{(i)})$, that is:

$$\begin{aligned}
\xi_j &= \xi_j^{(0)} + \sum_{i=1}^{N} q_{1j}^{(i)}, \\
\zeta_{j0} &= \zeta_{j0}^{(0)} + \sum_{i=1}^{N} \sum_{t=1}^{T_i} q_{tj}^{(i)} q_{tj0}^{(i)}, \\
\zeta_{jm} &= \zeta_{jm}^{(0)} + \sum_{i=1}^{N} \sum_{t=1}^{T_i} q_{tj}^{(i)} q_{tjm}^{(i)}, \\
\gamma_{jm} &= \gamma_{jm}^{(0)} + \sum_{i=1}^{N} \sum_{t=1}^{T_i} T q_{tj}^{(i)} q_{tjm}^{(i)}, \\
\delta_{jm} &= \delta_{jm}^{(0)} + \sum_{i=1}^{N} \sum_{t=1}^{T_i} q_{tj}^{(i)} q_{tjm}^{(i)} y_t^{(i)},
\end{aligned}$$

$$\alpha_{jk} = \alpha_{jk}^{(0)} + \sum_{i=1}^{N} \sum_{t=1}^{T_i - 1} q_{jk}^{(i)},$$

where $j, k = 1, \ldots, K$ and $m = 1, \ldots, M$. A consequence of multiple observation sequences is that the variational update for ξ_j is now based on *N* estimates of q_{1j} instead of just 1.

3.3 VB-HMM with Multivariate Emissions

We now consider the case where our data is *L*-dimensional, corresponding to precipitation at *L* locations. Let $\mathbf{y}_{1:T} = \{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$ be the precipitation time series of length *T*, with $\mathbf{y}_t' = (y_{t1}, \ldots, y_{tL})$ and $y_{tl} \geq 0$ for $l = 1, \ldots, L$. The data is generated by a set of underlying hidden states $s_{1:T} = \{s_1, \ldots, s_t, \ldots, s_T\}$, where each state $s_t \in \{1, \ldots, K\}$. For each state j, we define an indicator variable r_{tjlm} to connect the underlying state *j* to the emission distribution such that:

$$r_{tjlm} s_{tj} = \mathbb{I}\{y_{tl} \text{ comes from the } m^{th} \text{ mixture component and } s_t = j\},$$

with $m = 0, 1, \ldots, M$, $l = 1, \ldots, L$, and where $s_{tj} = \mathbb{I}(s_t = j)$ as before. The number of states (*K*), the number of locations (*L*), and the number of mixture components (*M*+1) in the HMM are assumed to be known. Note that for $j = 1 \ldots, K$, $m = 0, \ldots, M$ and $l = 1, \ldots, L$, we have:

$$\mathbb{E}[r_{tjlm} s_{tj}] = \mathbb{E}[r_{tjlm} | s_{tj}] \cdot \mathbb{E}[s_{tj}].$$

We define $r_{tjl} = (r_{tjl0}, \ldots, r_{tjlm})$. For each state *j* and location *l*, r_{tjl} follows a categorical distribution which corresponds to a single draw from a multinomial distribution, given by

$$p(r_{tjl} | c_{jl}, s_t = j) = \prod_{m=0}^{M} c_{jlm}^{r_{tjlm}}, \tag{3.10}$$

$$\text{and} \quad p(r_{tjl} | c_{jl}, s_t = j) \perp\!\!\!\perp p(r_{tjl'} | c_{jl'}, s_t = j) \quad \text{for } l \neq l', \tag{3.11}$$

with $m = 0, 1, \ldots, M$ and $l = 1, \ldots, L$; $c_{jl} = (c_{jl0}, \ldots, c_{jlM})$ are the mixture probabilities parameterizing r_{tjl}, with $c_{jlm} \geq 0$ for all *m*, and $\sum_{m=0}^{M} c_{jlm} = 1$. A consequence of the condi-

tional Independence assumption in (3.11) is that the correlation between precipitation at different locations is induced by the common state variable for all locations. Unless the emissions follow a multivariate distribution, e.g. a multivariate Normal distribution, this is the only source of correlation between the different emission chains. If we assume that positive rainfall at the l^{th} location from the m^{th} mixture component (where $m > 0$) arising from state *j* follows an Exponential distribution with rate λ_{jlm} with $\lambda_{jl} = \{\lambda_{jl1}, \ldots, \lambda_{jlm}\}$, the distribution of an observation from state *j* across all locations is given by

$$\begin{aligned}\prod_{l=1}^{L} p(y_{tl}, r_{tjl}|\lambda_{jl}, c_{jl}, s_t = j) &= \prod_{l=1}^{L} p(r_{tjl}|c_{jl}, s_t = j) \cdot p(y_{tl}|\lambda_{jl}, r_{tjl}, s_t = j) \\ &= \prod_{l=1}^{L} \left\{ c_{jl0}^{r_{tjl0}} \prod_{m=1}^{M} \left[c_{jlm} \lambda_{jlm} \exp\{-\lambda_{jlm} y_{tl}\} \right]^{r_{tjlm}} \right\}.\end{aligned} \tag{3.12}$$

As in the univariate case, the complete data likelihood is given by

$$p(y, s, r|\Theta) = p(y, r|s, \Theta) \cdot p(s|\Theta),$$

In this case, $A = ((a_{jk}))$ is the $K \times K$ transition probability matrix, and $C_l = ((c_{jlm}))$ is the $K \times (M + 1)$ matrix of mixture probabilities for each location *l*. Similarly, $\Lambda_l = ((\lambda_{jlm}))$ is a $K \times M$ matrix whose elements are the independently distributed rate parameters of the Exponential distributions which are part of the semi-continuous emissions in each state. We also define the tensors $C = (C_1, \ldots, C_L)$ and $\Lambda = (\Lambda_1, \ldots, \Lambda_L)$. Taken together, $\Theta = (A, C, \Lambda, \pi_1)$ parameterizes the HMM. We assign a prior on Θ which factorizes into a product over its components:

$$p(\Theta|\nu^{(0)}) = p(\pi_1) \cdot p(A) \cdot p(C) \cdot p(\Lambda),$$

where $\nu^{(0)}$ are known hyperparameters. The individual components of the prior are distributed as

$$\begin{aligned}p(\pi_1) &= Dirichlet(\pi_1|\xi^{(0)}), \\ p(A) &= \prod_{j=1}^{K} Dirichlet(a_j|\alpha_j^{(0)}),\end{aligned}$$

$$p(C) = \prod_{j=1}^{K}\prod_{l=1}^{L} Dirichlet(c_{jl}|\zeta_{jl}^{(0)}),$$
$$\text{and } p(\Lambda) = \prod_{j=1}^{K}\prod_{l=1}^{L}\prod_{m=1}^{M} Gamma(\lambda_{jlm}|\gamma_{jlm}^{(0)}, \delta_{jlm}^{(0)}),$$

where $a_j = (a_{j1}, \ldots, a_{jK})$, $\pi_1 = (\pi_{11}, \ldots, \pi_{1K})$, $\zeta_{jl}^{(0)} = (\zeta_{jl0}^{(0)}, \ldots, \zeta_{jlM}^{(0)})$, $\alpha_j^{(0)} = (\alpha_{j1}^{(0)}, \ldots, \alpha_{jK}^{(0)})$, and $\xi^{(0)} = (\xi_1^{(0)}, \ldots \xi_K^{(0)})$. $\gamma_{jlm}^{(0)}$ and $\delta_{jlm}^{(0)}$ are the shape and rate parameters of the Gamma distribution respectively.

The complete data likelihood can be expressed as

$$\begin{aligned} p(y, s, r|\Theta) &= \prod_{j=1}^{K}\{\pi_{1j}\}^{s_{1j}} \prod_{t=1}^{T}\prod_{j=1}^{K}\prod_{l=1}^{L}\{p_j(y_{tl}, r_{tjl}|\Theta)\}^{s_{tj}} \prod_{t=1}^{T-1}\prod_{j=1}^{K}\prod_{k=1}^{K}\{a_{jk}\}^{s_{tj}s_{t+1,k}} \\ &= \exp\Bigg\{\sum_{j=1}^{K} s_{1j}\log\pi_{1j} + \sum_{t=1}^{T}\sum_{j=1}^{K}\sum_{l=1}^{L}\Bigg[\sum_{m=1}^{M} s_{tj}r_{tjlm}(\log c_{jlm} + \log\lambda_{jlm} - y_t\lambda_{jlm}) \\ &\quad + s_{tj}r_{tjl0}\log c_{jl0}\Bigg] + \sum_{t=1}^{T-1}\sum_{j=1}^{K}\sum_{k=1}^{K} s_{tj}s_{t+1,k}\log a_{jk}\Bigg\}. \end{aligned} \tag{3.13}$$

Similarly, we write the prior as

$$\begin{aligned} p(\Theta|\nu^{(0)}) &= p(\pi_1)\cdot p(\lambda)\cdot p(C)\cdot p(A) \\ &= \exp\Bigg\{\sum_{j=1}^{K}\big\{(\xi_j^{(0)} - 1)\log\pi_{1j} + \sum_{l=1}^{L}\sum_{m=1}^{M}\big[-\delta_{jlm}^{(0)}\lambda_{jlm} + (\gamma_{jlm}^{(0)} - 1)\log\lambda_{jlm}\big] \\ &\quad + \sum_{l=1}^{L}(\zeta_{jl0}^{(0)} - 1)\log c_{jl0} + \sum_{l=1}^{L}\sum_{m=1}^{M}(\zeta_{jlm}^{(0)} - 1)\log c_{jlm} \\ &\quad + \sum_{k=1}^{K}(\alpha_{jk}^{(0)} - 1)\log a_{jk}\big\} - \log h^{(0)}\Bigg\}, \end{aligned} \tag{3.14}$$

where $h^{(0)} = h(\nu^{(0)})$ is the normalizing constant for the prior. Comparing this expression with the canonical form for the conjugate exponential family, we arrive at the following expressions for

the natural parameters $\phi(\Theta)$, their sufficient statistics $u(s, y, r)$, and the hyperparameters $\nu^{(0)}$:

$$\phi(\Theta) = \begin{bmatrix} \log \pi_{1j} \\ \log c_{jl0} \\ \log c_{jlm} \\ \log \lambda_{jlm} \\ \lambda_{jlm} \\ \log a_{jk} \end{bmatrix}, \qquad u(s, y, r) = \begin{bmatrix} s_{1j} \\ s_{tj} r_{tjl0} \\ s_{tj} r_{tjlm} \\ s_{tj} r_{tjlm} \\ y_{tl} s_{tj} r_{tjlm} \\ s_{tj} s_{t+1,k} \end{bmatrix}, \qquad \nu^{(0)} = \begin{bmatrix} \xi_j^{(0)} - 1 \\ \zeta_{jl0}^{(0)} - 1 \\ \zeta_{jlm}^{(0)} - 1 \\ \gamma_{jlm}^{(0)} - 1 \\ \delta_{jlm}^{(0)} \\ \alpha_{jk}^{(0)} - 1 \end{bmatrix}, \tag{3.15}$$

for $m = 1, \ldots, M, j = 1, \ldots, K, k = 1, \ldots, K$. The variational family $\mathbb{Q}$ is constrained to distributions which are separable in the following manner:

$$q_z(z) = q_\Theta(\Theta) \cdot q_{s,r}(s, r), \tag{3.16}$$

$$\text{where } q_\Theta(\Theta) = q(\pi_1) \cdot q(A) \cdot q(C) \cdot q(\Lambda). \tag{3.17}$$

The VBEM algorithm for multivariate emissions follows the univariate case closely.

VBM step: *With the variational posterior $q_{s,r}(s, r)$ of the latent variables fixed at their expected values, update $q_\Theta(\Theta)$, the variational posterior of the model parameters.*

Since $q_\Theta(\Theta)$ is conjugate to the prior, the posterior distribution for each component of $\phi(\Theta)$ in (3.15) is obtained by updating the coordinates of $\nu^{(0)}$ with the expected values of the corresponding sufficient statistics $u(s, y, r)$. To this end, we denote the expectations of the latent variables in (3.13) under $q_{s,r}(s, r)$ as

$$\begin{aligned} q_{1j} &= \mathbb{E}(s_{1j}), \\ q_{tj} &= \mathbb{E}(s_{tj}), \\ q_{tjlm} &= \mathbb{E}(r_{tjlm}), \\ \text{and } q_{jk} &= \mathbb{E}(s_{tj} s_{t+1,k}), \end{aligned}$$

where $j, k = 1, \ldots, K$, $l = 1, \ldots, L$, and $m = 0, 1, \ldots, M$. The variational updates at each

iteration of the VBM step are then given by

$$\begin{aligned}
\xi_j &= \xi_j^{(0)} + q_{1j}, \\
\zeta_{jl0} &= \zeta_{jl0}^{(0)} + \sum_{t=1}^{T} q_{tj} q_{tjl0}, \\
\zeta_{jlm} &= \zeta_{jlm}^{(0)} + \sum_{t=1}^{T} q_{tj} q_{tjlm}, \\
\gamma_{jlm} &= \gamma_{jlm}^{(0)} + \sum_{t=1}^{T} q_{tj} q_{tjlm}, \\
\delta_{jlm} &= \delta_{jlm}^{(0)} + \sum_{t=1}^{T} q_{tj} q_{tjlm} y_{tl}, \\
\alpha_{jk} &= \alpha_{jk}^{(0)} + \sum_{t=1}^{T-1} q_{jk},
\end{aligned}$$

where $j, k = 1, \ldots, K$, $l = 1, \ldots, L$, and $m = 1, \ldots, M$.

VBE step: *With the variational posterior $q_\Theta(\Theta)$ of the model parameters fixed at their expected values, update the variational posterior $q_{s,r}(s, r)$ of the latent variables.*

The variational posterior $q_{s,r}(s, r)$ has a form similar to the complete data likelihood as in the univariate case, i.e.

$$q_{s,r}(s, r) \propto \prod_{j=1}^{K} \{a_{1j}^*\}^{s_{1j}} \prod_{t=1}^{T} \prod_{j=1}^{K} \prod_{l=1}^{L} \prod_{m=0}^{M} \{b_{tjlm}^*\}^{s_{tj} r_{tjlm}} \prod_{t=1}^{T-1} \prod_{j=1}^{K} \prod_{k=1}^{K} \{a_{jk}^*\}^{s_{tj} s_{t+1,k}}, \tag{3.18}$$

with the natural parameters $\phi(\Theta)$ replaced by their expectations under $q_\Theta(\Theta)$. Comparing with (3.13), we get

$$\begin{aligned}
a_{1j}^* &= \exp\big\{\mathbb{E}_q \log \pi_{1j}\big\} = \exp\big\{\Psi(\xi_j) - \Psi(\xi_\cdot)\big\}, \\
\text{and} \quad a_{jk}^* &= \exp\big\{\mathbb{E}_q \log a_{jk}\big\} = \exp\big\{\Psi(\alpha_{jk}) - \Psi(\alpha_{j\cdot})\big\},
\end{aligned}$$

where $\xi_{\cdot} = \sum_{j=1}^{K} \xi_j$, $\alpha_{j\cdot} = \sum_{k=1}^{K} \alpha_{jk}$. Similarly,

$$b^*_{tjlm} = \begin{cases} \exp\big\{\mathbb{E}_q \log\big[c_{jl0}\big]\big\} & \text{if } m = 0, \\ \exp\big\{\mathbb{E}_q \log\big[c_{jlm} f(y_{tl}|\lambda_{jlm})\big]\big\} & \text{if } m > 0. \end{cases}$$

The expectations of the individual terms in b^*_{tjlm} are:

$$c^*_{jlm} = \exp\big\{\mathbb{E}_q \log c_{jlm}\big\} = \exp\big\{\Psi(\zeta_{jlm}) - \Psi(\zeta_{j\cdot})\big\}, \text{ where } \zeta_{j\cdot} = \sum_{m=0}^{M} \zeta_{jlm},$$

$$\lambda^*_{jlm} = \exp\big\{\mathbb{E}_q \log \lambda_{jlm}\big\} = \exp\big\{\Psi(\gamma_{jlm}) - \log \delta_{jlm}\big\},$$

$$\hat{\lambda}_{jlm} = \mathbb{E}_q \lambda_{jlm} = \gamma_{jlm}/\delta_{jlm}.$$

$$\text{Therefore, } b^*_{tjlm} = \begin{cases} \exp\big\{\Psi(\zeta_{jl0}) - \Psi(\zeta_{jl\cdot})\big\} & \text{if } m = 0, \\ \exp\big\{\Psi(\zeta_{jlm}) - \Psi(\zeta_{jl\cdot}) + \Psi(\gamma_{jlm}) - \log \delta_{jlm} - y_{tl}\frac{\gamma_{jlm}}{\delta_{jlm}}\big\} & \text{if } m > 0, \end{cases}$$

$$\text{and } b^*_{tj} = \prod_{l=1}^{L} \sum_{m=0}^{M} b^*_{tjlm}.$$

The quantities a^*_{1j}, a^*_{jk} and b^*_{tj} can be used as part of the Forward-Backward Algorithm to get our desired variational posterior estimates for the state probabilities as well as the cluster assignment probabilities. The updates to the variational posterior on the latent variables are q_{1j}, q_{tj} and q_{jk} are identical to the univariate case. The posterior for the mixture assignments for the l^{th} location is given by

$$q_{tjlm} \propto \begin{cases} 1 & \text{if } m = 0, y_{tl} = 0 \\ 0 & \text{if } m > 0, y_{tl} = 0 \text{ or } m = 0, y_{tl} > 0 \\ c^*_{jlm} f(y_{tl}|\lambda^*_{jlm}, \hat{\lambda}_{jlm}) & \text{if } m > 0, y_{tl} > 0 \end{cases}$$

where $c^*_{jlm} f(y_{tl}|\lambda^*_{jlm}, \hat{\lambda}_{jlm}) = \exp\big\{\Psi(\zeta_{jlm}) - \Psi(\zeta_{jl\cdot}) + \Psi(\gamma_{jlm}) - \log \delta_{jlm} - y_{tl}\frac{\gamma_{jlm}}{\delta_{jlm}}\big\}$. Note that the VBEM algorithm for multivariate data can be easily extended to accommodate multiple

observation sequences similar to the univariate case.

3.4 Model Selection using the DIC

Let Θ be our parameter vector, and $\tilde{\Theta}$ denote its posterior mean. The DIC for a VB-HMM can be expressed as

$$\text{DIC} = -2\log p(y|\tilde{\Theta}) + 2p_D, \tag{3.19}$$

$$\text{where } p_D \approx -2\int q_\Theta(\Theta)\log\left\{\frac{q_\Theta(\Theta)}{p(\Theta)}\right\}d\Theta + 2\log\left\{\frac{q_\Theta(\tilde{\Theta})}{p(\tilde{\Theta})}\right\} \tag{3.20}$$

is defined using a variational approximation. We present the DIC calculations for univariate emissions. The terms required to calculate p_D for our model are:

$$\begin{aligned}
\log q_\Theta(\Theta) &= \sum_{j=1}^{K}\Bigg\{\sum_{p=1}^{P}\big[-\delta_{jm}\lambda_{jm} + (\gamma_{jm}-1)\log\lambda_{jm}\big] + (\zeta_{j0}-1)\log c_{j0} \\
&\quad + \sum_{p=1}^{P}(\zeta_{jm}-1)\log c_{jm} + \sum_{l=1}^{K}(\alpha_{jk}-1)\log a_{jk}\Bigg\} + \text{constant}, \\
\log p(\Theta) &= \sum_{j=1}^{K}\Bigg\{\sum_{p=1}^{P}\big[-\delta_{jm}^{(0)}\lambda_{jm} + (\gamma_{jm}^{(0)}-1)\log\lambda_{jm}\big] + (\zeta_{j0}^{(0)}-1)\log c_{j0} \\
&\quad + \sum_{p=1}^{P}(\zeta_{jm}^{(0)}-1)\log c_{jm} + \sum_{l=1}^{K}(\alpha_{jk}^{(0)}-1)\log a_{jk}\Bigg\} + \text{constant}, \\
\log q_\Theta(\tilde{\Theta}) &= \sum_{j=1}^{K}\Bigg\{\sum_{p=1}^{P}\big[-\delta_{jm}\frac{\gamma_{jm}}{\delta_{jm}} + (\gamma_{jm}-1)\log\frac{\gamma_{jm}}{\delta_{jm}}\big] + (\zeta_{j0}-1)\log\frac{\zeta_{j0}}{\zeta_{j\cdot}} \\
&\quad + \sum_{p=1}^{P}(\zeta_{jm}-1)\log\frac{\zeta_{jm}}{\zeta_{j\cdot}} + \sum_{l=1}^{K}(\alpha_{jk}-1)\log\frac{\alpha_{jk}}{\alpha_{j\cdot}}\Bigg\} + \text{constant}, \\
\log p(\tilde{\Theta}) &= \sum_{j=1}^{K}\Bigg\{\sum_{p=1}^{P}\big[-\delta_{jm}^{(0)}\frac{\gamma_{jm}}{\delta_{jm}} + (\gamma_{jm}^{(0)}-1)\log\frac{\gamma_{jm}}{\delta_{jm}}\big] + (\zeta_{j0}^{(0)}-1)\log\frac{\zeta_{j0}}{\zeta_{j\cdot}} \\
&\quad + \sum_{p=1}^{P}(\zeta_{jm}^{(0)}-1)\log\frac{\zeta_{jm}}{\zeta_{j\cdot}} + \sum_{l=1}^{K}(\alpha_{jk}^{(0)}-1)\log\frac{\alpha_{jk}}{\alpha_{j\cdot}}\Bigg\} + \text{constant}.
\end{aligned}$$

We note that the prior hyperparameters and the posterior hyperparameters are linearly related through the variational updates; this can be used to simplify the expressions significantly. Also,

the constants all cancel out. Taking expectation of the first 2 terms under the variational posterior distribution we get

$$\begin{aligned} p_D = \sum_{j=1}^{K}\sum_{t=1}^{T}\Big\{ & q_{tj}q_{tj0}\big[\log\frac{\zeta_{j0}}{\zeta_{j\cdot}} - \Psi(\zeta_{j0}) + \Psi(\zeta_{j\cdot})\big] \\ & + \sum_{p=1}^{P} q_{tj}q_{tjm}\big[\log\frac{\zeta_{jm}}{\zeta_{j\cdot}} - \Psi(\zeta_{jm}) + \Psi(\zeta_{j\cdot}) + \log\gamma_{jm} - \Psi(\gamma_{jm})\big] \\ & + \sum_{l=1}^{K} q(s_j = j, s_{t+1} = l)\big[\log\frac{\alpha_{jk}}{\alpha_{j\cdot}} - \Psi(\alpha_{jk}) + \Psi(\alpha_{j\cdot})\big]. \end{aligned} \tag{3.21}$$

The density function $p(y|\tilde{\Theta})$ can be obtained from (2.36) using the scaling constants from the Forward recursion algorithm. We can then compute the DIC for the model.

3.5 Gamma Distribution for Positive Precipitation

Next, we consider a model identical to the previous section but replace the Exponential distribution for positive rainfall with a Gamma distribution. If the positive rainfall at the l^{th} location from the m^{th} mixture component (where $m > 0$) arising from state *j* follows a Gamma distribution with shape ω_{jlm} and rate λ_{jlm}, the density of an observation from state *j* across all locations is given by

$$\begin{aligned} \prod_{l=1}^{L} p(y_{tl}, r_{tjl}|\omega_{jl}, \lambda_{jl}, c_{jl}, s_t = j) &= \prod_{l=1}^{L} p(r_{tjl}|c_{jl}, s_t = j)\cdot p(y_{tl}|\omega_{jl}, \lambda_{jl}, r_{tjl}, s_t = j) \\ &= \prod_{l=1}^{L}\Big\{ c_{jl0}^{r_{tjl0}} \prod_{m=1}^{M}\big[c_{jlm}\frac{\lambda_{jlm}^{\omega_{jlm}}}{\Gamma(\omega_{jlm})}\exp\{-\lambda_{jlm}y_{tl}\}y_{tl}^{\omega_{jlm}-1}\big]^{r_{tjlm}}\Big\}. \end{aligned} \tag{3.22}$$

The optimization for all parameters other than ω_j and λ_j are identical to the previous section with Exponential distribution emissions. We therefore focus only on the emission distribution. We use

a modified Gamma conjugate prior of type II described in Section 2.6, denoted here by $\mathcal{GC}2$:

$$\begin{aligned} p(\Lambda) &= \prod_{j=1}^{K}\prod_{l=1}^{L}\prod_{m=1}^{M} \mathcal{GC}2(\omega_{jlm}, \lambda_{jlm}|\gamma_{jlm}^{(0)}, \delta_{jlm}^{(0)}, \theta_{jlm}^{(0)}, \beta_{jlm}^{(0)}) \\ &= \prod_{j=1}^{K}\prod_{l=1}^{L}\prod_{m=1}^{M} \exp\big\{(\omega_{jlm}\gamma_{jlm}^{(0)} - 1)\log\lambda_{jlm} - \delta_{jlm}^{(0)}\log\Gamma(\omega_{jlm}) \\ &\qquad + (\omega_{jlm} - 1)\log\beta_{jlm}^{(0)} - \theta_{jlm}^{(0)}\lambda_{jlm}\big\}, \end{aligned} \tag{3.23}$$

where Λ_l is a $K \times M \times 2$ array of $(\omega_{jlm}, \lambda_{jlm})$ pairs for all states and mixture components at the l^{th} location. The $\mathcal{GC}2$ prior and its corresponding posterior is discussed briefly in Section 2.6, and is based on Miller [1980]. The hyperparameters $(\gamma_{jlm}^{(0)}, \delta_{jlm}^{(0)}, \theta_{jlm}^{(0)}, \beta_{jlm}^{(0)})$ are known.

VBM step: The variational updates for the Gamma distribution parameters are given by:

$$\begin{aligned} \gamma_{jlm} &= \gamma_{jlm}^{(0)} + \sum_{t=1}^{T} q_{tj} q_{tjlm} \\ \delta_{jlm} &= \delta_{jlm}^{(0)} + \sum_{t=1}^{T} q_{tj} q_{tjlm} \\ \theta_{jlm} &= \theta_{jlm}^{(0)} + \sum_{t=1}^{T} y_{tl} q_{tj} q_{tjlm} \\ \log\beta_{jlm} &= \log\beta_{jlm}^{(0)} + \sum_{t=1}^{T} \log y_{tl} q_{tj} q_{tjlm} \end{aligned}$$

where $j, k = 1, \ldots, K$, $l = 1, \ldots, L$, and $m = 1, \ldots, M$. The remaining updates are identical to the previous section.

