Variational Bayes Estimation of Hidden Markov Models for Daily Precipitation with Semi-Continuous Emissions

Reetam Majumder[1], Matthias K. Gobbert[1], Amita Mehta[2], Nagaraj K. Neerchal[1]

[1]Department of Mathematics and Statistics, University of Maryland, Baltimore County, USA

[2]Joint Center for Earth Systems Technology, University of Maryland, Baltimore County, USA

Technical Report HPCF–2021–8, `hpcf.umbc.edu` > Publications

Abstract

Stochastic precipitation generators 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. Daily precipitation is specified as a semi-continuous distribution with a point mass at zero and a mixture of Exponential or Gamma distributions for positive precipitation. Our generators are obtained as hidden Markov models (HMM) where the underlying climate conditions form the states. Maximum likelihood estimation for HMMs has historically relied on the Baum-Welch algorithm. We implement variational Bayes as an alternative for parameter estimation in HMMs. In our simulation study for a 3-state HMM with positive rainfall specified as a mixture of 2 Exponential distributions, we get good posterior estimates when the model is initialized with the correct number of states and mixture components. We also fit a similar model to a single grid point within the Chesapeake Bay watershed based on GPM-IMERG remote sensing data for the wet season between July to September from 2000–2019. Synthetic data generated from the fitted model is able to replicate the monthly proportion of dry days at the location, as well as the total monthly precipitation.

1 Introduction

The modeling and forecasting of seasonal and inter-annual variation in precipitation is used to determine water allocation and resource management for regions dependent on precipitation as a primary water source. To this end, precipitation generators are constructed to produce time series of synthetic data representative of the general rainfall patterns within the region. In particular, stochastic precipitation generators aim to replicate key statistical properties of the historical data like dry and wet stretches, spatial correlations, and extreme weather events. Stochastic precipitation generators are used to downscale numerical weather models

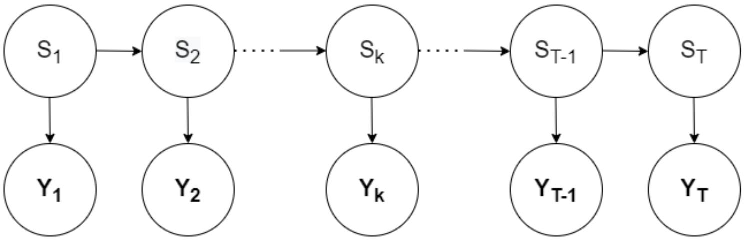

Figure 1: Graphical representation of a hidden Markov model.

and for climate projections, flood and drought assessments, and in studies pertaining to agriculture, food security, as well as public and veterinary health (Breinl et al., 2017).

Hidden Markov models (HMM) is one common approach for building a stochastic weather generator. An HMM is a doubly stochastic process $\{S_k, Y_k\}_{k\geq 0}$ where $\{S_k\}$ is a Markov chain, and conditional on it, $\{Y_k\}$ is a sequence of independent random variables such that the distribution of Y_k depends only on S_k. However, $\{S_k\}$ is unobservable, and only $\{Y_k\}$ is observed. $\{Y_k\}$ could be distributed as a discrete, continuous, or a mixture distribution, and could be multivariate. $\{S_k\}$ is known as the state process and $\{Y_k\}$ is called the emission process. A graphical representation of an HMM is shown in Figure 1.

HMMs have been applied to a wide class of problems such as speech recognition (Rabiner, 1989), DNA sequencing (Boys and Henderson, 2004), finance (Rydén et al., 1998), and precipitation modeling (Hughes and Guttorp, 1994). In modeling precipitation as an HMM, the underlying climate conditions form the state space. Furthermore, daily precipitation data usually contains a large number of zeros corresponding to dry days. We therefore consider a semi-continuous emission distribution with a point mass at 0 for no rainfall and multiple exponential distributions for positive rainfall. Gamma distributions have also been used as alternatives (Bellone et al., 2000). The groundwork for modeling precipitation using HMMs was laid out in Hughes and Guttorp (1994), and extended by Robertson et al. (2006). Parameter estimation for HMMs has historically relied on the Baum-Welch algorithm, a modification of the Expectation Maximization (EM) algorithm, which provides maximum likelihood estimates. However, maximum likelihood methods for graphical models like HMMs can lead to overfitting and tractability problems (Attias, 1999). Bayesian approaches are often used to counter these issues; while Markov chain Monte Carlo (MCMC) methods use sampling to find the posterior distribution, variational Bayes (VB) uses optimization to obtain an approximate posterior. The approximate posterior is referred to as the variational posterior, and is computed using an iterative EM-like algorithm which always converges (Attias, 1999). The variational posteriors have analytical forms and can be used to perform Bayesian inference. (Blei et al., 2017) provides a review of variational inference, and work focusing on the theoretical properties of the posteriors can be found in Zhang and Gao (2020); Wang and Blei (2019); Pati et al. (2018); Yang et al. (2020).

The rest of this article is organized as follows. A brief outline of variational inference is presented in Section 2. Section 3 introduces the HMM for precipitation at a single location with a semi-continuous emission distribution and its parameter estimation using variational

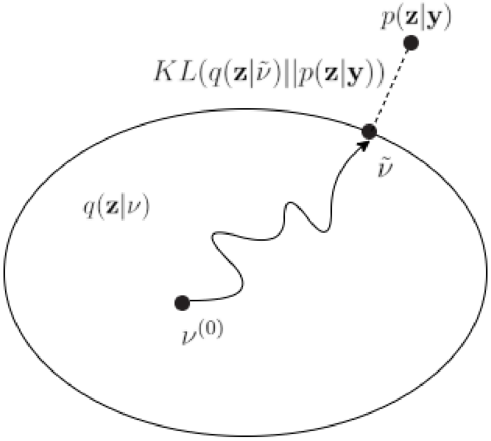
$p(\mathbf{z}|\mathbf{y})$
$KL(q(\mathbf{z}|\tilde{\nu})||p(\mathbf{z}|\mathbf{y}))$
$\tilde{\nu}$
$q(\mathbf{z}|\nu)$
$\nu^{(0)}$

Figure 2: Variational inference represented as an optimization problem

Bayes. Section 4 contains a simulation study, as well as a study for daily precipitation at a single grid point within the Chesapeake Bay watershed based on remote sensing data. Section 5 provides concluding remarks and future directions.

2 Variational Bayes for Approximating the Posterior Distribution

The variational Bayes (VB) algorithm provides an alternative to MCMC methods by approximating the posterior of the latent variables and parameters with a family of distributions $\mathbb{Q}$, referred to as the variational family. $\mathbb{Q}$ is indexed by its own variational parameters, and we look for the best approximation to the true posterior within $\mathbb{Q}$. The Kullback-Liebler (KL) divergence serves as a measure of how well the variational posterior approximates the true posterior, and the optimum satisfies

$$\tilde{q}(\cdot) = \arg\min_{q(\cdot)\in\mathbb{Q}} KL\big(q(\boldsymbol{z}) \parallel p(\boldsymbol{z}|\boldsymbol{y})\big). \tag{2.1}$$

The optimum values of the hyperparameters are found using an EM-like algorithm; Figure 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 (2.1) involves the posterior $p(\boldsymbol{z}|\boldsymbol{y})$ which is often difficult to compute in practice. However, the minimization in (2.1) is equivalent to the maximization of a quantity known as the evidence lower bound (ELBO), defined as

$$\text{ELBO}(q) = \mathbb{E}[\log p(\boldsymbol{z}, \boldsymbol{y})] - \mathbb{E}[\log q(\boldsymbol{z})]. \tag{2.2}$$

The equivalence follows from the identity

$$\log p(\boldsymbol{y}) = KL(q(\boldsymbol{z}) \parallel p\big(\boldsymbol{z}|\boldsymbol{y})\big) + \text{ELBO}(q), \tag{2.3}$$

and the fact that the evidence log-likelihood $\log p(\boldsymbol{y})$ is a function of only the data and therefore fixed. All expectations in (2.3) 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 inequality 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. Ji et al. (2006) have used the VB algorithm on HMMs where the emissions are continuous mixtures, and McGrory and Titterington (2009) have discussed model selection in variational HMMs using the Deviance Information Criterion (DIC) when the size of the model is unknown.

3 VB-HMM for Univariate Semi-Continuous Emissions

We now develop VB estimation for HMMs with a univariate semi-continuous emission distribution. We refer to it as a VB-HMM and will use it to estimate the parameters of an HMM for precipitation at a single location. Let $y_{1:T} = \{y_1, \dots, y_T\}$ be the precipitation time series of length T, with $y_t \geq 0$. The data is generated by a set of underlying hidden states $s_{1:T} = \{s_1, \dots, s_t, \dots, s_T\}$, where each state $s_t \in \{1, \dots, K\}$. Further, for each state j we define an indicator variable to connect the underlying state to the emission distribution:

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

where $r_{tj} = (r_{tj0}, r_{tj1}, \dots, r_{tjm})$ is encoded as a *one-hot* vector with r_{tj0} indicating no-rainfall events. We assume that the number of states (K) and mixture components (M+1) in the HMM are known. For each state j, r_{tj} follows a categorical distribution which corresponds to a single draw from a multinomial distribution, given by

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

where $p_j(\cdot|\cdot) \equiv p(\cdot|\cdot, s_t = j)$ corresponds to the distribution for state j, $c_j = (c_{j0}, \dots, c_{jM})$ are the 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

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

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 transition 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 independently distributed rate parameters of the exponential distributions which are part of the semi-continuous emissions in each state. 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. That is,

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

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

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

where $a_j = (a_{j1}, \ldots, a_{jK})$, $\pi_1 = (\pi_{11}, \ldots, \pi_{1K})$, $\zeta_j^{(0)} = (\zeta_{j0}^{(0)}, \ldots, \zeta_{jM}^{(0)})$, $\alpha_j^{(0)} = (\alpha_{j1}^{(0)}, \ldots, \alpha_{jK}^{(0)})$, 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 hyperparameters $(\gamma_j^{(0)}, \delta_j^{(0)}, \zeta_j^{(0)}, \alpha_j^{(0)}, \xi^{(0)})$ are known.

The complete data likelihood can be expressed as

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

where $s_{tj} = \mathbb{I}\{s_t = j\}$ denotes the daily state and $s_{tj}s_{t+1,k}$ denotes a typical state transition. Similarly, we write the prior as

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

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

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

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

Note that $q_\Theta(\Theta)$ and $q_{s,r}(s, r)$ are coupled in (3.6), and the optimization problem in (2.1) cannot be solved analytically. Instead, it is solved numerically by iteratively optimizing $q_\Theta(\Theta)$ and $q_{s,r}(s, r)$ using a variational Bayesian generalization to the EM algorithm (VBEM). The results in MacKay (1997) and Ghahramani and Beal (2000) for belief networks form the basis of our approach. In the variational M-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 we assume $q_\Theta(\Theta)$ to decompose as in (3.7), each of its components can be updated individually. In the variational E-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. McGrory and Titterington (2009) have described the variational Forward-Backward algorithm for univariate Normal distribution emissions. We lay out the steps for our VBEM algorithm which closely follow the estimation for mixture marginal emissions described in Rabiner (1989).

Variational M-step (VBM): *With the variational posteriors on hidden variables fixed at $q_{s,r}(s,r)$, update the variational posterior $q_\Theta(\Theta)$ on the model parameters.*

Since $q_\Theta(\Theta)$ is conjugate to the prior, the posterior distribution for each component of $\phi(\Theta)$ in (3.5) 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 $m = 1,\ldots,M$.

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

The variational posterior $q_{s,r}(s,r)$ has the same form as the known parameter posterior, i.e.

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

with the natural parameters $\phi(\Theta)$ replaced by their expectations under $q_\Theta(\Theta)$. Comparing with (3.3), we get

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

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

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

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

$$\begin{aligned} c^*_{jm} &= \exp\big\{\mathbb{E}_Q \log c_{jm}\big\} = \exp\big\{\Psi(\zeta_{jm}) - \Psi(\zeta_{j.})\big\}, \text{ where } \zeta_{j.} = \sum_{m=0}^{M} \zeta_{jm}, \\ \lambda^*_{jm} &= \exp\big\{\mathbb{E}_Q \log \lambda_{jm}\big\} = \exp\big\{\Psi(\gamma_{jm}) - \log \delta_{jm}\big\}, \\ \hat\lambda_{jm} &= \mathbb{E}_Q \lambda_{jm} = \gamma_{jm}/\delta_{jm}. \end{aligned}$$

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

Here a^*_{1j} estimates the initial state probabilities, a^*_{jk} estimates the transition probabilities from state j to state k, and $b^*_{tj} = \sum_{m=0}^M b^*_{tjm}$ estimates the emission probability distribution conditional on the system being in state j at time t. They can now be used as part of the Forward-Backward algorithm described in Appendix A 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

$$\begin{aligned} q_{1j} &= a^*_1, \\ q_{tj} &= \frac{\tilde F_{tj} \cdot \tilde B_{tj}}{\sum_{k=1}^K \tilde F_{tk} \cdot \tilde B_{tk}}, \\ q_{jk} &= \frac{\tilde F_{tj} \cdot a^*_{jk} \cdot b^*_{t+1,k} \cdot \tilde B_{t+1,k}}{\sum_{j=1}^K \sum_{k=1}^K \tilde F_{tj} \cdot a^*_{jk} \cdot b^*_{t+1,k} \cdot \tilde B_{t+1,k}}. \end{aligned}$$

where $\tilde F_{tj}$ and $\tilde B_{tj}$ are the scaled Forward and Backward variable respectively. 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 $c^*_{jm} f(y_t|\lambda^*_{jm}, \hat\lambda_{jm}) = \exp\big\{\Psi(\zeta_{jm}) - \Psi(\zeta_{j.}) + \Psi(\gamma_{jm}) - \log \delta_{jm} - y_t \frac{\gamma_{jm}}{\delta_{jm}}\big\}$.

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 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

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

where the first term on the right hand side is calculated as part of the Forward algorithm in (A.1). 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.

4 Applications for Simulated and Real Data

4.1 Simulation Study

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.

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. 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. Further, we order the exponential distributions within each state as follows:

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

The rate parameter for each of the Gamma priors is set to 2. This assignment follows 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 $m = 0$.

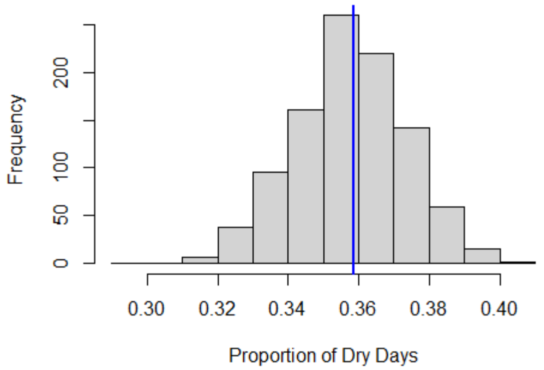

Figure 3: Histogram of the proportion of dry days in 1800 days of data simulated using the estimated parameters from each of the 1000 simulation studies. The blue line at 0.36 denotes the Monte Carlo estimate of the true proportion of dry days.

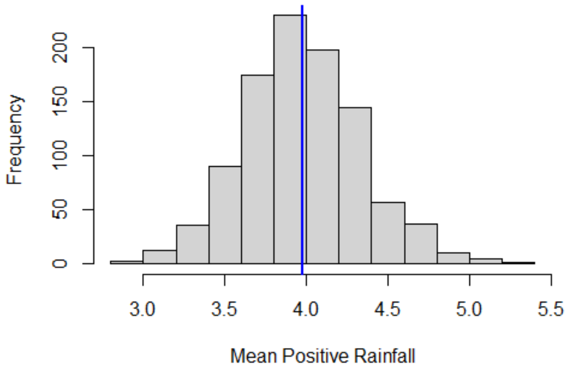

Figure 4: Histogram of mean positive rainfall (mm) in 1800 days of data simulated using the estimated parameters from each of the 1000 simulation studies. The blue line at 3.97 mm denotes the Monte Carlo estimate of the true mean of rainfall on wet days.

We average the posterior estimates obtained from the 1000 simulations of the VB-HMM. The posterior for the initial state probability is $\tilde{\pi}_1 = (0.38, 0.27, 0.35)$. 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 and the exponential rate parameters 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. We also 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 hyperparamters.

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 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 shows a histogram of the monthly proportion of dry days based on 1000 estimates. The blue line is an estimate of the true proportion 0.36. The 1000 estimates have a root mean square error (RMSE) of 0.02. We notice a slight negative skew in our histogram, suggesting the proportion of dry days is being underestimated. Similarly, Figure 4 plots a histogram of mean rainfall for wet days; the blue line at 3.97 mm

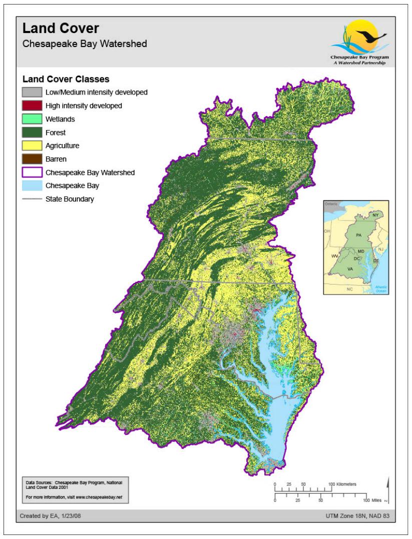

Figure 5: Land cover classes within the Chesapeake Bay watershed in the East coast of the USA.

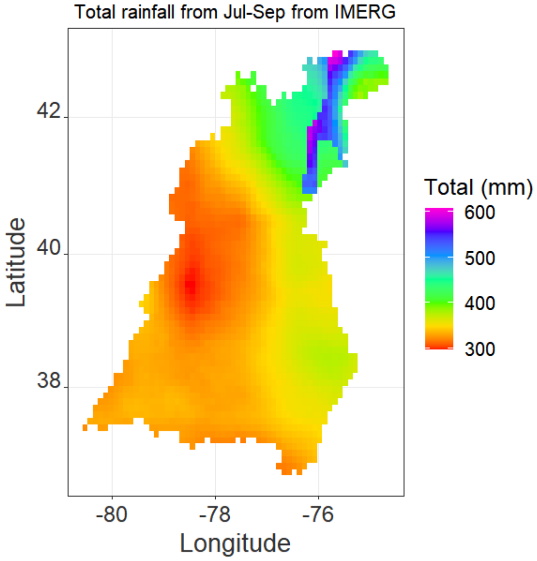

Figure 6: Total rainfall for Jul–Sep over the Chesapeake Bay watershed based on GPM-IMERG data for 2000–2019.

is an estimate of the true mean. The 1000 estimates have an RMSE of 0.36 mm, and have a slight positive skew. However, neither histogram shows noticeable bias.

4.2 Precipitation over the Chesapeake Bay watershed

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 5 shows the watershed and the different land cover classes within it. The watershed has a diverse, interconnected ecosystem which is affected by extreme weather potentially related to climate change (Chesapeake Bay Program, 2012), and has been targeted for restoration as an integrated watershed and ecosystem. Daily remote sensing precipitation data is available for the watershed from the Integrated Multi-satellitE Retrievals for Global Precipitation Measurement (GPM-IMERG) dataset (Huffman et al., 2019), and we focus on the months of July to September for 2000–2019 for our study. With a $0.1° \times 0.1°$ spatial resolution, the IMERG dataset covers the 64,000 square mile watershed with 1927 grid points. Figure 6 depicts seasonal rainfall for the months of Jul–Sep over the watershed based on GPM-IMERG data. We choose a grid point at random within the watershed for parameter estimation using the VB-HMM algorithm, and simulate synthetic data for the location.

The randomly selected grid point is located approximately at 38°16′ N, 76°27′ W. We have found that 4–6 state HMMs with 3 mixture components tend to be sufficient when modeling precipitation over the entire watershed (Majumder et al., 2020). Since we are working with a single location, we fit a 3 state HMM to the data. We assign a symmetric Dirichlet prior to the mixture probabilities with a concentration of 12 for each state. The shape parameters

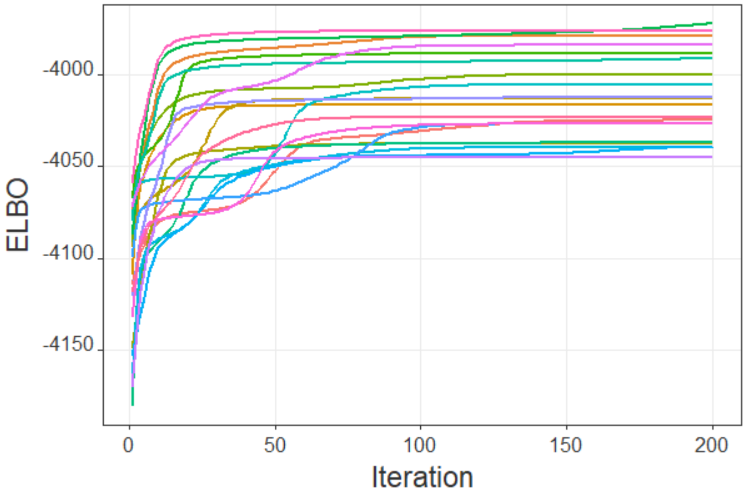

Figure 7: Convergence of the ELBO to local optima for 20 random restarts of the VBEM algorithm for a 3 state HMM with 3 mixture components.

for the Gamma priors are randomly chosen. Within each state, the first shape parameter is sampled from a Uniform(0,1) distribution, and the second shape parameter sampled from a Uniform(1,20) distribution. The remaining priors are identical to what we used in our simulation study.

One difference in our data compared to the simulation study is the presence of breaks. Since we use 3 months of data for 20 years, each year's precipitation is effectively an independent chain generated from the same HMM. We modify our VBEM algorithm to accommodate for this, following the outline in Rabiner (1989). Further, different initializations may lead to different local optima of the ELBO (Blei et al., 2017). Figure 7 demonstrates the convergence of the ELBO for 20 random restarts of the model by plotting the first 200 iterations of the ELBO for each restart. The restarts correspond to different $\gamma^{(0)}$ matrices which contains the shape parameters for the Gamma priors. We choose the solution which converges to the best local optimum in these 20 runs as our final model.

The VB-HMM provides the posterior estimate $\tilde{\pi}_1 = (0.51, 0.15, 0.34)$, and

$$\tilde{A} = \begin{bmatrix} 0.39 & 0.39 & 0.22 \\ 0.32 & 0.49 & 0.19 \\ 0.29 & 0.16 & 0.55 \end{bmatrix} \qquad \tilde{C} = \begin{bmatrix} 0.16 & 0.60 & 0.24 \\ 0.09 & 0.87 & 0.04 \\ 0.92 & 0.01 & 0.07 \end{bmatrix} \qquad \tilde{\Lambda} = \begin{bmatrix} 0.74 & 16.56 \\ 0.09 & 5.10 \\ 0.02 & 11.22 \end{bmatrix}$$

We see that the third state is the driest, with a probability 0.92 of zero precipitation. The second state is the wettest with a probability 0.87 of precipitation from an exponential distribution with a rate parameter of 0.09. The first state is an intermediate state which would have lower rainfall than state 2 and fewer dry days compared to state 3. It is also the initial state for our HMM.

We simulated 20 years of synthetic precipitation from this fitted model to compare with the historical IMERG data for 2000–2019. Our synthetic data also contains 3 months of data for each year. Figure 8 compares the distribution of the proportion of dry days for each

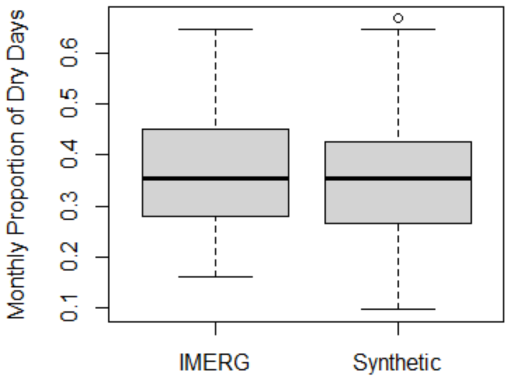

Figure 8: Boxplots of the monthly proportion of dry days in historical IMERG data (2000–2019) and synthetic HMM data.

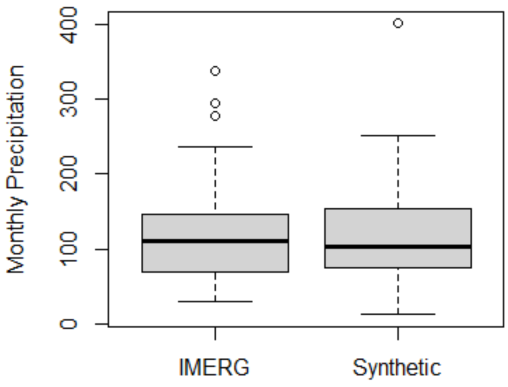

Figure 9: Boxplots of total monthly precipitation in historical IMERG data (2000–2019) and synthetic HMM data.

month in our IMERG and synthetic data. Even though both distributions have the same median of 0.3548, the synthetic data tends to underestimate the monthly proportion of dry days. Similarly, Figure 9 compares the distribution of monthly precipitation for IMERG and our synthetic data. The IMERG data has a median monthly precipitation of 110 mm, whereas the synthetic data has a median of 102 mm. In both figures, we notice that the variability in the synthetic data is higher than the IMERG data. The code and datasets for the simulation study as well as the HMM for IMERG data over the Chesapeake Bay watershed can be found at https://github.com/reetamm/VB-HMM.

5 Conclusions

The VB-HMM for precipitation at a single location can estimate the true parameters under general prior specifications. In particular, we found that as long as we assume an ordering of the positive rainfall 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 very low concentration or is asymmetric, it will dominate in the posterior.

In our preliminary study for remote sensing precipitation data at a single grid point over the Chesapeake Bay watershed, our estimated parameters showed clear dry, wet, and intermediate states, even when the model is initialized with very general priors. Further, synthetic data simulated using the estimated HMM parameters were able to replicate monthly precipitation as well as monthly proportions of dry days.

Future work will focus on HMMs with 4–7 states which are quite common in precipitation modeling. Under this setup, we want to explore the algorithm's sensitivity to initial values. Rabiner (1989) notes that the Baum-Welch algorithm is sensitive to the initial values of the emission distribution parameters, and suggests using a combination of Viterbi decoding (Viterbi, 1967) and k-means clustering to assign a state and a mixture component to each observation while initializing the algorithm. We believe a similar approach would benefit the VB-HMM as well. Additionally, we want to relax the assumption of known K and M, and initialize the model with a larger number of states or mixture components compared to the true model. Variational methods have been shown to identify the correct number of states in HMMs (McGrory and Titterington, 2009), and the correct number of mixture components for mixture models (McGrory and Titterington, 2007), but has not been tested for HMMs with semi-continuous emissions. The Deviance Information Criterion (DIC) can be used for model selection in such cases.

Finally, we will use precipitation over the Chesapeake Bay watershed during the rainy season as our demonstrative example for modeling multi-site rainfall. The current model specification extends naturally to multi-site precipitation, where precipitation at each location is considered independently distributed conditional on the daily state. EM based methods tend to underestimate spatial correlations for such models when working with remote sensing data over large areas, and Gaussian copulas can be used to capture the spatial correlation within the data (Majumder et al., 2020) if the VB-HMM cannot capture the spatial correlation adequately. Additionally, we will focus on efficient and scalable computation by replacing our current coordinate ascent optimization with stochastic gradient methods, and using the message passing interface (MPI) protocol for parallelizing our code wherever possible.

Acknowledgements

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. Reetam Majumder was supported by JCET and as HPCF RA.

Appendices

Appendix A The Forward-Backward Algorithm for VB

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

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

It is calculated for every time point using recursion. To prevent underflow errors, we scale the Forward Variable at every step. Rabiner (1989) has shown that scaling at each step is equivalent to scaling the entire sequence by the sum of all states at the end.

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

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

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

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

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

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

where $q(y|\tilde{\Theta})$ is the normalizing constant for the variational posterior $q_{s,r}(s, r)$ in (3.8).

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. Additionally, we use the same scaling factors that we derived in the Forward Algorithm.

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

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

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

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

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

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

References

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.

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

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.

K. Breinl, G. D. Baldassarre, M. G. 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.

Chesapeake Bay Program. Climate change, 2012. Accessed September 16, 2020, from https://www.chesapeakebay.net/issues/climate_change.

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.

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 × 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.

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.

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.

D. J. C. MacKay. Ensemble learning for hidden Markov models. Technical report, Department of Physics, University of Cambridge, 1997.

R. Majumder, A. Mehta, and N. K. Neerchal. Copula-based correlation structure for multivariate emission distributions in hidden Markov models. *JSM Proceedings, Section on Statistics and the Environment. Alexandria, VA: American Statistical Association*, 2020.

C. A. McGrory and D. M. 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.

D. Pati, A. Bhattacharya, and Y. Yang. On statistical optimality of variational Bayes. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1579–1588, 2018.

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

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.

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.

Y. Yang, D. Pati, and A. Bhattacharya. α-variational inference with statistical guarantees. *Ann. Statist.*, 48(2):886–905, 2020.

F. Zhang and C. Gao. Convergence rates of variational posterior distributions. *Ann. Statist.*, 48(4):2180–2207, 2020.