# Notes on Hidden Markov Models and Kalman Filters

*Hidden Markov models are versatile models for sequential data. The state at a given time encodes information about previous observations, and the next observation is predicted from the state. HMMs are mixture models with a very large state space, and can be fit with the EM algorithm and dynamic programming methods. The Kalman filter, or linear dynamical system, is the analogue of the hidden Markov model for a continous state space and Gaussian distributions.*

## 1. Introduction

The hidden Markov model is an important method that is widely used for modeling data having a natural sequential structure. Examples include speech recognition, gene finding in biological data, and linguistic analysis in natural language text. HMMs were first formally studied by Baum and Petrie (1966), who detailed the dynamic programming calculations needed for efficient inference, and presented results on consistency and asymptotic normality of the maximum likelihood estimator. Work on these models was carried out even earlier in classified US defense work, and the important properties of HMMs were recognized by coding theorists in the late 1960s and early 1970s. In fact, Claude Shannon probably appreciated the importance of HMMs early on; his seminal 1948 paper (Shannon, 1948) comments on graphical representations of Markov processes, describing how the state represents the "residue of influence" of previous observations. A closely related model is the Kalman filter, or linear dynamical system, which can be thought of as a hidden Markov model with a continous state space and Gaussian distributions. We treat both models together in these notes.

## 2. Basic Definitions and Examples

Let $X_1, X_2, X_3, \ldots, X_s, \ldots$ be a stationary Markov chain on $K$ states, with transition probabilities denoted by

$$\psi(a, b) = \mathbb{P}(X_{t+1} = b \,|\, X_t = a). \tag{1}$$

For each state $k = 1, \ldots, K$, let $f(y \,|\, k)$ denote a conditional density over $\mathcal{Y} \subset \mathbf{R}^d$. We let $Y_1, Y_2, Y_3 \ldots, Y_k, \ldots$ denote a $\mathcal{Y}$-valued sequence that is conditionally independent given $\{X_t\}_{t=1}^{\infty}$, and assume that $Y_t$ has a conditional density $f(Y_t \,|\, k)$ with respect to some $\sigma$-finite measure $\nu$ over $\mathcal{Y}$. Thus, the observations $\{Y_t\}$ satisfy

$$p\left(Y_t \,|\, \{X_s\}_{s=1}^{\infty}\right) = f(Y_t \,|\, X_t) \tag{2}$$

It is usually assumed that the conditional densities $f(\cdot \,|\, k)$ are specified by some parametric family

$$f_\theta(y \,|\, k) = f(y; \theta(k)) \tag{3}$$

usually taken to be an exponential family model. The set of parameters $\theta(1), \ldots, \theta(K)$ together with the transition probabilities $\psi(a, b)$ form the parameters $\theta \subset \mathbf{R}^p$ of the HMM. The transition probabilities may be parameterized as well, we denote them as

$$p_\theta(X_t = b \,|\, X_{t-1} = a) = \psi_\theta(a, b) \tag{4}$$

The joint distribution of $X$ and $Y$ is given by

$$p(X_1, X_2, \ldots, X_T; Y_1, Y_2, \ldots, Y_T) = \pi_1(X_1)\, p_\theta(Y_t \mid X_1) \prod_{t=2}^{T} p_\theta(X_t \mid X_{t-1}) p_\theta(Y_t \mid X_t) \tag{5}$$

where $\pi_1$ is a distribution for the first state $X_1$. The marginal density of $\{Y_t\}$ is given by

$$p(Y_1, Y_2, \ldots, Y_T) = \sum_{X_1, X_2, \ldots X_T} \pi_1(X_1)\, p_\theta(Y_1 \mid X_1) \prod_{t=2}^{T} p_\theta(X_t \mid X_{t-1}) p_\theta(Y_t \mid X_t) \tag{6}$$

$$= \sum_{s \in K^T} \lambda_s p_\theta(Y_1, \ldots, Y_T \mid s) \tag{7}$$

where the second expresses the model as a mixture model with $K^T$ components and mixing weights

$$\lambda_{s_1, s_2, \ldots, s_T} = \pi_1(s_1) \prod_{t=2}^{T} p_\theta(s_t \mid s_{t-1}) \tag{8}$$

and mixture components

$$p_\theta(Y_1, \ldots, Y_T \mid s_1, \ldots, s_T) = \prod_{t=1}^{T} p_\theta(Y_t \mid s_t). \tag{9}$$

The conditional distribution of $\{X_t\}$ given $\{Y_t\}$ is given by

$$p(X_1, \ldots, X_T \mid Y_1, \ldots, Y_T) = \frac{p(X_1, \ldots, X_T; Y_1, \ldots, Y_T)}{p(Y_1, \ldots, Y_T)} \tag{10}$$

$$= \frac{\pi_1(X_1)\, p_\theta(Y_t \mid X_1) \prod_{t=2}^{T} p_\theta(X_t \mid X_{t-1}) p_\theta(Y_t \mid X_t)}{p(Y_1, Y_2, \ldots, Y_t)} \tag{11}$$

$$= q_{1,\theta}(X_1 \mid Y) \prod_{t=2}^{T} q_{t,\theta}(X_t \mid X_{t-1}, Y) \tag{12}$$

where $q_{t,\theta}(X_t \mid X_{t-1}, Y) \propto p_\theta(X_t \mid X_{t-1}) p_\theta(Y_t \mid X_t)$.

We summarize these observations as follows:

A hidden Markov model of the form (5) is:

- a homogeneous Markov chain jointly on $(X, Y)$ and marginally on $X$;
- an inhomogeneous Markov chain conditionally on $X$ given $Y$;
- a mixture model marginally on $Y$.

*Example 2.1.* [Speech recognition] In a simple model of speech, we observe an acoustic signal $A_t$, for $t \in [0, 1]$, and the time interval is broken up into "frames"

$$[0, \delta), [\delta, 2\delta), \ldots, [1 - \delta, 1] \tag{13}$$

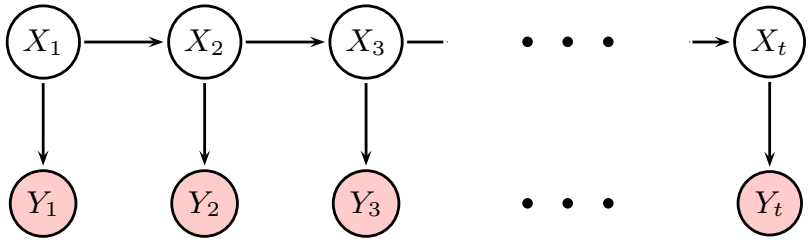


FIG 1. *Directed graph representing a hidden Markov model. The shaded nodes are observed, but the unshaded nodes, representing states in a latent Markov chain, are unobserved.*

The objective is to label each frame by a phonetic label, which can then be further processed into natural language; the states correspond to the phonetic labels. The $t$th frame of speech is processed into a feature vector $Y_t = Y_t(A)$, by various signal processing transformations, for example computing Fourier coefficients and their finite differences from frame to frame.

*Example 2.2.* [Gaussian Mixtures] Suppose that $K = 2$, $\mathcal{Y} = \mathbf{R}$, and

$$p_\theta(Y_t \,|\, k) = \sigma^{-1}\phi\left((Y_t - \mu_k)/\sigma\right) \tag{14}$$

where $\phi$ is the standard normal density. The parameters of the associated HMM are then

$$\theta = \left(\psi(1,2), \psi(2,1), \mu_1, \mu_2, \sigma^2\right) \in \mathbf{R}^5 \tag{15}$$

The marginal distribution of each $Y_t$ is then a mixture of two Gaussians. When there are $K$ states and each $f(\cdot \,|\, k)$ is a one-dimensional Gaussian with mean $\mu_k$ and variance $\sigma^2$ then there are $K(K-1) + K + 1 = K^2 + 1$ parameters. The measure $\nu$ is taken to be Lebesgue measure.

*Example 2.3.* [Poisson Mixtures] To model counts $Y_t$ in $\{0, 1, 2, 3, \ldots, \}$ we can replace the Gaussian models of the previous examples by Poissons:

$$p_\theta(Y_t \,|\, k) = \frac{\mu_k^{Y_t}}{Y_t!} e^{-\mu_k} \tag{16}$$

The parameters of the associated two-state HMM are then

$$\theta = (\psi(1,2), \psi(2,1), \mu_1, \mu_2) \in \mathbf{R}_+^4 \tag{17}$$

The marginal distribution of each $Y_t$ is then a mixture of two Poissons. When there are $K$ states there are $K(K-1) + K = K^2$ free parameters. The measure $\nu$ is the counting measure.

## 3. Computation in HMMs

The marginal probability of the outputs involves a sum over all state sequences:

$$p(Y_1, Y_2, \ldots, Y_T) \;=\; \sum_{X_1, X_2, \ldots X_T} \pi_1(X_1)\, p_\theta(Y_1 \,|\, X_1) \prod_{t=2}^{T} p_\theta(X_t \,|\, X_{t-1}) p_\theta(Y_t \,|\, X_t) \tag{18}$$

When there are $K$ states, $X_t \in \{1, 2, \ldots, K\}$, then expanding this summation explicitly yields $K^T$ terms. However, the sum can be computed in $O(TK^2)$ time using dynamic programming.

First, observe that the probability can be neatly expressed in terms of matrix multiplication. Assuming a finite set of $M$ output symbols, let $T$ be the $K \times K$ transition probability matrix $T_{a,b} = p_\theta(X_t = a \,|\, X_{t-1} = b)$, and let $O$ be the $M \times K$ emission matrix $O_{y,a} = p_\theta(Y_t = y \,|\, X_t = a)$. For each output symbol $y$ let $A(y)$ be the $K \times K$ matrix given by

$$A(y) = T\mathrm{diag}(O_y) = T\,\mathrm{diag}\big(p_\theta(y \,|\, X_t = 1), \ldots, p_\theta(y \,|\, X_t = K)\big). \tag{19}$$

The $(i, j)$ entry has the probabilistic meaning

$$A(y)_{ij} = p(X_t = i \text{ and } X_{t-1} = j \text{ and } Y_{t-1} = y). \tag{20}$$

Then the probability of an output sequence $(y_1, y_2, \ldots, y_T)$ can be written as

$$p_\theta(y_{1:T}) = \mathbf{1}^T A(y_T) A(y_{T-1}) \cdots A(y_2) A(y(1)) \pi. \tag{21}$$

Each matrix-vector operation requires $O(K^2)$ computation, and possibly less if the transition matrix for the Markov chain has special structure.

Computation in HMMs can also be viewed graphically, based on the Markov properties of the underlying graphical model. The Markov property for a Markov chain implies that the past and future are conditionally independent given the present.

