Hidden Markov Models

The Basic Building Block: Transition Function

- Suppose $X_t$ take values in \{0,1,2,...\}
- One-step transition function: $P_{i,j}^{(1)} = P(X_{t+1} = j | X_t = i)$
- Time-homogeneous MC: $P_{i,j}^{(n)} = P_{i,j}$
- Multi-step transition function
  $P_{i,j}^{(n)} = P(X_{t+n} = j | X_t = i)$
- Chapman-Kolmogorov equation:
  $$P_{i,j}^{(n)} = \sum_k P_{i,k}^{(n)} P_{k,j}$$
  $$P^{n+m} = P^n \cdot P^m$$

Basics of Markov chain

- **Definition:** $X_0, X_1, ..., X_n$ is called a Markov chain if
  $P(X_s = x_s | X_t = x_t, ..., X_n = x_n) = P(X_s = x_s | X_{s-1} = x_{s-1})$
- **Example** of Markov chains
  - Simple random walk with reflecting boundary: $X_t \in \{0,1,...,b\}$
    $P(X_t = j | X_{t-1} = i)$
    Written in a matrix form:

  \[
  \begin{pmatrix}
  \cdots & \cdots & \cdots & \cdots & \cdots \\
  \cdots & 0 & 1 & 0 & \cdots \\
  \cdots & 1 & 0 & 1 & \cdots \\
  \cdots & 0 & 1 & 0 & \cdots \\
  \cdots & \cdots & \cdots & \cdots & \cdots 
  \end{pmatrix}
  \]

  Properties of this matrix?

- **Suppose** $X_t$ take values in \{0,1,2,...\}
  - One-step transition function: $P_{i,j}^{(1)} = P(X_{t+1} = j | X_t = i)$
  - Time-homogeneous MC: $P_{i,j}^{(n)} = P_{i,j}$

A population genetics model:

Generation $n$:
$X_n$: Type $A$
$2N - X_n$: Type $a$

Generation $n+1$:
- randomly select $2N$ individuals from the previous generation

$$P(X_{n+1} = k | X_n = j) = \binom{2N}{k} p^k (1-p)^{2N-k}$$

Of interest: if $X_0 = i$, what is the fixation time?

\[
\text{time for which } X_n = 0 \text{ or } 2N
\]
Long-Run Behavior --- The Equilibrium Distribution

• General tendency of a Markov chain: “stabilize”
  – 2-state case: \[
  \begin{pmatrix}
  .7 & .3 \\
  .1 & .9 \\
  \end{pmatrix}
  \rightarrow
  \begin{pmatrix}
  .7 & .3 \\
  .1 & .9 \\
  \end{pmatrix}^n
  \rightarrow
  \begin{pmatrix}
  .25 & .75 \\
  .25 & .75 \\
  \end{pmatrix}
  \rightarrow
  \ldots
  \]
  – Its simulation
  
  
  In 100 steps, 77 ones and 23 zeros.
  
  – Another meaning of “equilibrium”
  
  \[
  \begin{pmatrix}
  .7 & .3 \\
  .1 & .9 \\
  \end{pmatrix}
  \rightarrow
  \begin{pmatrix}
  25 & 75 \\
  25 & 75 \\
  \end{pmatrix}
  \rightarrow
  \ldots
  \]
  – Summary: three faces of the “equilibrium distribution”

What conditions?

• The chain has to be “irreducible”, e.g.,
  \[
  \forall i, j, \exists n, \text{ so that } P(X_n = j | X_0 = i) > 0
  \]
  – Graphically, it means the states are all “connected.”
  
  • The chain has to be “aperiodic”
  
  – Thus, we can find a \( k_0 \) so that \( P^k \) has all the entries \( > 0 \) for \( k > k_0 \)
  
  • A chain is called “regular” if it is both irreducible and aperiodic.
  
  Now we try to answer why the Markov chain has a stationary distribution.
  
  – First
  
  \[
  P^n \rightarrow \pi
  \]
  
  \[\pi = (\pi_1, \ldots, \pi_n)\]
  
  – Second, \( \pi \) is the unique probability vector such that \( \pi = \pi P \)

Hidden Markov Models

• Markov:
  \[x_1 \rightarrow x_2 \rightarrow \cdots \rightarrow x_n\]
  
  • HMM:
  \[y_1 \rightarrow y_2 \rightarrow y_3 \rightarrow y_n\]
  
  \[z_1 \rightarrow z_2 \rightarrow z_3 \rightarrow \cdots \rightarrow z_n\]
  
  – We only observe the \( y \)’s;
  
  – The \( z \)’s governed by a Markov transition rule
  
  \[p_i = P(z_i = j | z_{i-1} = i)\]
  
  – The \( y \) can be viewed as a “noisy copy” of the \( z \)
  
  \[y_i | z_i \sim f_i(z_i)\]
  
  – Example: \( z_i \) indicates intron or exon; \( y_i \) observed sequence

An example

• Occasionally dishonest casino

Simulated a sequence of 100 “die usages”

\[
\begin{array}{cccccccccccccccc}
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 & 15 & 16 & 17 \\
\hline
1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
2 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
3 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
4 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
5 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
6 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
7 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
8 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
9 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
10 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
11 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
12 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
13 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
14 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
15 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
16 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
17 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \\
\end{array}
\]

Question: what’s the long-run fraction of each type?