VBE step: The variational posterior $q_{s,r}(s, r)$ has the same form as the known parameter posterior:

$$q_{s,r}(s, r) \propto \prod_{j=1}^{K}\{a_{1j}^*\}^{s_{1j}} \prod_{t=1}^{T}\prod_{j=1}^{K}\prod_{l=1}^{L}\prod_{m=0}^{M}\{b_{tjlm}^*\}^{s_{tj}r_{tjlm}} \prod_{t=1}^{T-1}\prod_{j=1}^{K}\prod_{k=1}^{K}\{a_{jk}^*\}^{s_{tj}s_{t+1,k}}, \tag{3.24}$$

with the natural parameters $\phi(\Theta)$ replaced by their expectations under $q_\Theta(\Theta)$. As before,

$$\begin{aligned} a_{1j}^* &= \exp\big\{\mathbb{E}_q \log\pi_{1j}\big\} = \exp\big\{\Psi(\xi_j) - \Psi(\xi_\cdot)\big\}, \\ \text{and} \quad a_{jk}^* &= \exp\big\{\mathbb{E}_q \log a_{jk}\big\} = \exp\big\{\Psi(\alpha_{jk}) - \Psi(\alpha_{j\cdot})\big\}, \end{aligned}$$

where $\xi_{\cdot} = \sum_{j=1}^{K} \xi_j$, $\alpha_{j\cdot} = \sum_{k=1}^{K} \alpha_{jk}$. Similarly,

$$b^*_{tjlm} = \begin{cases} \exp\{\mathbb{E}_q \log[c_{jl0}]\} & \text{if } m = 0, \\ \exp\{\mathbb{E}_q \log[c_{jlm} f(y_{tl}|\omega_{jlm}, \lambda_{jlm})]\} & \text{if } m > 0. \end{cases}$$

The expectations of the individual terms in b^*_{tjlm} are:

$$c^*_{jlm} = \exp\{\mathbb{E}_q \log c_{jlm}\} = \exp\{\Psi(\zeta_{jlm}) - \Psi(\zeta_{j\cdot})\}, \text{ where } \zeta_{j\cdot} = \sum_{m=0}^{M} \zeta_{jlm},$$

$$\begin{aligned} \mathbb{E}_q[\omega_{jlm} \log \lambda_{jlm}] &= \mathbb{E}_\omega \omega_{jlm} [\mathbb{E}_{\lambda|\omega} \log \lambda_{jlm}] \\ &= \mathbb{E}_\omega [\omega_{jlm} \Psi(\omega_{jlm}\gamma_{jlm}) - \omega_{jlm} \log \theta_{jlm}] \\ \mathbb{E}_q \lambda_{jlm} &= \mathbb{E}_\omega [\frac{\omega_{jlm}\gamma_{jlm}}{\theta_{jlm}}]. \end{aligned}$$

Functionals of the shape parameter ω_{jlm} need to be computed numerically at every iteration, as its normalizing constant does not have a closed form. This makes the VBE step computationally expensive, especially considering that $J \times L \times M$ of these need to be evaluated at every iteration. While there is no theoretical barrier to doing it, realistically, it makes more sense to assume the shape parameters are known, or use exponential distributions as described in the previous section. This also reiterates that conjugate priors are not always the best idea.

3.6 Gamma Shape Mixtures for Positive Precipitation

Venturini et al. [2008] proposed a Gamma shape mixture (GSM) for heavy tailed distributions in the context of medical expenditures which tend to have highly skewed distributions. For a positive random variable y, the GSM has a density of the form

$$p(y|c_1, \ldots c_M, \lambda) = \sum_{m=1}^{M} c_m \cdot p(y|\lambda, m), \tag{3.25}$$

$$\text{where} \quad p(y|\lambda, m) = \frac{\lambda^m}{\Gamma(m)} y^{m-1} \exp\{-\lambda y\}. \tag{3.26}$$

The GSM has several desirable properties. First, it addresses the identifiability issue of the individual components by assigning a natural ordering to the moments of the Gamma distributions. For example, the mean of the m^{th} mixture component is m/λ. Further, by assuming M to be known, all mixture components have only a shared unknown parameter λ. It is quite straightforward to assign a Gamma prior to λ, and thus the number of hyperparameters is significantly reduced. This simplification also allows the estimation of priors using empirical Bayes.

The approach is not without its limitations, however. The primary concern is that the shape components are too close to each other since they are just a sequence of natural numbers. Modeling emissions with a large range requires a high value for *M*, especially when using empirical Bayes to estimate priors. This is somewhat mitigated since it is possible for most of the mixture components to get negligible weights and eventually drop out as Venturini et al. [2008] demonstrate with examples. We propose a GSM which identifies the mixture components based on a more general sequence of shape parameters.

Multivariate emissions, as in the previous subsections, are also considered here. Let positive rainfall at the l^{th} location from the m^{th} mixture component (where $m > 0$) arising from state *j* follow a Gamma distribution with shape $f_m = f(m)$, a known function of *m*, and rate λ_{jl}. f_m is chosen such that $f_1 < f_2 < \ldots < f_M$. The distribution of precipitation on day *t* at location *l* for mixture component *m* is given by

$$p(y_{tl}|\lambda_{jl}, r_{tjlm}) = \begin{cases} 1 & \text{if } m = 0, \\ \frac{\lambda_{jl}^{f_m}}{\Gamma(f_m)} y_t^{f_m - 1} \exp\{-\lambda_{jl} y_t\} & \text{if } m > 0. \end{cases} \tag{3.27}$$

In Venturini et al. [2008], $f_m = m$. In a more general setting, f_m could be a polynomial or exponential function. The distribution of an observation from state *j* across all locations is

$$\begin{aligned} p(y_t, r_{tjl}|\lambda_j, c_{jl}, s_t = j) &= \prod_{l=1}^{L} p(r_{tjl}|c_{jl}, s_t = j) \cdot p(y_t|\lambda_j, r_{tjl}, s_t = j) \\ &= \prod_{l=1}^{L} \left\{ c_{jl0}^{r_{tjl0}} \prod_{m=1}^{M} \left[c_{jlm} \frac{\lambda_{jl}^{f_m}}{\Gamma(f_m)} y_t^{f_m - 1} \exp\{-\lambda_{jl} y_{tl}\} \right]^{r_{tjlm}} \right\}. \end{aligned} \tag{3.28}$$

The complete data likelihood is given by

$$p(y, s, r|\Theta) = p(y, r|s, \Theta) \cdot p(s|\Theta),$$

We define the parameters of the HMM using notation introduced earlier. Let $A = ((a_{jk}))$ be the $K \times K$ transition probability matrix. $C_l = ((c_{jlm}))$ is the $K \times (M+1)$ matrix of mixture probabilities for each location *l*, with $C = (C_1, \ldots, C_L)$. Similarly, $\Lambda = ((\lambda_{jl}))$ is a $K \times L$ matrix whose elements are the independently distributed rate parameters of the Gamma distributions which are part of the semi-continuous emissions in each state at every location. Finally, let π_1 be a *K*-vector of the initial distribution. Taken together, $\Theta = (A, C, \Lambda, \pi_1)$ parameterizes the HMM. We assign a prior on Θ which factorizes into a product over its components:

$$p(\Theta|\nu^{(0)}) = p(\pi_1) \cdot p(A) \cdot p(C) \cdot p(\Lambda),$$

where $\nu^{(0)}$ are the hyperparameters. We assign independent Dirichlet priors to the rows of *A*, and to the rows of C_l. Similarly, a Dirichlet prior is assigned to π_1, and independent Gamma priors are assigned to each element of Λ_l. That is,

$$\begin{aligned}
p(\pi_1) &= Dirichlet(\pi_1|\xi^{(0)}), \\
p(A) &= \prod_{j=1}^{K} Dirichlet(a_j|\alpha_j^{(0)}), \\
p(C) &= \prod_{j=1}^{K}\prod_{l=1}^{L} Dirichlet(c_{jl}|\zeta_{jl}^{(0)}), \\
\text{and } p(\Lambda) &= \prod_{j=1}^{K}\prod_{l=1}^{L} Gamma(\lambda_{jl}|\gamma_{jl}^{(0)}, \delta_{jl}^{(0)}),
\end{aligned}$$

where $a_j = (a_{j1}, \ldots, a_{jK})$, $\pi_1 = (\pi_{11}, \ldots, \pi_{1K})$, $\zeta_{jl}^{(0)} = (\zeta_{jl0}^{(0)}, \ldots, \zeta_{jlM}^{(0)})$, $\alpha_j^{(0)} = (\alpha_{j1}^{(0)}, \ldots, \alpha_{jK}^{(0)})$, and $\xi^{(0)} = (\xi_1^{(0)}, \ldots \xi_K^{(0)})$. $\gamma_{jl}^{(0)}$ and $\delta_{jl}^{(0)}$ are the shape and rate parameters of the Gamma distribution respectively. The hyperparameters are known.

The complete data likelihood can be expressed as

$$\begin{aligned} p(y,s,r|\Theta) &= \prod_{j=1}^{K}\{\pi_{1j}\}^{s_{1j}} \prod_{t=1}^{T}\prod_{j=1}^{K}\prod_{l=1}^{L}\{p_j(y_{tl},r_{tjl}|\Theta)\}^{s_{tj}} \prod_{t=1}^{T-1}\prod_{j=1}^{K}\prod_{k=1}^{K}\{a_{jk}\}^{s_{tj}s_{t+1,k}} \\ &= \exp\Bigg\{\sum_{j=1}^{K} s_{1j}\log\pi_{1j} + \sum_{t=1}^{T}\sum_{j=1}^{K}\sum_{l=1}^{L}\Bigg[\sum_{m=1}^{M} s_{tj}r_{tjlm}(\log c_{jlm} + f_m\log\lambda_{jl} - \log\Gamma(f_m) \\ &\quad + (f_m-1)\log y_t - y_t\lambda_{jl}) + s_{tj}r_{tjl0}\log c_{jl0}\Bigg] + \sum_{t=1}^{T-1}\sum_{j=1}^{K}\sum_{k=1}^{K} s_{tj}s_{t+1,k}\log a_{jk}\Bigg\}, \end{aligned} \tag{3.29}$$

where $s_{tj} = \mathbb{I}\{s_t = j\}$ denotes the daily state and $s_{tj}s_{t+1,k}$ denotes a typical state transition. The prior is given by:

$$\begin{aligned} p(\Theta|\nu^{(0)}) &= p(\pi_1)\cdot p(\lambda)\cdot p(C)\cdot p(A) \\ &= \exp\Bigg\{\sum_{j=1}^{K}\Big\{(\xi_j^{(0)}-1)\log\pi_{1j} + \sum_{l=1}^{L}\big[-\delta_{jl}^{(0)}\lambda_{jl} + (\gamma_{jl}^{(0)}-1)\log\lambda_{jl}\big] \\ &\quad + \sum_{l=1}^{L}(\zeta_{jl0}^{(0)}-1)\log c_{jl0} + \sum_{l=1}^{L}\sum_{m=1}^{M}(\zeta_{jlm}^{(0)}-1)\log c_{jlm} \\ &\quad + \sum_{k=1}^{K}(\alpha_{jk}^{(0)}-1)\log a_{jk}\Big\} - \log h^{(0)}\Bigg\}, \end{aligned} \tag{3.30}$$

where $h^{(0)} = h(\nu^{(0)})$ is the normalizing constant for the prior. Comparing this expression with the canonical form for the conjugate exponential family, we arrive at the following expressions for the natural parameters $\phi(\Theta)$, their sufficient statistics $u(s,y,r)$, and the hyperparameters $\nu^{(0)}$:

$$\phi(\Theta) = \begin{bmatrix} \log\pi_{1j} \\ \log c_{jl0} \\ \log c_{jlm} \\ \log\lambda_{jl} \\ \lambda_{jl} \\ \log a_{jk} \end{bmatrix}, \quad u(s,y,r) = \begin{bmatrix} s_{1j} \\ s_{tj}r_{tjl0} \\ s_{tj}r_{tjlm} \\ s_{tj}\sum_{m=1}^{M} f_m r_{tjlm} \\ y_{tl}s_{tj}\sum_{m=1}^{M} r_{tjlm} \\ s_{tj}s_{t+1,k} \end{bmatrix}, \quad \nu^{(0)} = \begin{bmatrix} \xi_j^{(0)}-1 \\ \zeta_{jl0}^{(0)}-1 \\ \zeta_{jlm}^{(0)}-1 \\ \gamma_{jl}^{(0)}-1 \\ \delta_{jl}^{(0)} \\ \alpha_{jk}^{(0)}-1 \end{bmatrix}, \tag{3.31}$$

for $m = 1,\ldots,M, j = 1,\ldots,K, k = 1,\ldots,K$. The variational family $\mathbb{Q}$ is constrained to

distributions which are separable in the following manner:

$$q_z(z) = q_\Theta(\Theta) \cdot q_{s,r}(s,r), \tag{3.32}$$

$$\text{where } q_\Theta(\Theta) = q(\pi_1) \cdot q(A) \cdot q(C) \cdot q(\Lambda). \tag{3.33}$$

VBM step: *With the variational posterior of the latent variables $q_{s,r}(s,r)$ fixed at their expected value, update $q_\Theta(\Theta)$, the variational posterior of the model parameters.*

The variational updates at each iteration of the VBM step are then given by:

$$\begin{aligned}
\xi_j &= \xi_j^{(0)} + q_{1j}, \\
\zeta_{jl0} &= \zeta_{jl0}^{(0)} + \sum_{t=1}^{T} q_{tj} q_{tjl0}, \\
\zeta_{jlm} &= \zeta_{jlm}^{(0)} + \sum_{t=1}^{T} q_{tj} q_{tjlm}, \\
\gamma_{jl} &= \gamma_{jl}^{(0)} + \sum_{t=1}^{T} q_{tj} (\sum_{m=1}^{M} f_m q_{tjlm}), \\
\delta_{jl} &= \delta_{jl}^{(0)} + \sum_{t=1}^{T} q_{tj} y_{tl} (\sum_{m=1}^{M} q_{tjlm}), \\
\alpha_{jk} &= \alpha_{jk}^{(0)} + \sum_{t=1}^{T-1} q_{jk},
\end{aligned}$$

where $j, k = 1, \ldots, K$, $l = 1, \ldots, L$, $m = 1, \ldots, M$.

VBE step: *With the variational posterior on the model parameters $q_\Theta(\Theta)$ fixed, update the variational posterior $q_{s,r}(s,r)$ of the latent variables.*

The variational posterior $q_{s,r}(s,r)$ has the form

$$q_{s,r}(s,r) \propto \prod_{j=1}^{K} \{a_{1j}^*\}^{s_{1j}} \prod_{t=1}^{T} \prod_{j=1}^{K} \prod_{l=1}^{L} \prod_{m=0}^{M} \{b_{tjlm}^*\}^{s_{tj} r_{tjlm}} \prod_{t=1}^{T-1} \prod_{j=1}^{K} \prod_{k=1}^{K} \{a_{jk}^*\}^{s_{tj} s_{t+1,k}}, \tag{3.34}$$

with the natural parameters $\phi(\Theta)$ replaced by their expectations under $q_\Theta(\Theta)$. Comparing with (3.29), we get

$$a_{1j}^* = \exp\{\mathbb{E}_q \log \pi_{1j}\} = \exp\{\Psi(\xi_j) - \Psi(\xi_\cdot)\},$$

$$\text{and } a^*_{jk} = \exp\big\{\mathbb{E}_q \log a_{jk}\big\} = \exp\big\{\Psi(\alpha_{jk}) - \Psi(\alpha_{j.})\big\},$$

where $\xi_. = \sum_{j=1}^K \xi_j$, $\alpha_{j.} = \sum_{k=1}^K \alpha_{jk}$. Similarly,

$$b^*_{tjlm} = \begin{cases} \exp\big\{\mathbb{E}_q \log\big[c_{jl0}\big]\big\} & \text{if } m = 0, \\ \exp\big\{\mathbb{E}_q \log\big[c_{jlm} f(y_{tl}|f_m, \lambda_{jl})\big]\big\} & \text{if } m > 0. \end{cases}$$

The expectations of the individual terms in b^*_{tjlm} are:

$$c^*_{jlm} = \exp\big\{\mathbb{E}_q \log c_{jlm}\big\} = \exp\big\{\Psi(\zeta_{jlm}) - \Psi(\zeta_{j.})\big\}, \text{ where } \zeta_{j.} = \sum_{m=0}^{M} \zeta_{jlm},$$

$$\lambda^*_{jl} = \exp\big\{\mathbb{E}_q \log \lambda_{jl}\big\} = \exp\big\{\Psi(\gamma_{jl}) - \log \delta_{jl}\big\},$$

$$\hat{\lambda}_{jl} = \mathbb{E}_q \lambda_{jl} = \gamma_{jl}/\delta_{jl}.$$

Therefore,

$$b^*_{tjlm} = \begin{cases} \exp\big\{\Psi(\zeta_{jl0}) - \Psi(\zeta_{jl.})\big\} & \text{if } m = 0, \\ \exp\big\{\Psi(\zeta_{jlm}) - \Psi(\zeta_{jl.}) + f_m[\Psi(\gamma_{jl}) - \log \delta_{jl}] \\ \qquad\qquad - \log\Gamma(f_m) + (f_m - 1)\log y_{tl} - y_{tl}\frac{\gamma_{jl}}{\delta_{jl}}\big\} & \text{if } m > 0, \end{cases}$$

$$\text{and } \; b^*_{tj} = \prod_{l=1}^{L} \sum_{m=0}^{M} b^*_{tjlm}.$$

As before, a^*_{1j}, a^*_{jk}, and b^*_{tj} can be used as part of the Forward-Backward Algorithm to get our desired variational posterior estimates for the state probabilities as well as the cluster assignment probabilities. The posterior for the mixture assignments for the l^{th} location is given by

$$q_{tjlm} \propto \begin{cases} 1 & \text{if } m = 0, y_{tl} = 0, \\ 0 & \text{if } m > 0, y_{tl} = 0 \text{ or } m = 0, y_{tl} > 0, \\ c^*_{jlm} f(y_{tl}|\lambda^*_{jlm}, \hat{\lambda}_{jlm}) & \text{if } m > 0, y_{tl} > 0, \end{cases}$$

where

$$c^*_{jlm} f(y_{tl}|\lambda^*_{jlm}, \hat{\lambda}_{jlm}) = \exp\big\{\Psi(\zeta_{jlm}) - \Psi(\zeta_{jl\cdot}) + f_m[\Psi(\gamma_{jl}) - \log\delta_{jl}] \\ - \log\Gamma(f_m) + (f_m - 1)\log y_{tl} - y_{tl}\frac{\gamma_{jl}}{\delta_{jl}}\big\}.$$

3.7 Assigning Priors using Empirical Bayes

Rabiner [1989] notes that good initial estimates of the emission distribution parameters are essential for rapid and proper convergence of the Baum-Welch algorithm. He suggests the following procedure as a way to obtain good initial estimates:

Algorithm 1 Initial estimates for the Baum-Welch algorithm [Rabiner, 1989].

1: Subset a section of the data to use as a training dataset
2: Choose initial model estimates for $\Theta^{(0)} = (A, B, \pi_1)$ at random
3: Segment the training dataset using the Viterbi algorithm and assign a state to each observation
4: Use *k*-means clustering to assign each state's observations to mixture components
5: Update the estimates of the emission distribution parameters based on this assignment for each state and each mixture component
6: Update estimates for the state distribution parameters based on the state segmentation in Step 3

In our Bayesian context, three possible approaches for assigning priors are:

1. Use historical data if available. The historical data might not be on the same spatial grid as our current data. In that case, one option is to use a nearest-neighbor approach to connect the historical spatial grid to our current spatial grid.

2. Assign vague priors to all parameters. This means symmetric Dirichlet priors for the components of *A* and *C*, and randomly chosen priors to elements of Λ, under some identifiability constraints. This has the risk of running into the issues documented by Rabiner [1989].

3. Adapt the procedure described in Algorithm 1.

The final option, in particular, works well when we use a GSM for positive precipitation. This is because Step 4 from Algorithm 1 is no longer necessary. The empirical Bayes procedure described

in Venturini et al. [2008] can be modified for this purpose:

Algorithm 2 Empirical Bayes priors for positive precipitation.

1: Subset the first year's data to use as training set
2: Assign states to each day's emissions. This can be done using the Viterbi algorithm, or by clustering the data
3: Choose $\tilde{\lambda}_{jl}$, an estimate of λ_{jl}, in a manner such that the mixture distribution forms a 90% interval for positive precipitation.
4: Use $\tilde{\lambda}_{jl}$ to assign $\gamma_{jl}^{(0)}$ and $\delta_{jl}^{(0)}$ using empirical Bayes

Estimating λ_{jl}: Resorting to a slight abuse in notation, we denote the GSM components for each state and location combination as $y_{jl}^{(m)}$ such that $\mathbb{E}(y_{jl}^{(m)}) = f_m/\lambda_{jl}$ for $m = 1, \ldots, M$. The 90% interval should satisfy:

$$Pr[y_{jl}^{(1)} < y_{jl(1)}] < 0.05, \tag{3.35}$$

$$\text{and} \quad Pr[y_{jl}^{(M)} > y_{jl(n)}] < 0.05, \tag{3.36}$$

where $y_{jl(1)}$ and $y_{jl(n)}$ are the smallest and largest order statistics for observed positive precipitation arising from state j at the l^{th} location. We consider these conditions sufficient to ensure that the proposed emission distribution can model the observed data. Since f_m are fixed for all m and $y_{jl(1)}$ and $y_{jl(n)}$ are observed from the data, all quantities except λ_{jl} are known in (3.35) and (3.36). Since the CDF of a Gamma distribution does not have a closed form, we can use a root-finding algorithm to solve the 2 inequalities and get a range of estimates of λ_{jl} as a function of f_1 and f_m. Let us denote this range as $(\tilde{\lambda}_{jl}^{min}, \tilde{\lambda}_{jl}^{max})$, and any value in this range can be used as a point estimate for λ_{jl}. For simplicity, we choose

$$\tilde{\lambda}_{jl} = \frac{\tilde{\lambda}_{jl}^{min} + \tilde{\lambda}_{jl}^{max}}{2}. \tag{3.37}$$

Note that we can pool over all locations to get a single estimate for each state should we want to simplify things further. Venturini et al. [2008] derive a similar estimate but use the means m/λ_{jl} instead of percentiles. We believe this to be too restrictive since a large number of mixture components are needed to satisfy their criteria.

Assigning the prior hyperparameters $(\gamma_{jl}^{(0)}, \delta_{jl}^{(0)})$: The conditional distribution of positive

precipitation has the form

$$p(y_{tl}|\lambda_{jl}, r_{tjlm}, s_t = j) = \frac{\lambda_{jl}^{f_m r_{tjlm}}}{\Gamma(f_m)} y_{tl}^{f_m - 1} \exp\{-\lambda_{jl} y_{tl}\} \qquad \text{when } m > 0.$$

Using this expression and the prior for λ_{jl}, we get the complete conditional for λ_{jl}:

$$p(\lambda_{jl}|y, r, s) \propto \text{Gamma}(\sum_{t=1}^{T}\sum_{m=1}^{M} f_m s_{tj} r_{tjlm} + \gamma_{jl}^{(0)}, \sum_{t=1}^{T} s_{tj} y_{tl} + \delta_{jl}^{(0)}).$$

This gives us

$$\begin{aligned}
\mathbb{E}(\lambda_{jl}|y, r, s) &= \frac{\sum_{t=1}^{T}\sum_{m=1}^{M} f_m s_{tj} r_{tjlm} + \gamma_{jl}^{(0)}}{\sum_{t=1}^{T} s_{tj} y_{tl} + \delta_{jl}^{(0)}} \\
&= \frac{\delta_{jl}^{(0)}}{\sum_{t=1}^{T} s_{tj} y_{tl} + \delta_{jl}^{(0)}} \cdot \frac{\gamma_{jl}^{(0)}}{\delta_{jl}^{(0)}} + \frac{\sum_{t=1}^{T} s_{tj} y_{tl}}{\sum_{t=1}^{T} s_{tj} y_{tl} + \delta_{jl}^{(0)}} \cdot \frac{\sum_{t=1}^{T}\sum_{m=1}^{M} f_m s_{tj} r_{tjlm}}{\sum_{t=1}^{T} s_{tj} y_{tl}} \\
&= \tilde{\omega} \cdot \frac{\gamma_{jl}^{(0)}}{\delta_{jl}^{(0)}} + (1 - \tilde{\omega}) \cdot \frac{\sum_{t=1}^{T}\sum_{m=1}^{M} f_m s_{tj} r_{tjlm}}{\sum_{t=1}^{T} s_{tj} y_{tl}} \\
\text{where } \tilde{\omega} &= \frac{\delta_{jl}^{(0)}}{\sum_{t=1}^{T} s_{tj} y_{tl} + \delta_{jl}^{(0)}}.
\end{aligned}$$

The posterior expectation is thus a weighted average of the prior mean and the data. We assign a weight to the prior information $\tilde{\omega}$; Venturini et al. [2008] suggests values between 0.2–0.5 as being reasonable. We can then solve for the hyperparameters, which gives us:

$$\delta_{jl}^{(0)} = \frac{\tilde{\omega}}{1 - \tilde{\omega}} \sum_{t=1}^{T} s_{tj} y_{tl}, \tag{3.38}$$

$$\gamma_{jl}^{(0)} = \tilde{\lambda}_{jl} \cdot \delta_{jl}^{(0)}. \tag{3.39}$$

Note that we use $\tilde{\omega}$ and not $\tilde{\omega}_{jl}$ since we assume the prior to have the same weight across all states and locations. This assumption can be relaxed.

3.8 Stochastic Variational Bayes for HMMs

As discussed in Chapter 2, most VB applications make the mean field assumption, which specifies the approximate variational posterior of all parameters and latent variables as a product of the distributions of its individual components. Parameters can then be estimated using a VB version of the Expectation Maximization (VBEM) algorithm. However, each iteration of the VBEM requires computing means over the complete data and results in a performance bottleneck for large datasets. Stochastic optimization methods provide us a way around this, and stochastic variational Bayes (SVB) [Hoffman et al., 2013] is one such approach which implements VBEM as a stochastic gradient ascent algorithm for each parameter. Instead of computing gradients based on the entire data, SVB uses an unbiased estimate of the gradient at each iteration. The SVB algorithm converges to a local optimum as long as the step sizes for the gradient ascent satisfy the Robbins-Monro conditions [Robbins and Monro, 1951].

Under the mean field assumption, an unbiased estimator of the gradient can be constructed using a single observation. At each iteration, the updates take the form described in (2.34) – (2.35). The mean field assumption does not hold for an HMM $\{S_t, Y_t\}_{t\geq 0}$, and a single sample point (S_i, Y_i) cannot be used to estimate the transition probabilities of $\{S_t\}$. Consequently, a sample consisting of a sequence of observations is required to estimate the variational posterior of the parameters of $\{S_t\}$. We denote this sequential sample, or minibatch, as y^*. The nature of the dataset dictates the procedure for selecting y^*. If the data consists of a single long chain, Foti et al. [2014] proposed subsampling from the chain and buffering the beginning and end with extra observations to preserve the Markov properties of the states. If, however, the data is seasonal or cyclical in nature that can be represented as N blocks each of size D, then a minibatch is constructed at each optimization iteration by randomly sampling blocks with replacement and selecting all D time points within the block. This approach is discussed in Johnson and Willsky [2014]. In both cases, the variational E-step employs the Forward-Backward algorithm [Rabiner, 1989], and the variational M-step often takes advantage of conjugate priors and provides parameter updates through stochastic gradient ascent.

For the second case where the data admits a block structure, a concern arises from having

to run a large number of optimization iterations using a relatively small number of blocks of data. If we want to select a minibatch of D time points from N blocks of data at each iteration, this is equivalent to sampling with replacement from $1, \ldots, N$. The value of N is often not large in practice, and the approach in Johnson and Willsky [2014] is not ideal since there might not be a lot of variability within the samples between different iterations. We propose an alternative method which leverages the exchangeability inherent in HMMs.