Let $\mathcal{P}_t = (X_{1:(t-1)}, Y_{1:(t-1)})$ denote the past at time $t$; this is the collection of all observed and latent variables before $t$. Similarly, let $\mathcal{F}_t = (X_{(t+1):T}, Y_{(t+1):T})$ denote the collection of variables after time $t$. Then the Markov property of the graph is expressed in the equation

$$p(X_{1:T}, Y_{1:T}) = p(\mathcal{P}_t)\, p(X_t, Y_t \,|\, \mathcal{P}_t)\, p(\mathcal{F}_t \,|\, X_t, Y_t) \tag{22}$$

Let $\mathcal{P}_t(Y) = Y_{1:(t-1)}$ and $\mathcal{F}_t(Y) = Y_{(t+1):T}$ denote the observed past and future. Then we have that

$$p(X_t, Y_{1:T}) = p(\mathcal{P}_t(Y), X_t)\, p(Y_t \,|\, X_t)\, p(\mathcal{F}_t(Y) \,|\, X_t) \tag{23}$$

The conditional probability of being in state $X_t$ at time $t$, given the observed sequence $Y_{1:T}$, is therefore given by

$$p(X_t \,|\, Y_{1:T}) = \frac{p(\mathcal{P}_t(Y), X_t)\, p(Y_t \,|\, X_t)\, p(\mathcal{F}_t(Y) \,|\, X_t)}{p(Y_{1:T})} \tag{24}$$

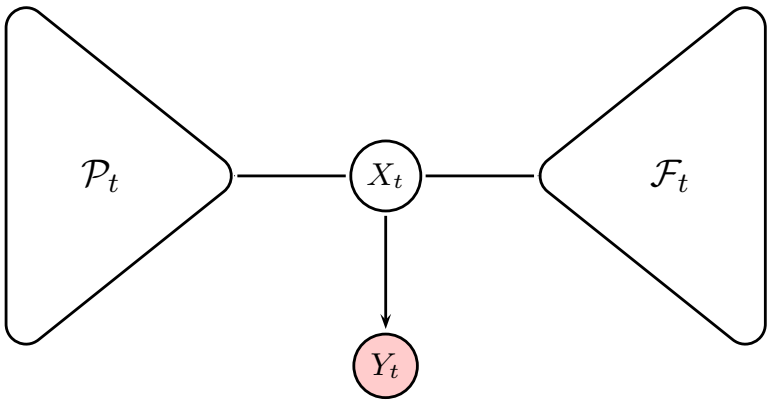


FIG 2. *Graphical view of the forward-backward algorithm. The "past" $\mathcal{P}_t$ at time $t$ is a function of $X_{1:t-1}$ and $Y_{1:t-1}$, and the "future" $\mathcal{F}_t$ at time $t$ is a function of $X_{t+1:T}$ and $Y_{t+1:T}$. The past and future are independent conditioned on the present $(X_t, Y_t)$.*

and the denominator can be computed as the marginal probability

$$\begin{aligned} p(Y_{1:T}) &= \sum_{X_t} p(X_t, Y_{1:T}) && (25) \\ &= \sum_{X_t} p(\mathcal{P}_t(Y), X_t)\, p(Y_t \,|\, X_t)\, p(\mathcal{F}_t(Y) \,|\, X_t) && (26) \end{aligned}$$

To simplify notation, let

$$\begin{aligned} \alpha_t(X_t) &= p(\mathcal{P}_t(Y), X_t)\, p(Y_t \,|\, X_t) && (27) \\ &= p(Y_{1:t}, X_t) && (28) \\ \beta_t(X_t) &= p(\mathcal{F}_t(Y) \,|\, X_t) && (29) \\ &= p(Y_{(t+1):T} \,|\, X_t) && (30) \end{aligned}$$

Then equations (24) and (25) are reexpressed as

$$\begin{aligned} p(X_t \,|\, Y_{1:T}) &= \frac{\alpha_t(X_t)\beta_t(X_t)}{p(Y_{1:T})} && (31) \\ p(Y_{1:T}) &= \sum_{X_t} \alpha_t(X_t)\beta_t(X_t) && (32) \end{aligned}$$

Now observe that these quantities can be computed recursively. Using the Markov property we

have

$$\begin{aligned}
\alpha_t(X_t) &= p(\mathcal{P}_t(Y), X_t)\, p(Y_t \mid X_t) && (33)\\
&= p(Y_t \mid X_t) \sum_{X_{t-1}} p(\mathcal{P}_{t-1}(Y), X_{t-1}, X_t) && (34)\\
&= p(Y_t \mid X_t) \sum_{X_{t-1}} p(\mathcal{P}_{t-1}(Y), X_{t-1})\, p(X_t \mid X_{t-1}) && (35)\\
&= p(Y_t \mid X_t) \sum_{X_{t-1}} \alpha_{t-1}(X_{t-1})\, p(X_t \mid X_{t-1}) && (36)
\end{aligned}$$

This recursion is forward in time, with the initial conditions

$$\alpha_1(X_1) = \pi(X_1)\, p(Y_1 \mid X_1). \tag{37}$$

Similarly, we the recursion backward in time

$$\begin{aligned}
\beta_t(X_t) &= p(\mathcal{F}_t(Y) \mid X_t) && (38)\\
&= \sum_{X_{t+1}} p(\mathcal{F}_t(Y), X_{t+1} \mid X_t) && (39)\\
&= \sum_{X_{t+1}} p(X_{t+1}, Y_{t+1} \mid X_t)\, p(\mathcal{F}_{t+1}(Y) \mid X_{t+1}) && (40)\\
&= \sum_{X_{t+1}} p(X_{t+1}, Y_{t+1} \mid X_t)\, \beta_{t+1}(X_{t+1}) && (41)\\
&= \sum_{X_{t+1}} p(X_{t+1} \mid X_t)\, p(Y_{t+1} \mid X_{t+1})\, \beta_{t+1}(X_{t+1}) && (42)
\end{aligned}$$

with the initial conditions

$$\beta_T(X_T) = 1 \quad \text{for all } X_T. \tag{43}$$

The marginal probability of the entire observed sequence can then be computed in several different, but mathematically equivalent ways:

$$\begin{aligned}
p(Y_{1:T}) &= \sum_{X_t} \alpha_t(X_t)\, \beta_t(X_t) \quad \text{for any } t && (44)\\
&= \sum_{X_T} \alpha_T(X_T) && (45)\\
&= \sum_{X_1} \pi(X_1)\, p(Y_1 \mid X_1)\, \beta_1(X_1) && (46)
\end{aligned}$$

These calculations make use of a slightly different form of the Markov property, as shown in Figure 3. The past and future is divided by a bridge that spans two time steps. Using this view the marginal and conditional probability of $(X_{t-1}, X_t)$ can be computed as

$$\begin{aligned}
p(X_{t-1}, X_t, Y_{1:T}) &= \alpha_{t-1}(X_{t-1})\, p(X_t \mid X_{t-1})\, p(Y_t \mid X_t)\, \beta_t(X_t) && (47)\\
p(X_{t-1}, X_t \mid Y_{1:T}) &= \frac{\alpha_{t-1}(X_{t-1})\, p(X_t \mid X_{t-1})\, p(Y_t \mid X_t)\, \beta_t(X_t)}{p(Y_{1:T})}. && (48)
\end{aligned}$$

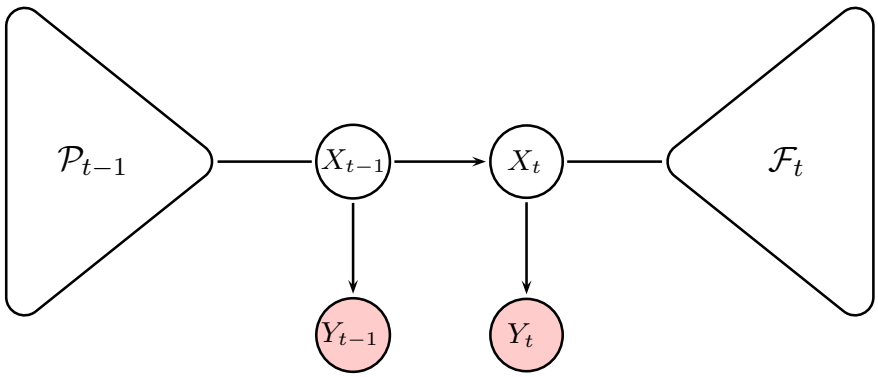


FIG 3. *The forward-backward calculations split the past and future by a bridge that spans two time steps.*

## 4. Avoiding Underflow

When computing the forward and backward probabilities $\alpha_t(X_t) = p(X_t, Y_{1:t})$ and $\beta_t(X_t) = p(Y_{(t+1):T} \,|\, X_t)$ for a long sequence, the numbers can get very small and underflow the numerical precision. There are two basic schemes for avoiding underflow: log-domain arithmetic and scaling.

Underflow occurs when the probabilities $p(Y_1, Y_2, \ldots, Y_t)$ become too small. The idea behind scaling is to instead compute the product of conditional probabilities

$$p(Y_1)\, p(Y_2 \,|\, Y_1)\, p(Y_3 \,|\, Y_1, Y_2) \cdots p(Y_t \,|\, Y_1, Y_2, \ldots, Y_{t-1}) = \prod_{s=1}^{t} c_s \tag{49}$$

where $c_1 = p(Y_1)$ and $c_s \equiv p(Y_s \,|\, Y_{1:(s-1)})$ for $s > 1$; the forward and backward probabilities are scaled in terms of these. Define $\widehat{\alpha}_t$ to be the conditional probability

$$\widehat{\alpha}_t(X_t) \;=\; p(X_t \,|\, Y_{1:t}) \tag{50}$$

Then we have that

$$\begin{aligned} \alpha_t(X_t) &= p(Y_{1:t}, X_t) && (51)\\ &= p(Y_{1:t})\, p(X_t \,|\, Y_{1:t}) && (52)\\ &= \left( \prod_{s=1}^{t} c_s \right) \widehat{\alpha}_t(X_t) && (53) \end{aligned}$$

so that

$$\widehat{\alpha}_t(X_t) \;=\; \frac{\alpha_t(X_t)}{\prod_{s=1}^{t} c_s} \tag{54}$$

FORWARD-BACKWARD ALGORITHM

- In a forward pass through the sequence, compute the forward probabilities

$$\begin{aligned}\alpha_1(X_t) &= \pi(X_1)\, p(Y_1 \mid X_1) \\ \alpha_t(X_t) &= p(Y_t \mid X_t) \sum_{X_{t-1}} \alpha_{t-1}(X_{t-1})\, p(X_t \mid X_{t-1}),\ t > 1\end{aligned}$$

  The probability of the sequence is then

$$p(Y_{1:T}) = \sum_{X_T} \alpha_T(X_T)$$

- In a backward pass through the sequence, compute the backward probabilities

$$\begin{aligned}\beta_T(X_T) &= 1 \\ \beta_t(X_t) &= \sum_{X_{t+1}} p(X_{t+1} \mid X_t)\, p(Y_{t+1} \mid X_{t+1})\, \beta_{t+1}(X_{t+1})\end{aligned}$$

