Ch2. Hidden Markov models: definition and properties

HyeRim Park

July 21, 2023

Table of contents

Table of contents

Definition and notation

- A **hidden Markov model** $\{X_t : t \in \mathbb{N}\}$ is a particular kind of dependent mixture.

- The model consists of two parts:
  - Unobserved 'parameter process' $\{C_t : t = 1, 2, \cdots\}$ satisfying the Markov property.
  - The 'state-dependent process' $\{X_t : t = 1, 2, \cdots\}$

- With $\mathbf{X}^{(t)}$ and $\mathbf{C}^{(t)}$ representing the histories from time 1 to time t,

$$\Pr(C_t \mid \mathbf{C}^{(t-1)}) = \Pr(C_t \mid C_{t-1}), \quad t = 2, 3, \cdots \tag{1}$$

$$\Pr(X_t \mid \mathbf{X}^{(t-1)}, \mathbf{C}^{(t)}) = \Pr(X_t \mid C_t), \quad t \in \mathbb{N} \tag{2}$$

Definition and notation

- The distribution of X_t depends only on the current state C_t.
- If the Markov chain $\{C_t\}$ has m states, we call $\{X_t\}$ an m-state HMM.

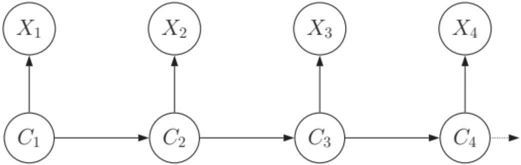

Figure 1: Directed graph of basic HMM.

Definition and notation

- Introduce some notation which cover both discrete- and continuous-valued observations.

- In the case of discrete observations, we define

$$p_i(x) = \Pr(X_t = x \mid C_t = i) \qquad for \quad i = 1, 2, \cdots, m$$

- p_i is the probability mass function of X_t if the Markov chain is in state *i* at time *t*.

- The m distributions p_i as the **state-dependent distributions** of the model.

Homogeneous and Stationary Markov chain

- Define **transition probabilities**:

$$\gamma_{ij}(t) = \Pr(C_{s+t} = j | C_s = i)$$

- If these probabilities do not depend on s, the Markov chain is called **homogeneous**.

- A Markov chain with transition probability matrix Γ is said to have **stationary distribution** δ if $\delta\Gamma = \delta$ and $\delta\mathbf{1}' = 1$.

Homogeneous and Stationary Markov chain

- Denote these by the row vector

$$u(t) = \ (\Pr(C_t = 1),\ \cdots,\ \Pr(C_t = m)), \quad t \in \mathbb{N}$$

- Homogeneity alone would not be sufficient to render the Markov chain a stationary process.
- Stationary for homogeneous Markov chains that have the additional property that the initial distribution $u(1)$ is the stationary distribution.

Marginal distributions

- We shall often need the marginal distribution of X_t and $(X_t,\ X_{t+k})$.
- Assume that the Markov chain is homogeneous but not necessarily stationary.
- For convenience the derivation is given only for discrete state-dependent distributions.

Marginal distributions: Univariate distributions

- For discrete-valued observations X_t,

$$\begin{aligned} \Pr(X_t = x) &= \sum_{i=1}^{m} \Pr(C_t = i) \Pr(X_t = x \mid C_t = i) \\ &= \sum_{i=1}^{m} u_i(t)\, p_i(x) \end{aligned}$$

where $u_i(t) = \Pr(C_t = i) \quad for\ t = 1, \cdots, T$.

Marginal distributions: Univariate distributions

- This expression can conveniently be rewritten in matrix notation:

$$
\begin{aligned}
\Pr(X_t = x) &= (u_1(t), \cdots, u_m(t)) \begin{pmatrix} p_1(x) & & 0 \\ & \ddots & \\ 0 & & p_m(x) \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \\
&= \mathbf{u}(t)\mathbf{P}(x)\mathbf{1},
\end{aligned}
$$

where $\mathbf{P}(x)$ is defined as the diagonal matrix with ith diagonal element $p_i(x)$.