For an HMM to model daily precipitation over the course of a season, we note a break in the collection of data between the end of the season in a year and the beginning of the season the next year. Each year's data therefore constitutes blocks with exchangeable distributions [Rabiner, 1989]. Further, for N years' data with D days in each year, data for the d^{th} day of every year has the same distribution. In particular, if we think of the D days as coming from C months, the c^{th} month has the same distribution for all years. With this exchangeability of days between years in mind, we propose the following algorithm for constructing a minibatch y^* from the data:

Algorithm 3 Minibatch sampling in Stochastic Variational Bayes for HMMs

1: Divide the D days in each year into C months
2: Draw a sample $s_1, \ldots, s_C$ of size C from $\{1, \ldots, N\}$ with replacement
3: If $s_c = i$, the c^{th} month of the i^{th} year provides data for the c^{th} month of the minibatch

Instead of N possible unique minibatches of size D, our approach allows for N^C unique minibatches of size D. An extreme extension of this would be sampling each of the D days from the N different years with replacement; however, we found that to be detrimental in estimating the transition probability parameters. The minibatch y^* can now be used for SVB. When positive precipitation follows a mixture of Exponential distributions, the hyperparameter updates in the VBM step for the i^{th} iteration is a natural gradient step of size ρ_i:

$$\xi_j^{(i)} = (1-\rho_i)\xi_j^{(i-1)} + \rho_i(\xi_j^{(0)} + q_{1j}),$$
$$\zeta_{jl0}^{(i)} = (1-\rho_i)\zeta_{jl0}^{(i-1)} + \rho_i(\zeta_{jl0}^{(0)} + N \cdot \sum_{t=1}^{D} q_{tj} q_{tjl0}),$$
$$\zeta_{jlm}^{(i)} = (1-\rho_i)\zeta_{jlm}^{(i-1)} + \rho_i(\zeta_{jlm}^{(0)} + N \cdot \sum_{t=1}^{D} q_{tj} q_{tjlm}),$$

$$\gamma_{jlm}^{(i)} = \big(1-\rho_i\big)\gamma_{jlm}^{(i-1)} + \rho_i\big(\gamma_{jlm}^{(0)} + N\cdot\sum_{t=1}^{D} q_{tj}q_{tjlm}\big),$$
$$\delta_{jlm}^{(i)} = \big(1-\rho_i\big)\delta_{jlm}^{(i-1)} + \rho_i\big(\delta_{jlm}^{(0)} + N\cdot\sum_{t=1}^{D} q_{tj}q_{tjlm}y_{tl}\big),$$
$$\alpha_{jk}^{(i)} = \big(1-\rho_i\big)\alpha_{jk}^{(i-1)} + \rho_i\big(\alpha_{jk}^{(0)} + N\cdot\sum_{t=1}^{D-1} q_{jk}\big),$$

where $j,k = 1,\ldots,K$, $l = 1,\ldots,L$, and $m = 1,\ldots,M$. When the emissions follow a GSM, $(\gamma_{jlm},\delta_{jlm})$ is replaced by $(\gamma_{jl},\delta_{jl})$ whose updates are:

$$\gamma_{jl}^{(i)} = \big(1-\rho_i\big)\gamma_{jl}^{(i-1)} + \rho_i\big(\gamma_{jl}^{(0)} + N\cdot\sum_{t=1}^{D} q_{tj}q_{tjl\cdot}\big),$$
$$\delta_{jl}^{(i)} = \big(1-\rho_i\big)\delta_{jl}^{(i-1)} + \rho_i\big(\delta_{jl}^{(0)} + N\cdot\sum_{t=1}^{D} q_{tj}q_{tjl\cdot}y_{tl}\big),$$

where $j,k = 1,\ldots,K$, $l = 1,\ldots,L$, $m = 1,\ldots,M$, and $q_{tjl\cdot} = 1 - q_{tjl0} = \sum_{m=1}^{M} q_{tjlm}$.

3.9 Simulation Studies

We present 4 simulation studies testing variational Bayes estimation under different scenarios. This includes coordinate ascent variational Bayes (CAVI) which uses the entire data for the estimation process, as well as stochastic variational Bayes (SVB) which uses stochastic optimization for speedup. The first study corresponds to the model presented in Section 3.1, corresponding to precipitation at a single location with a mixture of Exponential distributions used to specify positive precipitation. Next, we present a study with Exponential distribution mixtures replaced by Gamma shape mixtures (GSM) coinciding with the model in Section 3.6. For the GSM case, we use empirical Bayes to obtain priors for the emission distribution parameters. We then expand our scope to a model which consists of precipitation at 3 locations and uses Exponential mixtures, coinciding with the model in Section 3.4. These 3 studies use CAVI for parameter estimation. The final simulation study uses SVB as in Section 3.8 for precipitation at a single location with Exponential mixtures for positive precipitation.

Along with comparing parameter estimates from these studies, we also look at two precipi-

tation statistics that we want to replicate using synthetic data - the proportion of dry days, and the mean positive precipitation corresponding to the days when it rains. To construct the distributions of these 2 statistics, we run 1000 simulation studies. In each study, we first generate a dataset from the data generation process. We call this the original data, and estimate the parameters using variational Bayes. We then generate a fresh dataset using these estimated parameters, which we consider as synthetic data. The proportion of dry days as well as the mean positive precipitation are estimated from this second dataset. The 1000 simulation studies provide 1000 estimates, which are used to construct distributions of these two statistics. For comparison, the true proportion of dry days and the true mean positive precipitation are computed from the original datasets.

3.9.1 CAVI for single-site precipitation with Exponential mixtures

We simulated 1800 time steps from an HMM with 3 states (K=3), each with a dry component and 2 wet components (M=2), corresponding to 1800 days of daily precipitation data. For the simulation, we consider the initial probability vector to be $\pi_1 = (0.7, 0.2, 0.1)$ and

$$A = \begin{bmatrix} 0.45 & 0.35 & 0.20 \\ 0.30 & 0.40 & 0.30 \\ 0.30 & 0.30 & 0.40 \end{bmatrix}, \qquad C = \begin{bmatrix} 0.3 & 0.5 & 0.2 \\ 0.3 & 0.3 & 0.4 \\ 0.5 & 0.2 & 0.3 \end{bmatrix}, \qquad \Lambda = \begin{bmatrix} 0.08 & 1 \\ 0.60 & 5 \\ 1.00 & 8 \end{bmatrix},$$

where A, C, and Λ are the matrices of transition probabilities, mixture assignment probabilities, and exponential rate parameters for precipitation respectively. The rows of C and Λ correspond to parameter values for each state.

We keep our prior specifications as broad as possible, and assign symmetric Dirichlet priors for π_1 and A. A symmetric Dirichlet distribution is one where all the parameters are set to the same value; the sum total of the parameters is known as the concentration of the distribution. In our setup, $p(\pi_1)$ has a concentration of 1, and each row of $p(A)$ has a concentration of 10. Low concentration values are preferred since we do not want the prior to dominate the data. Without loss of generality, we order the states to correspond to heavy, medium, and low rainfall respectively. We denote the priors for the mixture assignment probabilities as $\zeta^{(0)}$. The elements of Λ

have Gamma distribution priors, and the shape and rate parameters of the Gamma distributions are given by $\gamma^{(0)}$ and $\delta^{(0)}$ respectively. These are set to the following values:

$$\zeta^{(0)} = \begin{bmatrix} 3.0 & 4.0 & 3.0 \\ 3.0 & 3.5 & 3.5 \\ 4.0 & 3.0 & 3.0 \end{bmatrix}, \qquad \gamma^{(0)} = \begin{bmatrix} 0.5 & 2 \\ 1.5 & 9 \\ 2.0 & 16 \end{bmatrix}, \qquad \delta^{(0)} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \\ 2 & 2 \end{bmatrix}.$$

The components are ordered to ensure that wetter states will have lower exponential rates and higher mixture probabilities for the exponential components, while drier states will have higher rates and more weight placed on the dry component corresponding to $m = 0$. This makes the model identifiable.

We perform 1000 simulation studies based on this setup, estimating the parameters using CAVI for each simulation. Note that the only thing that varies across the 1000 studies is the dataset used; CAVI is deterministic once we fix the dataset and the priors. We average the posterior means obtained across all the simulations. The posterior for the initial state probability is $\tilde{\pi}_1 = (0.38, 0.27, 0.35)$. Similarly,

$$\tilde{A} = \begin{bmatrix} 0.43 & 0.30 & 0.27 \\ 0.31 & 0.33 & 0.36 \\ 0.30 & 0.33 & 0.37 \end{bmatrix}, \qquad \tilde{C} = \begin{bmatrix} 0.29 & 0.50 & 0.21 \\ 0.32 & 0.29 & 0.39 \\ 0.47 & 0.21 & 0.32 \end{bmatrix}, \qquad \tilde{\Lambda} = \begin{bmatrix} 0.08 & 0.92 \\ 0.60 & 4.62 \\ 1.00 & 8.09 \end{bmatrix}.$$

We see that for the mixture probabilities in $\tilde{C}$ and the Exponential rate parameters in $\tilde{\Lambda}$ where we weigh our prior concentrations based on how weather states tend to be, the posteriors are quite close to the true values. But for the initial probability and the state transitions which have symmetric priors, the posteriors are not as close to the true values. In general, we found that while we can make the Dirichlet prior for the mixture probabilities symmetric without significant loss of accuracy in the posterior, the model is sensitive to the Gamma prior's hyperparameters. Details of studies using other priors are omitted.

For each of the 1000 simulations, we also generated 1800 days of data based on that iteration's estimated parameters to verify whether some of the key statistical characteristics of the

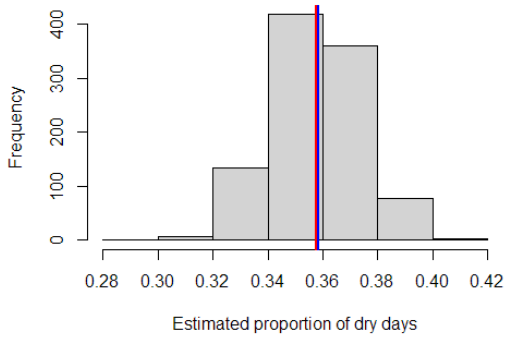

Figure 3.1: Proportion of dry days in 1800 days of data simulated using estimated parameters from 1000 simulation studies.

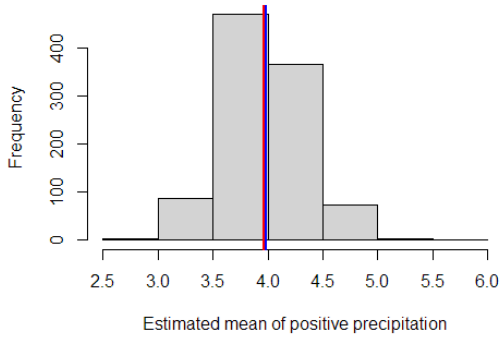

Figure 3.2: Mean positive rainfall (mm) in 1800 days of data simulated using estimated parameters from 1000 simulation studies.

HMM are captured. We compute the proportion of dry days and mean rainfall for wet days from each of these 1000 datasets. They are compared with Monte Carlo estimates derived from the true parameters. Figure 3.1 shows a histogram of the monthly proportion of dry days based on 1000 estimates. The red line at 0.358 is the mean of the data presented in the histogram, and the blue line is an estimate of the true proportion 0.359. The 1000 estimates have a root mean square error (RMSE) of 0.01. Similarly, Figure 3.2 plots a histogram of mean rainfall for wet days. The red line at 3.97 mm is the mean of the histogram data, and the blue line at 3.99 mm is an estimate of the true mean. The 1000 estimates have an RMSE of 0.26 mm. We do not notice any significant bias in our estimates for precipitation at a single location. As long as the model is well specified and the priors are reasonable, we are able to get good posterior estimates, as well as able to capture key statistics of daily precipitation of interest.

3.9.2 CAVI for single-site precipitation with Gamma shape mixtures

We simulated 1800 time steps from an HMM with 3 states (K=3), each with a dry component and 3 wet components (M=3), corresponding to 1800 days of daily precipitation data. For

the simulation, we consider the initial probability vector to be $\pi_1 = (0.7, 0.2, 0.1)$ and

$$A = \begin{bmatrix} 0.45 & 0.45 & 0.10 \\ 0.40 & 0.20 & 0.40 \\ 0.20 & 0.30 & 0.50 \end{bmatrix}, \qquad C = \begin{bmatrix} 0.1 & 0.2 & 0.3 & 0.4 \\ 0.2 & 0.2 & 0.4 & 0.2 \\ 0.5 & 0.2 & 0.2 & 0.1 \end{bmatrix}, \qquad \Lambda = \begin{bmatrix} 0.10 \\ 0.50 \\ 1.00 \end{bmatrix},$$

where *A* and *C* are the matrices of transition probabilities and mixture assignment probabilities respectively. Positive precipitation is assumed to follow a GSM distribution. As discussed in Section 3.6, GSM distributions will have a single rate parameter for each state. The elements of Λ are thus the three rate parameters for the 3 states. The vector of shape parameters is given by $f_m = (0.5, 3.5, 8.5)$ and is shared between the states. f_m is assumed to be known.

We assign symmetric Dirichlet priors for π_1 and *A*. $p(\pi_1)$ has a concentration of 1, and each row of $p(A)$ has a concentration of 10. C is also assumed to have Dirichlet priors. The parameters associated with the emission process are assigned empirical Bayes (EB) priors as discussed in Section 3.7. For this, we first need to know the state each data point comes from, since the emission distribution parameters are state specific. For real datasets, crude estimates can be obtained by clustering the data directly. Alternatively, we can fit an initial model under general priors and run the Viterbi Algorithm on the fitted model to obtain the most likely sequence of states. However, for the simulation study, we assume the true sequence of states to be known. The first 90 days of data is used to estimate the EB priors. Once we segment the data based on their states, we first obtain prior estimates for the mixture component assignments. Since it is a Dirichlet distribution with 4 components, the prior for the first component is assigned a weight proportionate to the number of dry days in the state, and the remaining weight is divided up among the three other components corresponding to positive precipitation. The weights sum up to a concentration of 10. We use (3.35) and (3.36) to obtain prior estimates of λ_{jl}, the Gamma distribution rates for each state. Algorithm 2 then provides us with EB estimates of $\gamma_{jl}^{(0)}$ and $\delta_{jl}^{(0)}$.

Like in the previous study, 1000 datasets are generated from the true model and parameter estimation is carried out for each of them. Afterwards, we average the estimates obtained from

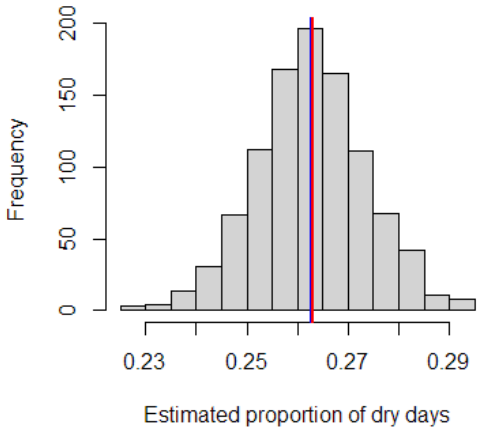

Figure 3.3: Proportion of dry days in 1800 days of data simulated using estimated parameters from 1000 simulation studies.

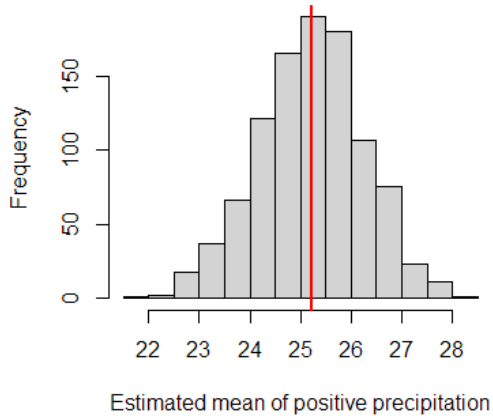

Figure 3.4: Mean positive rainfall (mm) in 1800 days of data simulated using estimated parameters from 1000 simulation studies.

each of them. The posterior for the initial state probability is $\tilde{\pi}_1 = (0.38, 0.32, 0.30)$. Similarly,

$$\tilde{A} = \begin{bmatrix} 0.41 & 0.37 & 0.22 \\ 0.34 & 0.27 & 0.39 \\ 0.23 & 0.28 & 0.49 \end{bmatrix}, \quad \tilde{C} = \begin{bmatrix} 0.07 & 0.22 & 0.42 & 0.29 \\ 0.17 & 0.22 & 0.39 & 0.22 \\ 0.51 & 0.17 & 0.20 & 0.12 \end{bmatrix}, \quad \tilde{\Lambda} = \begin{bmatrix} 0.18 \\ 0.55 \\ 1.07 \end{bmatrix}.$$

We see that the estimates for state 1, the wettest state, have the most errors, whereas the parameters associated with states 2 and 3 are much better estimated. We believe this to be an effect of the way we set the EB priors for state 1. State 1 has the widest range in precipitation values as it is the wettest, and our criteria based around (3.35)–(3.36) might not be enough to accurately set the priors for it. This is backed by the the pattern of errors we see in the first row of $\tilde{C}$, where the last two components have weights quite different from their true values.

For each of the 1000 simulations, we generated 1800 days of data based on that iteration's estimated parameters to verify whether the key statistical characteristics of daily precipitation arising from this model are captured. Figure 3.3 shows a histogram of the monthly proportion of dry days based on 1000 estimates. The red line at 0.263 is the mean of the data presented in the histogram, and the blue line is an estimate of the true proportion 0.263. While the proportions

are very close to each other, the 1000 estimates have a root mean square error (RMSE) of 0.02. Similarly, Figure 3.4 plots a histogram of mean rainfall for wet days. The red line at 25.21 mm is the mean of the histogram data, and the blue line at 25.215 mm is an estimate of the true mean. The 1000 estimates have an RMSE of 1.46 mm. We point out that this model has a much higher range of values compared to using exponential mixtures. While the data simulated using fitted parameters accurately replicates the statistics from the true data, care needs to be taken when setting the EB priors since they depend only on the smallest and largest GSM components.

3.9.3 CAVI for multi-site precipitation with Exponential mixtures

We consider precipitation at 3 locations, and the simulation setup is otherwise very similar to its single-site counterpart. The HMM has 3 states ($K = 3$). Precipitation at the 3 locations is considered distributed independently of each other conditional on the state. At every location, positive precipitation is distributed as a mixture of two exponential distributions, i.e. $M = 2$. We simulated 1800 time steps from this HMM. For the simulation, we consider the initial probability vector to be $\pi_1 = (0.38, 0.34, 0.28)$ and transition matrix

$$A = \begin{bmatrix} 0.60 & 0.30 & 0.10 \\ 0.20 & 0.50 & 0.30 \\ 0.30 & 0.20 & 0.50 \end{bmatrix}.$$

Each location would have a matrix for mixture component assignments, which we denote as C_1, C_2, and C_3. Similarly, the matrices for the Exponential distribution rates are denoted by

Λ_1, Λ_2, and Λ_3. They are set to the following values:

$$C_1 = \begin{bmatrix} 0.10 & 0.60 & 0.30 \\ 0.20 & 0.40 & 0.40 \\ 0.30 & 0.40 & 0.30 \end{bmatrix}, \quad C_2 = \begin{bmatrix} 0.20 & 0.70 & 0.10 \\ 0.40 & 0.20 & 0.40 \\ 0.50 & 0.20 & 0.30 \end{bmatrix}, \quad C_3 = \begin{bmatrix} 0.20 & 0.60 & 0.20 \\ 0.50 & 0.30 & 0.20 \\ 0.60 & 0.20 & 0.20 \end{bmatrix},$$

$$\Lambda_1 = \begin{bmatrix} 0.08 & 1 \\ 0.60 & 5 \\ 1.00 & 8 \end{bmatrix}, \qquad \Lambda_2 = \begin{bmatrix} 0.05 & 1 \\ 0.50 & 4 \\ 1.00 & 10 \end{bmatrix}, \qquad \Lambda_3 = \begin{bmatrix} 0.10 & 1 \\ 0.10 & 5 \\ 0.90 & 6 \end{bmatrix}.$$

The values of the mixture component assignments and the exponential rates are ordered such that state 1 corresponds to the wettest rainfall regime, and state 3 corresponds to the driest rainfall regime. For $l = 1, 2,$ and 3, the rows of C_l and Λ_l correspond to the parameter values for each state. We keep our prior specifications as broad as possible, and assign symmetric Dirichlet priors for π_1 and A. $p(\pi_1)$ has a concentration of 1, and each row of $p(A)$ has a concentration of 10. For each location l, the rows of C_l have Dirichlet priors, and elements of Λ_l have Gamma priors. The parameters for each location are assigned identical priors. We denote the prior for C_l using $\zeta_l^{(0)}$. Similarly, the Gamma priors of the elements of Λ_l have shape parameters $\gamma^{(0)}$ and rate parameters $\delta^{(0)}$. They are assigned the following values:

$$\zeta_l^{(0)} = \begin{bmatrix} 3.0 & 4.0 & 3.0 \\ 3.0 & 3.5 & 3.5 \\ 4.0 & 3.0 & 3.0 \end{bmatrix} \qquad \gamma_l^{(0)} = \begin{bmatrix} 0.5 & 2 \\ 1.5 & 9 \\ 2.0 & 16 \end{bmatrix} \qquad \delta_l^{(0)} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \\ 2 & 2 \end{bmatrix}$$

These assignments follow the reasoning that wetter states will have lower exponential rates and higher mixture probabilities for exponential components, while drier states will have higher rates and more weight placed on the dry component corresponding to $m = 0$.

We average the posterior estimates obtained from the 1000 simulations of the VB-HMM. The posterior for the initial state probability is $\tilde{\pi}_1 = (0.32, 0.24, 0.4)$. The posterior for the

transition probability matrix is

$$\tilde{A} = \begin{bmatrix} 0.59 & 0.28 & 0.13 \\ 0.21 & 0.41 & 0.38 \\ 0.29 & 0.28 & 0.43 \end{bmatrix},$$

and the posterior distributions of the mixture components and exponential rates are

$$\tilde{C}_1 = \begin{bmatrix} 0.10 & 0.59 & 0.31 \\ 0.20 & 0.38 & 0.42 \\ 0.30 & 0.40 & 0.30 \end{bmatrix} \quad \tilde{C}_2 = \begin{bmatrix} 0.20 & 0.68 & 0.12 \\ 0.39 & 0.23 & 0.38 \\ 0.49 & 0.20 & 0.31 \end{bmatrix} \quad \tilde{C}_3 = \begin{bmatrix} 0.20 & 0.60 & 0.20 \\ 0.48 & 0.32 & 0.20 \\ 0.60 & 0.21 & 0.19 \end{bmatrix},$$

$$\tilde{\Lambda}_1 = \begin{bmatrix} 0.08 & 0.99 \\ 0.65 & 4.92 \\ 0.91 & 7.63 \end{bmatrix} \quad \tilde{\Lambda}_2 = \begin{bmatrix} 0.05 & 0.91 \\ 0.63 & 4.35 \\ 0.99 & 9.02 \end{bmatrix} \quad \tilde{\Lambda}_3 = \begin{bmatrix} 0.10 & 1.07 \\ 0.10 & 4.76 \\ 1.02 & 7.32 \end{bmatrix}$$

We see that while the elements of the transition probability matrix are well estimated, the posterior of the initial distribution is quite far from the true parameters. Furthermore, the emission distribution parameters in $\tilde{C}_l$ and $\tilde{\Lambda}_l$ are well estimated for the most part. The largest errors are seen in the two extreme weather regimes - the first row of $\tilde{\Lambda}_1$ which corresponds to the highest rainfall among all states and locations, and the last row of $\tilde{\Lambda}_3$ which corresponds to the the lowest rainfall across all the states and locations. These are also the two cases where the prior means are farthest from the true parameter values, and demonstrate that the model can still estimate reasonable posteriors in such cases.

For each of the 1000 simulations, we generated 1800 days of data based on that iteration's estimated parameters. We compute the proportion of dry days and mean rainfall for wet days from each of these 1000 datasets. They are compared with Monte Carlo estimates derived from the true parameters. Figures 3.5, 3.7, and 3.9 show histograms of the monthly proportion of dry days based on the 1000 estimates for each of the 3 locations. The red lines are the means of the data presented in the histogram, and the blue lines are estimates of the true proportion of dry days at those locations. Similarly, Figures 3.6, 3.8, and 3.10 plot histograms of mean rainfall for wet

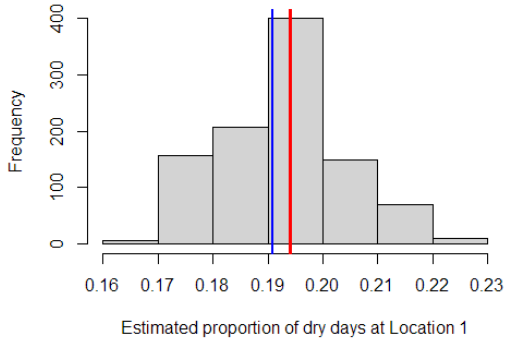

Figure 3.5: The proportion of dry days in 1800 days of data simulated using estimated parameters from 1000 simulation studies.

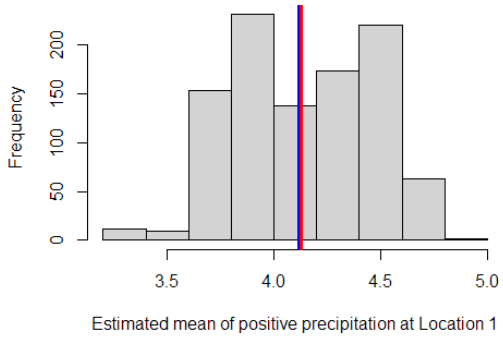

Figure 3.6: Mean positive rainfall (mm) in 1800 days of data simulated using estimated parameters from 1000 simulation studies.

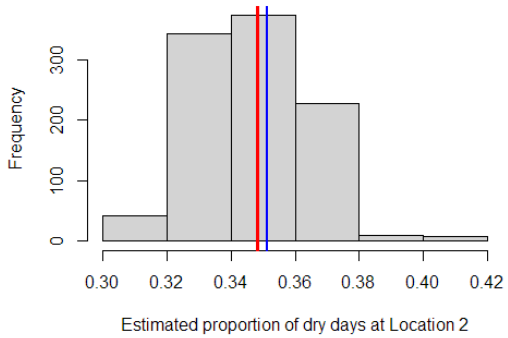

Figure 3.7: The proportion of dry days in 1800 days of data simulated using estimated parameters from 1000 simulation studies.

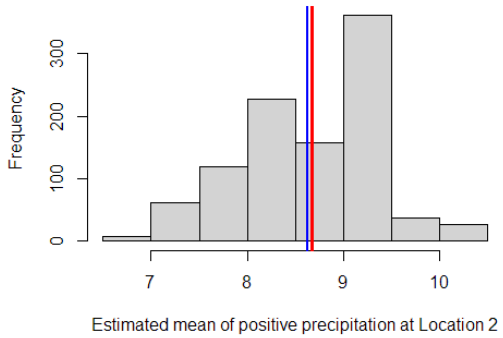

Figure 3.8: Mean positive rainfall (mm) in 1800 days of data simulated using estimated parameters from 1000 simulation studies.

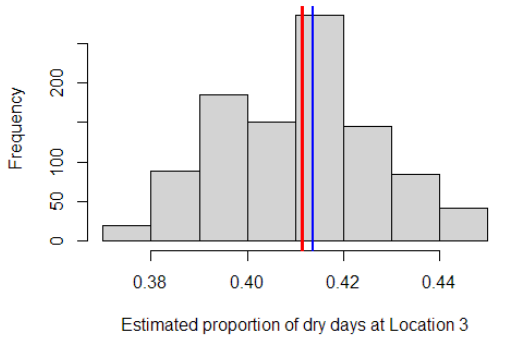

Figure 3.9: The proportion of dry days in 1800 days of data simulated using estimated parameters from 1000 simulation studies.