- The conditional probability of a state $X_t$ and state pair $(X_t, X_{t+1})$ are then given by

$$\begin{aligned}p(X_t \mid Y_{1:T}) &= \frac{\alpha_t(X_t)\beta_t(X_t)}{p(Y_{1:T})} \\ p(X_t, X_{t+1} \mid Y_{1:T}) &= \frac{\alpha_t(X_t)\, p(X_{t+1} \mid X_t)\, p(Y_{t+1} \mid X_{t+1})\beta_{t+1}(X_{t+1})}{p(Y_{1:T})}\end{aligned}$$

This leads to the following modified recursion for the normalized forward probabilities:

$$\widehat{\alpha}_t(X_t) = \frac{1}{c_t} \frac{\alpha_t(X_t)}{\prod_{s=1}^{t-1} c_s} \tag{55}$$

$$= \frac{1}{c_t} p(Y_t \mid X_t) \sum_{X_{t-1}} \frac{\alpha_{t-1}(X_{t-1})}{\prod_{s=1}^{t-1} c_s} p(X_t \mid X_{t-1}) \tag{56}$$

$$= \frac{1}{c_t} p(Y_t \mid X_t) \sum_{X_{t-1}} \widehat{\alpha}_{t-1}(X_{t-1})\, p(X_t \mid X_{t-1}) \tag{57}$$

The term $c_t$ is then calculated by constraint that $\widehat{\alpha}_t$ is a conditional probability:

$$c_t = \sum_{X_t} p(Y_t \mid X_t) \sum_{X_{t-1}} \widehat{\alpha}_{t-1}(X_{t-1})\, p(X_t \mid X_{t-1}) \tag{58}$$

The initial conditions in this recursion are

$$\widehat{\alpha}_1(X_1) = \frac{\pi(X_1)\, p(Y_1 \mid X_1)}{\sum_{X_1} \pi(X_1)\, p(Y_1 \mid X_1)}. \tag{59}$$

Similarly, for the backward recursion, define

$$\widehat{\beta}_t(X_t) = \frac{\beta_t(X_t)}{\prod_{s=t+1}^{T} c_s} \tag{60}$$

Then the backward recursion takes the form

$$\widehat{\beta}_t(X_t) \;=\; \frac{1}{c_{t+1}} \sum_{X_{t+1}} p(X_{t+1} \,|\, X_t)\, p(Y_{t+1} \,|\, X_{t+1})\, \widehat{\beta}_{t+1}(X_{t+1}) \tag{61}$$

The marginal probability $p(Y_{1:T})$ can be expressed now as

$$p(Y_1, Y_2, \ldots, Y_T) = \prod_{s=1}^{T} c_s \tag{62}$$

and the conditional probability of $X_t$ is computed in terms of the scaled forward and backward probabilities as

$$p(X_t \,|\, Y_{1:T}) = \widehat{\alpha}_t(X_t)\, \widehat{\beta}_t(X_t) \tag{63}$$

This is a special case of the sum-product algorithm, where the graph is a simple chain, and where the messages sent between adjacent nodes are normalized to be probability distributions; see the notes on variational inference.

SCALED FORWARD-BACKWARD ALGORITHM

- In a forward pass through the sequence, compute the forward probabilities

$$\begin{aligned} \widehat{\alpha}_1(X_t) &= \frac{1}{c_1} \pi(X_1)\, p(Y_1 \,|\, X_1) \\ \widehat{\alpha}_t(X_t) &= \frac{1}{c_t} p(Y_t \,|\, X_t) \sum_{X_{t-1}} \widehat{\alpha}_{t-1}(X_{t-1})\, p(X_t \,|\, X_{t-1}),\; t > 1 \end{aligned}$$

  with $c_t$ calculated so that $\widehat{\alpha}_t(X_t)$ sums to one. The probability of the sequence is then

$$p(Y_{1:T}) = \prod_{s=1}^{T} c_s$$

- In a backward pass through the sequence, compute the backward probabilities

$$\begin{aligned} \widehat{\beta}_T(X_T) &= 1 \\ \widehat{\beta}_t(X_t) &= \frac{1}{c_{t+1}} \sum_{X_{t+1}} p(X_{t+1} \,|\, X_t)\, p(Y_{t+1} \,|\, X_{t+1})\, \widehat{\beta}_{t+1}(X_{t+1}) \end{aligned}$$

- The conditional probability of a state $X_t$ and state pair $(X_t, X_{t+1})$ are then given by

$$\begin{aligned} p(X_t \,|\, Y_{1:T}) &= \widehat{\alpha}_t(X_t) \widehat{\beta}_t(X_t) \\ p(X_t, X_{t+1} \,|\, Y_{1:T}) &= \widehat{\alpha}_t(X_t)\, \frac{p(X_{t+1} \,|\, X_t)\, p(Y_{t+1} \,|\, X_{t+1})}{c_{t+1}} \widehat{\beta}_{t+1}(X_{t+1}) \end{aligned}$$

## 5. Estimation in HMMs

If both $X_{1:T} = x_{1:T}$ and $Y_{1:T} = y_{1:T}$ are observed in training data, and the joint density $p_\theta(x_{1:T}, y_{1:T})$ lies in the exponential family, then the maximum likelihood estimates are obtained using standard

techniques. If $X_{1:T}$ is unobserved, but $Y_{1:T} = y_{1:T}$ is observed, then the parameters can be estimated by maximizing the marginal probability $p_\theta(y_{1:T})$, using the EM algorithm.

The incomplete data is the observation sequence $y_{1:T}$, and the complete data is the observation and state sequence together, $(X_{1:T}, y_{1:T})$. The E-step is implemented efficiently using the forward and backward dynamic programs.

The complete data likelihood is

$$p_\theta(y_{1:T}; X_{1:T}) = \pi(X_1) p_\theta(y_1 \,|\, X_1) \prod_{s=2}^{T} p_\theta(X_s \,|\, X_{s-1})\, p_\theta(y_s \,|\, X_s). \tag{64}$$

The auxiliary function is

$$\begin{aligned}
\bar{Q}(\theta; \theta') &= \mathbb{E}_{\theta'}\big[\log p_\theta(X_{1:T}, Y_{1:T}) \,|\, Y_{1:T} = y_{1:T}\big] \\
&= \sum_{X_{1:T}} p_{\theta'}(X_{1:T} \,|\, y_{1:T}) \log\Big\{\pi(X_1) p_\theta(X_2 \,|\, X_1) \prod_{s=2}^{T} p_\theta(X_s \,|\, X_{s-1})\, p_\theta(y_s \,|\, X_s)\Big\} \qquad (65) \\
&= \sum_{X_{1:T}} p_{\theta'}(X_{1:T} \,|\, y_{1:T}) \log\Big\{\pi(X_1) \prod_{a,b} \psi(a,b)^{c(a,b;X_{1:T})} \prod_{b,y} f(y \,|\, \theta_b)^{c(b,y;X_{1:T},y_{1:T})}\Big\}
\end{aligned}$$

where $c(a, b; X_{1:T})$ denotes the number of times state $b$ follows state $a$ in the sequence $X_1, \ldots, X_T$, and $c(b, y; X_{1:T}, y_{1:T})$ denotes the number of times output symbol $y$ is emitted from state $b$ in the complete data sequence $(X_1, \ldots, X_T; y_1, \ldots, y_T)$. That is,

$$c(a, b; X_{1:T}) = \sum_{t=2}^{T} \delta(X_{t-1}, a)\, \delta(X_t, b) \tag{66}$$

$$c(b, y; X_{1:T}, y_{1:T}) = \sum_{t=1}^{T} \delta(X_t, b)\, \delta(y_t, y). \tag{67}$$

Disregarding the terms involving the initial state probability $\pi(X_1)$, we can write this as

$$\bar{Q}(\theta; \theta') = \sum_{a,b} \mathbb{E}_{\theta'}[c(a, b; X_{1:T}) \,|\, Y_{1:T} = y_{1:T}] \log \psi(a, b) + \tag{68}$$

$$\sum_{b,y} \mathbb{E}_{\theta'}[c(b, y; X_{1:T}, y_{1:T}) \,|\, Y_{1:T} = y_{1:T}] \log f(y \,|\, \theta_b). \tag{69}$$

The sum on $y$ can be restricted to only the observed values $y_1, \ldots, y_T$. The expected counts are computed using the forward-backward calculations. For example,

$$\mathbb{E}_{\theta'}[c(b, y; X_{1:T}, y_{1:T}) \,|\, Y_{1:T} = y_{1:T}] = \sum_{t=1}^{T} p_\theta(X_t = b \,|\, y_{1:T})\, \delta(y_t, y) \tag{70}$$

$$= \sum_{t=1}^{T} \widehat{\alpha}_t(b)\, \widehat{\beta}_t(b)\, \delta(y_t, y). \tag{71}$$

Without any restrictions on the state transition structure, the transition probabilities $\psi(a, b)$ are just multinomial; that is, $\psi(a, b) \geq 0$ and $\sum_{b=1}^{K} \psi(a, b) = 1$, for each state $a$. Therefore, the $M$-step reestimates $\psi(a, b)$ as the normalized expected counts:

$$\psi(a, b) = \frac{\mathbb{E}_{\theta'}[c(a, b; X_{1:T}) \,|\, y_{1:T}]}{\sum_{b'} \mathbb{E}_{\theta'}[c(a, b'; X_{1:T}) \,|\, y_{1:T}]}. \tag{72}$$

If the output space $\mathcal{Y}$ is finite, and the emission models are multinomial with parameters $\gamma(b, y)$, so that $\gamma(b, y) \geq 0$ and $\sum_y \gamma(b, y) = 1$ for each state $b$, then the emission probabilities are similarly reestimated by just normalizing the expected counts:

$$\gamma(b, y) = \frac{\mathbb{E}_{\theta'}[c(b, y; X_{1:T}, y_{1:T}) \,|\, y_{1:T}]}{\sum_{y'} \mathbb{E}_{\theta'}[c(b, y'; X_{1:T}, y_{1:T}) \,|\, y_{1:T}]}. \tag{73}$$

More generally, if the emission probabilities are from an exponential family model

$$f_\theta(y \,|\, b) = \eta_b(y) \exp\{\theta_b^T \phi_b(y) - \Psi_b(\theta_b)\}, \tag{74}$$

with sufficient statistic $\phi_b(y)$, then $\partial\Psi(\theta_b)/\partial\theta_b = \mathbb{E}_\theta[\phi_b(Y)]$, and a calculation shows that the M-step is to set $\theta_b$ according to the moment condition satisfied:

$$\mathbb{E}_{\theta_b}[\phi_b(Y)] = \frac{\sum_{t=1}^{T} \mathbb{E}_{\theta'}\left[\delta(X_t, b) \,|\, y_{1:T}\right] \, \phi_b(y_t)}{\sum_{t=1}^{T} \mathbb{E}_{\theta'}\left[\delta(X_t, b) \,|\, y_{1:T}\right]} \tag{75}$$

$$= \frac{\sum_{t=1}^{T} p_{\theta'}\left(X_t = b \,|\, y_{1:T}\right) \, \phi_b(y_t)}{\sum_{t=1}^{T} p_{\theta'}\left(X_t = b \,|\, y_{1:T}\right)} \tag{76}$$

$$= \frac{\sum_{t=1}^{T} \widehat{\alpha}_t(b) \, \widehat{\beta}_t(b) \, \phi_b(y_t)}{\sum_{t=1}^{T} \widehat{\alpha}_t(b) \, \widehat{\beta}_t(b)}. \tag{77}$$

For example, if $f_\theta(y \,|\, b)$ is the Gaussian density with mean $\theta_b$ and known variance, then the sufficient statistic is $\phi_b(y) = y$, and the update is simply the weighted sample mean:

$$\theta_b = \mathbb{E}_{\theta_b}[Y] = \frac{\sum_{t=1}^{T} p_{\theta'}\left(X_t = b \,|\, y_{1:T}\right) \, y_t}{\sum_{t=1}^{T} p_{\theta'}\left(X_t = b \,|\, y_{1:T}\right)}. \tag{78}$$

## 6. Instability of HMMs

The parameter estimates for HMMs can be very sensitive to initial conditions. The EM algorithm can also be very slow to converge. In this section we illustrate these facts with some simple examples.