Marginal distributions: Univariate distributions

- From equation $\mathbf{u}(t+1) = \mathbf{u}(t)\,\boldsymbol{\Gamma}$, $\mathbf{u}(t) = \mathbf{u}(1)\boldsymbol{\Gamma}^{t-1}$ and then

$$\Pr(X_t = x) = \mathbf{u}(1)\boldsymbol{\Gamma}^{t-1}\mathbf{P}(x)\mathbf{1} \tag{3}$$

- If the Markov chain is stationary, with stationary distribution $\boldsymbol{\delta}$, in that case $\boldsymbol{\delta}\boldsymbol{\Gamma}^{t-1} = \boldsymbol{\delta}$ for all $t \in \mathbb{N}$, and so

$$\Pr(X_t = x) = \boldsymbol{\delta}\mathbf{P}(x)\mathbf{1} \tag{4}$$

Marginal distributions: Bivariate distributions

- In any directed graphical model, the joint distribution of a set of random variables V_i is given by

$$\Pr(V_1, V_2, \cdots, V_n) = \prod_{i=1}^{n} \Pr(V_i \mid \mathrm{pa}(V_i)), \tag{5}$$

  where $\mathrm{pa}(V_i)$ denotes the set of all 'parents' of V_i in the set $V_1, V_2, \cdots, V_n$.

- In the directed graph of the four random variables X_t, X_{t+k}, C_t, C_{t+k} for positive integer k, $\mathrm{pa}(X_t) = \{C_t\}$, $\mathrm{pa}(C_{t+k}) = \{C_t\}$, $\mathrm{pa}(X_{t+k}) = \{C_{t+k}\}$.

Marginal distributions: Bivariate distributions

- $$\Pr(X_t, X_{t+k}, C_t, C_{t+k}) \\ = \Pr(C_t)\Pr(X_t \mid C_t)\Pr(C_{t+k} \mid C_t)\Pr(X_{t+k} \mid C_{t+k})$$

- $$\begin{aligned} \Pr(X_t = v, X_{t+k} = w) &= \sum_{i=1}^{m}\sum_{j=1}^{m} \Pr(X_t = v, X_{t+k} = w, C_t = i, C_{t+k} = j) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m} \underbrace{\Pr(C_t = i)}_{u_i(t)}\, p_i(v)\, \underbrace{\Pr(C_{t+k} = j \mid C_t = i)}_{\gamma_{ij}(k)}\, p_j(w) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{m} u_i(t) p_i(v) \gamma_{ij}(k) p_j(w) \end{aligned}$$

where $\gamma_{ij}(k)$ denotes the $(i,\, j)$ element of $\mathbf{\Gamma}^k$.

Marginal distributions: Bivariate distributions

- Writing the above double sum as a product of matrices

$$\Pr(X_t = v, X_{t+k} = w) = \mathbf{u}(t)\mathbf{P}(v)\mathbf{\Gamma}^k\mathbf{P}(w)\mathbf{1} \tag{6}$$

- If the Markov chain is stationary,

$$\Pr(X_t = v, X_{t+k} = w) = \boldsymbol{\delta}\mathbf{P}(v)\mathbf{\Gamma}^k\mathbf{P}(w)\mathbf{1} \tag{7}$$

Moments

- We note that

$$\begin{aligned}\mathrm{E}(X_t) &= \sum_{i=1}^{m} \mathrm{E}(X_t \mid C_t = i)\,\Pr(C_t = i)\\ &= \sum_{i=1}^{m} u_i(t)\,\mathrm{E}(X_t \mid C_t = i)\end{aligned}$$

- In the stationary case, $\mathrm{E}(X_t) = \sum_{i=1}^{m} \delta_i\,\mathrm{E}(X_t \mid C_t = i)$

Moments

- More generally, for any functions g in the stationary case

$$\mathrm{E}(g(X_t)) = \sum_{i=1}^{m} \delta_i \, \mathrm{E}(g(X_t) \,|\, C_t = i) \tag{8}$$

$$\mathrm{E}(g(X_t, X_{t+k})) = \sum_{i,j=1}^{m} \mathrm{E}(g(X_t, X_{t+k}) \,|\, C_t = i, C_{t+k} = j) \delta_i \gamma_{ij}(k) \tag{9}$$

- If a function g factorizes as

$$g(X_t, X_{t+k}) = g_1(X_t) g_2(X_{t+k}),$$

$$\mathrm{E}(g(X_t, X_{t+k})) = \sum_{i,j=1}^{m} \mathrm{E}(g_1(X_t) \,|\, C_t = i) \, \mathrm{E}(g_2(X_{t+k}) \,|\, C_{t+k} = j) \delta_i \gamma_{ij}(k) \tag{10}$$

Table of contents

The likelihood

- The aim of this section is to develop a convenient formula for the likelihood L_T of T consecutive observations $x_1, x_2, \ldots, x_T$ assumed to be generated by an m-state HMM.