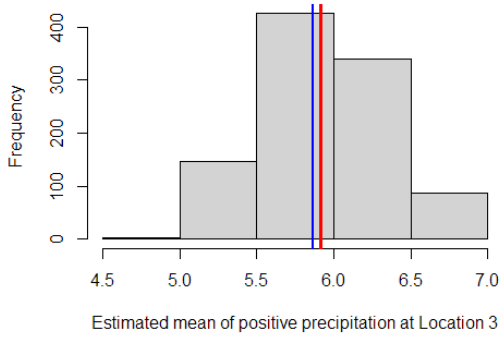

Figure 3.10: Mean positive rainfall (mm) in 1800 days of data simulated using estimated parameters from 1000 simulation studies.

Table 3.1: Proportion of dry days and mean positive rainfall at three locations estimated from data generated from the true and fitted models, along with the root mean square error (RMSE) between the two estimates.

Statistic	**Parameters Used and RMSE**	**Locations**		
		Loc. 1	Loc. 2	Loc. 3
Proportion of dry days	True Parameters	0.19	0.35	0.41
	Estimated Parameters	0.19	0.35	0.41
	Root Mean Square Error	0.01	0.01	0.01
Mean positive precipitation in mm	True Parameters	4.12	8.62	5.87
	Estimated Parameters	4.12	8.67	5.92
	Root Mean Square Error	0.25	0.49	0.34

days. The red lines correspond to the mean of the histogram data, and the blue lines are estimates of the true means. Table 3.1 lists the mean values from Figures 3.5 – 3.10 corresponding to the blue and red lines, as well as the RMSE based on the 1000 sets of estimates. We see that the monthly precipitation statistics from the true and the fitted models at each location are very close to each other. The RMSE values are also within 6% of the estimates based on the true model in all cases. We conclude that the model can replicate the marginal distributions of precipitation at least in cases where the number of locations is not too high.

3.9.4 SVB for single-site precipitation with Exponential mixtures

Our simulation study has the same setup as Section 3.9.2, including model parameters and prior specifications. The parameter space is given by $\Theta = (\alpha, \zeta, \Lambda)$ corresponding to the 3 variables, where α, ζ and Λ are matrices. We first compare the computational cost and accuracy of the old and our new approach for minibatch selection. Our data comprises 90 days of data for 20 years. The $D = 90$ days are divided into $C = 3$ months with each month being 30 days. We constructed minibatches as described in Algorithm 3 using samples of size 3. The difference in computational time between the new and old methods was below the measurement threshold, pointing to no additional computational cost.

Like in the previous study, 1000 datasets are generated from the true model and parameter estimation is carried out for each of them. Afterwards, we average the estimates obtained from each of them. The estimates from SVB are denoted by $\tilde{\Theta} = (\tilde{\alpha}, \tilde{\zeta}, \tilde{\Lambda})$. The posterior for the initial

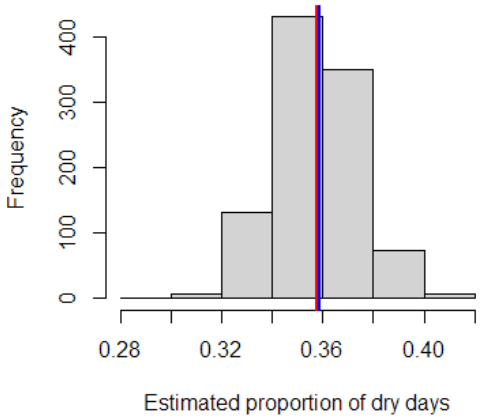

Figure 3.11: Proportion of dry days in 1800 days of data simulated using estimated parameters from 1000 simulation studies.

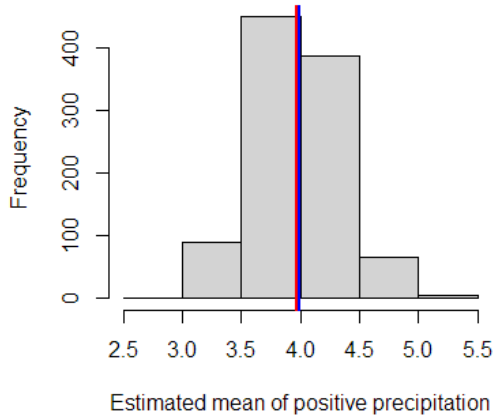

Figure 3.12: Mean positive rainfall (mm) in 1800 days of data simulated using estimated parameters from 1000 simulation studies.

state probability is $\hat{\pi}_1 = (0.34, 0.33, 0.33)$. Similarly,

$$\hat{A} = \begin{bmatrix} 0.40 & 0.30 & 0.30 \\ 0.32 & 0.33 & 0.35 \\ 0.32 & 0.33 & 0.35 \end{bmatrix}, \quad \hat{C} = \begin{bmatrix} 0.29 & 0.50 & 0.21 \\ 0.32 & 0.28 & 0.40 \\ 0.46 & 0.22 & 0.32 \end{bmatrix}, \quad \hat{\Lambda} = \begin{bmatrix} 0.08 & 0.93 \\ 0.57 & 4.68 \\ 0.95 & 7.98 \end{bmatrix}.$$

We see that the state parameters are not estimated well; $\hat{A}$ and $\hat{\pi}_1$ are actually very close to their prior values. In general, we found that using minibatches were detrimental to the accuracy of the parameter estimates for the state process. despite this, due to the structured mean field assumption, $\hat{C}$ and $\hat{\Lambda}$ are well estimated. They are at least as good as $\tilde{C}$ and $\tilde{\Lambda}$, their counterparts estimated by CAVI. In many cases, the SVB estimates for the emission distribution are better than the estimates from CAVI. Since different minibatches are chosen for every iteration of the SVB optimization, it is likely that the effect of extreme values and outliers in the observed data stay under control compared to CAVI which uses all data at every step.

For each of the 1000 simulations, we generated 1800 days of data based on that iteration's estimated parameters to verify whether the key statistical characteristics of daily precipitation arising from this model are captured. Figure 3.11 shows a histogram of the monthly proportion

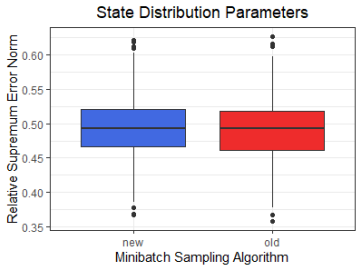

Figure 3.13: Distribution of relative supremum error norms when estimating A based on the old and new method.

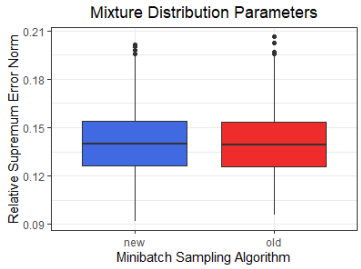

Figure 3.14: Distribution of relative supremum error norms when estimating C based on the old and new method.

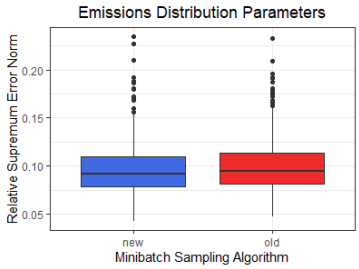

Figure 3.15: Distribution of relative supremum error norms when estimating Λ based on the old and new method.

of dry days based on 1000 estimates. The red line at 0.358 is the mean of the data presented in the histogram, and the blue line is an estimate of the true proportion 0.359. While the proportions are very close to each other, the 1000 estimates have a root mean square error (RMSE) of 0.01. Similarly, Figure 3.12 plots a histogram of mean rainfall for wet days. The red line at 3.97 mm is the mean of the histogram data, and the blue line at 3.99 mm is an estimate of the true mean. The 1000 estimates have an RMSE of 0.25 mm. These numbers are consistent with those obtained using $\tilde{\Theta}$, the CAVI estimate.

We also quantify the effect of using our modified minibatch sampling algorithm over regular minibatches without any subsampling in terms of the the errors in the estimates. For each of the 1000 studies, we compute 2 sets of $\hat{\Theta}$ estimates - one using the original minibatch sampling method, and one using our modified minibatch sampling procedure. For the matrices $\hat{A}, \hat{C},$ and $\hat{\Lambda}$, we define the relative supremum error norms as the metric for estimation error:

$$A_{err} = \frac{\| \hat{A} - A \|_\infty}{\| A \|_\infty}$$
$$C_{err} = \frac{\| \hat{C} - C \|_\infty}{\| C \|_\infty}$$
$$\Lambda_{err} = \frac{\| \hat{\Lambda} - \Lambda \|_\infty}{\| \Lambda \|_\infty}$$

Figures 3.13–3.15 show the distribution of the relative supremum error norms for the 1000 estimates, with the old minibatch algorithm represented in red and the new one in blue. While there is more error in $\hat{A}$ if we use our modified minibatch sampling procedure, we get better estimates for

the emission distribution parameters compared to the regular minibatch sampling approach. Our work seems to indicate that there is merit in devising a customized minibatch sampling method that takes advantage of the data structure to strike a balance between greater variability in our samples for SVB and the need to reproduce the estimates obtained from the entire data. SVB in general seems to result in sub-optimal estimates of the state process parameters. In real data problems, we recommend augmenting SVB with a few iterations of CAVI at the end to get better overall estimates at a reasonable computational cost.

3.10 Conclusions

The primary focus of this chapter has been to lay out a variational Bayesian approach to daily precipitation modeling that is flexible, scalable, and easy to implement. Historically, the statistical modeling of daily precipitation has relied on the Wilks method [Wilks, 1998] or on HMMs [Hughes and Guttorp, 1994], and made use of weather station data which is often available for long durations but for irregular locations. Our interest is in HMMs since they provide a rich model specification. We develop parameter estimation using variational Bayes for daily precipitation distributed as a semi-continuous distribution. Most literature on variational Bayes and HMMs tends to focus on Normal distribution emissions due to their simplicity. But the work by Rabiner [1989] for maximum likelihood estimation in HMMs provides an outline which we use to implement the VB-HMM for semi-continuous emissions.

A second focus in this chapter has been to develop and compare methods for different emission distributions. Gamma distributions or mixtures of Exponential distributions are the two most common emission distributions used for positive precipitation. We have also employed a mixture of two Gamma distributions in our previous work [Kroiz et al., 2020a,b, Majumder et al., 2020]. We consider VB-HMMs for both mixtures of Exponential and Gamma distributions. Gamma distributions prove challenging to work with due to the VB-HMMs' reliance on conjugacy. While the Gamma distribution does have a conjugate prior, the prior for the shape parameter does not have a closed form. If we have multi-site data, a total of $K \times L \times M$ numerical integrations are needed at each iteration of the VBEM algorithm to compute posterior estimates of the shape parameter,

where K is the number of states, L is the number of locations, and M is the number of Gamma mixture components for positive precipitation at each location. This adds a large computational burden, which is not feasible in practical cases. This is one of the weaknesses of the VB-HMM, since Gamma mixtures provides better model fit for remote sensing precipitation data in our experience [Kroiz et al., 2020a]. We also develop Gamma shape mixtures (GSM) as a candidate distribution for positive precipitation. As far as we can tell, the use of the GSM distribution for a VB-HMM is novel, as is its usage for precipitation modeling. We modify the original GSM specification of Venturini et al. [2008] which now requires fewer components to capture the range of positive precipitation. The GSM distribution has a richer specification than using exponential mixtures while having fewer parameters. We also develop empirical Bayes priors for the modified GSM distribution. Empirical Bayes priors are not common in variational literature, and would be especially helpful for studies over large areas where setting broad priors might not result in the best posterior estimates. However, they require further investigation beyond what is presented in this thesis to get estimates that are better than when we use a mixture of Exponential distributions.

In simulation studies, the VB-HMM for precipitation can estimate the true parameters using CAVI under general prior specifications. In particular, we found that as long as we establish an order in the amount of rainfall each state will receive and set reasonable priors for the mixture components and exponential rates, our corresponding posteriors are quite close to the true values. The posterior is farthest from the true values for the initial probability distribution and some entries of the transition probability matrix. For the initial distribution, the variational update depends only on one data point, and unless the Dirichlet prior has a very low concentration or is asymmetric, it will dominate in the posterior. Previous studies often assumed the initial distribution to be known when using variational Bayes, and that might be a practical solution for this case as well.

The real strength of the VB-HMM is in its stochastic implementation. SVB becomes critical for remote sensing data which is available on a much denser grid compared to weather station data. SVB uses only a subset of the data known as a minibatch at each iteration of the VBEM, which speeds up both the VBE and VBM steps. While both CAVI and the Baum-Welch algorithm can be parallelized for computational efficiency, the Baum-Welch algorithm does not have an obvious counterpart to SVB. We employ a modified minibatch sampling algorithm for SVB, which

adds more variability to the minibatches and could prove valuable if the SVB is run for a large number of iterations for large datasets. In simulation studies, the modified minibatch sampling algorithm provides slight improvements in the emission distribution parameter estimates. However, we observe that the parameter estimates from SVB are not as good as those coming from CAVI, especially the parameters for the state process. We recommend running a few iterations of CAVI at the end of the SVB to balance the computational efficiency with the quality of parameter estimates.

Chapter 4

Parameterizing Correlation in HMMs using Gaussian Copulas

Hidden Markov models were originally developed for univariate discrete data and then extended to accommodate multivariate data which could be discrete, continuous, or a mixture. All HMMs considered in this thesis so far have had a univariate state process in the form of a single Markov chain. Under this setup, both the current state and information regarding the previous states are propagated through this Markov chain even when we have multiple data streams. While this assumption is convenient for parameter estimation and therefore fairly ubiquitous, it is not a necessary assumption. For many examples of interest, it is reasonable to consider that multiple streams of data are generated by not one, but multiple interacting state processes. For example, large geographical domains where different areas can have local weather patterns, or the financial market where a single process for the 'state' of the market is not enough to capture the dynamics of hundreds of stocks. In such cases, an HMM with a multivariate state process can be considered as a natural extension of a univariate state process HMM. The individual component Markov chains of the multivariate state process are not independent of each other and admit a correlation structure. Similarly, the components of the emission process are also correlated with each other. In this chapter, we employ copulas to develop explicit correlation parameterizations for both a multivariate state process and a multivariate emission process.

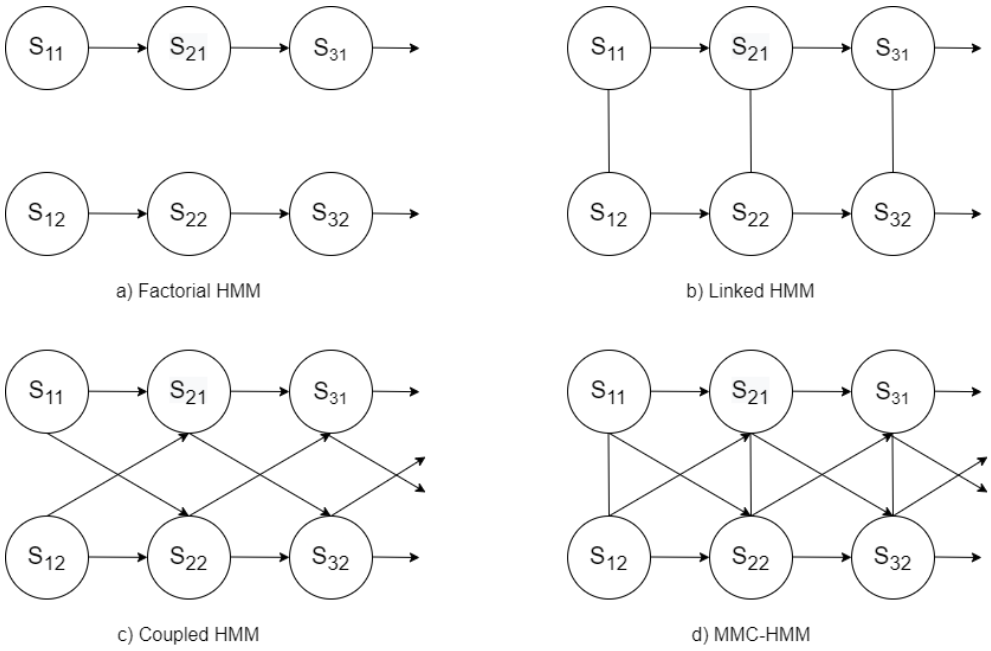

Figure 4.1: Graphical representation of 3 time slices of the different ways to specify an HMM with 2 state processes. The circles denote state nodes; emission nodes at each time point is assumed to depend on all state nodes at that time point, and are omitted for clarity.

4.1 A Multivariate State Process for HMMs

HMMs with multiple state processes can be specified in a few different ways depending on how the state processes evolve, and how each state process acts upon the emission processes. These two things taken together decide how the state processes share and propagate information about the system. The first question about how the state processes evolve relative to each other has been widely studied in the literature, and Figure 4.1 contains graphical representations of four ways HMMs with multiple state processes can evolve over time. They are all represented using 3 time slices for an HMM with 2 state processes. We have omitted the emission process nodes for ease of representation, and it is assumed that the emission at each time point depends on all state processes at that time point. Directed edges between nodes are denoted by arrows and represent a conditional distribution, e.g., $p(s_{21}|s_{11})$ in 4.1(b). On the other hand, undirected edges between nodes are denoted by lines and represent a joint distribution between variables, e.g., $p(s_{11}, s_{12})$ in 4.1(b). The simplest form is a factorial HMM (FHMM) shown in Figure 4.1(a), where the state processes are independent and are coupled only through the emission process. FHMMs with C

chains, each of which have K states, describe a state process with CK states. This formulation is tractable in space, but the CK possible combinations of states are not guaranteed to occur in the observed data. This leads to the lack of sufficient data even for relatively modest state spaces, and exact solutions tend to be infeasible. Ghahramani and Jordan [1996] have developed a structured mean field approximation for FHMMs with exact solutions that can be tractably estimated. In their approach, the entire graph is partitioned up into independent sub-graphs using a mean field assumption. Parameters for each subgraph are estimated individually in a way that minimizes an energy function defined over the entire graph. However, the mean field approximation proves to be poor if there are strong and varied interactions across the subgraphs. Figure 4.1(b) depicts state processes that evolve in lockstep and do not have temporal influences on each other. In graphical model terminology, this implies that nodes of different state processes are connected by an edge if and only if the nodes are at the same time point. The resulting HMM is known as a linked HMM (LHMM). When multiple state processes influence each other over time, Brand [1997] proposed the coupled HMM (CHMM) which considers pairwise cross-transition probabilities between state processes. This is depicted in Figure 4.1(c). As with the other methods discussed here, exact algorithms for parameter estimation are not feasible, and all methods rely on approximations of some form or the other. Finally, Ching et al. [2013] proposed a multivariate Markov chain HMM (MMC-HMM), depicted in Figure 4.1(d). They incorporated cross-transition probabilities and the result is a combination of LHMMs and CHMMs that has the highest complexity among the four approaches described here. Parameter estimation for their model involves solving systems of linear equations and tends to be computationally challenging. For meteorological processes like precipitation, an LHMM structure is a reasonable assumption and is what we base our own model on going forward.

A common feature of HMMs with multiple state processes is that each state process affects the entire L dimensional emission process. This has remained an underlying assumption in all literature we have come across; irrespective of the dependence structure between the state processes, they affect each coordinate of the emission process. Figure 4.2 is a graphical representation of an LHMM with 2 state processes, with each state process affecting the entire emission process. In this thesis, we propose a dependence structure for LHMMs which takes advantage of the geo-

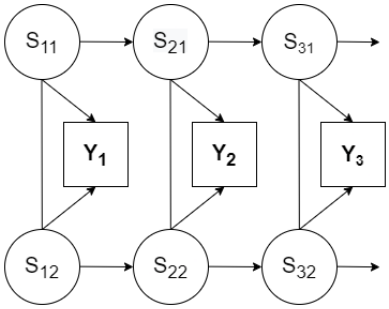

Figure 4.2: Graphical representation of 3 time slices of an LHMM with 2 state processes, where each Markov chain affects the entire emission process.

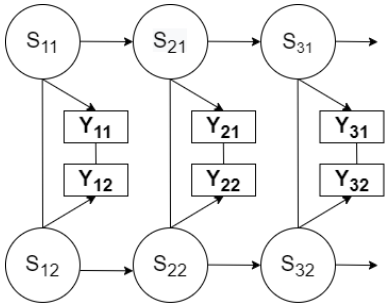

Figure 4.3: Graphical representation of 3 time slices of an LHMM with 2 state processes, where each Markov chain affects only a partition of the emission process.

statistical context of remote sensing data for meteorological processes. Since the L-dimensional emission process is on a gridded spatial map, we partition it into C clusters, each representing a local weather regime. In such a scenario, each cluster's precipitation will be driven by a different univariate state process. This is shown in Figure 4.3, where each state process affects only a subset of the coordinates of the emission process. The subsets form partitions of the emission process and can be identified in practice using a clustering algorithm. Under these assumptions, we define a clustered LHMM corresponding to the graphical structure of Figure 4.3, where a C-dimensional multivariate state process has a correlation structure specified through a copula.

4.1.1 Gaussian copulas for the state process of a clustered LHMM

Let $\mathbf{y}_{1:T} = \{\mathbf{y}_1, \ldots, \mathbf{y}_T\}$ be the precipitation time series of length T as before. Let the L locations be partitioned into D clusters with their own local weather patterns. Each cluster d consists of l_d locations with $d = 1, \ldots, D$, and $\sum_{d=1}^{D} l_d = L$. The precipitation at time point t, $\mathbf{y}_t$, can thus be expressed as $\mathbf{y}_t = (\mathbf{y}_{tl_1}, \ldots, \mathbf{y}_{tl_D})$. The data is generated by a set of underlying hidden states $\mathbf{s}_{1:T} = \{\mathbf{s}_1, \ldots, \mathbf{s}_t, \ldots, \mathbf{s}_T\}$, with $\mathbf{s}'_t = (s_t^{(1)}, \ldots, s_t^{(D)})$ where each state $s_t^{(d)} \in \mathcal{K} = \{1, \ldots, K\}$. r_{tjl_dm} is similarly defined for $m = 0, \ldots, M$. The complete data likelihood is written as

$$p(y, s, r|\Theta) = p(s|\Theta) \cdot \prod_{d=1}^{D} p(y_{l_d}, r_{l_d}|s^{(d)}, \Theta_d). \tag{4.1}$$

The likelihood of this clustered LHMM for L locations can thus be factorized into D independent HMM likelihoods parameterized by Θ_d, the parameter set for cluster $d \in \mathcal{D} = \{1, \ldots, D\}$. Let F be the D-dimensional joint CDF of $s = (s^{(1)}, \ldots, s^{(D)})$. Let $F_1, \ldots, F_D$ be the marginal CDFs of $s^{(1)}, \ldots, s^{(D)}$ respectively. We define a Gaussian copula for the state processes as:

$$\begin{aligned} F(s^{(1)}, \ldots, s^{(d)}) &= \mathcal{C}\big(F_1(s^{(1)}; \Theta_1), \ldots, F_D(s^{(D)}; \Theta_D)\big) \\ &= \Phi_D\big(\Phi^{-1}(u^{(1)}), \ldots, \Phi^{-1}(u^{(D)}); \Sigma\big) \\ &= \Phi_D\big(z^{(1)}, \ldots, z^{(D)}; \Sigma\big), \end{aligned} \tag{4.2}$$

where $u^{(1)}, \ldots, u^{(D)}$ are $Uniform(0, 1)$ variates, and $z^{(1)}, \ldots, z^{(D)}$ are standard Normal variates. Φ_D is a D-dimensional multivariate Normal CDF, while Φ^{-1} is the inverse CDF of a univariate standard Normal distribution. The copula augmented model has a clustered LHMM structure similar to Figure 4.3. We choose a Gaussian copula due to its simplicity of formulation. The copula is parameterized only by a correlation matrix Σ, and its individual components follow standard Normal distributions.

The elements of Σ are Pearson correlation coefficients, and Pearson correlation is not preserved under monotone transforms. Further, since $s^{(d)}$ is a discrete variable and $z^{(d)}$ is a continuous variable for all $d = 1, \ldots, D$, the Pearson correlation between two state processes is not the same as the corresponding element of Σ. Further, since the state processes are ordinal data as per our definition, Pearson correlation is not the best measure of association between two Markov chains anyway. Spearman's ρ and Kendall's τ are both better options for ordinal data, and are invariant under monotone transformations. Kruskal [1958] provided a relationship between the Pearson correlation ρ and the Spearman correlation ρ^* for bivariate Normal variables (X_1, X_2):

$$\rho = 2 \sin\left[\pi \frac{\rho^*}{6}\right]. \tag{4.3}$$

For two state processes $s^{(d_1)}$ and $s^{(d_2)}$, (4.3) gives us a way to connect their Spearman correlation $\rho^*(s^{(d_1)}, s^{(d_2)})$ to the corresponding Pearson correlation for the Gaussian copula, $\rho(z^{(d_1)}, z^{(d_2)})$, by way of $\rho^*(z^{(d_1)}, z^{(d_2)})$. Once all marginal parameters associated with the likelihood in (4.1) are

estimated, the Viterbi algorithm provides us the most likely sequence of states for each of the D components of the clustered LHMM. These state sequences are used to construct the copula. Note that marginal parameter estimation can be carried out either using variational Bayes or the Baum-Welch algorithm. The steps for constructing the Gaussian copula afterwards remain unchanged.

A final assumption is required as a consequence of using pairwise Spearman correlations as the measure of association between Markov chains. While Σ in (4.2) is a $D \times D$ matrix of Pearson correlations whose elements can be obtained by transforming Spearman correlations, a $D \times D$ matrix of Spearman correlations cannot be interpreted in a manner congruent with Σ. Furthermore, element-wise transformations do not take into account any higher order correlations except pairwise correlations. With that in mind, we simplify the form of the Gaussian copula and rewrite (4.2) as:

$$\Phi_D\big(z^{(1)}, \ldots, z^{(D)}; \Sigma\big) \approx \prod_{d_1=1}^{D} \prod_{d_2=1}^{D} \Phi_2(z^{(d_1)}, z^{(d_2)}; \rho_{d_1 d_2}) \tag{4.4}$$

where $\rho_{d_1 d_2} = \rho(z^{(d_1)}, z^{(d_2)})$ are the Pearson correlations which can be computed for each $(d_1, d_2) \in \mathcal{D}^2$ pair using (4.3). This formulation can be interpreted in a manner similar to a pairwise simplified regular vine (R-vine) copula [Brechmann et al., 2012], with all pair-copula terms involving a conditioning set replaced by bivariate Gaussian copulas. We refer to this as the pair-copula approximation. The copula density associated with (4.4) can also be interpreted as a composite likelihood [Varin et al., 2011]. In practice, this will allow us to estimate the $\rho_{d_1 d_2}$ using the right hand side of (4.4), but simulate from the copula using the left hand side of (4.4), as long as we can ensure that Σ is a positive-definite matrix.