Figure 4 shows the dynamics of the EM algorithm on a simple hidden Markov model for text with two states. The output alphabet is the twenty-six letters A-Z, together with the space character. The model is trained 20,000 characters of English text, with initial parameters chosen to be close to

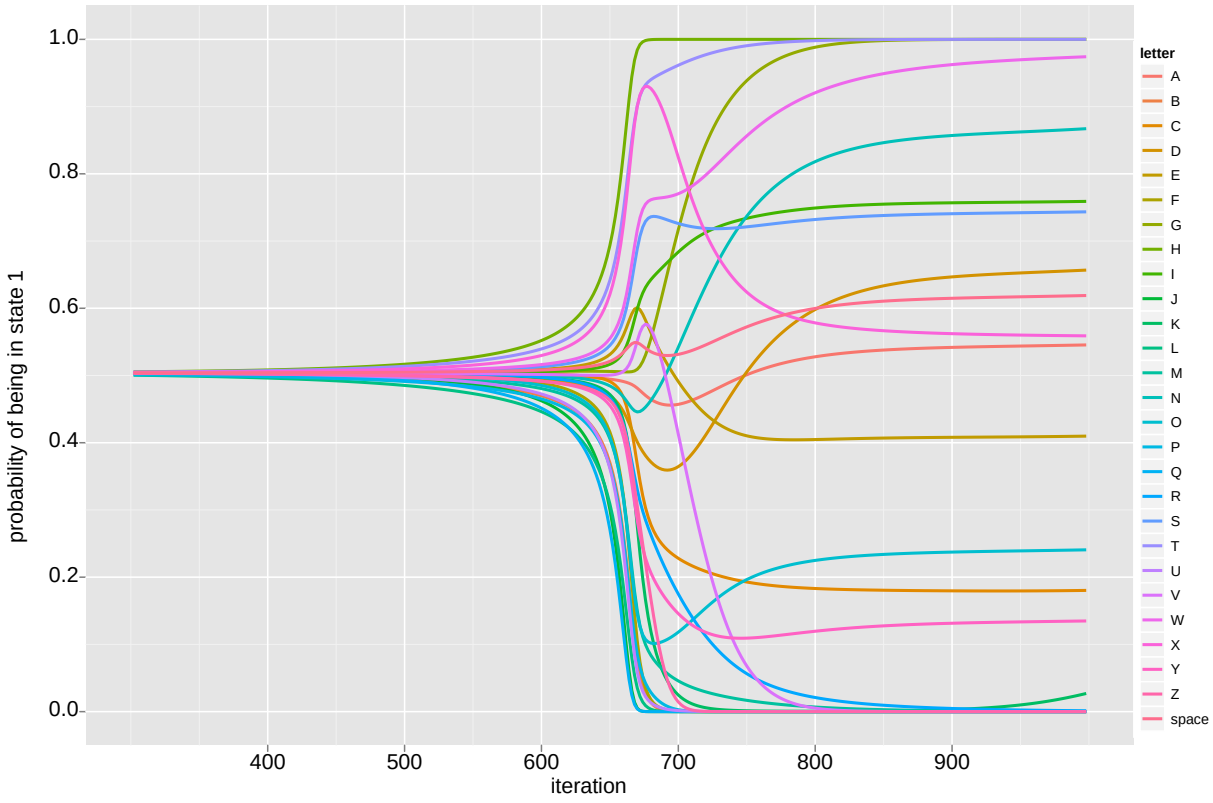


FIG 4. *The dynamics of the EM algorithm on a simple hidden Markov model. The HMM has two states, and emits one of the twenty-six letters A-Z, or a space. The model is trained 20,000 characters of English text, with initial parameters chosen to be close to uniform. The curves show the posterior probability* $\mathbb{P}(\textit{state} = 1 \,|\, \textit{letter}, \mathcal{D})$ *of the probability each letter is generated from state one, over the training corpus* $\mathcal{D}$*. There is a "phase transition" around 650 iterations.*

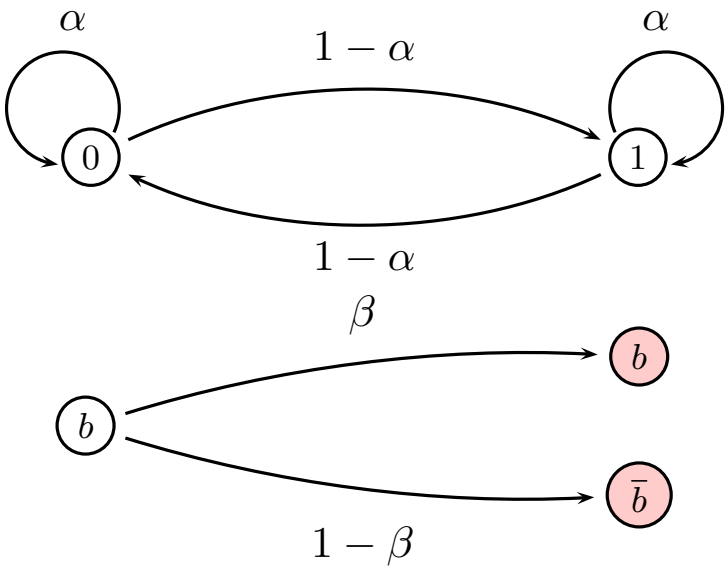


FIG 5. *A two-parameter HMM, with two states and two output symbols, parameterized by* $p(X_t = b \,|\, X_{t-1} = b) = \alpha$ *and* $p(Y_t = b \,|\, X_t = b) = \beta$.

uniform. The algorithm requires more than 1,000 EM iterations to converge, and the parameters change very little during the first few hundred iterations. The curves in Figure 4 show the conditional probabilities $\mathbb{P}(\text{state} = 1 \,|\, \text{letter}, \mathcal{D})$; that is, the probability each letter is generated from state one. There is a "phase transition" around 650 iterations.

The next figures show the sensitivity of HMMs to initialization. We fit a two-state HMM where the output space also has two symbols. The state transition parameters are "tied" so that $p(X_t = 0 \,|\, X_{t-1} = 0) = p(X_t = 1 \,|\, X_{t-1} = 1) = \alpha$. Similarly, the output probabilities are tied so that $p(Y_t = 0 \,|\, X_t = 0) = p(Y_t = 1 \,|\, X_t = 1) = \beta$. The model is shown in Figure 5.

The dependence of the model fit on the string $y_{1:T} = 0100010000111$ to the initial parameters $\alpha_0$ and $\beta_0$ is shown in Figure 6. Points $(\alpha_0, \beta_0)$ that converge to models having the same likelihood are colored the same. Figure 10 shows the analogous plots of likelihood versus initial parameters $(\alpha_0, \beta_0)$, using a variant of EM called *Viterbi training*, where the M-step is computed using the most probable state sequence, rather than using the expectation over all states.

## 7. Asymptotic Properties of HMMs

When estimated using maximum likelihood, HMMs share the properties of consistency and asymptotic normality of maximum likelihood estimators for standard parametric models, although care needs to be taken in specifying appropriate conditions on the underlying Markov chain.

Suppose that the data are generated from an HMM with parameter $\theta^*$, and

$$\widehat{\theta}_T = \operatorname{argmax}\, p_\theta(y_{1:T}) \tag{79}$$

satisfies $\widehat{\theta}_T \xrightarrow{P} \theta^*$ as $T \longrightarrow \infty$. Assume the following conditions:

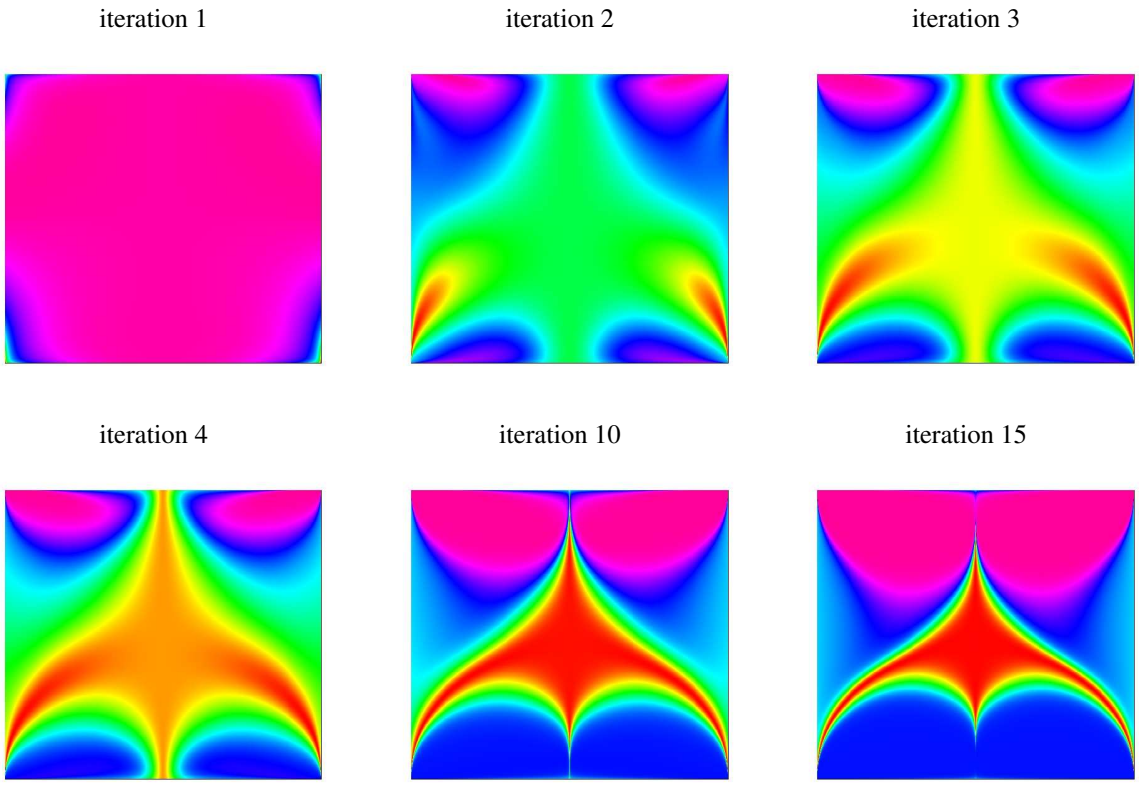


FIG 6. *Convergence regions of the EM algorithm for the simple hidden Markov model shown in Figure 5, trained on the sequence* $y_{1:T} = 0100010000111$. *The horizontal and vertical dimensions correspond to the initial transition and emission parameters* $\alpha$ *and* $\beta$*; the color indicates the log-likelihood of the model on the training data.*