- The computation of the likelihood appears to require $O(Tm^T)$ operations.

$$
\begin{aligned}
&\Pr(X_1 = x_1, \ldots, X_T = x_T) \\
&\quad = \sum_{c_1,\ldots,c_T=1}^{m} \Pr(X_1 = x_1, \ldots, X_T = x_T, C_1 = c_1, \ldots, C_T = c_T)
\end{aligned}
$$

$$
\begin{aligned}
&\Pr(X_1, \ldots, X_T, C_1, \ldots, C_T) \\
&\quad = \underbrace{\Pr(X_1|C_1) \cdots \Pr(X_T|C_T)}_{product\, of\ T\ factors} \underbrace{\Pr(C_1)\Pr(C_2|C_1) \cdots \Pr(C_T|C_{T-1})}_{product\, of\ T\ factors}
\end{aligned}
$$

The likelihood

- It is our purpose here to demonstrate that L_T can in general be computed relatively simply in $O(Tm^2)$ operations.

- First the likelihood of a two-state model will be explored.

- Then the general formula will be presented.

The likelihood of a two-state Bernoulli–HMM

- **Example (Bernoulli–HMM)**
  Consider the stationary two-state HMM with t.p.m.

$$\mathbf{\Gamma} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{4} & \frac{3}{4} \end{pmatrix}$$

  and state-dependent distributions given by

$$\Pr\left(X_t = x \mid C_t = 1\right) = \frac{1}{2} \quad (for\ x = 0, 1),$$
$$\Pr\left(X_t = 1 \mid C_t = 2\right) = 1$$

- The stationary distribution of the Markov chain is $\boldsymbol{\delta} = \frac{1}{3}(1, 2)$

The likelihood of a two-state Bernoulli–HMM

- Note that

$$\begin{aligned}
&\Pr(X_1, X_2, X_3, C_1, C_2, C_3) \\
&= \Pr(C_1)\Pr(X_1|C_1)\Pr(C_2|C_1)\Pr(X_2|C_2)\Pr(C_3|C_2)\Pr(X_3|C_3)
\end{aligned}$$

$$\begin{aligned}
&\Pr(X_1 = 1, X_2 = 1, X_3 = 1) \\
&\quad = \sum_{i,j,k=1}^{2} \Pr(X_1 = 1, X_2 = 1, X_3 = 1, C_1 = i, C_2 = j, C_3 = k) \\
&\quad = \sum_{i,j,k=1}^{2} \delta_i p_i(1)\gamma_{ij} p_j(1)\gamma_{jk} p_k(1) \qquad (11)
\end{aligned}$$

- Notice that the triple sum (11) has $m^T = 2^3$ terms, each of which is a product of $2T = 2 \times 3$ factors.

The likelihood of a two-state Bernoulli–HMM

Table 1: Example of a likelihood computation.

i	j	k	$p_i(1)$	$p_j(1)$	$p_k(1)$	δ_i	γ_{ij}	γ_{jk}	Product
1	1	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{2}{4}$	$\frac{2}{4}$	$\frac{1}{96}$
1	1	2	$\frac{1}{2}$	$\frac{1}{2}$	1	$\frac{1}{3}$	$\frac{2}{4}$	$\frac{2}{4}$	$\frac{1}{48}$
1	2	1	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{2}{4}$	$\frac{1}{4}$	$\frac{1}{96}$
1	2	2	$\frac{1}{2}$	1	1	$\frac{1}{3}$	$\frac{2}{4}$	$\frac{3}{4}$	$\frac{1}{16}$
2	1	1	1	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{48}$
2	1	2	1	$\frac{1}{2}$	1	$\frac{2}{3}$	$\frac{1}{4}$	$\frac{2}{4}$	$\frac{1}{24}$
2	2	1	1	1	$\frac{1}{2}$	$\frac{2}{3}$	$\frac{3}{4}$	$\frac{1}{4}$	$\frac{1}{16}$
2	2	2	1	1	1	$\frac{2}{3}$	$\frac{3}{4}$	$\frac{3}{4}$	$\mathbf{\frac{3}{8}}$
									$\mathbf{\frac{29}{48}}$

The likelihood of a two-state Bernoulli–HMM

- The state sequence that maximizes the joint probability

$$\Pr\left(X_1 = 1, X_2 = 1, X_3 = 1, C_1 = i, C_2 = j, C_3 = k\right)$$

is therefore the sequence $i = 2, j = 2, k = 2$.