4.1.2 Estimation of the copula parameters

The construction of the copula for state processes is preceded by the estimation of the marginal distribution parameters of each state process. The marginal distributions and the copula together describe a multivariate Markov chain (MMC) which generates correlated states for the clustered LHMM. For a *D*-dimensional MMC, each chain provides the daily states for a partition of the emission process, with $s^{(d)} = (s_1^{(d)}, \ldots, s_T^{(d)})$. The state space can thus be written as a

$D \times T$ matrix. The inherent challenge of estimating Φ_D and F using $F_1, \ldots,$ and F_D arises from the discrete-to-continuous transformation that is required in the process. Since each element of $s = (s^{(1)}, \ldots, s^{(D)})$ has a finite state space by definition, $F_1, \ldots, F_D$ are all step functions with a small number of steps. Both the inversion method [Nelsen, 2010] and the inference functions for margins method [Joe and Xu, 1996] for estimating Σ require evaluating the terms $F_d(s^{(d)}; \Theta_d)$ in (4.2) for all $d \in \mathcal{D}$. However, evaluating the CDF of Markov chains is not straightforward. We propose an alternate approach by modifying and extending the Wilks method [Wilks, 1998], originally used to model multi-site daily precipitation occurrence using a 2-state Markov chain of dry and wet days. Note that the Wilks method was developed specifically for fully-observed 2-state Markov chains whose states are assumed to follow Binomial distributions. An expression for the correlation between a pair of Markov chains can be derived under this assumption, denoted by $\xi(k, l)$ for two arbitrary Markov chains s_k and s_l. Further, for (X_k, X_l) which follow a bivariate normal distribution with Pearson correlation $\omega(k, l)$, Wilks transformed $X_k \to s_k$ and $X_l \to s_l$ and found empirically that there is a monotonic relationship between $\omega(k, l)$ and $\xi(k, l)$, and that $\xi(k, l) < \omega(k, l)$. This empirical relationship has been confirmed for 2-state Markov chains under the aforementioned assumptions with a variety of different parameter values by several authors; see Mhanna and Bauwens [2012] for a review. Note that the relationship between $\xi(k, l)$ and $\omega(k, l)$ does not have a closed form expression. This, combined with the fact that $\xi(k, l) < \omega(k, l)$, means that a line search is needed to find the value of $\omega(k, l)$ which can generate $\xi^*(k, l)$, the observed value of $\xi(k, l)$ from the available data. The relationship $\xi(k, l) < \omega(k, l)$ means that $\xi(k, l)$ is not guaranteed to reach ± 1. The maximum values that can be attained depend on the discretizing function, i.e., on the parameters of the Markov process. It can be intuitively interpreted as the association between two continuous variables being captured better than their association after they have both been discretized and transformed into Markov chains.

For our clustered LHMM, let $\{r_{d_1 d_2}\}$ denote the observed Spearman correlations between the states of each $(d_1, d_2) \in \mathcal{D}^2$ pair of the D component HMMs. The states for each HMM are obtained using the Viterbi algorithm once Θ_d has been estimated. Given the $D \times T$ matrix of states, the initial distributions π_{d1}, and the transition matrix A_d for $s^{(d)}$, we want to construct a Gaussian copula that can generate an MMC whose pairwise Spearman correlations $r^*_{d_1 d_2}$ coincide

with the observed Spearman correlations $\{r_{d_1d_2}\}$. Let $\hat{\rho}^*_{d_1d_2}$ be the estimate of the population Spearman correlation between $(s^{(d_1)}, s^{(d_2)})$, and let $\hat{\rho}_{d_1d_2}$ be the corresponding estimate of the Pearson correlation using (4.3); $\hat{\rho}_{d_1d_2}$ is used in the Gaussian copula to generate correlated states. If all distributions were continuous, we could expect $r_{d_1d_2}$ to coincide with $\hat{\rho}^*_{d_1d_2}$. However, the attenuation that happens when we transform the continuous $z^{(d)}$ into the discrete $s^{(d)}$ means that in practice, $r_{d_1d_2}$ is less than both $\hat{\rho}_{d_1d_2}$ and $\hat{\rho}^*_{d_1d_2}$. Since the relationship between $\hat{\rho}^*_{d_1d_2}$ and $r_{d_1d_2}$ cannot be expressed in closed form, we resort to a simulation approach to compute our estimate $\hat{\rho}^*_{d_1d_2}$. We initialize $\hat{\rho}^*_{d_1d_2}$ with $r_{d_1d_2}$ for each pair of Markov chains $(s^{(d_1)}, s^{(d_2)})$ and simulate an MMC from the Gaussian copula. We compute the pairwise Spearman correlations between the Markov chains in the MMC and denote it by $r^*_{d_1d_2}$. If $r^*_{d_1d_2} < r_{d_1d_2}$, we increment $\hat{\rho}^*_{d_1d_2}$ by a step size τ and repeat the process. We stop when $0 < r^*_{d_1d_2} - r_{d_1d_2} \leq \epsilon$, for some predefined tolerance ϵ. The entire procedure is formalized in the algorithm below.

Algorithm 4 Algorithm to construct a Gaussian copula for a clustered LHMM.

Cluster $y_{1:L}$ into D clusters corresponding to local weather regimes
Estimate marginal HMM parameters π_{1d} and A_d for clusters $d = 1, \ldots, D$
Estimate $s_1^{(d)}, \ldots s_T^{(d)}$ using the Viterbi algorithm for clusters $d = 1, \ldots, D$
Set step size τ and tolerance ϵ
for *clusters* $(d_1, d_2) \in \mathcal{D}^2$ **do**
    Compute the observed Spearman correlation $r_{d_1d_2}$
    Initialize $\hat{\rho}^*_{d_1d_2} = r_{d_1d_2}$
    Initialize $r^*_{d_1d_2} = 0$
    while $|r_{d_1d_2} - r^*_{d_1d_2}| > \epsilon$, **do**
        Increment $\hat{\rho}^*_{d_1d_2}$ by τ
        Compute Pearson correlation $\hat{\rho}_{d_1d_2}$ from $\hat{\rho}^*_{d_1d_2}$ using (4.3)
        Generate correlated bivariate sequence from $N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \hat{\rho}_{d_1d_2}\\ \hat{\rho}_{d_1d_2} & 1\end{pmatrix}\right)$
        Use $\pi_{1d_1}, \pi_{1d_2}, A_{d_1}, A_{d_2}$ and the correlated sequences to generate synthetic states
        Calculate $r^*_{d_1d_2}$ as the Spearman correlation of the synthetic state sample
    end
end
Construct correlation matrix $\hat{\Sigma}$ whose diagonals are 1 and off-diagonals are $\hat{\rho}_{d_1d_2}$
if $\hat{\Sigma}$ *is not positive definite* **then**
    Eigendecompose $\hat{\Sigma}$ as $\hat{\Sigma} = VRV^T$
    Replace negative and zero eigenvalues in R with 10^{-6}; call the new matrix R^*
    Recalculate $\hat{\Sigma} = VR^*V^T$
end

Since the entries of $\hat{\Sigma}$ are constructed independently, the resultant matrix is not guaranteed to be positive definite. The final steps of our algorithm ensures the positive-definiteness of $\hat{\Sigma}$.

Simulating a D-dimensional MMC from the Gaussian copula requires us to transform a D-dimensional Normal dataset into an MMC. We follow the algorithm described in Serfozo [2009] for constructing a univariate Markov chain whose parameters are known using a sequence of $Uniform(0,1)$ observations. Ji [2019] expanded on this for multivariate Markov chains. Let $x_1, \dots x_T$ be T independent observations generated from the left hand side expression of (4.4), where $x_t = (x_t^{(1)}, \dots, x_t^{(D)})^T$. Let

$$u_t^{(d)} = \Phi(x_t^{(d)}),$$

for $t = 1, \dots, T$ and $d = 1, \dots, D$. Since $(x_t^{(1)}, \dots, x_t^{(D)})^T$ are correlated, $(u_t^{(1)}, \dots, u_t^{(D)})^T$ are also correlated. Consider a Markov chain with initial distribution $\pi_1 = (\pi_{11}, \dots, \pi_{1K})$ and a $K \times K$ transition matrix $A = ((a_{jk}))$. We can transform a sequence of independent uniform random variables $u_1, \dots, u_T$ into a Markov chain using its marginal parameters. Let $h(u)$ and $f(j,u)$ be functions transforming continuous values into $\mathcal{K} = \{1, \dots, K\}$, given by

$$h(u) = j \text{ if } u \in I_j \text{ for some } j \in \mathcal{K} \tag{4.5}$$

where $I_1 = [0, \pi_{11})$ and $I_j = \left[\sum_{l=1}^{j-1} \pi_{1l}, \sum_{l=1}^{j} \pi_{1l}\right)$ for $1 < j \leq K$, and

$$f(i,u) = j \text{ if } u \in I_{ij} \text{ for some } j \in \mathcal{K} \tag{4.6}$$

where $I_{i1} = [0, a_{i1})$, and $I_{ij} = \left[\sum_{l=1}^{j-1} a_{il}, \sum_{l=1}^{j} a_{il}\right)$ for $1 < j \leq K$. Let $u_t = u_t^{(d)}$ for $t = 1, \dots, T$. Similarly, let $A = A_d$ and $\pi_1 = \pi_{1d}$ be the marginal parameters of the d^{th} Markov chain. Using these in (4.5)–(4.6), we denote $h(u_1^{(d)})$ as $s_1^{(d)}$ and $f(s_{t-1}^{(d)}, u_t^{(d)})$ as $s_t^{(d)}$ for $t > 1$. Then $\{s_1^{(d)}, \dots, s_T^{(d)}\}$ is a Markov chain with initial distribution π_{1d} and transition matrix A_d. This transformation can be applied to $u^{(1)}, \dots, u^{(D)}$, and the resulting $D \times T$ matrix would be our desired MMC that can now be used to generate a linked HMM.

4.2 A Gaussian Copula for Semi-Continuous Emissions

In regard to a multivariate emission process, most studies concern themselves with a multivariate Normal distribution, or similar distributions which have multivariate specifications. However, most univariate exponential family distributions either do not have a corresponding multivariate specification, or have multivariate forms which are challenging to work with from a practical perspective. This is also true for our own motivating example of semi-continuous mixtures with Exponential distributions which do not admit a natural multivariate extension. For low dimensional emission processes which share a common state process, specifying a univariate distribution for each coordinate of the emission process can be sufficient to capture the dependencies between them. From the graphical depictions of HMMs in Figures 2.1 and 4.3, we see that even if two univariate chains are conditionally independent conditioned on the state, unconditionally they are dependent through the state process. Most studies in the precipitation literature that we have come across [Hughes and Guttorp, 1994, Robertson et al., 2006, Greene et al., 2008] use data observed by a small number of irregularly located weather stations. The largest study we encountered consists of data from 52 stations [Holsclaw et al., 2016]. HMMs adequately capture spatial correlations in these situations through the shared daily state, but that is not necessarily the case with gridded remote sensing data over large areas. For example, Robertson et al. [2004] considered daily precipitation from a network of 10 weather stations in NE Brazil from the February–April wet season between 1975–2002 for their study, and reported the mean of observed Pearson correlation coefficients between stations as 0.248. However, IMERG data for the Potomac basin in Eastern USA comprises 387 grid points at $0.1^\circ \times 0.1^\circ$ resolution. If we use the same model for the 2001–2018 IMERG data from July to September for the Potomac river basin, the mean and maximum of observed Pearson correlation between grid points are 0.642 and 0.986 respectively, far higher than most weather station based studies [Kroiz et al., 2020a]. The Chesapeake Bay watershed is even larger with 1927 IMERG grid points. To employ HMMs for these high dimensional problems, there is a need to explicitly specify a correlation structure. One way to achieve this is by constructing Gaussian copulas for each state's emission distribution. We develop a Gaussian copula for the precipitation model in Section 3.3, where the joint distribution for the emissions

conditional on the state is a product of their marginals.

4.2.1 Parameter estimation for the Gaussian copula

The HMM described in Section 3.3 is assumed to have a univariate state process and a multivariate emission process, corresponding to multi-site daily precipitation data. We have used variational Bayes to estimate the model parameters. We want to construct a copula for the emission process $y_{1:T}$, where $y_t^{'} = (y_{t1}, \ldots, y_{tL})$. It is obvious that we would need to construct a separate copula corresponding to each state, since the emission distribution parameters are state specific. The Viterbi Algorithm can be used to obtain the most likely sequence of states to have generated the observations, which identifies the observations corresponding to each of the states. Since the interpretations for the states do not differ beyond the rainfall intensity they are associated with, the construction of the copula is carried out identically for each state. The marginal parameters of the HMM are assumed to be known, and we plug their estimates into the likelihood. Then the distribution of an observation from state j for location l is given by

$$p(y_{tl}|\lambda_{jl}, c_{jl}, s_t = j) = c_{jl0} + \sum_{m=1}^{M} c_{jlm}\lambda_{jlm} \exp\{-\lambda_{jlm} y_{tl}\}.$$

Its CDF is denoted by $F_{jl}(y)$ and has the form

$$F_{jl}(y) = c_{jl0} + \sum_{m=1}^{M} \mathbb{I}\{y > 0\} c_{jlm}(1 - \exp\{\lambda_{jlm} y\}).$$

Further, we denote by $F^{(j)}$ the L-dimensional joint CDF of $y_{1:T}$ for state j. Then, we can represent the joint distribution by a Gaussian copula as follows:

$$\begin{aligned} F^{(j)}(y_1, \ldots, y_L|\Theta_j) &= \mathcal{C}\big(F_{j1}(y_1; \Theta_{j1}), \ldots, F_{jL}(y_L; \Theta_{jL})\big) \\ &= \Phi_L\big(\Phi^{-1}(u^{(1)}), \ldots, \Phi^{-1}(u^{(L)}); \Sigma_j\big) \\ &= \Phi_L\big(z^{(1)}, \ldots, z^{(L)}; \Sigma_j\big), \end{aligned} \tag{4.7}$$

where $u^{(1)}, \ldots, u^{(L)}$ are $Uniform(0,1)$ variates and $z^{(1)}, \ldots, z^{(L)}$ are $\mathcal{N}(0,1)$ variates. For each state $j = 1, \ldots, K$, Σ_j is an $L \times L$ correlation matrix whose off-diagonals represent the Pearson correlation coefficients of the copula for positive precipitation at 2 different locations when the data arises from the j^{th} state. Φ^{-1} is the inverse CDF of a univariate $\mathcal{N}(0,1)$ variable, and Φ_L is the CDF of an L-variate Normal distribution with a mean vector $\mathbf{0}$ and correlation matrix Σ_j. Let $(y_1, \ldots, y_L)$ represent the L chains of the emission process, with $y_l = (y_{1l}, \ldots, y_{Tl})$.

Since the univariate CDFs can be evaluated for the emissions at every value of y_{tl} where $y_{tl} > 0$, we can directly estimate Σ_j using (4.7). It can be shown that this is equivalent to the inference functions for margins (IFM) estimate of Joe and Xu [1996]. We can also use the pair copula approximation as stated in (4.4). For the pair-copula approximation, we resort to the relation between the Spearman correlation and Pearson correlation from (4.3), which bypasses the need to evaluate the CDF of the mixture marginals. This is formalized in the algorithm below.

Algorithm 5 Gaussian copula for positive emissions

for *state j in 1:K* **do**
  Use the Viterbi Algorithm to subset the days corresponding to state *j*
  for *locations* l_1 *and* l_2 *in* $1, \ldots, L, l_1 \neq l_2$ **do**
    Classify observations into mixture components they arise from
    Subset the observations for days where it rains at both locations and the data arises from the same mixture component; denote them as y_{l_1} and y_{l_2}
    Compute estimates of the Spearman correlations, $\hat{\rho}_j^*(y_{l_1}, y_{l_2})$
    Estimate Pearson correlations for the copula, $\hat{\rho}_j(y_{l_1}, y_{l_2})$, using (4.3)
  end
  Set diagonal elements of $\hat{\Sigma}_j$ to 1
  if $\hat{\Sigma}_j$ *is not positive definite* **then**
    Eigendecompose $\hat{\Sigma}_j$ as $\hat{\Sigma}_j = VRV^T$
    Replace negative and zero eigenvalues in R with 10^{-6}; call the new matrix R^*
    Recalculate $\hat{\Sigma}_j = VR^*V^T$
  end
end

The correlation matrix $\hat{\Sigma}_j$ estimated in Algorithm 5 is not guaranteed to be positive definite; the final steps ensure positive-definiteness. The pair-copula approximation in this case does not suffer from the attenuation issues as the copula for the states of a clustered LHMM. Data can be generated from this model using the following steps:

Algorithm 6 Generating multi-site daily precipitation using a Gaussian copula

Generate daily states $s_{1:T}$ either from an HMM or from an LHMM
for *day t in 1:T* **do**
  if $s_t = j$ **then**
    Generate $r_{tjl} \sim Cat(c_{jl})$ independently for each location $l = 1, \ldots, L$
    Generate $(z^{(1)}, \ldots, z^{(L)}) \sim \mathcal{N}_L(\mathbf{0}, \hat{\Sigma}_j)$
    for *location l in 1:L* **do**
      if $r_{tjlm} = 1$ *for* $m > 0$ **then**
        $y_{tl} \sim Exp(\lambda_{jlm})$
      else
        if $r_{tjl0} = 1$ **then**
          $y_{tl} \leftarrow 0$
        end
      end
    end
  end
end

Note that the correlation structure does not take into account which mixture distribution the data comes from. When simulating data using Algorithm 6, the mixture component assignments r_{tjl} at each location l are generated from independent Categorical distributions parameterized by c_{jl}. Only the positive precipitation amounts are correlated. In principle, we can construct another Gaussian copula for correlated mixture component assignments. However, it will inflate the parameter space and add a significant computational burden. We also make a final note regarding extending this copula for the emissions of a clustered LHMM. In practice, the Gaussian copula for the emissions will get simplified if the underlying state process is a clustered LHMM. For each cluster, we can estimate the correlations of locations belonging to the cluster either by the IFM or pair-copula approximation approach as described in this section. Locations belonging to different clusters can be assumed to have 0 correlation in the emission copula, which would still make them correlated through their respective state processes. We can also set it to a constant value representative of the correlation between their respective clusters instead of being the correlation between the individual locations. A clustered LHMM with a Gaussian copula for both states and the emissions can thus be described using fewer copula parameters compared to an HMM with a univariate state process.

4.2.2 Case study: the Potomac river basin

We present here part of a case study involving Gaussian copulas for remote sensing precipitation data over the Potomac river basin in Eastern USA. This study is originally a part of Kroiz et al. [2020a], and serves to answer 2 important questions:

1. Why a Gaussian copula for emissions is necessary for gridded remote sensing data,
2. Why we choose to subset data by the mixture components they arise from.

We fitted a 4-state HMM with 2 gamma distributions to daily IMERG data over the Potomac river basin for the wet season months of July to September, 2001–2018. The HMM is considered to have a univariate state process. The data over the basin is distributed as 387 IMERG grids. One of the goals of this study was to ascertain how well the HMM replicates spatial associations, and if adding a Gaussian copula (HMM-GC) to the emission process would provide better performance. Marginal parameter estimation was carried out using the Baum-Welch algorithm. The software used for the majority of the hidden Markov model computations is the MVNHMM toolbox [Kirshner, 2005] developed by Sergey Kirshner and Padhraic Smyth and available at `http://www.sergeykirshner.com/software/mvnhmm`. We then fit a Gaussian copula for each state's data. This model with Gaussian copulas is referred to in this section as an HMM-GC, whereas the marginal model is referred to as an HMM. A crucial difference in the copula construction for this study from Algorithm 5 lies in the treatment of the mixture components. In Kroiz et al. [2020b], we considered data from all days when it rained at both locations for computing the pairwise Spearman correlations for a particular state. However, in Algorithm 5, we assign an additional filter and only consider data points which arise from the same mixture components.

Figure 4.4 shows box plots of the daily average precipitation amount for the HMM, HMM-GC, and the IMERG data. The low median and interquartile range of HMM and HMM-GC compared to IMERG suggest that both models struggle with capturing spatial correlation to different degrees. We see that the classical HMM for precipitation tends to severely underestimate the correlations between precipitation amounts since it does not have an explicit parameterization for the

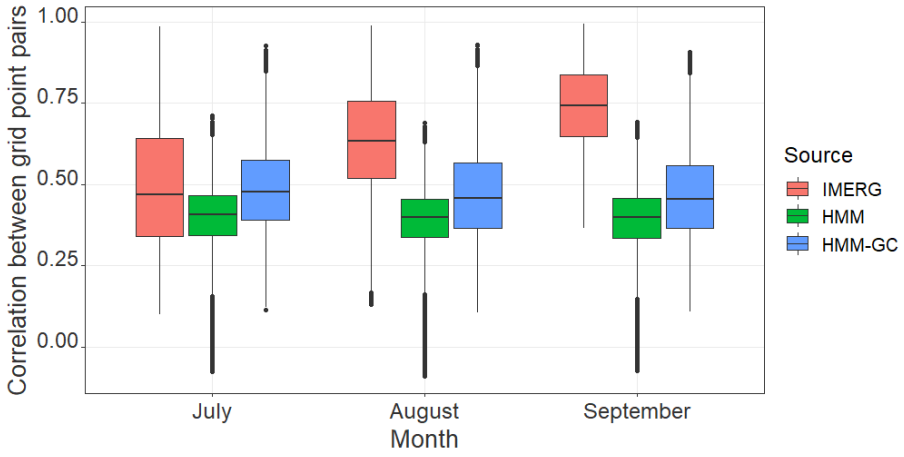

Figure 4.4: Pairwise spatial correlation between grid points for historical IMERG data (2001–2018) compared with synthetic data from HMM and HMM-GC models based on 18 years of data

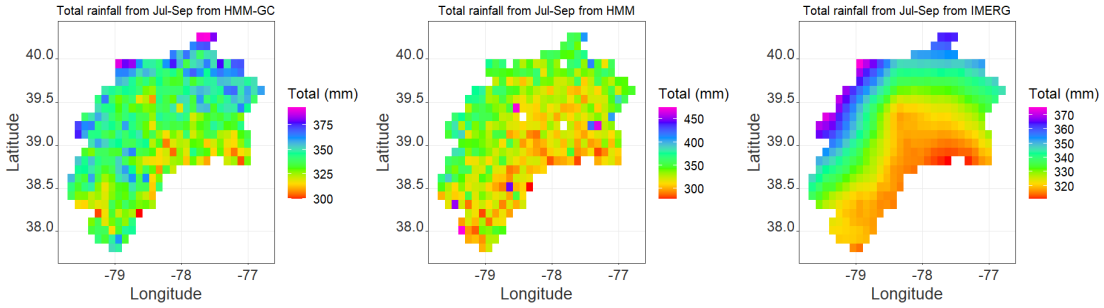

Figure 4.5: Spatial patterns in the total rainfall over the basin from July to September averaged over 18 years of data

correlations. The HMM-GC performs better but due to computing unconditional correlations instead of conditioning them on the mixture component assignments, they are also fairly low. There are also negative values which are artifacts of trying to estimate zero correlation in the simulated data. The HMM-GC does a significantly better job of estimating the spatial correlations.

Figure 4.5 shows the total basin rainfall for the wet season averaged over 18 years of data at the 387 grid points. A visual inspection suggests that the HMM-GC does a better job of simulating spatial patterns within the basin than the classical HMM. However, it is nowhere as smooth as the historical data. We found another issue regarding extreme values in the regular HMM; 5 of the simulated values were greater than 500 mm, with the largest being over 1500 mm. They have been left out of the plot in the interest of legibility, and are denoted by 5 white grid points within

the plot. These values are far higher than the historical data, and probably caused due to the underestimated correlation of the HMM. The HMM-GC is not affected by this problem.

Figures 4.6 and 4.7 plot the distributions of daily maximum and mean basin precipitation for IMERG, HMM and HMM-GC data. We notice in Figure 4.6 that the classical HMM tends to overestimate the daily maximum precipitation, as shown through the long upper tail. However, the short upper tail for the HMM in Figure 4.7 shows that the HMM also underestimates the daily mean precipitation. This can be attributed to the lack of spatial correlation, where some locations simulated very high values, but since they were being generated independently of the other locations, there was no way to simulate basin-wide consistent behaviour. This has largely been mitigated by the HMM-GC approach based on the relative similarities between the HMM-GC and IMERG plots in both Figures 4.6 and 4.7. The daily mean is, however, still slightly underestimated, suggesting the presence of local extreme weather events within the basin influenced by factors not currently captured by our states.

4.3 Numerical Studies on Simulated Data

Our simulation studies in this chapter are based around a data generating process (DGP) which consists of 3 state processes distributed as first order Markov processes, and 4 emission processes distributed as semi-continuous Exponential distribution mixtures. We note here again an assumption that has been made throughout the course of this chapter - the marginal parameters of the HMM are assumed to be known. To that end, the studies in this section focus on estimating the correlation structure in HMMs given data from the HMM, its states, and the marginal parameters.

4.3.1 Estimating copula parameters of a multivariate state process

Consider the 3-dimensional state processes corresponding to a clustered LHMM, denoted by $s = (s^{(1)}, s^{(2)}, s^{(3)})$. For $d = 1, 2, 3$, the state process $s^{(d)}$ is parameterized by an initial distribution $\pi_1^{(d)}$ and a transition matrix $A^{(d)}$. Furthermore, they are also parameterized by a Gaussian copula with a mean vector $\mathbf{0}$ and a correlation matrix Σ. We set the parameters to the

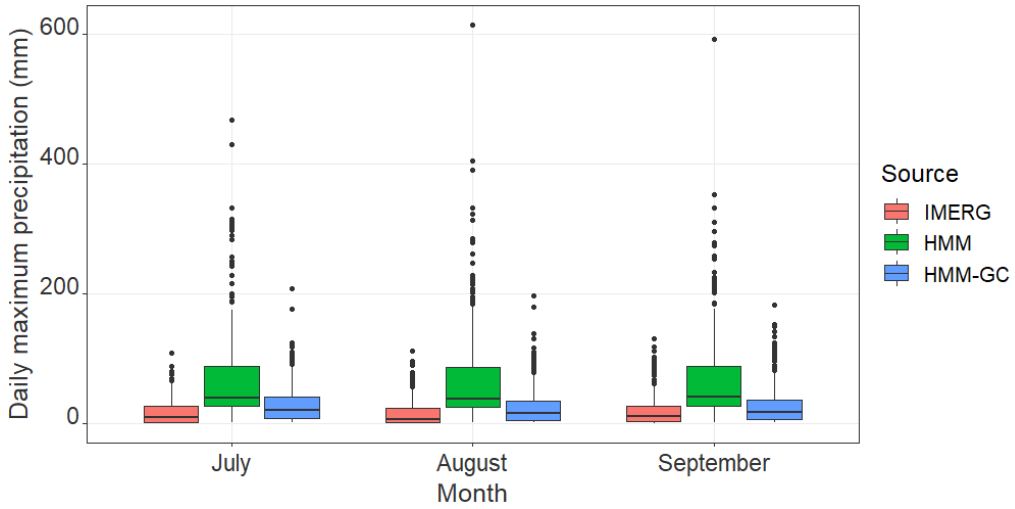

Figure 4.6: Distribution of the maximum daily basin precipitation for historical data from 2001–2018 compared against 18 years of HMM and HMM-GC simulated data

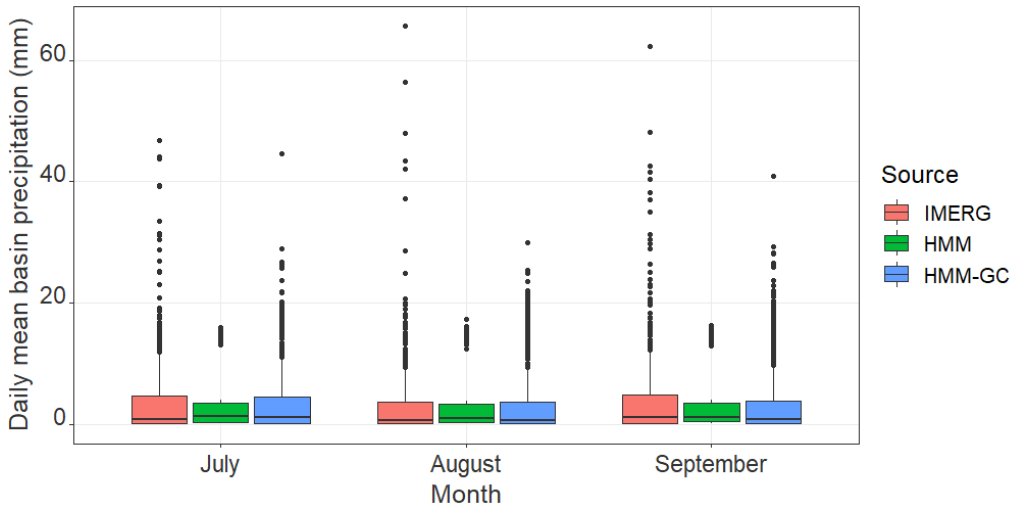

Figure 4.7: Distribution of the average daily basin precipitation for historical data from 2001–2018 compared against 18 years of HMM and HMM-GC simulated data

following values:

$$A^{(1)} = \begin{bmatrix} 0.60 & 0.30 & 0.10 \\ 0.20 & 0.50 & 0.30 \\ 0.30 & 0.20 & 0.50 \end{bmatrix}, \quad A^{(2)} = \begin{bmatrix} 0.40 & 0.40 & 0.20 \\ 0.30 & 0.40 & 0.30 \\ 0.40 & 0.20 & 0.40 \end{bmatrix}, \quad A^{(3)} = \begin{bmatrix} 0.20 & 0.30 & 0.50 \\ 0.20 & 0.60 & 0.20 \\ 0.20 & 0.40 & 0.40 \end{bmatrix},$$

$$\pi_1^{(1)} = (0.38, 0.34, 0.28), \qquad \pi_1^{(2)} = (0.36, 0.34, 0.30), \qquad \pi_1^{(3)} = (0.20, 0.54, 0.26),$$

where the initial distributions are obtained as the steady states for the transition probability matrices. Our first order of business is to verify whether the value of the copula correlation has a monotone relationship with the Spearman correlation of states generated from this DGP. To that end, we take each pair of Markov chains, vary their copula correlation from -1 to 1 in increments of 0.001, and generate 90000 state pairs from the model. Figure 4.8 plots the copula correlations on the x-axis and the observed Spearman correlation $r_{d_1 d_2}$ on the y-axis for each pair (d_1, d_2) of state processes. We notice a monotone relationship for each pair of state processes, with correlations for the pairs (1,2), (1,3), and (2,3) represented using red, green and blue lines. Based on the $y = x$ line in black running through the middle of the graph, we see that the absolute value of $r_{d_1 d_2}$ is less than the copula correlation, confirming the compression in the range of observed correlation we expect due to the discretization that happens as part of the data generation process. The range of $r_{d_1 d_2}$ depends on the marginal parameters of the two state processes. Unless the two state processes are identical, this value will not reach -1 or 1. In Figure 4.9, the Pearson copula correlations in the x-axis are transformed to Spearman correlations using (4.3), and plotted against the observed $r_{d_1 d_2}$ of the generated states. We see the same behaviour which is expected since (4.3) is a deterministic function. The only changes we observe are in the slopes of the curves, especially around the extremes.