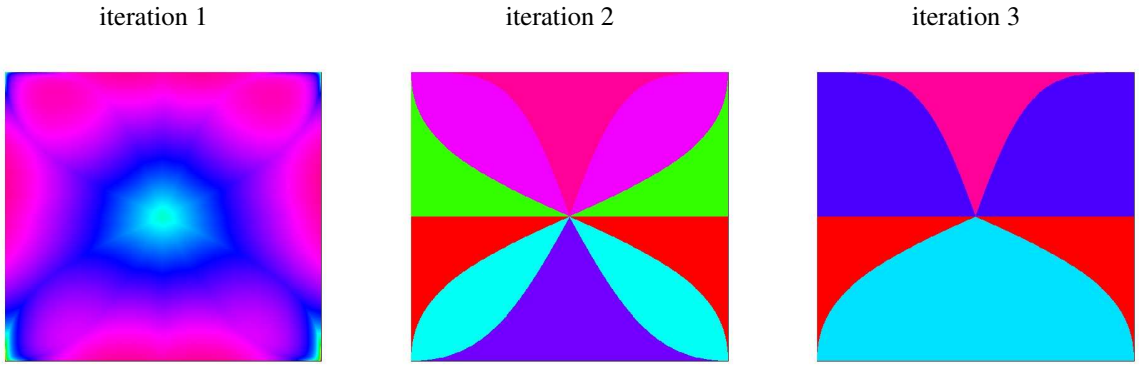


FIG 7. *Convergence regions of the EM algorithm for the simple hidden Markov model shown in Figure 5, trained on the sequence* $y_{1:T} = 0100010000111$ *using the Viterbi state sequence in the E-step. The parameters do not change after the third iteration.*

(A1) The Markov chain governing $X_t$, which has transition matrix

$$p_{\theta^*}(X_t = a \mid X_{t-1} = b),$$

is irreducible and aperiodic (e.g., ergodic).

(A2) For some $\delta > 0$, the quantity

$$\rho(y) = \sup_{\|\theta - \theta^*\| < \delta} \max_{s_1, s_2} \frac{p_\theta(y \mid X_t = s_1)}{p_\theta(y \mid X_t = s_2)} \tag{80}$$

satisfies

$$\max_c \mathbb{P}_{\theta^*}(\rho(y) = \infty \mid X_t = c) < 1 \tag{81}$$

Under these and other technical conditions, which we do not list here, Bickel et al. (1998) show the following. **Theorem 7.4.** *Suppose that the Fisher information matrix*

$$I(\theta^*) = -\mathbb{E}_{\theta^*} \nabla^2 \log p_{\theta^*}(y_{1:T}) \tag{82}$$

*is nonsingular. Then*

$$\sqrt{T}\left(\widehat{\theta}_T - \theta^*\right) \to \mathcal{N}\left(0, I^{-1}(\theta^*)\right). \tag{83}$$

The Fisher information $I(\theta^*)$ can be thought of as the limiting covariance matrix of the score

$$\frac{1}{\sqrt{T}} \nabla \ell_T(\theta^*) = \frac{1}{\sqrt{T}} \nabla \log p_\theta(y_{1:T}) \tag{84}$$

as for standard parametric models, or as the limiting covariance of the log conditionals $\log p(y_T \mid y_{1:T-1})$, as a consequence of a central limit theorem for the sum of random variables

$$\nabla \log p_{\theta^*}(Y_{1:T}) = \sum_{i=1}^{T} \log p_{\theta^*}(Y_i \mid Y_{1:(i-1)}). \tag{85}$$

## 8. Equivalence of HMMs

Two hidden Markov models with the same output space could have different numbers of states and different parameters, but still assign the same probability to all output sequences. Such HMMs could be naturally be thought of as being equivalent.

**Definition 8.5.** *Two hidden Markov models $M_1$ and $M_2$ defined over a finite alphabet $\mathcal{Y}$ are equivalent in case $p_{M_1}(y_{1:T}) = p_{M_2}(y_{1:T})$ for any $T$ and $y_{1:T} \in \mathcal{Y}^T$.*

Recall from (21) that the marginal probabilities of the outputs $p_{M_i}(y_{1:T})$ can be expressed in terms of matrix multiplication. So, it may not come as no surprise that the the equivalence of HMMs can be recast in linear algebraic terms, and decided in polynomial time. In particular, results of Balasubramanian (1993) include an algorithm that tests equivalence of two HMMs in time

$$O\Big(\max(K_1, K_2) M (K_1 + K_2)^2 + (K_1 K_2)^{5.5}\Big) \tag{86}$$

where $|\mathcal{Y}| = M$ and $K_1$ and $K_2$ are the numbers of states in the two HMMs. Related results were obtained by Ito et al. (1992).

## 9. State Space Models: Gaussian HMMs

The state space model can be thought of as a form of hidden Markov model where the state variable $X_t \in \mathbf{R}^d$ is real-valued, the observation $Y_t \in \mathbf{R}^p$ is also real-valued, and the transition and emission probabilities are Gaussian. The joint distribution of $X_{1:T}$ and $Y_{1:T}$ is then Gaussian, so that the conditionals $X_t \,|\, Y_{1:t}$ and $X_t \,|\, Y_{1:T}$ are also Gaussian. Thus, the forward-backward calculations can be expressed in terms of recursions that compute the mean and variance of these Gaussians. The mean and variance of $X_t$ given the previous and current observations $Y_{1:t}$ are calculated with forward recursions that are referred to as Kalman filtering. The mean and variance of $X_t$ given all of the observations $Y_{1:T}$ are computed with backward recursions that are called Kalman smoothing. As the term "smoothing" suggests, these estimates will tend to be smoother, or less noisy, since they make use of more data than the forward filtering estimates.

### *9.1. Intuition: A Simple Example*

This form of model is used in many different settings that involve noisy data that evolves over time. For example, suppose that an animal is tagged with a radio collar that emits a tracking signal; the unobserved state is the animal's position $X_t \in \mathbf{R}^2$ at each time. The observed data are the signals $Y_t \in \mathbf{R}^2$ at a sequence of times; if these signals are very noisy, then the data can be smoothed across time to better infer the position $X_t$. To gain some intuition for the smoothing effects of the Kalman filter, consider the very special case where the animal's movements are governed by

$$X_{t+1} = X_t + c + \epsilon_t \tag{87}$$

where $\epsilon_t \sim \mathcal{N}(0, \gamma^2 I)$ and $c$ is a fixed direction. Also, suppose that the signals are observed according to

$$Y_t = X_t + \delta_t \tag{88}$$

where $\delta_t \sim \mathcal{N}(0, \sigma^2 I)$. If, based on measurements $Y_{1:t}$ we have an estimate $n_t$ of the position, with variance $v_t$, then at the next time step, in the absence of any further data, we would estimate the position to be

$$n_{t+1} = n_t + c \tag{89}$$

However, our uncertainty in the estimate would increase—the variance would now be

$$v_{t+1} = v_t + \gamma^2 \tag{90}$$

If, on the other hand, we get another data point $y_{t+1}$, then our estimate is the weighted sum of the previous estimate and the measurement:

$$\begin{aligned} n_{t+1} &= \frac{\sigma^2}{v_{t+1} + \sigma^2} (n_t + c) + \frac{v_{t+1}}{v_{t+1} + \sigma^2} y_{t+1} && (91) \\ &= n_t + c + K_{t+1} (y_{t+1} - (n_t + c)) && (92) \end{aligned}$$

where we define the *Kalman gain* as

$$K_{t+1} = \frac{v_{t+1}}{v_{t+1} + \sigma^2} \tag{93}$$

If there is little uncertainty in the measurement ($\sigma^2$ small), then the weight on the observation increases; similarly, if there is little uncertainty on the next position given the current position ($\gamma^2$ small), then more weight is placed on the previous estimate. Moreover, our uncertainty in the new estimate becomes

$$\begin{aligned} v_{t+1} &= \frac{(v_t + \gamma^2)\sigma^2}{v_t + \gamma^2 + \sigma^2} && (94) \\ &= (1 - K_{t+1})(v_t + \gamma^2) && (95) \end{aligned}$$

If $\sigma^2$ is small, then the variance is greatly reduced by using the next measurement.