- More convenient way to present the sum is to use matrix notation. Let $\mathbf{P}(u)$ be defined (as before) as $\mathrm{diag}\left(p_1(u), p_2(u)\right)$. Then

$$\mathbf{P}(0) = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad \mathbf{P}(1) = \begin{pmatrix} \frac{1}{2} & 0 \\ 0 & 1 \end{pmatrix},$$

and the triple sum (11) can be written as a matrix product:

$$\Pr(X_1 = 1, X_2 = 1, X_3 = 1) = \delta\mathbf{P}(1)\boldsymbol{\Gamma}\boldsymbol{P}(1)\boldsymbol{\Gamma}\boldsymbol{P}(1)\mathbf{1}'$$

The likelihood in general

- Consider the likelihood of an HMM in general.
- Suppose there is an observation sequence $x_1, x_2, \ldots, x_T$ generated by such a model.
- Seek the probability L_T of observing that sequence under an m-state HMM that has initial distribution $\boldsymbol{\delta}$ and t.p.m. $\boldsymbol{\Gamma}$ for the Markov chain, and state-dependent probability (density) functions p_i.

The likelihood in general

Proposition 1

The likelihood is given by

$$L_T = \boldsymbol{\delta}\mathbf{P}(x_1)\,\mathbf{\Gamma P}(x_2)\,\mathbf{\Gamma P}(x_3)\cdots\mathbf{\Gamma P}(x_T)\,\mathbf{1}. \tag{12}$$

If $\boldsymbol{\delta}$, the distribution of C_1, is the stationary distribution of the Markov chain, then in addition

$$L_T = \boldsymbol{\delta}\mathbf{\Gamma}\boldsymbol{P}(x_1)\,\mathbf{\Gamma P}(x_2)\,\mathbf{\Gamma P}(x_3)\cdots\mathbf{\Gamma}\boldsymbol{P}(x_T)\,\mathbf{1}. \tag{13}$$

Proof

Only the case of discrete observations. First, note that

$$L_T = \Pr\left(\mathbf{X}^{(T)} = \mathbf{x}^{(T)}\right) = \sum_{c_1,c_2,\ldots,c_T=1}^{m} \Pr\left(\mathbf{X}^{(T)} = \mathbf{x}^{(T)}, \mathbf{C}^{(T)} = \mathbf{c}^{(T)}\right),$$

The likelihood in general

Proof (Cont.)

and that, by equation (5),

$$\Pr\left(\mathbf{X}^{(T)}, \mathbf{C}^{(T)}\right) = \Pr\left(C_1\right) \prod_{k=2}^{T} \Pr\left(C_k \mid C_{k-1}\right) \prod_{k=1}^{T} \Pr\left(X_k \mid C_k\right). \tag{14}$$

It follows that

$$\begin{aligned} L_T &= \sum_{c_1,\dots,c_T=1}^{m} \left(\delta_{c_1}\gamma_{c_1,c_2}\gamma_{c_2,c_3}\cdots\gamma_{c_{T-1},c_T}\right)\left(p_{c_1}\left(x_1\right)p_{c_2}\left(x_2\right)\cdots p_{c_T}\left(x_T\right)\right) \\ &= \sum_{c_1,\dots,c_T=1}^{m} \delta_{c_1}p_{c_1}\left(x_1\right)\gamma_{c_1,c_2}p_{c_2}\left(x_2\right)\gamma_{c_2,c_3}\cdots\gamma_{c_{T-1},c_T}p_{c_T}\left(x_T\right) \\ &= \boldsymbol{\delta}\mathbf{P}\left(x_1\right)\boldsymbol{\Gamma}\mathbf{P}\left(x_2\right)\boldsymbol{\Gamma}\mathbf{P}\left(x_3\right)\cdots\boldsymbol{\Gamma}\mathbf{P}\left(x_T\right)\mathbf{1}, \end{aligned}$$

If $\boldsymbol{\delta}$ is the stationary distribution, $\boldsymbol{\delta}\mathbf{P}\left(x_1\right) = \boldsymbol{\delta}\boldsymbol{\Gamma}\mathbf{P}\left(x_1\right)$ □

The likelihood in general

- A consequence of the matrix expression for the likelihood is the 'forward algorithm' for recursive computation of the likelihood.

- The recursive nature of likelihood evaluation via either (12) is computationally much more efficient than brute-force summation over all possible state sequences.

- To state the forward algorithm, define the vector $\boldsymbol{\alpha}_t$, for $t = 1, 2, \ldots, T$, by

$$\boldsymbol{\alpha}_t = \boldsymbol{\delta} \mathbf{P}(x_1) \boldsymbol{\Gamma} \mathbf{P}(x_2) \boldsymbol{\Gamma} \mathbf{P}(x_3) \cdots \boldsymbol{\Gamma} \mathbf{P}(x_t) = \boldsymbol{\delta} \mathbf{P}(x_1) \prod_{s=2}^{t} \boldsymbol{\Gamma} \mathbf{P}(x_s), \tag{15}$$

The likelihood in general

- Then, in the likelihood formula (12) :

$$\begin{aligned} \boldsymbol{\alpha}_1 &= \boldsymbol{\delta}\mathbf{P}(x_1); \\ \boldsymbol{\alpha}_t &= \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\mathbf{P}(x_t) \qquad \text{for } t = 2, 3, \ldots, T; \\ L_T &= \boldsymbol{\alpha}_T \mathbf{1}. \end{aligned}$$

- That the number of operations involved is of order Tm^2 can be deduced thus.

- The elements of the vector $\boldsymbol{\alpha}_t$ are usually referred to as **forward probabilities**.

- The multiple-sum expression for the likelihood $\Longrightarrow$ the matrix expression $\Longrightarrow$ the forward recursion.

The likelihood when data are missing

- If some of the data are missing, the likelihood computation turns out to be a simple one.

- **Example**
  Suppose that one has available the observations $x_1, x_2,$ $x_4, x_7, x_8, \ldots, x_T$ of an HMM, but x_3, x_5 and x_6 are missing. Then the likelihood of the observations is given by

$$
\begin{aligned}
&\Pr\left(X_1 = x_1, X_2 = x_2, X_4 = x_4, X_7 = x_7, \ldots, X_T = x_T\right) \\
&= \sum \delta_{c_1} \gamma_{c_1,c_2} \gamma_{c_2,c_4}(2) \gamma_{c_4,c_7}(3) \gamma_{c_7,c_8} \cdots \gamma_{c_{T-1},c_T} \\
&\qquad \times p_{c_1}\left(x_1\right) p_{c_2}\left(x_2\right) p_{c_4}\left(x_4\right) p_{c_7}\left(x_7\right) \cdots p_{c_T}\left(x_T\right),
\end{aligned}
$$

  where the sum is taken over all indices c_t other than c_3, c_5 and c_6.

The likelihood when data are missing

- This is just

$$\sum \delta_{c_1} p_{c_1}(x_1)\, \gamma_{c_1,c_2} p_{c_2}(x_2)\, \gamma_{c_2,c_4}(2) p_{c_4}(x_4)\, \gamma_{c_4,c_7}(3) p_{c_7}(x_7) \\ \times \cdots \times \gamma_{c_{T-1},c_T} p_{c_T}(x_T) \\ = \boldsymbol{\delta}\mathbf{P}(x_1)\,\boldsymbol{\Gamma}\mathbf{P}(x_2)\,\boldsymbol{\Gamma}^2\mathbf{P}(x_4)\,\boldsymbol{\Gamma}^3\mathbf{P}(x_7)\cdots\boldsymbol{\Gamma}\mathbf{P}(x_T)\,\mathbf{1}$$

- In general, in the expression for the likelihood the diagonal matrices $\mathbf{P}(x_t)$ corresponding to missing observations x_t are replaced by the identity matrix.

- Thus, even if some observations are missing, the likelihood of an HMM can be computed easily.