We now consider estimating the correlation matrix of the Gaussian copula. We fix Σ at:

$$\Sigma = \begin{bmatrix} 1.00 & 0.80 & 0.20 \\ 0.80 & 1.00 & 0.70 \\ 0.20 & 0.70 & 1.00 \end{bmatrix}.$$

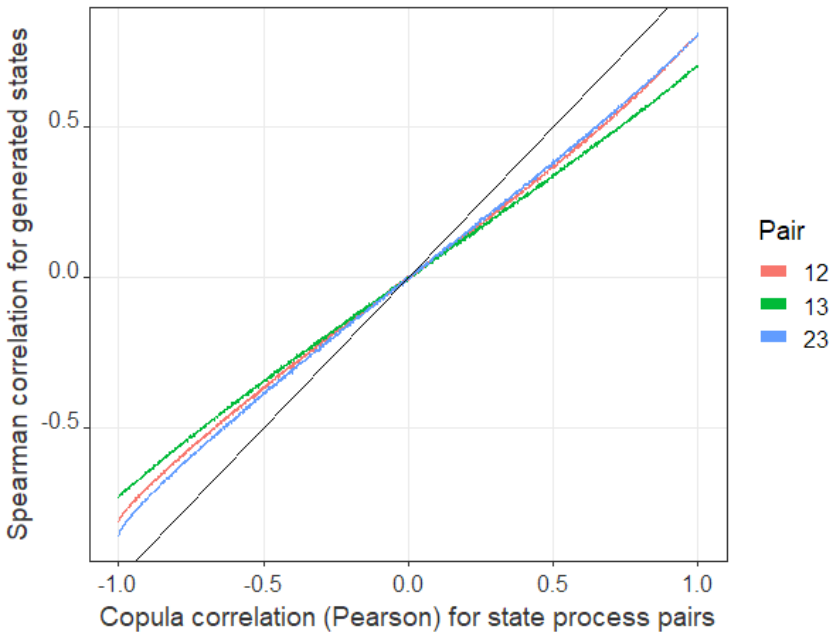

Figure 4.8: Relationship between the copula correlation (Pearson) between pairs of state processes and the Spearman correlation from 90000 states generated from the models.

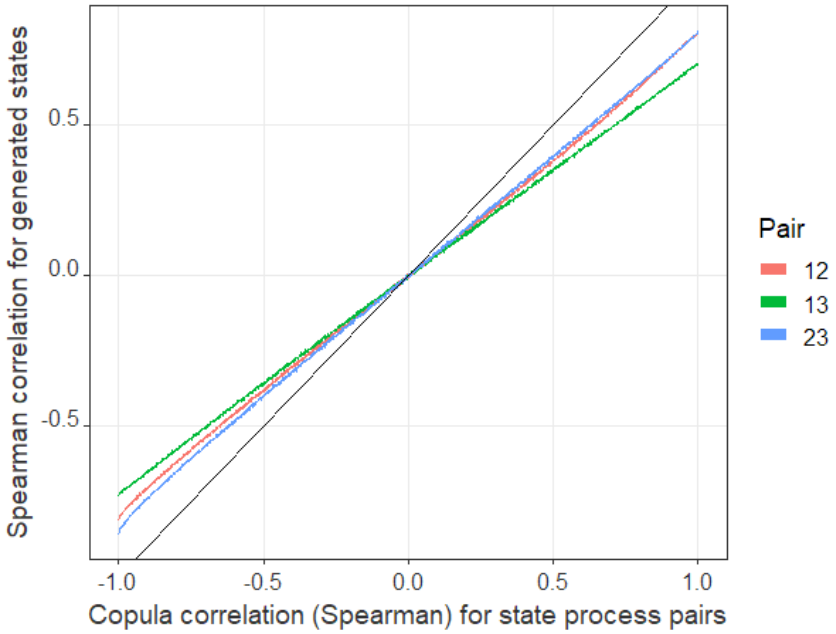

Figure 4.9: Relationship between the copula correlation (Spearman) between pairs of state processes and the Spearman correlation from 90000 states generated from the models.

with the remaining marginal parameters as defined previously. We generate 90000 states from this 3-dimensional Markov process using Serfozo's algorithm, corresponding to 3 months' daily states for 1000 years. The Spearman correlations between each pair of states are:

$$r_{12} = 0.6095, \qquad r_{13} = 0.1403, \qquad r_{23} = 0.5538.$$

We then run Algorithm 5 to try and recover Σ, assuming the marginal parameters and states are known. The tolerance ϵ is set to be 0.001 and the step size τ is set to be 0.01. The estimated copula correlation matrix is:

$$\hat{\Sigma} = \begin{bmatrix} 1.00 & 0.79 & 0.21 \\ 0.79 & 1.00 & 0.71 \\ 0.21 & 0.71 & 1.00 \end{bmatrix}.$$

$\hat{\Sigma}$ is also found to be positive definite. The number of iterations required for the line search for each pair are:

$$\hat{\rho}_{12} : 25 \text{ iterations}, \qquad \hat{\rho}_{13} : 13 \text{ iterations}, \qquad \hat{\rho}_{23} : 37 \text{ iterations}.$$

4.3.2 Gaussian copula for emissions of an HMM

In this study, we consider a data generating process consisting of a 3-state HMM with a univariate state process and a three-dimensional emission process. Let $s = s^{(1)}$ with the transition probability matrix

$$A = \begin{bmatrix} 0.60 & 0.30 & 0.10 \\ 0.20 & 0.50 & 0.30 \\ 0.30 & 0.20 & 0.50 \end{bmatrix},$$

with the initial distribution $\pi_1 = (0.38, 0.34, 0.28)$. Let the matrix of mixture distribution probabilities be denoted by $C_1, C_2,$ and C_3, and the corresponding matrix of Exponential distribution

rates for positive precipitation be denoted by $\Lambda_1, \Lambda_2,$ and Λ_3. They are set to the following values:

$$C_1 = \begin{bmatrix} 0.10 & 0.60 & 0.30 \\ 0.20 & 0.40 & 0.40 \\ 0.30 & 0.40 & 0.30 \end{bmatrix}, \quad C_2 = \begin{bmatrix} 0.20 & 0.60 & 0.20 \\ 0.40 & 0.20 & 0.40 \\ 0.50 & 0.20 & 0.30 \end{bmatrix}, \quad C_3 = \begin{bmatrix} 0.20 & 0.60 & 0.20 \\ 0.50 & 0.30 & 0.20 \\ 0.60 & 0.20 & 0.20 \end{bmatrix},$$

$$\Lambda_1 = \begin{bmatrix} 0.08 & 1.00 \\ 0.20 & 5.00 \\ 0.50 & 8.00 \end{bmatrix}, \qquad \Lambda_2 = \begin{bmatrix} 0.02 & 1.00 \\ 0.30 & 6.00 \\ 0.50 & 10.0 \end{bmatrix}, \qquad \Lambda_3 = \begin{bmatrix} 0.05 & 1.00 \\ 0.10 & 5.00 \\ 0.50 & 8.00 \end{bmatrix}.$$

Now, we add Gaussian copulas for each state's data. The correlation matrices for the Gaussian copulas are denoted by Σ_1, Σ_2, and Σ_3 corresponding to the 3 states, and set to be:

$$\Sigma_1 = \begin{bmatrix} 1.00 & 0.30 & 0.60 \\ 0.30 & 1.00 & 0.90 \\ 0.60 & 0.90 & 1.00 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} 1.00 & 0.20 & 0.50 \\ 0.20 & 1.00 & 0.80 \\ 0.50 & 0.80 & 1.00 \end{bmatrix}, \quad \Sigma_3 = \begin{bmatrix} 1.00 & 0.10 & 0.40 \\ 0.10 & 1.00 & 0.70 \\ 0.40 & 0.70 & 1.00 \end{bmatrix}.$$

We first generate 1000 independent datasets from this model using Algorithm 6; each dataset is of 1800 days and can thus be represented as 1800×3 matrices. Assuming all marginal parameters, i.e., all parameters except $\Sigma_1, \Sigma_2, \Sigma_3$ to be known, we use the pair-copula approximation of Algorithm 5 to estimate the parameters of the Gaussian copulas for each dataset. Averaging over the 1000 estimates, we get:

$$\tilde{\Sigma}_1 = \begin{bmatrix} 1.00 & 0.29 & 0.58 \\ 0.29 & 1.00 & 0.89 \\ 0.58 & 0.89 & 1.00 \end{bmatrix}, \quad \tilde{\Sigma}_2 = \begin{bmatrix} 1.00 & 0.19 & 0.48 \\ 0.19 & 1.00 & 0.78 \\ 0.48 & 0.78 & 1.00 \end{bmatrix}, \quad \tilde{\Sigma}_3 = \begin{bmatrix} 1.00 & 0.09 & 0.38 \\ 0.09 & 1.00 & 0.67 \\ 0.38 & 0.67 & 1.00 \end{bmatrix}.$$

We notice that the estimates are very close to their true values, but are slightly lower than the true values. This is expected due to the way we are generating data from the copula for this simulation study. On every day that it rains, correlated uniform variates are used to generate Exponential observations to ensure correlated precipitation amounts. This is the case even when different mixture components are selected at different locations. However, when we are estimating the

parameters, we only consider the days when it rains at both locations and furthermore the rainfall is generated from the same mixture component. Thus, we are inherently leaving some information on the table when estimating the copula correlations. As seen in Section 4.2.2, this is necessary since using all days' of data actually confounds the correlation.

In practice, the mixture component assignments will not be known. There are two ways to estimate them from the data. If we use variational Bayes for parameter estimation, the quantity q_{tjlm} corresponds to the posterior probabilities of the data arising from a particular mixture component, conditional on the most likely state for the day. This information can be used to assign mixture components to each day's data. Alternatively, we can resort to a maximum likelihood approach, where we plug in the precipitation values into the complete data likelihood with the parameters replaced by their estimates. Since r_{tjlm}, the variable for mixture component assignments, is encoded as a one-hot vector, this will give us a crude estimate of the likelihood of the observation arising from a particular mixture component. This is the approach we now use to pair the observed precipitation with mixture components. Using that as the information to subset and compute the Spearman correlations from data for days with matching mixture components, we get the following estimates of copula correlation:

$$\hat{\Sigma}_1 = \begin{bmatrix} 1.00 & 0.27 & 0.56 \\ 0.27 & 1.00 & 0.87 \\ 0.56 & 0.87 & 1.00 \end{bmatrix}, \quad \hat{\Sigma}_2 = \begin{bmatrix} 1.00 & 0.22 & 0.45 \\ 0.22 & 1.00 & 0.76 \\ 0.45 & 0.76 & 1.00 \end{bmatrix}, \quad \hat{\Sigma}_3 = \begin{bmatrix} 1.00 & 0.14 & 0.39 \\ 0.14 & 1.00 & 0.66 \\ 0.39 & 0.66 & 1.00 \end{bmatrix}.$$

The copula correlations when use use estimated mixture component assignments are similar to the ones obtained when we use the true mixture component assignments. However, there is one important difference that we would like to point out. We notice that low correlation values are sometimes overestimated, whereas high correlation values tend to be slightly underestimated. Similarly, the correlation from the dry states with fewer wet days tends to be higher than that from the wet states. We believe this to be an artifact of computing rank correlations from limited data.

Figures 4.10-4.12 consist of histograms depicting the distribution of the copula correlation estimates, corresponding to data from states 1, 2, and 3. For each state, we plot the distributions of the 3 pairwise correlations over the course of 1000 simulations, with the mixture assignments be-

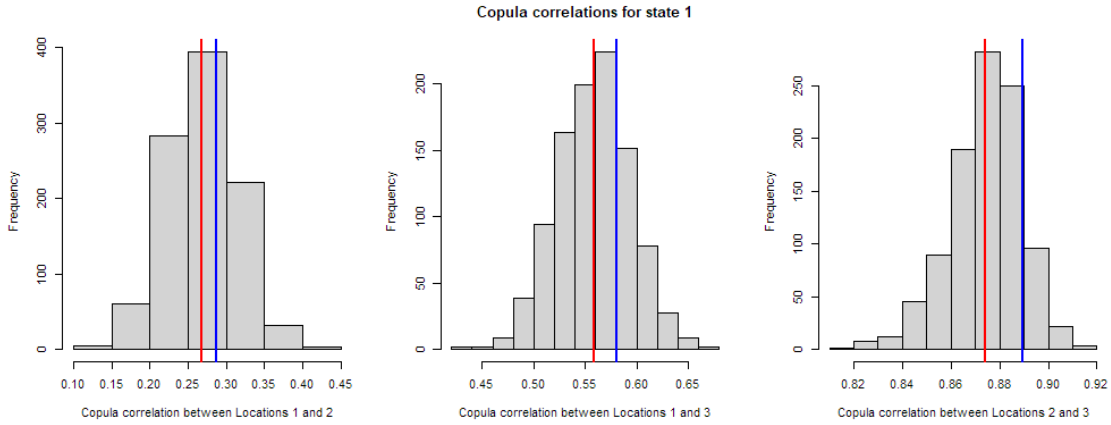

Figure 4.10: Estimated copula correlations for the emission process between pairs of locations for data arising from State 1 of the HMM.

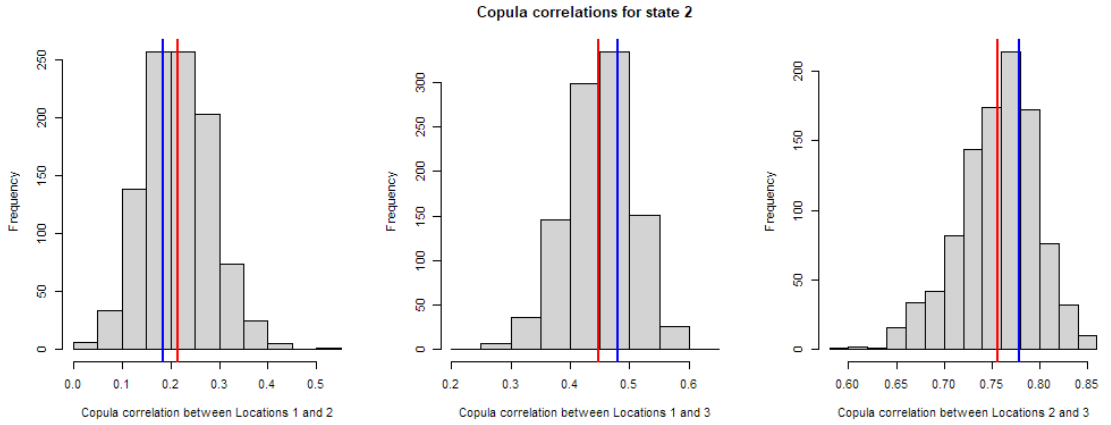

Figure 4.11: Estimated copula correlations for the emission process between pairs of locations for data arising from State 2 of the HMM.

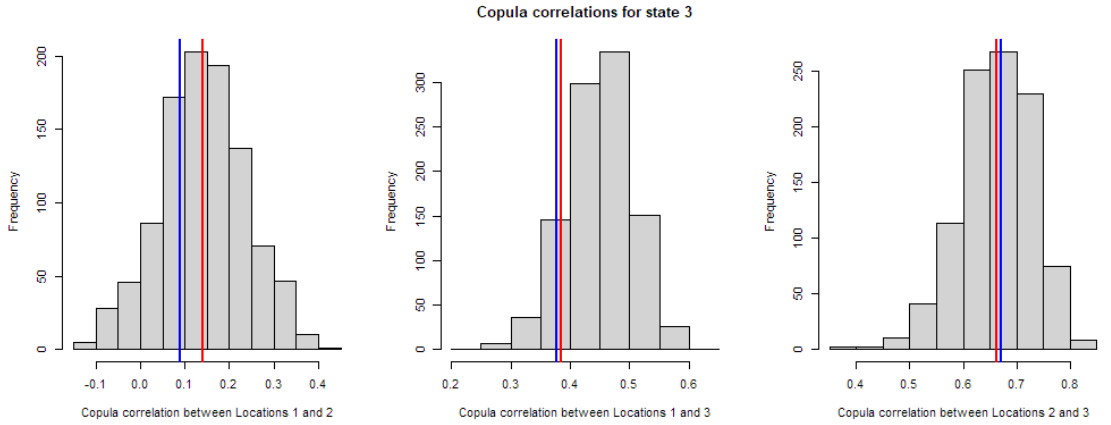

Figure 4.12: Estimated copula correlations for the emission process between pairs of locations for data arising from State 3 of the HMM.

ing estimated using their likelihoods. The red lines are the means of the histograms corresponding to elements of $\hat{\Sigma}_j$ for each state j, and the blue lines are the values obtained when we use the true mixture component assignments, i.e. the values in $\tilde{\Sigma}_j$. We find our estimates of copula correlation to be reasonable for the different weather regimes and correlation values.

4.4 Conclusions

As the case study in Section 4.2.2 demonstrates, the correlations between the emission distributions in a classical HMM that implicitly arise through the state process are not enough to describe highly correlated geostatistical data. HMM literature tends to work with multivariate Normal emissions as a way around this, since the multivariate Normal distribution has an explicit correlation structure. However, no explicit multivariate parameterizations exist for semi-continuous emission distributions. The Gaussian copula approach allows us to parameterize the pairwise correlations between locations explicitly and provides a way to generate synthetic data from it as well. Copulas constructed using Algorithm 5 are able to replicate a wide variety of pairwise correlation values with acceptable error margins. This would allow simulation of basin-wide weather patterns in a manner consistent with historical data.

While the aim of the copula for emissions is to explicitly specify a correlation structure for observed precipitation, the aim of a clustered LHMM is not to imbue the HMM with a more complex correlation structure. As a matter of fact, the implicit correlation transmitted to the emissions via a clustered LHMM's state processes is bound to be less than if it were an HMM with a single state process. However, a clustered LHMM allows more variety in how the latent processes evolve. A single state process forces the entire basin to have a uniform weather regime; while increasing the number of states mitigates this to an extent, this is still an unrealistic assumption. Local weather regimes are a far more realistic setup for large watersheds. Some HMM formulations quantify this using hierarchical state processes - a univariate global state process and a multivariate local state process which depends on the global state process. Our clustered LHMM follows this same idea, and it will allow different sections of the emission process to evolve more freely and thus describes a richer, more flexible model. A copula for discrete distributions is chal-

lenging to begin with, and things are made more complicated by the Markov property of the state processes. However, Algorithm 4 works around many of these problems by using Spearman correlation, which takes advantage of the ordinal nature of the states and sidesteps the need to evaluate the marginal CDFs of each Markov process.

Chapter 5

Application to Daily Precipitation Data over the Chesapeake Bay Watershed

5.1 An HMM without Clusters

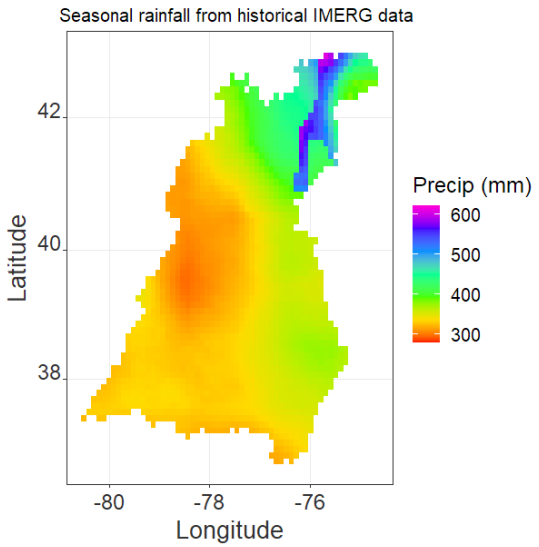

Figure 5.1: Historical precipitation for Jul–Sep over the Chesapeake Bay watershed from GPM-IMERG data.

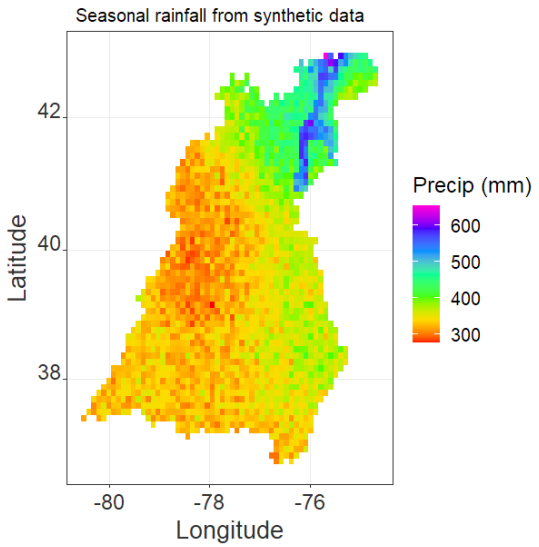

Figure 5.2: Synthetic precipitation for Jul–Sep over the Chesapeake Bay watershed from a base HMM.

We return now to our application introduced in Section 1.1 - modeling daily precipitation data from GPM-IMERG over the Chesapeake Bay watershed in Eastern USA. Figure 5.1 shows a spatial map of total rainfall from July–September at each IMERG grid point of the watershed, averaged over 20 years of data. We want to be able to replicate not only this spatial map using synthetic data, but also the monthly and seasonal statistics of precipitation at each grid point. To

demonstrate why modifications such as an LHMM for the state process or a Gaussian copula for the emission process are necessary, we first fit a 3-state base HMM to the data corresponding to the model presented in 3.3. Our priors for the model are very similar to what we have used in simulation studies so far. We assign symmetric Dirichlet priors for π_1 and A. The prior $p(\pi_1)$ has a concentration of 1, and each row of the prior $p(A)$ has a concentration of 10. Without loss of generality, we order the states to correspond to heavy, medium, and low rainfall respectively. For each location $l = 1, \ldots, 1927$, precipitation is specified as a mixture with a point mass at zero and two Exponential distributions for positive precipitation. We denote the priors for the mixture assignment probabilities as $\zeta_l^{(0)}$. The elements of the matrix Λ_l correspond to the rate parameters of the Exponential distributions for positive precipitation. They are assigned Gamma distribution priors, and the shape and rate parameters of the Gamma distributions are given by the matrices $\gamma_l^{(0)}$ and $\delta_l^{(0)}$ respectively. These are set to the following values:

$$\zeta_l^{(0)} = \begin{bmatrix} 3.0 & 4.0 & 3.0 \\ 3.0 & 3.5 & 3.5 \\ 4.0 & 3.0 & 3.0 \end{bmatrix}, \qquad \gamma_l^{(0)} = \begin{bmatrix} 0.5 & 2 \\ 1.5 & 5 \\ 2.0 & 10 \end{bmatrix}, \qquad \delta_l^{(0)} = \begin{bmatrix} 2 & 2 \\ 2 & 2 \\ 2 & 2 \end{bmatrix}.$$

The components are ordered to ensure that wetter states will have lower exponential rates and higher mixture probabilities for the exponential components, while drier states will have higher rates and more weight placed on the dry component corresponding to $m = 0$. This makes the model identifiable.

For fitting the model, we ran SVB for 300 iterations with step sizes $\tau_i = (1 + i)^{-0.9}$. This was followed by 30 iterations of CAVI since we noticed in the simulations that the state distribution parameters did not converge very well under SVB. The fitted model has a posterior initial probability $\tilde{\pi}_1 = c(0.20, .35, 0.45)$ and the transition probability matrix

$$\tilde{A} = \begin{bmatrix} 0.41 & 0.34 & 0.25 \\ 0.34 & 0.37 & 0.29 \\ 0.13 & 0.31 & 0.56 \end{bmatrix}.$$

We note that the two lowest probabilities in the transition matrix occur when the driest state tran-

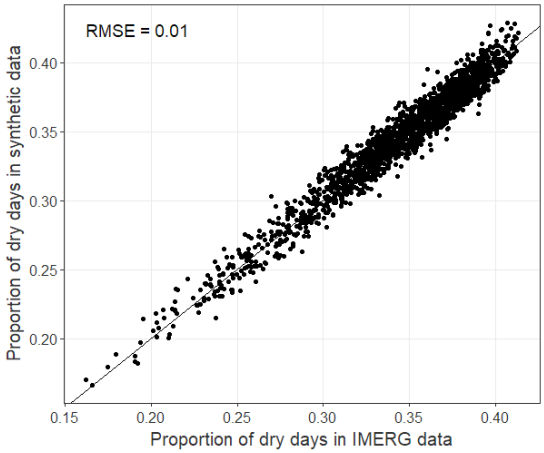

Figure 5.3: Historical and synthetic proportion of dry days at each location of the watershed based on base HMM.

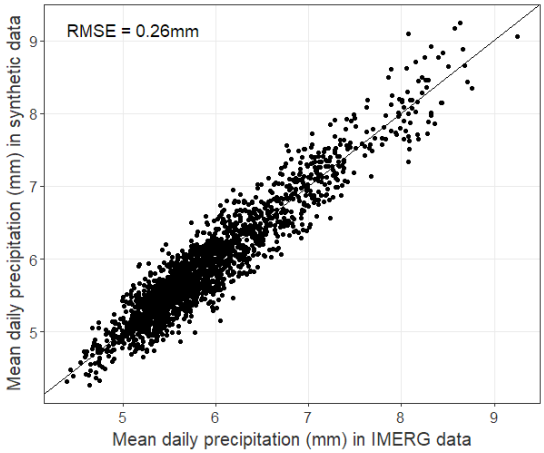

Figure 5.4: Historical and synthetic daily mean precipitation (mm) at each location of the watershed based on base HMM.

sitions to the wettest state (0.13), and vice versa (0.25). States 1 and 3 tend to transition between each other through state 2 most of the time.

We generated 1840 days of synthetic data from this 3-state model that we refer to as the base HMM in this chapter. Figure 5.2 shows a plot of total precipitation at each grid point over the 3 months of Jul–Sep averaged over 20 years, based on the synthetic data simulated from the model. We note that while the plot is much noisier compared to the corresponding plot of the historical data in Figure 5.1, it recreates seasonal precipitation at individual locations to a certain degree. Figure 5.3 plots the proportion of dry days averaged over 20 years at each location based on historical IMERG data on the x-axis and synthetic data from the base HMM on the y-axis. Similarly, Figure 5.4 plots the mean daily precipitation averaged over 20 years at each location based on historical IMERG data on the x-axis and synthetic data from the base HMM on the y-axis. In both cases, the line through the middle of the plot corresponds to $y = x$. We see that even the base HMM with VB parameter estimation is good at reproducing seasonal characteristics of precipitation. In both plots, the points show a linear pattern and their spread along the $y = x$ line points to a lack of any bias in seasonal data. However, this model does not represent the spatial characteristics of the data very well. Figure 5.5 shows boxplots of the pairwise correlations in daily precipitation between grid points for the historical IMERG data and synthetic data generated by the base HMM. The distribution of the correlations in the synthetic data has both a smaller spread

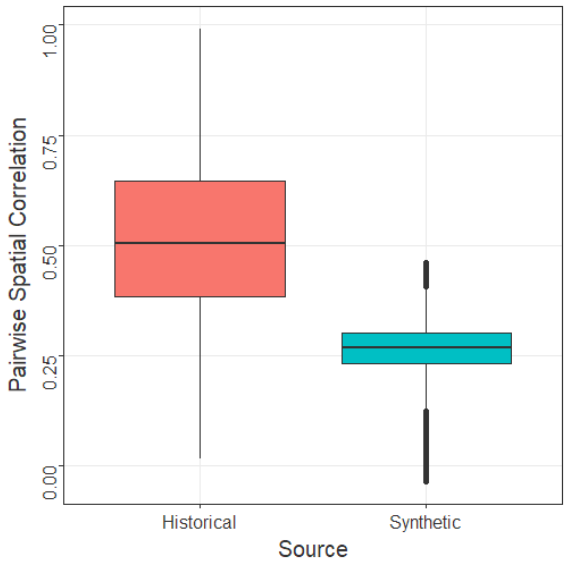

Figure 5.5: Spatial correlation in daily precipitation between pairs of grid points for historical IMERG data and synthetic data from base HMM.

and lower quantiles compared to the distribution based on the historical data. This necessitates the requirement of our copula approach.

We first fit a Gaussian copula to the emission distribution using Algorithm 5, reusing the marginal parameter estimates we obtained earlier. The most likely sequence of states are found using the Viterbi algorithm, and the mixture component assignment at each time point is done using the the posterior distribution q_{tjlm} as obtained at the final iteration of our optimization process. For 1927 grid points, the copula correlation matrix consists of 1.86×10^6 pairwise correlations for each state's data. Figure 5.6 shows the total precipitation at each grid point based on the synthetic data from this model with a Gaussian copula for emissions. We immediately notice that this model overestimates precipitation, as the entire range of precipitation in the image is around 100mm higher than the corresponding historical range in Figure 5.1. Figure 5.7 shows the distribution of the pairwise spatial correlations when correlated emissions are generated based on a Gaussian copula. We notice two differences from Figure 5.5. The first is that adding the copula has increased the range as well as the quantile values of the distribution. Further, the distribution is now positively skewed. However, the values are still quite low compared to correlations in historical data, in part due to the uncertainty added to our copula distribution from using marginal estimates to obtain copula parameters.

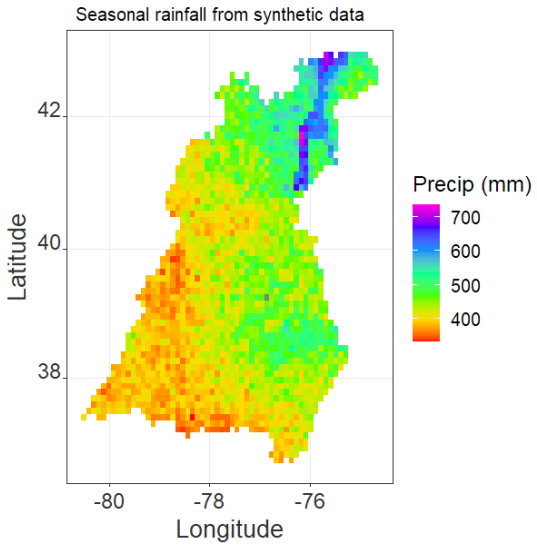

Figure 5.6: Synthetic precipitation for Jul–Sep over the Chesapeake Bay from a base HMM with a Gaussian copula for emissions.

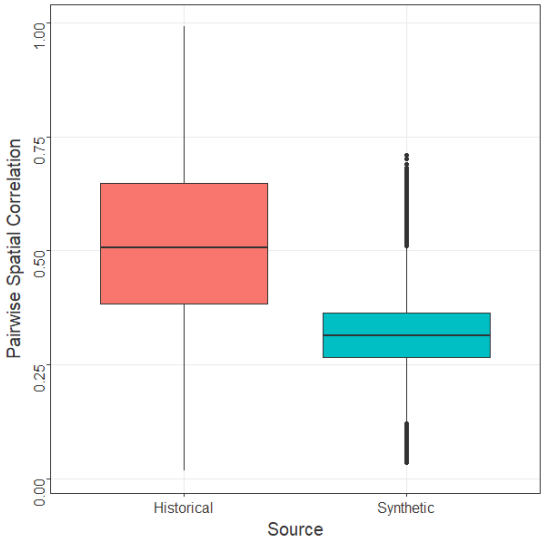

Figure 5.7: Spatial correlation in daily precipitation for historical IMERG data and synthetic data from an HMM with a Gaussian copula for emissions.

5.2 Constructing a Cluster LHMM for the Chesapeake Bay Dataset

5.2.1 Clustering the region by local weather regime

Next, we split the 1927 grid points into clusters based on their precipitation patterns, to fit an LHMM to this data. We use the data at hand to find the optimum number of clusters; for each location, the following variables are considered:

1. Latitude and Longitude
2. Proportion of dry days in a season
3. Total seasonal precipitation
4. Maximum seasonal precipitation.

Ideally, we want to use more information to inform the selection of clusters. Meteorological data like temperature or wind speed can be used, as well as terrain information. However, each of

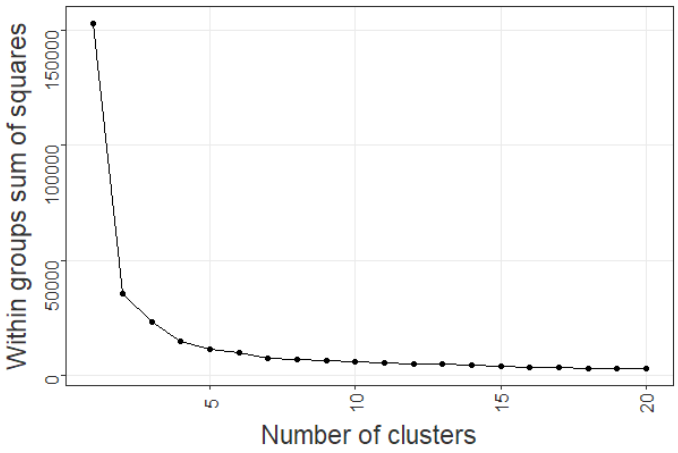

Figure 5.8: Scree plot of within group sum of squares for 1–20 cluster solutions for the 1927 IMERG grid points.

these extraneous variables would be at different spatial resolutions and will need to be aligned to IMERG's spatial grid.

Based on the variables listed above, we compute the within groups sum of squares (WSS) for a range of k-means clustering solutions to find the optimum number of clusters in the data. Figure 5.8 plots the WSS values for 1–20 cluster solutions. The optimum number of clusters is usually chosen around the elbow of the graph - where the reduction in WSS when the number of clusters are increased starts to flatten out. In this case, it coincides with a 3 or a 4 cluster solution.

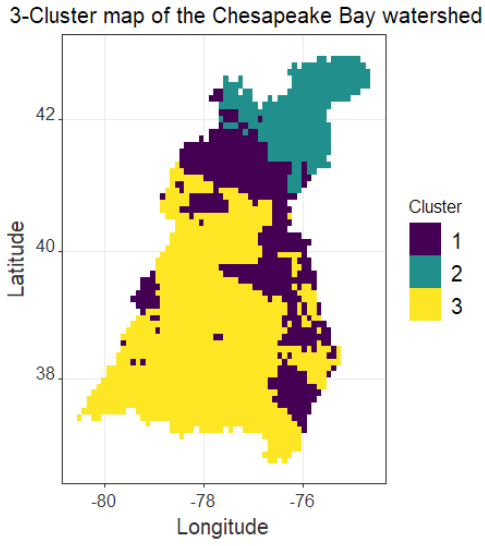

Figure 5.9: Grid points of the Chesapeake Bay watershed divided into 3 clusters using k-means clustering.

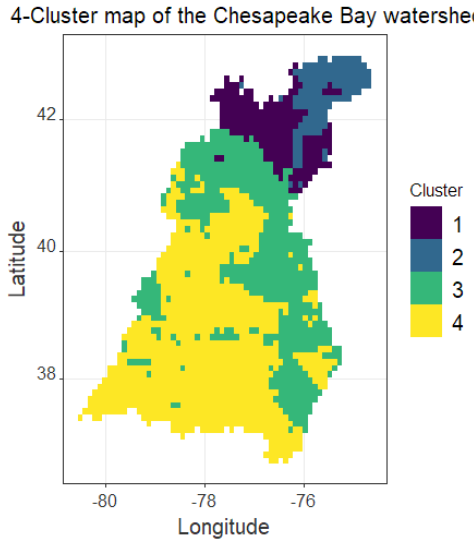

Figure 5.10: Grid points of the Chesapeake Bay watershed divided into 4 clusters using k-means clustering.

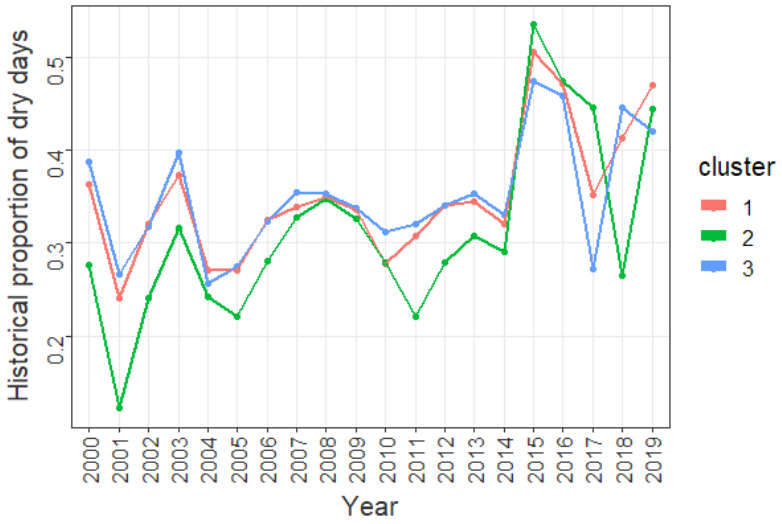

Figure 5.11: Proportion of dry days during the wet season for each cluster based on historical IMERG data.

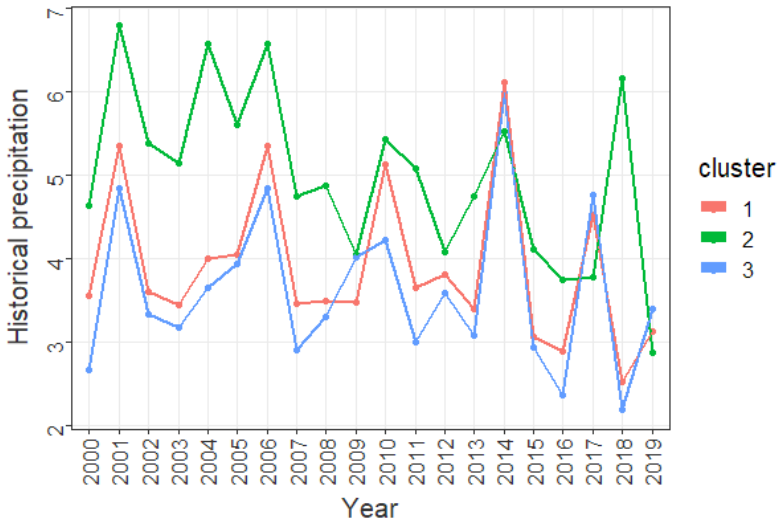

Figure 5.12: Mean daily precipitation during the wet season for each cluster based on historical IMERG data.

Figures 5.9 and 5.10 plot the k-means clustering solutions for 3 and 4 clusters respectively. We note that the main difference going from 3 to 4 clusters is that cluster 2 in the northern part of the watershed splits into 2 clusters. Since that area's features are captured well in our base model so far, we decide to go with a 3 cluster solution to develop the cluster LHMM.

Under this segmentation, cluster 1 contains 512 grid points, cluster 2 contains 286 grid points, and cluster 3 contains 1927 grid points. Figure 5.11 shows the proportion of dry days each wet season from 2000–2019 for each of the 3 clusters' data. We see that cluster 2 has the lowest

proportion of dry days up until 2014, beyond which the 3 clusters' statistics seem to intersect. Similarly, Figure 5.12 plots the mean daily precipitation each wet season from 2000–2019 for each of the clusters' data. Cluster 2 has the highest precipitation until 2014, and cluster 3 has the lowest. These 2 plots show clear differences between the 3 clusters' precipitation patterns across the 20 years of data that we have. Cluster 2 in the northern half of the bay is the wettest, while cluster 3 in the west and southern areas of the bay is the driest cluster. More importantly, the clusters' precipitation patterns are much closer to each other since 2014, suggesting that the distribution of precipitation has become more spatially uniform in recent years.

5.2.2 Estimating marginal HMM parameters using VB

We fit a 3-state HMM to each cluster's data. Every cluster is assigned the same priors as the base HMM in Section 5.1. Since the marginal parameter estimation is independent for each cluster, it is done in parallel. 300 SVB iterations are run with step sizes $\tau_i = (1+i)^{-0.9}$ followed by 30 iterations of CAVI. Note that the 30 iterations of CAVI take longer to execute than the 300 iterations of SVB in all these cases, confirming the necessity of stochastic optimization for large datasets. We obtain the following estimates of the state processes for each cluster of the LHMM:

$$\tilde{A}^{(1)} = \begin{bmatrix} 0.44 & 0.38 & 0.18 \\ 0.34 & 0.45 & 0.31 \\ 0.13 & 0.30 & 0.57 \end{bmatrix}, \quad \tilde{A}^{(2)} = \begin{bmatrix} 0.46 & 0.35 & 0.19 \\ 0.35 & 0.36 & 0.29 \\ 0.12 & 0.27 & 0.61 \end{bmatrix}, \quad \tilde{A}^{(3)} = \begin{bmatrix} 0.42 & 0.35 & 0.23 \\ 0.32 & 0.35 & 0.33 \\ 0.18 & 0.29 & 0.53 \end{bmatrix},$$

$$\tilde{\pi}_1^{(1)} = (0.30, 0.44, 0.25), \qquad \tilde{\pi}_1^{(2)} = (0.35, 0.30, 0.35), \qquad \tilde{\pi}_1^{(3)} = (0.06, 0.49, 0.45),$$

where $\tilde{A}^{(d)}$ and $\tilde{\pi}_1^{(d)}$ are estimates of the transition matrix and initial distribution for cluster d, $d = 1, 2,$ and 3.

5.2.3 Constructing a Gaussian copula for the LHMM

The Viterbi Algorithm identifies the most likely sequence of states that generated each cluster's rainfall. Like the estimation of the marginal parameters, we ran the Viterbi Algorithm in

parallel for computational efficiency. The matrix of pairwise Spearman correlations between the states obtained from the Viterbi algorithm was

$$R = \begin{bmatrix} 1.00 & 0.66 & 0.68 \\ 0.66 & 1.00 & 0.31 \\ 0.68 & 0.31 & 1.00 \end{bmatrix}.$$

We see that cluster 1 and cluster 3 have high correlations in their state processes, but clusters 2 and 3 do not. Looking at the cluster positions in Figure 5.9, we can see that this is possibly because cluster 1 neighbors both clusters 2 and 3, and thus its states are correlated with the states from both clusters. Using Algorithm 4, we estimate a Gaussian copula for the joint distribution of the states parameterized by the 3×3 correlation matrix $\tilde{\Sigma}$, given by:

$$\tilde{\Sigma} = \begin{bmatrix} 1.00 & 0.80 & 0.81 \\ 0.80 & 1.00 & 0.41 \\ 0.81 & 0.41 & 1.00 \end{bmatrix}.$$

Based on this Gaussian copula, we generated synthetic states for the LHMM, in the form of an 1840×3 matrix. The Spearman correlations of the synthetic chains was

$$R^* = \begin{bmatrix} 1.00 & 0.69 & 0.68 \\ 0.69 & 1.00 & 0.33 \\ 0.68 & 0.33 & 1.00 \end{bmatrix},$$

whose elements are fairly close to the values in R.

5.2.4 Constructing a Gaussian copula for emissions

When constructing the Gaussian copula for emissions, we need to take into account two sets of dependencies: the correlation between locations belonging to the same cluster and the correlation between locations belonging to different clusters. We already assume that the state processes of 2 different clusters are correlated. Therefore, it is reasonable to further assume that

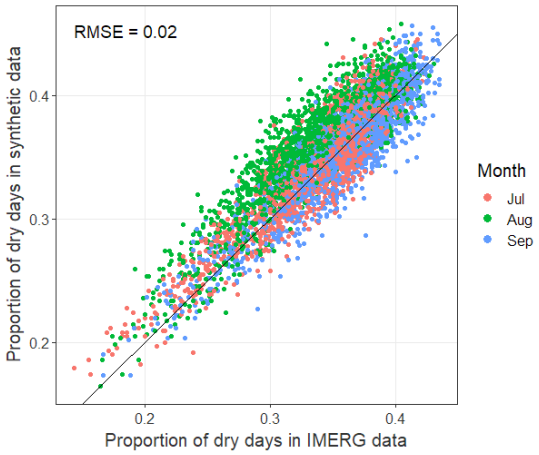

Figure 5.13: Historical and synthetic proportion of dry days for every month at each location of the watershed based on LHMM.

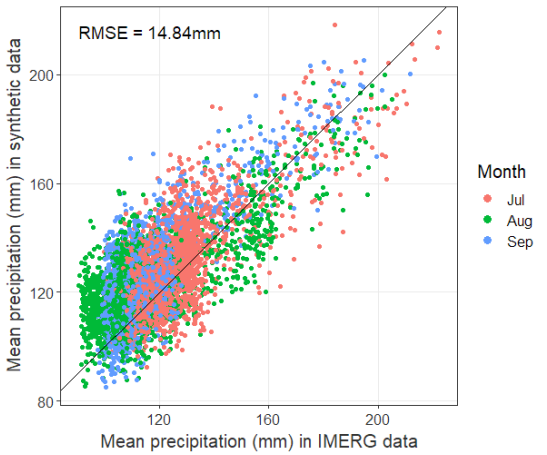

Figure 5.14: Historical and synthetic daily mean precipitation (mm) for every month at each location of the watershed based on LHMM.

the correlation between locations belonging to 2 different clusters is 0. An immediate benefit of this assumption is that it cuts down the number of pairwise correlations that need to be estimated for each state from 1.86×10^6 to 8.08×10^5. Copula matrices for each cluster were constructed in parallel. A total number of 9 matrices are estimated, one for each cluster and state.

With the estimation of the Gaussian copulas complete, the fitted 3-cluster LHMM was used to generate 20 years of daily precipitation data for the wet season over the entire watershed. In the next section, we compare the distribution of the synthetic data with the historical data.

5.3 Performance for Synthetic Data

We begin by comparing the monthly statistics for each grid point. Figure 5.13 shows the proportion of dry days for each of the 3 months at every grid point of the watershed. The root mean square error (RMSE) between the historical and the synthetic data is 0.02. Comparing with the $y = x$ line through the middle, we see that the proportion is slightly overestimated in the synthetic data for August. Similarly, Figure 5.14 shows the total monthly precipitation at each grid point for each month. The synthetic data in this plot has an RMSE of 14.84 mm. Comparing with the $y = x$ line through the middle, we see that the mean precipitation is slightly overestimated across the board. However, in both cases, there exists a linear relationship between the historical statistics

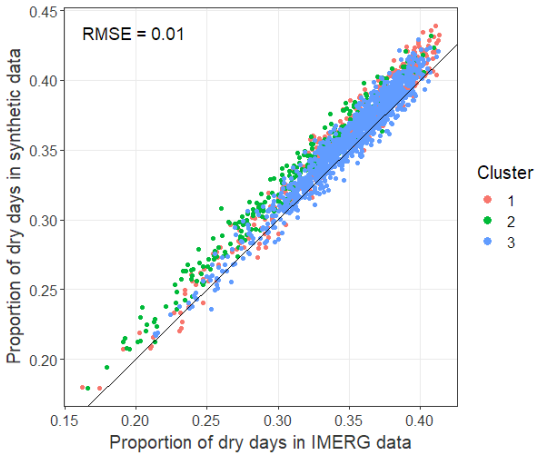

Figure 5.15: Historical and synthetic proportion of dry days for the wet season at each location of the watershed based on LHMM.

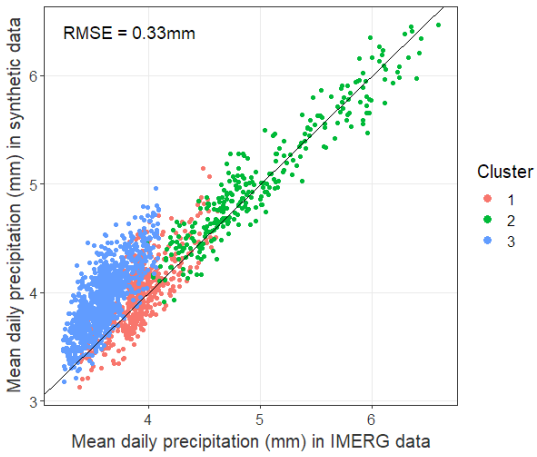

Figure 5.16: Historical and synthetic daily mean precipitation (mm) for the wet season at each location of the watershed based on LHMM.

and those derived from synthetic data.

A similar pattern is seen in Figures 5.15 and 5.16, which has seasonal proportions of dry days and total precipitation respectively for historical and synthetic data. In these plots, the information is grouped by the cluster the data arises from. Looking at Figure 5.16 in particular, we notice that cluster 3 has some of the lowest precipitation values and yet overestimates daily mean precipitation values the most. For the other two clusters, however, the synthetic data is well representative of the historical data in the context of mean precipitation values. This implies that the emission distribution parameters for this cluster are not well estimated. This could be due to the dry nature of cluster 3 which might need more informative priors.

Figure 5.17 depicts the spatial map of seasonal precipitation over the watershed averaged over 20 years of synthetic data. The data is smoother compared to Figure 5.2, and the range of values is better aligned with the historical data than, say, Figure 5.6. Looking at the distribution of spatial correlations for each cluster in Figure 5.18, we see that the synthetic data from the LHMM has higher correlations that the HMM without any clusters, as depicted in Figure 5.7. However, the correlation in the synthetic data is still low compared to the historical data.

To compare how the states differ for each cluster, Tables 5.1, 5.2, and 5.3 present key statistics of interest for locations belonging to clusters 1, 2, and 3 respectively. Within each table, the information is divided by the states. Since there are 3 clusters and 3 states for each cluster of

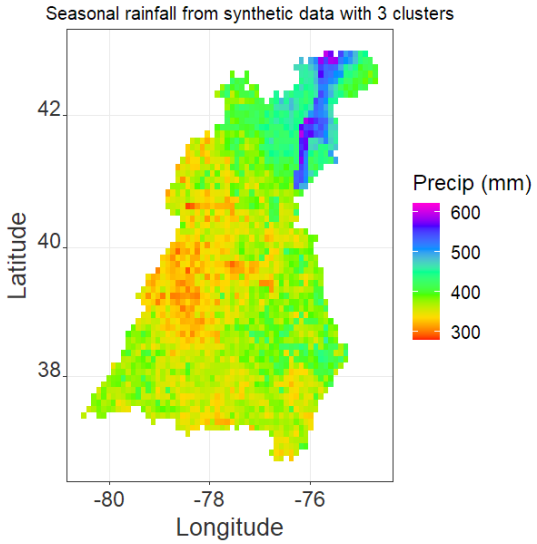

Figure 5.17: Synthetic precipitation for Jul–Sep over the Chesapeake Bay watershed from a 3-cluster LHMM with a Gaussian copula for emissions.

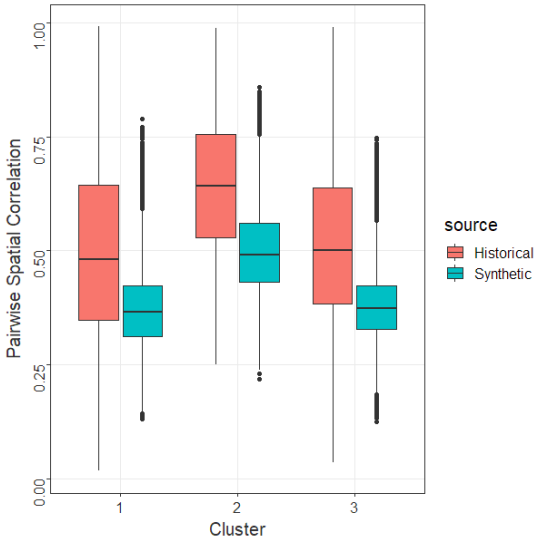

Figure 5.18: Pairwise Spatial correlation in daily precipitation for historical IMERG data and synthetic data from a 3-cluster LHMM with a Gaussian copula for emissions.

the LHMM, we have a total of 9 subgroups to compare. The state assignment for the historical data is done on the basis of the most likely sequence of states obtained using the Viterbi Algorithm. Looking within each cluster, we see that the states maintain the ordering established in our prior; the first state is the wettest with high rainfall and a low proportion of dry days, and the last state is the wettest with the lowest rainfall and the highest proportion of dry days. Comparing how the states differ from cluster to cluster, we note that the states are not equivalent for the three clusters. While the states divide the precipitation in each cluster into three regimes, or bins, the precipitation statistics of those regimes are cluster specific.