## 10. Kalman Filtering

The Kalman filter is the extension of the previous simple derivation for a more general linear state space model. Consider the model

$$\begin{aligned} X_t \,|\, X_{t-1} &\sim \mathcal{N}(AX_{t-1}, \Gamma) && (96) \\ Y_t \,|\, X_t &\sim \mathcal{N}(BX_t, \Sigma) && (97) \end{aligned}$$

Expressed as a linear dynamical system, this becomes

$$\begin{aligned} X_t &= AX_{t-1} + W_t, \quad W_t \sim \mathcal{N}(0, \Gamma) && (98) \\ Y_t &= BX_t + V_t, \quad V_t \sim \mathcal{N}(0, \Sigma) && (99) \end{aligned}$$

The Kalman filtering equations are precisely the scaled forward-backward calculations under these Gaussian assumptions. To derive the recursions, it will be helpful to recall some facts about transforming and conditioning multivariate Gaussians. First, if

$$\begin{aligned} \xi &\sim \mathcal{N}(\mu, \Xi) && (100) \\ \zeta \,|\, \xi &\sim \mathcal{N}(A\xi, \Gamma) && (101) \end{aligned}$$

then the marginal distribution of $\zeta$, obtained by integrating out $\xi$, is given by

$$\zeta \sim \mathcal{N}(A\mu, A\Xi A^T + \Gamma) \tag{102}$$

Moreover, if $(\xi, \zeta)$ are jointly Gaussian with distribution

$$\begin{pmatrix} \xi \\ \zeta \end{pmatrix} \sim \mathcal{N}\left( \begin{pmatrix} \mu \\ \nu \end{pmatrix}, \begin{pmatrix} A & C \\ C^T & B \end{pmatrix} \right) \tag{103}$$

then the conditional distributions are given by

$$\begin{aligned} \xi \,|\, \zeta &\sim \mathcal{N}\left(\mu + CB^{-1}(\zeta - \nu), A - CB^{-1}C^T\right) && (104) \\ \zeta \,|\, \xi &\sim \mathcal{N}\left(\nu + C^TA^{-1}(\xi - \mu), B - C^TA^{-1}C\right) && (105) \end{aligned}$$

Now, to derive the forward recursion, suppose that we have

$$X_{t-1} \,|\, Y_{1:t-1} \;\sim\; \mathcal{N}(m_{t-1}, V_{t-1}) \tag{106}$$

In other words, the forward probability $\widehat{\alpha}_{t-1}(X_{t-1})$ is given by

$$\widehat{\alpha}_{t-1}(x) = \varphi(x; m_{t-1}, V_{t-1}) \tag{107}$$

where $\varphi$ is the Gaussian density function

$$\varphi(x; m, V) = \frac{1}{|2\pi V|^{1/2}} \exp\left(-\frac{1}{2}(x-m)^T V^{-1}(x-m)\right) \tag{108}$$

Then we have that

$$\begin{aligned} p(X_t, Y_t \,|\, Y_{1:t-1}) &= \varphi(Y_t; BX_t, \Sigma) \int \widehat{\alpha}_{t-1}(z)\, \varphi(X_t; Az, \Gamma)\, dz & (109)\\ &= \varphi(Y_t; BX_t, \Sigma)\, \varphi(X_t; Am_{t-1}, AV_{t-1}A^T + \Gamma) & (110) \end{aligned}$$

Using (102), the conditional distribution of the current observation and state is

$$\begin{aligned} (X_t, Y_t)^T \,|\, Y_{1:t-1} &\sim \mathcal{N}(\mu_t, \Xi_t) & (111)\\ \mu_t &= \begin{pmatrix} Am_{t-1} \\ BAm_{t-1} \end{pmatrix} & (112)\\ \Xi_t &= \begin{pmatrix} AV_{t-1}A^T + \Gamma & \left(AV_{t-1}A^T + \Gamma\right) B^T \\ B\left(AV_{t-1}A^T + \Gamma\right) & B\left(AV_{t-1}A^T + \Gamma\right) B^T + \Sigma \end{pmatrix} & (113)\\ &\equiv \begin{pmatrix} Q_{t-1} & Q_{t-1}B^T \\ BQ_{t-1} & BQ_{t-1}B^T + \Sigma \end{pmatrix} & (114) \end{aligned}$$

Applying (104), the conditional distribution of the current state given the current and previous observations is

$$\begin{aligned} X_t \,|\, Y_t, Y_{1:t-1} &\sim \mathcal{N}(m_t, V_t) & (115)\\ m_t &= Am_{t-1} + K_{t-1}(Y_t - BAm_{t-1}) & (116)\\ V_t &= (I - K_{t-1}B)Q_{t-1} & (117)\\ \text{where} \qquad K_{t-1} &= Q_{t-1}B^T(BQ_{t-1}B^T + \Sigma)^{-1} & (118)\\ Q_{t-1} &= AV_{t-1}A^T + \Gamma & (119) \end{aligned}$$

The normalizing constant $c_t$ is then given by

$$c_t = \varphi(Y_t; BAm_{t-1}, BQ_{t-1}B^T + \Sigma) \tag{120}$$

The initial conditions for the recursion are determined by starting off the latent Markov process in state $X_0$, with distribution

$$X_0 \sim \mathcal{N}(m_0, V_0) \tag{121}$$

Kalman Filtering Under the Gaussian model

$$X_t \,|\, X_{t-1} \sim \mathcal{N}(AX_{t-1}, \Gamma) \tag{126}$$
$$Y_t \,|\, X_t \sim \mathcal{N}(BX_t, \Sigma) \tag{127}$$

the probability of state $X_t$ given the observation sequence $Y_{1:t}$ up to time $t$ has a Gaussian distribution

$$X_t \,|\, Y_{1:t} \sim \mathcal{N}(m_t, V_t) \tag{128}$$

where the mean and covariance are computed recursively according to

$$m_t = Am_{t-1} + K_{t-1}(Y_t - BAm_{t-1}) \tag{129}$$
$$V_t = (I - K_{t-1}B)Q_{t-1} \tag{130}$$
$$\text{where} \quad K_{t-1} = Q_{t-1}B^T(BQ_{t-1}B^T + \Sigma)^{-1} \tag{131}$$
$$Q_{t-1} = AV_{t-1}A^T + \Gamma. \tag{132}$$

Note from (130) that the variance of the state $X_t$ does not depend on the observation sequence $Y_{1:t}$.

The update of the covariance, given in equation (130), may not be numerically stable, in the sense that since it is a difference of positive definite matrix it may not give a positive-definite matrix due to numerical error. But note that from the definition of $K_t$ it follows that

$$K_{t-1}\Sigma K_{t-1}^T = (I - K_{t-1}B)Q_{t-1}B^T K_{t-1}^T \tag{122}$$

and therefore we can write the update of the covariance as a *sum* of two positive definite matrices as

$$V_t = (I - K_{t-1}B)Q_{t-1} \tag{123}$$
$$= (I - K_{t-1}B)Q_{t-1}(I - K_{t-1}B)^T + (I - K_{t-1}B)Q_{t-1}B^T K_{t-1}^T \tag{124}$$
$$= (I - K_{t-1}B)Q_{t-1}(I - K_{t-1}B)^T + K_{t-1}\Sigma K_{t-1}^T. \tag{125}$$

## 11. Kalman Smoothing

The Kalman filtering recursions derived above compute the distribution over the current state, given the observations up to the current time. These calculations can be carried out in an online manner, using the observations as they come in one at a time. It is also possible to go back and update the state estimates at previous times.

If we have observed $Y_{1:T}$, the conditional probability of a state $X_t$ at time $t \leq T$ is given by the product of forward and backward probabilities:

$$p(X_t \,|\, Y_{1:T}) = \widehat{\alpha}_t(X_t)\widehat{\beta}_t(X_t) \tag{133}$$

Since $X_{1:T}$ and $Y_{1:T}$ are jointly Gaussian, $X_t$ is conditionally Gaussian given $Y_{1:T}$; suppose that

$$X_t \,|\, Y_{1:T} \sim \mathcal{N}(n_t, W_t) \tag{134}$$

Kalman Smoothing Under the Gaussian model

$$X_t \,|\, X_{t-1} \sim \mathcal{N}(AX_{t-1}, \Gamma) \tag{142}$$
$$Y_t \,|\, X_t \sim \mathcal{N}(BX_t, \Sigma) \tag{143}$$

the probability of state $X_t$ given the observation sequence $Y_{1:T}$ up to time $T \geq t$ has a Gaussian distribution

$$X_t \,|\, Y_{1:T} \sim \mathcal{N}(n_t, W_t). \tag{144}$$

After computing the foward Kalman filtering means $m_t$ and covariances $V_t$, a recursion backward in time computes $n_t$ and $W_t$ according to

$$n_{t-1} = m_{t-1} + J_{t-1}(n_t - Am_{t-1}) \tag{145}$$
$$W_{t-1} = V_{t-1} + J_{t-1}(W_t - Q_{t-1})J_{t-1}^T \tag{146}$$

where $J_{t-1} = V_{t-1}A^T(AV_{t-1}A^T + \Gamma)^{-1} = V_{t-1}A^TQ_{t-1}^{-1}$. The initial conditions are

$$n_T = m_T \tag{147}$$
$$W_T = V_T. \tag{148}$$

To derive a backward recurrence for $n_t$ and $W_t$, we use the continuous verison of the scaled backward recursion (61), in the form

$$\widehat{\alpha}_{t-1}(X_{t-1})\widehat{\beta}_{t-1}(X_{t-1}) \tag{135}$$
$$= \frac{1}{c_t}\widehat{\alpha}_{t-1}(X_{t-1}) \int p(X_t = z \,|\, X_{t-1})\; p(Y_t \,|\, X_t = z)\widehat{\beta}_t(z)\, dz \tag{136}$$
$$= \varphi(X_{t-1}; m_{t-1}, V_{t-1}) \int \varphi(z; AX_{t-1}, \Gamma)\, \varphi(Y_t; Bz, \Sigma)\, \frac{\varphi(z; n_t, W_t)}{c_t\widehat{\alpha}_t(z)}\, dz \tag{137}$$

where the last equality uses the definition of the model and (134). Now, using (102) and (104) we have

$$c_t\widehat{\alpha}_t(z) = \varphi(Y_t; Bz, \Sigma)\, \varphi(z; Am_{t-1}, AV_{t-1}A^T + \Gamma) \tag{138}$$

and

$$\varphi(X_{t-1}; m_{t-1}, V_{t-1})\, \varphi(z; AX_{t-1}, \Gamma) = \varphi(z; Am_{t-1}, AV_{t-1}V_{t-1}A^T + \Gamma) \times \varphi(X_{t-1}, m_{t-1} + J_{t-1}(z - Am_{t-1}), (I - J_{t-1}A)V_{t-1}) \tag{139}$$

where $J_{t-1} = V_{t-1}A^TQ_{t-1}^{-1}$. Therefore, we obtain

$$\widehat{\alpha}_{t-1}(X_{t-1})\widehat{\beta}_{t-1}(X_{t-1}) \tag{140}$$
$$= \int \varphi(X_{t-1}, m_{t-1} + J_{t-1}(z - Am_{t-1}), (I - J_{t-1}A)V_{t-1})\, \varphi(z; n_t, W_t)\, dz$$
$$= \varphi(X_{t-1}, m_{t-1}J_{t-1}(n_t - Am_{t-1}), V_{t-1} + J_{t-1}(W_t - Q_{t-1})J_{t-1}^T) \tag{141}$$

This yields the Kalman smoothing equations.

*Example 11.6.* [Ozone Concentration Levels] The data for this example are daily maximum 8-hour ozone concentrations (in parts-per-billion) at 153 sites in the US midwest near Lake Michigan, for 89 days during the summer of 1987. The observed values $Z_t$ are transformed to $Y_t = \log Z_t - \mu$.

We use a discretized version of a latent random field model. The latent (log) ozone level $X_t(u)$ at location $u$ on day $t$ is governed by

$$X_t \,|\, X_{t-1} \sim N(\alpha X_{t-1}, \gamma^2 I) \tag{149}$$

where $\alpha < 1$. The observed values $Y_t$ are distributed as

$$\begin{aligned} Y_t(s) \,|\, X_t &\sim N(\mu_t(s), \sigma^2) && (150)\\ \mu_t(s) &= \int K(s,u) X_t(u)\, du && (151) \end{aligned}$$

where $K$ is a smoothing kernel. Thus, the mean of the measurement on day $t$ at site $s$ is a spatial smooth of the latent field ozone levels. Similar models are given by *Kriging* and Gaussian process regression.

Discretizing to a grid of spatial locations $u \in \mathcal{G}$, this becomes the Kalman filter

$$\begin{aligned} X_t \,|\, X_{t-1} &\sim N(AX_{t-1}, \gamma^2 I) && (152)\\ Y_t \,|\, X_t &\sim N(BX_t, \sigma^2 I) && (153) \end{aligned}$$

where $A = \alpha I$ and

$$B_{su} = \frac{K(s,u)}{\sum_{u \in \mathcal{G}} K(s,u)}. \tag{154}$$

We use a Gaussian kernel $K(s,u)$ on the spatial locations, in terms of latitude and longitude.

Figure 8 shows the posterior mean $\mathbb{E}[X_t \,|\, Y_{1:T}]$ of this Kalman filter together with the observed values $Y_t$ for four days, June 18–21. On the first three of the four days, the estimated ozone levels are highest in the Chicago area and southeastern Wisconsin. On the fourth day, the high ozone areas have shifted toward the south.

## 12. State Space Models for Discrete Data

In this section we show how the Kalman filter can be used to estimate a time series model for discrete data, using variational methods. See **??** for a treatment of variational methods.

Suppose our data counts over some vocabulary in a time series of text documents. Consider a model where at each time $t$ we have a multinomial model $\beta_t$ in the natural parameterization. Let $w_{t,n}$ denote a vector of $n$ word observations in a document at time $t$. Using the natural parameterization, we can consider a state space model

$$\begin{aligned} \beta_t \,|\, \beta_{t-1} &\sim \mathcal{N}(\beta_{t-1}, \sigma^2 I) && (155)\\ w_{t,n} \,|\, \beta_t &\sim \textit{Mult}(\pi(\beta_t)) && (156) \end{aligned}$$

where $\pi$ maps the multinomial natural parameters to the mean parameters:

$$\pi(\beta_t)_w = \frac{\exp(\beta_{t,w})}{\sum_w \exp(\beta_{t,w})}. \tag{157}$$

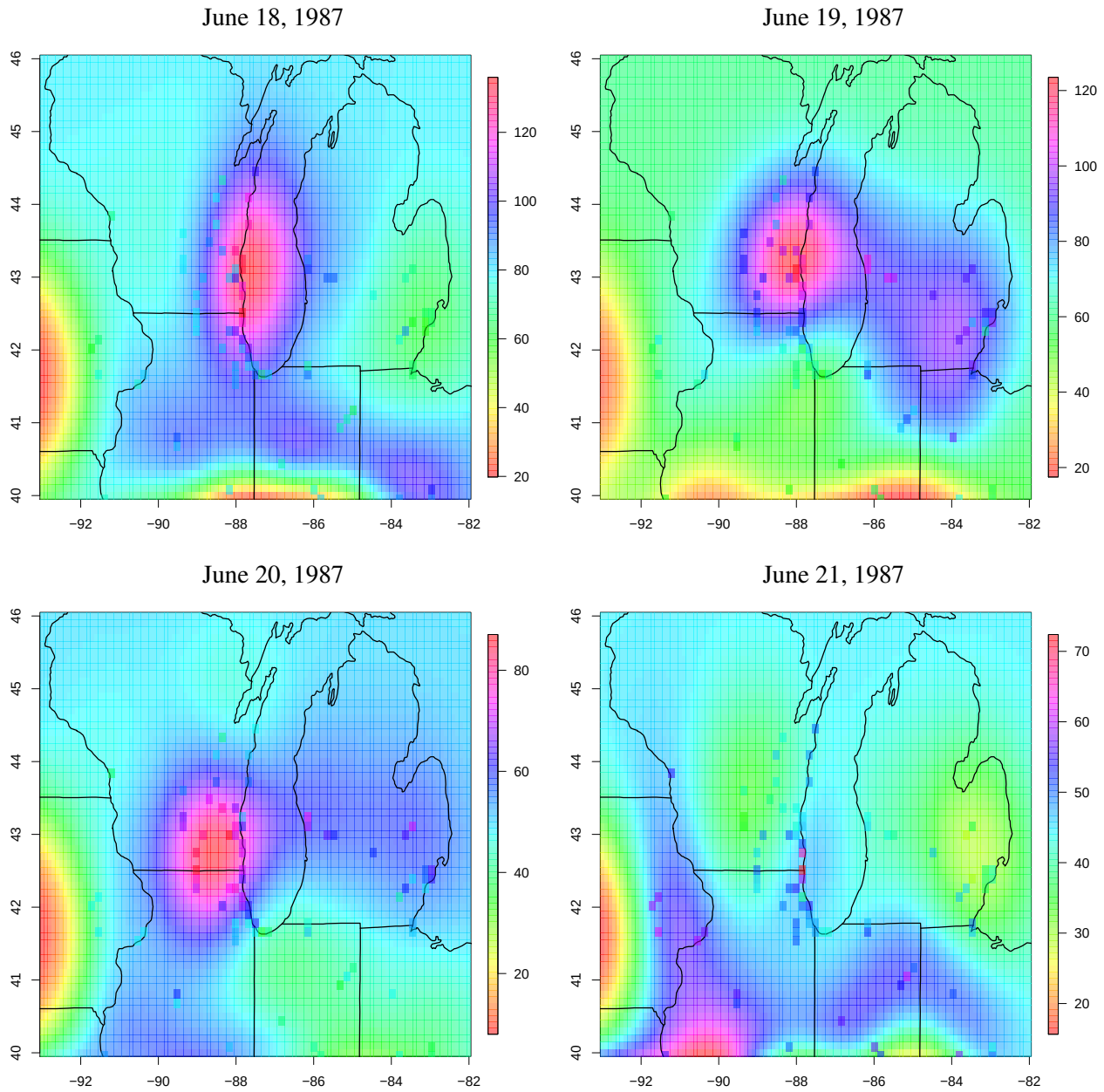


FIG 8. *Estimated daily maximum 8-hour ozone concentrations (in parts-per-billion) for the US midwest near Lake Michigan, for four days during the summer of 1987. The observed data are shown as colored rectangles; the Kalman smooth of these observed values is shown on a grid. Lake Michigan is near the center of each image; Chicago lies on its south-western shore.*

Since the observation model is multinomial and not Gaussian, it is not possible to use standard Kalman filtering and smoothing. However, we can use variational inference, and use the Kalman filter as a variational approximation.

To do so, we form a *variational state space model* where

$$\widehat{\beta}_t \,|\, \beta_t \;\sim\; \mathcal{N}(\beta_t, \widehat{\nu}_t^2 I). \tag{158}$$

The variational parameters are $\widehat{\beta}_t$ and $\widehat{\nu}_t$. Using the Kalman filtering calculations from Section 10, the forward mean and variance of the variational posterior are given by

$$\begin{aligned} m_t &\equiv \mathbb{E}(\beta_t \,|\, \widehat{\beta}_{1:t}) && (159)\\ &= \left(\frac{\widehat{\nu}_t^2}{V_{t-1}+\sigma^2+\widehat{\nu}_t^2}\right) m_{t-1} + \left(1 - \frac{\widehat{\nu}_t^2}{V_{t-1}+\sigma^2+\widehat{\nu}_t^2}\right)\widehat{\beta}_t && (160)\\ V_t &\equiv \mathbb{E}((\beta_t - m_t)^2 \,|\, \widehat{\beta}_{1:t}) && (161)\\ &= \left(\frac{\widehat{\nu}_t^2}{V_{t-1}+\sigma^2+\widehat{\nu}_t^2}\right)(V_{t-1}+\sigma^2) && (162) \end{aligned}$$

with initial conditions specified by fixed $m_0$ and $V_0$. The backward Kalman smoothing recursion then calculates the marginal mean and variance of $\beta_t$ given $\widehat{\beta}_{1:T}$ as

$$\begin{aligned} \widetilde{m}_{t-1} &\equiv \mathbb{E}(\beta_{t-1} \,|\, \widehat{\beta}_{1:T}) && (163)\\ &= \left(\frac{\sigma^2}{V_{t-1}+\sigma^2}\right) m_{t-1} + \left(1 - \frac{\sigma^2}{V_{t-1}+\sigma^2}\right)\widetilde{m}_t && (164)\\ \widetilde{V}_{t-1} &\equiv \mathbb{E}((\beta_{t-1} - \widetilde{m}_{t-1})^2 \,|\, \widehat{\beta}_{1:T}) && (165)\\ &= V_{t-1} + \left(\frac{V_{t-1}}{V_{t-1}+\sigma^2}\right)^2 \left(\widetilde{V}_t - (V_{t-1}+\sigma^2)\right) && (166)\\ &= \left(1 - \left(\frac{V_{t-1}}{V_{t-1}+\sigma^2}\right)^2\right) V_{t-1} + \left(\frac{V_{t-1}}{V_{t-1}+\sigma^2}\right)^2 \left(\widetilde{V}_t - \sigma^2\right), && (167) \end{aligned}$$

with initial conditions $\widetilde{m}_T = m_T$ and $\widetilde{V}_T = V_T$. We approximate the posterior $p(\beta_{1:T} \,|\, w_{1:T})$ using the state space posterior $q(\beta_{1:T} \,|\, \widehat{\beta}_{1:T})$. From Jensen's inequality, the log-likelihood is bounded from below as

$$\begin{aligned} \log p(d_{1:T}) &\geq \int q(\beta_{1:T} \,|\, \widehat{\beta}_{1:T}) \log \left(\frac{p(\beta_{1:T})\, p(d_{1:T} \,|\, \beta_{1:T})}{q(\beta_{1:T} \,|\, \widehat{\beta}_{1:T})}\right) d\beta_{1:T} && (168)\\ &= \mathbb{E}_q \log p(\beta_{1:T}) + \sum_{t=1}^{T} \mathbb{E}_q \log p(d_t \,|\, \beta_t) + H(q). && (169) \end{aligned}$$

Details of optimizing this bound are given in Blei and Lafferty (2006). Examples of the resulting Kalman smooths are shown in Figure 9. Here we model the probability of occurence of selected words in the journal *Science* between 1880 and 2003, fitting a simple unigram state space model to the data.