However, we only observe the result of the rolls of a die: can we tell?
How do we find the most probable path?

• Optimal probability of the observed rolls $Y_1, Y_2, \cdots, Y_n$

$$f(z) = P(Y_1 | z_i = z) \times \max_{z_{i+1}} f_{i+1}(z')P(z_i = z | z_{i+1} = z')$$

$f(z)$ is the optimal prob value for the first $i$ observations with $z_i = z$.

• Tracing backward to find the optimal path

$$z_i = \arg \max_z \left[ f(z)P(z_{i+1} | z_i = z) \right]$$

• Total probability and Monte Carlo sampling

Forward-sum and Backward sampling

• Goal: sample $(z_1, \ldots, z_n)$ from $P(z_1, \ldots, z_n | y_1, \ldots, y_n)$

• Idea: sample

$$z_1 \sim P(z_1 | y_1, \ldots, y_n)$$
$$z_{i+1} \sim P(z_{i+2} | y_1, \ldots, y_n, z_{i+1})$$
$$\vdots$$
$$z_n \sim P(z_n | y_1, \ldots, y_n, z_{i+1}, \ldots, z_n)$$

• Recursive summation to get marginal probability

$$g(z) = P(y_i | z_i = z) \times \sum z_{i+1} g_{i+1}(z')P(z_i = z | z_{i+1} = z')$$

• Backward sampling:

Draw $z_n \sim P(z_n = z) = g_n(z) \sum z_{n+1} g_{n+1}(z')$

Draw $z_i \sim P(z_i = z) = \frac{g_i(z)}{\sum z_{i+1} g_{i+1}(z')}$

Hidden Markov Model

Simulated a sequence of 100 die tosses

Posterior mean Marginal posterior mode

Joint posterior mode

Forward-sum and Backward Sampling

• Recursive summation to get marginal probability

$$g(z) = P(y_i | z_i = z) \times \sum z_{i+1} g_{i+1}(z')P(z_i = z | z_{i+1} = z')$$

• Backward sampling:

Draw $z_n \sim P(z_n = z) = g_n(z) \sum z_{n+1} g_{n+1}(z')$

Draw $z_i \sim P(z_i = z) = \frac{g_i(z)}{\sum z_{i+1} g_{i+1}(z')}$
A HMM for Sequence Alignment

• A generative model

\[
P(\text{red path}) = \delta(1-e-d)\delta(1-\delta(1-e-d))
\]

\[
P(\text{ATCTTG | red path}) = p_{a1}p_{a2}p_{a3}p_{a4}p_{a5}
\]

What is hidden?

Is it Markovian?

The Sequence Alignment problem

Given a pair of sequences, how do we view them?

A better representation?

Or

Alignment Recursion

\[
F(i, j) = \max \left\{ F(i-1, j-1) + s(x_i, y_j), F(i-1, j) - \gamma, F(i, j-1) - \gamma \right\}
\]
The block-motif model

- Only a “block” of residues are conserved

\[ M_0 \rightarrow M_1 \rightarrow M_2 \rightarrow \ldots \rightarrow M_w \]

Data:

\[ a_1 \rightarrow a_2 \rightarrow \ldots \rightarrow a_{n_k} \]

The Hidden Markov Model

- For given \( z_s \), \( y_s \sim f(y_s | z_s, f) \), and the \( z_s \) follow a Markov process with transition \( p_s(z_s | z_{s-1}, q) \).

\[ y_1 \rightarrow y_2 \rightarrow y_3 \rightarrow \ldots \rightarrow y_s \rightarrow \ldots \rightarrow y_t \]

\[ \pi_s(z_s) = p(z_1, \ldots, z_s | y_1, \ldots, y_s ; \phi, \theta) \]

“The State Space Model”

What Are Hidden in Sequence Alignment?

- The Architecture

\[ G_{ij} = \begin{cases} 0 & \text{if generated by insertion} \\ 1 & \text{if generated by a match} \end{cases} \]

The Path

\[ \delta_i = \# \text{ of deletions} \]

Solid path: (0,1) → (2,0) → (2,0) → (2,1) → (2,0) → (2,0)

Dashed path: (0,1) → (0,0) → (0,0) → (2,1) → (2,0) → (2,0)
An Equivalent Model

- No Deletion Model
- Deletion Indicators

More Restricted Models

- Block Motif Propagation Model: limit the total number of gaps but no deletions allowed.

- With Deletion Indicators: