Variational Bayes Estimation of Hidden Markov Models for Daily Precipitation at a Single Location

Reetam Majumder[1], Nagaraj K. Neerchal[1,2], Amita Mehta[3]

[1]Department of Mathematics and Statistics, University of Maryland, Baltimore County, USA
[2]Chinmaya Vishwavidyapeeth, Kerala, India
[3]Joint Center for Earth Systems Technology, University of Maryland, Baltimore County, USA

Abstract

Stochastic precipitation generators are able to 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 Gamma or Exponential 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.

1 VB-HMM for Univariate Semi-Continuous 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 $s_t \in \{1, \ldots, K\}$. Further, for each state j we define an indicator variable to connect the underlying state to the emission distribution:

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

where $r_{tj} = (r_{tj0}, r_{tj1}, \ldots, r_{tjm})$ 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

$$p_j(r_{tj}|c_j, s_t = j) = \prod_{m=0}^{M} c_{jm}^{r_{tjm}}, \ m = 0, 1, \ldots, M, \tag{1.1}$$

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

$$\begin{aligned} p(y_t, r_{tj}|\lambda_j, c_j, s_t = j) &= p(r_{tj}|c_j, s_t = j) \cdot p(y_t|\lambda_j, r_{tj}, s_t = j) \\ &= c_{j0}^{r_{tj0}} \prod_{m=1}^{M} \left[c_{jm}\lambda_{jm} \exp\{-\lambda_{jm} y_t\}\right]^{r_{tjm}}. \end{aligned} \tag{1.2}$$

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,

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

where $a_j = (a_{j1}, \ldots, a_{jK})$, $\pi_1 = (\pi_{11}, \ldots, \pi_{1K})$, $\zeta_j^{(0)} = (\zeta_{j0}^{(0)}, \ldots, \zeta_{jM}^{(0)})$, $\alpha_j^{(0)} = (\alpha_{j1}^{(0)}, \ldots, \alpha_{jK}^{(0)})$, 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

$$\begin{aligned} p(y, s, r|\Theta) &= \prod_{j=1}^{K} \pi_{1j}^{s_{1j}} \prod_{t=1}^{T}\prod_{j=1}^{K} \{p_j(y_t, r_{tj}|\Theta)\}^{s_{tj}} \prod_{t=1}^{T-1}\prod_{j=1}^{K}\prod_{k=1}^{K} \{a_{jk}\}^{s_{tj}s_{t+1,k}} \\ &= \exp\Bigg\{\sum_{j=1}^{K} s_{1j}\log\pi_{1j} + \sum_{t=1}^{T}\sum_{j=1}^{K}\Bigg[\sum_{m=1}^{M} s_{tj}r_{tjm}(\log c_{jm} + \log\lambda_{jm} - y_t\lambda_{jm}) \\ &\quad + s_{tj}r_{tj0}\log c_{j0}\Bigg] + \sum_{t=1}^{T-1}\sum_{j=1}^{K}\sum_{k=1}^{K} s_{tj}s_{t+1,k}\log a_{jk}\Bigg\}, \end{aligned} \tag{1.3}$$

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

$$
\begin{aligned}
p(\Theta|\nu^{(0)}) &= p(\pi_1)\cdot p(\lambda)\cdot p(C)\cdot p(A) \\
&= \exp\Bigg\{\sum_{j=1}^{K}\big\{(\xi_j^{(0)}-1)\log\pi_{1j} + \sum_{m=1}^{M}\big[-\delta_{jm}^{(0)}\lambda_{jm} + (\gamma_{jm}^{(0)}-1)\log\lambda_{jm}\big] \\
&\quad + (\zeta_{j0}^{(0)}-1)\log c_{j0} + \sum_{m=1}^{M}(\zeta_{jm}^{(0)}-1)\log c_{jm} + \sum_{k=1}^{K}(\alpha_{jk}^{(0)}-1)\log a_{jk}\big\} - \log h^{(0)}\Bigg\},
\end{aligned}
\tag{1.4}
$$

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 $\nu^{(0)}$:

$$
\phi(\Theta) = \begin{bmatrix} \log\pi_{1j} \\ \log c_{j0} \\ \log c_{jm} \\ \log\lambda_{jm} \\ \lambda_{jm} \\ \log a_{jk} \end{bmatrix}
\qquad
u(s,y,r) = \begin{bmatrix} s_{1j} \\ s_{tj}r_{tj0} \\ s_{tj}r_{tjm} \\ s_{tj}r_{tjm} \\ y_t s_{tj} r_{tjm} \\ s_{tj}s_{t+1,k} \end{bmatrix}
\qquad
\nu^{(0)} = \begin{bmatrix} \xi_j^{(0)}-1 \\ \zeta_{j0}^{(0)}-1 \\ \zeta_{jm}^{(0)}-1 \\ \gamma_{jm}^{(0)}-1 \\ \delta_{jm}^{(0)} \\ \alpha_{jk}^{(0)}-1 \end{bmatrix}
\tag{1.5}
$$

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:

$$
q_z(z) = q_\Theta(\Theta)\cdot q_{s,r}(s,r), \tag{1.6}
$$

$$
\text{where } q_\Theta(\Theta) = q(\pi_1)\cdot q(A)\cdot q(C)\cdot q(\Lambda). \tag{1.7}
$$

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 (1.5) 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 (1.3)

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.

$$q_{s,r}(s,r) \propto \prod_{j=1}^{K} {a_{1j}^*}^{s_{1j}} \prod_{t=1}^{T}\prod_{j=1}^{K}\prod_{m=0}^{M} {b_{tjm}^*}^{s_{tj}r_{tjm}} \prod_{t=1}^{T-1}\prod_{j=1}^{K}\prod_{k=1}^{K} {a_{jk}^*}^{s_{tj}s_{t+1,k}}, \tag{1.8}$$

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

(1.3), we get

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

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

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

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

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

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

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

$$q_{1j} = a^*_1,$$

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

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 $c^*_{jm} f(y_t | \lambda^*_{jm}, \hat{\lambda}_{jm}) = \exp\big\{\Psi(\zeta_{jm}) - \Psi(\zeta_{j\cdot}) + \Psi(\gamma_{jm}) - \log \delta_{jm} - y_t \frac{\gamma_{jm}}{\delta_{jm}}\big\}$.

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 Equations (1.3)–(1.8), 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

$$\begin{aligned} ELBO(q) = \log q(y|\tilde{\Theta}) - KL\big(q(\pi_1) \parallel p(\pi_1)\big) - KL\big(q(A) \parallel p(A)\big) \\ - KL\big(q(C) \parallel p(C)\big) - KL\big(q(\Lambda) \parallel p(\Lambda)\big), \end{aligned} \tag{1.9}$$

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.

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 (1.8).

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

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

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

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

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

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

References

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

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.

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