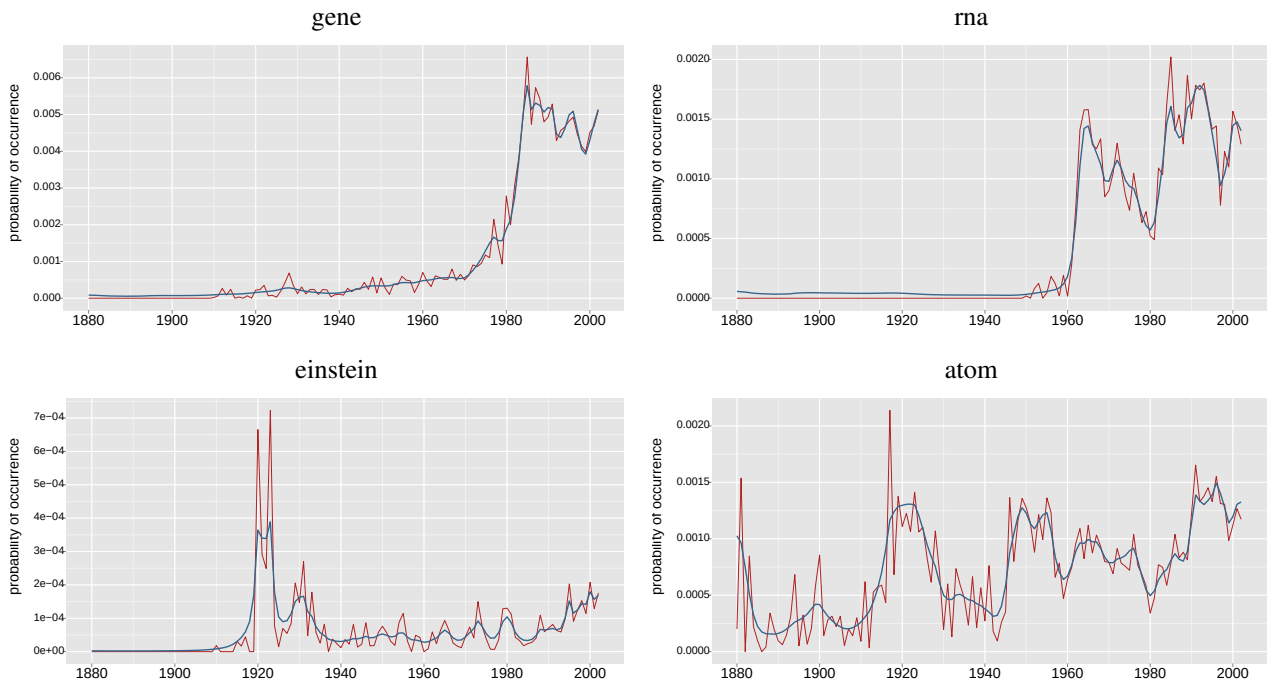


FIG 9. *Probability of occurence of selected words in the journal Science between 1880 and 2003. A simple unigram state space model is fit to the data. Since multinomial data is not Gaussian, the model is fit using a variational algorithm that uses Kalman filtering. The smoothed curves are the variational posterior means.*

## 13. Spectral Methods for HMMs

Recently hidden Markov models have been approached using spectral methods. This approach is based on viewing an HMM as a kind of linear state space model, as explained briefly in this section.

Using the matrix view of HMMs we can write

$$p(y_{1:T}) = \mathbf{1}^T A(y_T) A(y_{T-1}) \cdots A(y_1)\, \pi \tag{170}$$

with the notation of Section 3. Now let

$$H_t = \begin{pmatrix} \delta(X_t, 1) \\ \delta(X_t, 2) \\ \vdots \\ \delta(X_t, K) \end{pmatrix} \tag{171}$$

be the binary vector indicating the state $X_t$ at time $t$. Then $\mathbb{E}(H_{t+1} \mid H_t = h_t) = T h_t$. Similarly, let

$$Z_t = \begin{pmatrix} \delta(Y_t, 1) \\ \delta(Y_t, 2) \\ \vdots \\ \delta(Y_t, M) \end{pmatrix} \tag{172}$$

be the binary vector indicating the emission $Y_t$ at time $t$. Then $\mathbb{E}(Z_t \,|\, H_t = h_t) = Oh_t$. This suggests how an HMM can be viewed as a linear state space model, and this representation is behind the spectral approach to estimation.

Assume that $T$ and $O$ both have rank $K < M$. Define the $M \times M$ matrix $P^{(2)} \in \mathbf{R}^{M \times M}$ by

$$P^{(2)}_{ij} = p(Y_t = j, Y_{t+1} = i). \tag{173}$$

The matrix is easily estimated from data by simply counting bigram. Now consider the SVD of $P^{(2)}$:

$$P^{(2)} = U \Sigma V^T \tag{174}$$

where $U \in \mathbf{R}^{M \times K}$ and $V \in \mathbf{R}^{M \times K}$ are orthogonal matrices. Now define

$$B(y) = (U^T O) A(y) (U^T O)^{-1}. \tag{175}$$

Then

$$\begin{aligned} p(y_{1:T} &= \mathbf{1}^T A(y_T) \cdots A(y_1) \pi & (176) \\ &= \mathbf{1}^T (U^T O)^{-1} B(y_T) \cdots B(y_1) (U^T O) \pi. & (177) \end{aligned}$$

Some linear algebra shows that

$$B(y) = (U^T P^{(3)}(y))(U^T P^{(2)})^\dagger \tag{178}$$

where $P^{(3)}(y) \in \mathbf{R}^{M \times M}$ is a matrix of trigram probabilities

$$P^{(3)}(y)_{ij} = p(Y_{t-1} = j, Y_t = y, Y_{t+1} = i), \tag{179}$$

and $X^\dagger$ denotes the pseudo-inverse. Thus, to estimate probabilities it suffices to estimate bigram and trigram probabilities of output symbols, and to compute the SVD of the bigram matrix. The end factors $(U^T O)^{-1}$ and $(U^T O)\pi$ are estimated as $(U^T P^{(2)})^\dagger$ and $U^T P^{(2)} \mathbf{1}_M$.

**Theorem 13.7.** *Suppose that $\widehat{P}^{(2)}_n$ and $\widehat{P}^{(3)}_n$ are estimated from a sequence of length $n$, and the joint probability of an observation sequence is estimated as*

$$\widehat{p}_n(y_{1:T} = \mathbf{1}^T (U^T \widehat{P}^{(2)}_n)^{-1} B(y_T) \cdots B(y_1) U^T \widehat{P}^{(2)}_n \mathbf{1}_M \tag{180}$$

*Suppose the sequence length $n$ satisfies*

$$n > C \frac{t^2 K M}{\epsilon^2 \sigma^2_K(O) \sigma^2_K(P^{(2)})} \log \frac{1}{\delta} \tag{181}$$

*where $C$ is a constant and $\sigma_K$ denotes the $K$th largest singular value. Then with probability at least $1 - \delta$,*

$$\sum_{y_{1:T}} |p(y_{1:T}) - \widehat{p}_n(y_{1:T})| < \epsilon. \tag{182}$$

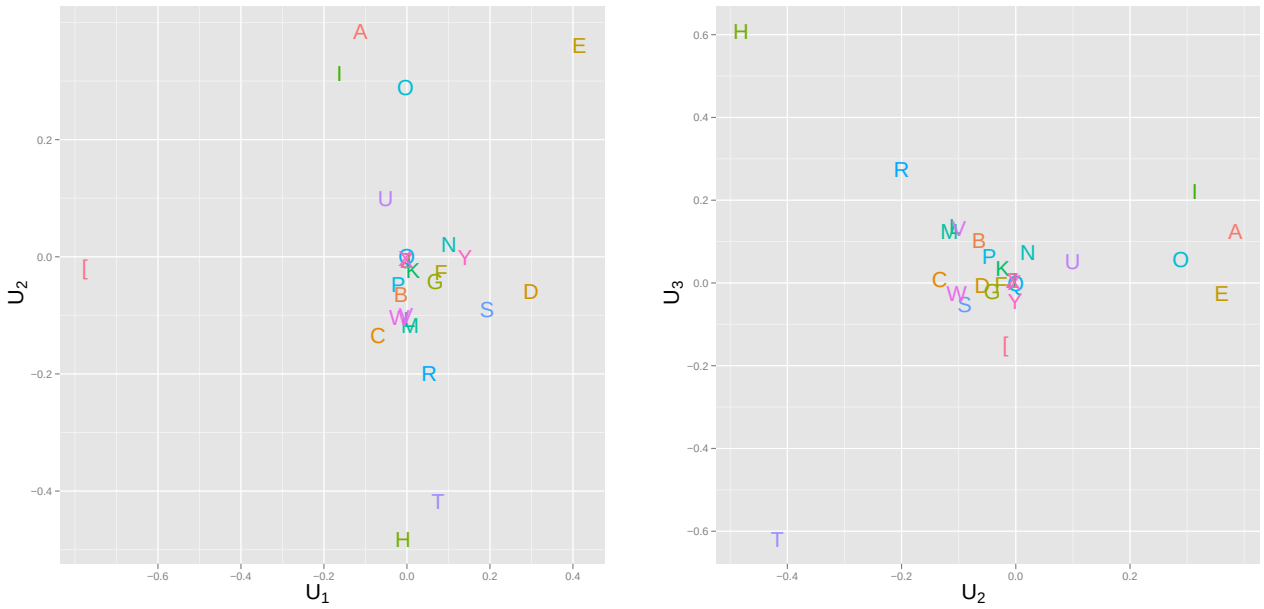


FIG 10. *The spectral approach to HMMs is based on the singular value decomposition of the bigram matrix* $p(x_{t-1}, x_t)$. *These plots show the first three singluar vectors* $U_1, U_2, U_3$ *for the alphabet* `A`–`Z` *together with space* `[`, *with probabilities estimated on six million characters of text. The vectors primarily separate the vowels, space, and* `T` *and* `H` *from the other letters.*

## 14. Notes

The calculations in Sections 10 and 11 follow the presentation of Minka (1999).

## References


Balasubramanian, V. (1993). Equivalence and reduction of hidden Markov models. Technical report, MIT Artificial Intelligency Laboratory, TR 1370.

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

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

Bickel, P. J. and Ritov, Y. (1996). Inference in hidden Markov models I: Local asymptotic normality in the stationary case. *Bernoulli*, 2:199–228.

Bickel, P. J., Ritov, Y., and Tydén, T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *The Annals of Statistics*, 26(4):1614–1635.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *ICML-06, 23rd International Conference on Machine Learning*, pages 113–120.

Ito, H., Amari, S.-I., and Kobayashi, K. (1992). Identifiability of hidden Markov information sources and their minimum degrees of freedom. *IEEE Trans. Information Theory*, 38(2):324–333.

Minka, T. (1999). From hidden Markov models to linear dynamical systems. Technical report, MIT Media Lab, TR-531.

Nádas, A. and Mercer, R. L. (1991). Hidden Markov models and some connections to artificial neural networks. In Smolensky, P., Mozer, M., and Rumelhart, D., editors, *Mathematical Perspectives on Neural Networks*, pages 603–649.

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

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal*.