Looking at the final 2 columns in each of the tables, we see that the synthetic data provides

Table 5.1: Cluster 1 statistics for historical IMERG and synthetic data for each state of the LHMM averaged across all locations within the cluster.

State	**Mean Daily Positive Precipitation (mm)**		**Median Daily Precipitation (mm)**		**Maximum Daily Precipitation (mm)**		**Proportion of Dry Days**	
	Hist.	Synth.	Hist.	Synth.	Hist.	Synth.	Hist.	Synth.
1	11.15	11.87	6.25	5.71	103.33	126.80	0.03	0.03
2	2.46	2.60	0.29	0.25	36.01	33.78	0.21	0.21
3	0.51	0.56	0.0006	0.0009	12.15	9.46	0.74	0.73

Table 5.2: Cluster 2 statistics for historical IMERG and synthetic data for each state of the LHMM averaged across all locations within the cluster.

State	Mean Daily Positive Precipitation (mm)		Median Daily Precipitation (mm)		Maximum Daily Precipitation (mm)		Proportion of Dry Days	
	Hist.	Synth.	Hist.	Synth.	Hist.	Synth.	Hist.	Synth.
1	15.11	15.71	9.73	9.48	153.40	128.46	0.01	0.02
2	2.14	2.33	0.34	0.33	39.85	28.56	0.12	0.12
3	0.27	0.31	0	0	9.14	5.89	0.70	0.70

Table 5.3: Cluster 3 statistics for historical IMERG and synthetic data for each state of the LHMM averaged across all locations within the cluster.

State	Mean Daily Positive Precipitation (mm)		Median Daily Precipitation (mm)		Maximum Daily Precipitation (mm)		Proportion of Dry Days	
	Hist.	Synth.	Hist.	Synth.	Hist.	Synth.	Hist.	Synth.
1	10.34	11.65	5.97	5.73	89.64	127.78	0.02	0.03
2	2.27	2.57	0.32	0.28	33.84	34.06	0.20	0.21
3	0.50	0.57	0	0	12.39	9.62	0.73	0.73

good estimates of the proportion of dry days for all clusters and states. This implies that the marginal mixture component assignment parameters are well estimated. However, the synthetic data consistently overestimates the mean of positive precipitation amounts, as visible in the first 2 columns for each cluster. This is potentially a side effect of the Gaussian copula for emissions. Some interesting patterns also show up in the median and maximum precipitation as a consequence of the number of dry and wet days being better estimated than the amount of precipitation. We see that the median is slightly underestimated in all the subgroups, while maximum precipitation values are generally not well estimated in any of the subgroups. All these together point to the need for improving upon the Gaussian copula for emissions.

5.4 Discussion

We fit a 3-cluster LHMM, each with 3 states, to the Chesapeake Bay watershed data. Marginal parameter estimation is carried out using variational Bayes, and Gaussian copulas are constructed both for the state processes of each cluster as well as the emission processes within each cluster. Synthetic data generated from the model has monthly statistics comparable to the

historical data. Furthermore, it can partially replicate spatial correlations present in the data. However, the model tends to overestimate the statistics associated with positive precipitation. While these are the statistics that we have chosen to focus on, there are a variety of other measures that can be used to compare the synthetic data with the historical data depending on the needs of the study. For example, we can look at run lengths of dry and wet days and similar temporal characteristics at each location, as well as the capability of the model to simulate extreme weather events simultaneously for large stretches of the watershed.

This case study presented validates our approach to parameter estimation in HMMs for geostatistical data, while pointing out areas of potential improvement. One of our basic assumptions made for the simulation studies carries over to the case study - general priors. We assign near identical priors from our simulation studies to this case study. The priors for the transition matrix are symmetric in nature, and the priors for the emission distribution parameters only seeks to order the components and ensure identifiability. To that end, three conditions are enforced on the priors:

1. The first state has the lowest mixture assignment probability for the dry component, and the last state has the highest mixture assignment probability for the dry component.

2. For each state, there is one Exponential distribution with rate < 1 and one with rate ≥ 1.

3. The first state has the lowest rates of precipitation, and the final state has the highest.

Conditions 1 and 3 ensure that the first state is the wettest, with the lowest probability of being a dry day, and highest rainfall. Similarly, the third state ends up the driest, with the highest probability of being a dry day and low rainfall amounts. Condition 2 ensures that a wide range of precipitation values can be captured by each state. In particular, the component with rate < 1 allows the modeling of high values of precipitation, and the component with rate ≥ 1 allows the modeling of low positive values. This is especially relevant for our IMERG data - since the grids are fairly large, we observe a large range of positive daily precipitation values, from sub 1mm to well over 100mm. If our dataset has a smaller range, the emission distribution priors can be modified accordingly.

We note the necessity of augmenting the SVB optimization with a few CAVI iterations at the end. This might not be necessary if we have better prior information about the state process

priors, since those are the ones that have the most trouble converging under SVB. Despite the expense of CAVI iterations, it helps with algorithm convergence and we deem it necessary for HMMs.

The simplest model fitted in the beginning of the chapter contains no copulas, and it does quite well in replicating marginal statistics without any noticeable bias. However, it fares very poorly when it comes to replicating the spatial correlation present in the historical data. This might not be a problem if the data came from a small number of locations spaced far apart. The base HMM will prove adequate for such cases. However, working with densely gridded remote sensing data requires us to consider correlation structures in the data. This leads to some bias in the estimates. In particular, since the point of the copula for emissions is to generate correlated positive precipitation values, it reinforces high precipitation regimes by design. Further, it still cannot capture the correlations between locations adequately as the construction of the copula relies on marginal parameter estimates. However, we consider the emissions copula a critical part of the model in generating spatially consistent precipitation.

One way to address the overestimation of positive precipitation is by allowing for a larger number of weather regimes in the model. This way, there is more granularity available when switching weather regimes. We could increase the number of states in the data, but it has a few practical downsides. Increasing the state space requires a proportionate increase in the number of emission distribution parameters. It also increases the computational cost of parameter estimation. Clustering our data into an LHMM is an alternative approach which has a similar effect at a lower computational cost. The LHMM has the same number of emission distribution parameters as the base HMM, since the partitioning is not done to the temporal component of the data, but along the spatial dimension. The increase in the number of state process parameters is negligible compared to the number of emission process parameters. Since the intermediate weather regime for the wettest cluster has similar properties as the wettest weather state for a drier cluster, it allows for a similar granularity in the weather regimes.

A question also remains on how the clusters should be defined. If we use only precipitation statistics to cluster the data, then we are double dipping in the data to an extent as we are using the same information to divide the data into clusters as would later be used to divide it into states.

Ideally, we would want the clusters to be defined based on at least some additional data which is independent of precipitation. In our case, the only extra information we use are the latitude and longitude values of each grid point, which the HMM otherwise does not have access to. We could also be using information like terrain or other meteorological data.

This brings us back to the copula for emissions. Using an LHMM allows us an intuitive way to reduce the parameter space for the emission copula. In our case study, we have assumed zero copula correlation for emissions from points belonging to different clusters. This can also be set a fixed positive value for all cases. We believe this to be key to simulating more realistic basin-wide precipitation values. An alternative approach is to have emission copulas only for some of the states. For example, drier states which have a lot of days without precipitation could just be assigned an independence copula. This would have the added benefit of preventing the estimation of spurious correlation values based on a very small number of points. This approach can also be applied in conjunction with the LHMM.

Chapter 6

Summary and Future Work

The thesis introduces methods to model geostatistical data using hidden Markov models with an emphasis on variational Bayesian estimation. Our motivating and demonstrative example pertains to daily precipitation data observed over the Chesapeake Bay watershed in Eastern USA. The data is obtained from GPM-IMERG, which is on a $0.1^\circ \times 0.1^\circ$ grid with high spatial correlations present in the data. Furthermore, the precipitation at each location is described as a semi-continuous distribution with a point mass at zero for zero precipitation and a mixture of Exponential distributions for positive precipitation. The GPM-IMERG data covers the watershed using 1927 grid points - the densely gridded structure makes it important for the spatial characteristics of the data to be captured. The large spatial stretch of the basin also requires efficient computational algorithms that scale well and can be parallelized. In this concluding chapter, we summarize our contributions towards each facet of this modeling problem, and then outline future research directions.

6.1 List of Contributions

We present our contributions in the context of three broad questions pertaining to modeling geostatistical data using HMMs.

How to estimate the marginal parameters using Bayesian methods?

Our choice of a Bayesian approach instead of a maximum likelihood one is largely motivated by data size and model complexity. Since the data contains multiple latent variables, a

spatiotemporal structure, and a large dimension, a Bayesian framework where we can use prior knowledge to direct our model is an attractive proposition. However, regular MCMC approaches are usually too computationally intensive for an HMM. We instead resort of variational Bayes (VB) which is computationally more efficient. While previous literature on VB for HMMs has focused primarily on multivariate Normal emission distributions, we extend that in the following ways:

- **Variational Bayes parameter estimation for HMMs with semi-continuous emissions** [Majumder et al., 2021, in preparation (2021]. In particular, we focus on developing parameter estimation for emissions where the positive part of precipitation is a mixture of two Exponential distributions. We derive parameter estimates for a single chain of univariate emissions, multiple chains of univariate emissions, as well as for multivariate emissions data.

- **Developing model convergence metrics**. We derived expressions for the evidence lower bound (ELBO) which is used to assess model convergence in VB. We also derived expressions for the deviance information criterion (DIC) which can be used for model selection. However, we assume in this thesis that the model size is known; the DIC thus does not play a significant role in our own numerical studies.

- **Developing parameter estimation for Gamma distribution emissions**. We derived the variational posterior estimates for emission distributions where positive precipitation is either specified as a mixture of Gamma distributions or a modified Gamma shape mixture (GSM) distribution. Using Gamma distributions was deemed infeasible due to its computational cost. However, GSM provides a viable alternative to Exponential distributions and required a sparser parameter set. Our formulation for the GSM distribution generalizes the existing approach and requires fewer mixture components. We developed empirical Bayes prior estimation for the GSM distribution; however, we concluded that GSM for positive precipitation requires more development until it can be a good alternative to using Exponential distributions.

- **Stochastic variational Bayes (SVB) for computational efficiency** [Majumder et al., ac-

cepted (2021]. SVB is developed for this model since using the entire data for optimizing parameters using CAVI proves unfeasible in real-life scenarios. The SVB uses minibatches for fast optimization which takes advantage of our data structure where there is a gap between the end of the wet season for a year and the beginning of the wet season for the next year. We also modified the minibatch sampling algorithm to add more variability to our minibatches. The modified minibatch sampling method led to better estimates of the emission distribution parameters. However, SVB struggles to estimate the state parameters, and we recommend running a few iterations of CAVI after SVB.

Is a univariate state process sufficient to capture the underlying weather regimes for large spatial domains?

Proposing a clustered LHMM approach to model large geostatistical datasets [Majumder et al., in preparation (2021]. The spatial domain is divided into partitions, each with their own state processes. For large areas like the Chesapeake Bay, it is naive to assume that a single state process can drive the emission process for the entire watershed. It is much more reasonable to consider a group of correlated state processes driving precipitation over different parts of the basin. The state processes are connected by means of a Gaussian copula which allows the generation of correlated states across the basin. Emissions at any given location depend on exactly one of the state processes. This model increases the state space of the latent process that is driving precipitation over an area; however, it does so in a way which does not affect the number of parameters in the emission process. Since it is not easy to evaluate the CDF of a Markov chain, we used line-search to estimate the copula correlation parameters.

How to generate correlated precipitation amounts over a large area?

Developing Gaussian copulas for each state's emissions which can then simulate correlated precipitation amounts [Majumder et al., 2020]. A separate copula is constructed for each state and cluster combination. Locations in different clusters of the LHMM are assumed to have zero correlation, significantly lowering the number of pairwise correlations that need to

be estimated. The copula has a two-stage estimation process, where the marginal parameters are estimated in the first stage and the copula parameters estimated in the second stage. While in our simulation studies we assumed that most of the true marginal parameters are known, this is of course not the case when it comes to real life data. The use of marginal estimates results in underestimating the copula correlation values; as a result, synthetic data generated from the fitted model tended to have lower correlations than in historical data. However, not using a copula for emissions results in synthetic data which has very little correlation between locations and cannot be used for understanding the long term properties of precipitation over the entire basin. Since multivariate specifications of semi-continuous mixture marginals are not a well-defined concept, we believe copulas to be the way forward when developing correlation structures for HMMs with arbitrary exponential family emission distributions.

6.2 Future Work

- **Non-homogeneous hidden Markov models (NHMM)**. The HMMs that are considered in this thesis are homogeneous, which means that its parameters are not time dependent. However, it is common in precipitation literature to relax this assumption. Since the latent variables are assumed to be local variables, its parameters can be made to vary over time. Robertson et al. [2006] have done this for the state process parameters, while Holsclaw et al. [2016] made the mixture assignment variable change over time. Neither of these approaches have employed VB parameter estimation as far as we know. In both these examples, the latent variables are categorical and this results in a generalized linear regression problem. A Bayesian setup of this, however, cannot take advantage of conjugacy. This makes it challenging to implement VB parameter estimation. Jordan et al. [1999] provides some approaches which we hope to explore further.

- **Estimating copula parameters simultaneously with marginal parameters using VB**. Our copula is likelihood based to a large degree at this point and its parameters are estimated only after the marginal parameters have been estimated. However, being able to estimate all the parameters simultaneously could provide better copula parameter estimates. Doing so in

a variational context would require incorporating the copula likelihood into the expressions for the ELBO, introducing more structure in the variational posterior. Tran et al. [2015] proposed copula variational inference for estimating correlations between latent variables using copulas. While it is difficult to implement for HMMs due to the modified VBE step, we want to develop a similar approach for the emission process parameters. Finally, Grazian and Liseo [2017] provides some novel approaches for the Bayesian estimation of copula parameters which perform better than the IFM approach.

- **Bayesian Deep Learning for geostatistical data**. We want to expand the scope of our current model and look at Deep Learning models. Variational Autoencoders (VAE) is a class of models we want to explore. Generative Adverserial Networks (GAN) have also found success in generating synthetic data. There are almost always complications when we are trying to explore a spatiotemporal latent space. However, we believe that this is a crucial problem we need to tackle if we want to deploy synthetic precipitation generator models at scale and at arbitrary resolutions.

Bibliography

K. Aas, C. Czado, A. Frigessi, and H. Bakken. Pair-copula constructions of multiple dependence. *Insur. Math. Econ.*, 44(2):182–198, 2009.

S. Amari. Natural gradient works efficiently in learning. *Neural Comput.*, 10(2):251–276, 1998.

H. Attias. Inferring parameters and structure of latent variable models by variational Bayes. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, UAI'99, page 21–30. Morgan Kaufmann Publishers Inc., 1999.

L. E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. In *Inequalities III: Proceedings of the Third Symposium on Inequalities*, pages 1–8, University of California, Los Angeles, 1972.

L. E. Baum and J. A. Eagon. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bull. Amer. Math. Soc.*, 73(3):360–363, 1967.

L. E. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Statist.*, 37(6):1554–1563, 1966.

L. E. Baum and G. R. Sell. Growth transformations for functions on manifolds. *Pacific J. Math.*, 27(2):211–227, 1968.

L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.*, 41(1):164–171, 1970.

M. J. Beal. Variational algorithms for approximate Bayesian inference. Ph.D. Thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

E. Bellone. Nonhomogeneous hidden Markov models for downscaling synoptic atmospheric patterns to precipitation amounts. Ph.D. Thesis, Department of Statistics, University of Washington, 2000.

E. Bellone, J. Hughes, and P. Guttorp. A hidden Markov model for downscaling synoptic atmospheric patterns to precipitation amounts. *Clim. Res.*, 15(1):1–12, 2000.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *J. Am. Stat. Assoc.*, 112(518):859–877, 2017.

R. J. Boys and D. A. Henderson. A Bayesian approach to DNA sequence segmentation. *Biometrics*, 60(3):573–581, 2004.

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

E. C. Brechmann, C. Czado, and K. Aas. Truncated regular vines in high dimensions with application to financial data. *Can. J. Stat.*, 40(1):68–85, 2012.

K. Breinl, G. Di Baldassarre, M. Girons Lopez, M. Hagenlocher, G. Vico, and A. Rutgersson. Can weather generation capture precipitation patterns across different climates, spatial scales and under data scarcity? *Sci. Rep.-UK*, 7, 2017.

O. Cappé, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models (Springer Series in Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2005.

Chesapeake Bay Program. Climate change, 2012. Accessed September 16, 2020, from https://www.chesapeakebay.net/issues/climate_change.

W.-K. Ching, X. Huang, M. K. Ng, and T.-K. Siu. *Markov Chains: Models, Algorithms, and Applications*. Springer, 2013.

E. Damsleth. Conjugate classes for Gamma distributions. *Scand. J. Stat.*, 2(2):80–84, 1975.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. B*, 39(1):1–22, 1977.

A. B. Dieng, D. Tran, R. Ranganath, J. Paisley, and D. Blei. Variational inference via χ upper bound minimization. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

F. Durante, J. Fernández-Sánchez, and C. Sempi. A topological proof of Sklar's theorem. *Appl. Math. Lett.*, 26(9):945–948, 2013.

N. Foti, J. Xu, D. Laird, and E. Fox. Stochastic variational inference for hidden Markov models. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Z. Ghahramani and M. J. Beal. Propagation algorithms for variational Bayesian learning. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS'00, page 486–492. MIT Press, 2000.

Z. Ghahramani and M. I. Jordan. Factorial hidden Markov models. In *Machine Learning*. MIT Press, 1996.

C. Grazian and B. Liseo. Approximate Bayesian inference in semiparametric copula models. *Bayesian Anal.*, 12(4):991–1016, 2017.

A. M. Greene, A. W. Robertson, and S. Kirshner. Analysis of Indian monsoon daily rainfall on subseasonal to multidecadal time-scales using a hidden Markov model. *Q. J. Roy. Meteor. Soc.*, 134(633):875–887, 2008.

J. J. Hamman, B. Nijssen, T. J. Bohn, D. R. Gergel, and Y. Mao. The Variable Infiltration Capacity model version 5 (VIC-5): infrastructure improvements for new applications and reproducibility. *Geosci. Model Dev.*, 11(8):3481–3496, 2018.

J. Hernandez-Lobato, Y. Li, M. Rowland, T. Bui, D. Hernandez-Lobato, and R. Turner. Black-box alpha divergence minimization. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1511–1520, New York, New York, USA, 2016.

M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *J.Mach. Learn. Res.*, 14(4):1303–1347, 2013.

T. Holsclaw, A. M. Greene, A. W. Robertson, and P. Smyth. A Bayesian hidden Markov model of daily precipitation over South and East Asia. *J. Hydrometeorol.*, 17(1):3–25, 2016.

G. J. Huffman, E. F. Stocker, D. T. Bolvin, E. J. Nelkin, and J. Tan. GPM IMERG final precipitation L3 1 day 0.1 degree $\times$ 0.1 degree V06, 2019. Edited by Andrey Savtchenko, Greenbelt, MD, Goddard Earth Sciences Data and Information Services Center (GES DISC), https://disc.gsfc.nasa.gov/datasets/GPM_3IMERGDF_06/summary, accessed on Aug 28, 2020.

J. P. Hughes and P. Guttorp. Incorporating spatial dependence and atmospheric data in a model of precipitation. *J. Appl. Meteorol.*, 33:1503–1515, 1994.

Q. Ji. Computational methods for hidden Markov models with applications. Ph.D. Thesis, Department of Mathematics and Statistics, University of Maryland, Baltimore County, 2019.

S. Ji, B. Krishnapuram, and L. Carin. Variational Bayes for continuous hidden Markov models and its application to active learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:522–532, 2006.

H. Joe and J. J. Xu. The estimation method of inference functions for margins for multivariate models. Technical Report No. 166, Department of Statistics, University of British Columbia, Vancouver, 1996.

M. J. Johnson and A. S. Willsky. Stochastic variational inference for Bayesian time series models. In *ICML*, pages 1854–1862, 2014.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, 1999.

B.-H. Juang, S. Levinson, and M. Sondhi. Maximum likelihood estimation for multivariate mixture observations of Markov chains. *IEEE T. Inform. Theory*, 32(2):307–309, 1986.

S. Kirshner. Modeling of multivariate time series using hidden Markov models. Ph.D. Thesis, University of California, Irvine, 2005.

S. Kirshner, P. Smyth, and A. W. Robertson. Conditional Chow-Liu tree structures for modeling discrete-valued vector time series. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, page 317–324. AUAI Press, 2004.

G. C. Kroiz, J. N. Basalyga, U. Uchendu, R. Majumder, C. A. Barajas, M. K. Gobbert, K. Markert, A. Mehta, and N. K. Neerchal. Stochastic precipitation generation for the Potomac river basin using hidden Markov models. Technical Report HPCF–2020–11, UMBC High Performance Computing Facility, University of Maryland, Baltimore County, 2020a. URL http://hpcf.umbc.edu.

G. C. Kroiz, R. Majumder, M. K. Gobbert, N. K. Neerchal, K. Markert, and A. Mehta. A hidden Markov model with correlated emissions for daily precipitation generation for the Potomac river basin. *Proc. Appl. Math. Mech. (PAMM)*, 20(1):e202000117, 2020b.

W. H. Kruskal. Ordinal measures of association. *J. Am. Stat. Assoc.*, 53(284):814–861, 1958.

A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. Automatic differentiation variational inference. *J. Mach. Learn. Res.*, 18(1):430–474, 2017.

S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Stat.*, 22(1): 79–86, 1951.

Y. Li and R. E. Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

L. Liporace. Maximum likelihood estimation for multivariate observations of Markov sources. *IEEE T. Inform. Theory*, 28(5):729–734, 1982.

D. J. C. MacKay. Ensemble learning for hidden Markov models. Technical report, Department of Physics, University of Cambridge, 1997.

M. Maechler. *Rmpfr: R MPFR - Multiple Precision Floating-Point Reliable*, 2021. URL https://CRAN.R-project.org/package=Rmpfr. R package version 0.8-5.

R. Majumder, R. Walid, J. Zheng, C. Barajas, P. Guo, C. Rajapakshe, A. Gangopadhyay, M. K. Gobbert, J. Wang, Z. Zhang, K. Markert, A. Mehta, and N. K. Neerchal. Assessing water budget sensitivity to precipitation forcing errors in Potomac river basin using the VIC hydrologic model. Technical Report HPCF–2019–11, UMBC High Performance Computing Facility, University of Maryland, Baltimore County, 2019. URL http://hpcf.umbc.edu.

R. Majumder, A. Mehta, and N. K. Neerchal. Copula-based correlation structure for multivariate emission distributions in hidden Markov models. In *JSM Proceedings, Section on Statistics and the Environment*. VA: American Statistical Association, 2020.

R. Majumder, M. K. Gobbert, A. Mehta, and N. K. Neerchal. Variational Bayes estimation of hidden Markov models for daily precipitation with semi-continuous emissions. Technical Report HPCF–2021–8, UMBC High Performance Computing Facility, University of Maryland, Baltimore County, 2021.

R. Majumder, M. K. Gobbert, and N. K. Neerchal. A modified minibatch sampling method for parameter estimation in hidden Markov models using stochastic variational Bayes. *Proc. Appl. Math. Mech. (PAMM)*, accepted (2021).

R. Majumder, N. K. Neerchal, M. K. Gobbert, and A. Mehta. Parameter estimation using stochastic variational bayes for hidden Markov models with semi-continuous emission distributions, in preparation (2021)a.

R. Majumder, N. K. Neerchal, and A. Mehta. A linked hidden Markov model for daily precipitation generation over the chesapeake bay watershed, in preparation (2021)b.

C. McGrory and D. Titterington. Variational approximations in Bayesian model selection for finite mixture distributions. *Comput. Stat. Data An.*, 51(11):5352–5367, 2007.

C. A. McGrory and D. M. Titterington. Variational Bayesian analysis for hidden Markov models. *Aust. NZ J. Stat.*, 51(2):227–244, 2009.

M. Mhanna and W. Bauwens. A stochastic space-time model for the generation of daily rainfall in the Gaza Strip. *Int. J. Climatol.*, 32:1098–1112, 2012.

R. B. Miller. Bayesian analysis of the two-parameter Gamma distribution. *Technometrics*, 22(1):65–69, 1980.

C. Naesseth, F. Lindsten, and D. Blei. Markovian score climbing: Variational inference with KL(p || q). In *Advances in Neural Information Processing Systems*, volume 33, pages 15499–15510. Curran Associates, Inc., 2020.

R. B. Nelsen. *An Introduction to Copulas*. Springer Publishing Company, Incorporated, 2010.

D. Pati, A. Bhattacharya, and Y. Yang. On statistical optimality of variational Bayes. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1579–1588, 2018.

S. K. Popuri. Prediction methods for semi-continuous data with applications in climate science. Ph.D. Thesis, Department of Mathematics and Statistics, University of Maryland, Baltimore County, 2019.

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

R. Ranganath, S. Gerrish, and D. M. Blei. Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 814–822, Reykjavik, Iceland, 2014. PMLR.

H. Robbins and S. Monro. A stochastic approximation method. *Ann. Math. Stat.*, 22(3): 400–407, 1951.

A. W. Robertson, S. Kirshner, and P. Smyth. Downscaling of daily rainfall occurrence over Northeast Brazil using a hidden Markov model. *J. Climate*, 17:4407–4424, 2004.

A. W. Robertson, S. Kirshner, P. Smyth, S. P. Charles, and B. C. Bates. Subseasonal-to-interdecadal variability of the Australian monsoon over North Queensland. *Q. J. Roy. Meteor. Soc.*, 132:519–542, 2006.

T. Rydén, T. Teräsvirta, and S. Åsbrink. Stylized facts of daily return series and the hidden Markov model. *J. Appl. Econom.*, 13(3):217–244, 1998.

S. L. Scott. Bayesian methods for hidden Markov models. *J. Am. Stat. Assoc.*, 97(457): 337–351, 2002.

R. Serfozo. *Basics of Applied Stochastic Processes*. Springer, 2009.

M. Sklar. Fonctions de repartition an dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, 8:229–231, 1959.

D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. Van Der Linde. Bayesian measures of model complexity and fit. *J. R. Stat. Soc. B*, 64(4):583–639, 2002.

D. Tran, D. M. Blei, and E. M. Airoldi. Copula variational inference. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 3564–3572, Cambridge, MA, USA, 2015. MIT Press.

C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Stat. Sinica*, 21(1):5–42, 2011.

S. Venturini, F. Dominici, and G. Parmigiani. Gamma shape mixtures for heavy-tailed distributions. *Ann. Appl. Stat.*, 2(2):756–776, 2008.

A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE T. Inform. Theory*, 13(2):260–269, 1967.

Y. Wang and D. M. Blei. Frequentist consistency of variational Bayes. *J. Am. Stat. Assoc.*, 114(527):1147–1161, 2019.

H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. URL https://ggplot2.tidyverse.org.

D. S. Wilks. Multisite generalization of a daily stochastic precipitation generation model. *J. Hydrol.*, 210(1–4):178–191, 1998.

Y. Yang, D. Pati, and A. Bhattacharya. α-variational inference with statistical guarantees. *Ann. Statist.*, 48(2):886–905, 2020.

H. Yu. Rmpi: Parallel statistical computing in r. *R News*, 2(2):10–14, 2002. URL https://cran.r-project.org/doc/Rnews/Rnews_2002-2.pdf.

A. Y. Zhang and H. H. Zhou. Theoretical and computational guarantees of mean field variational inference for community detection, 2017.

F. Zhang and C. Gao. Convergence rates of variational posterior distributions. *Ann. Statist.*, 48(4):2180–2207, 2020.