Time reversibility

Let P be the transition matrix of an irreducible Markov chain with stationary distribution v.

- The chain is "time reversible" if, when running from its stationary distribution, it looks the same moving foreard as backwards, i.e., $\pi(X_k = i, X_{k+1} = j) = \pi(X_{k+1} = i, X_k = j)$.
- This may also be written as $v_i P_{ij} = v_j P_{ji}$ for all i, j: The *detailed balance condition*.
- Show: If x is a probability vector satisfying $x_i P_{ij} = x_j P_{ji}$ for all i, j, then necessarily x is the stationary distribution, so that $x = v$.
- Show: If a Markov chain is defined as a random walk on a weighted undirected graph, then it is time reversible.
- Show: If a finite Markov chain is time reversible, it can be represented as a random walk on a weighted undirected graph.

Canonical decomposition (assume a finite state space)

- The states of a Markov chain can be subdivided into communication classes, each consisting only of transient or recurrent states.
- Let T denote the union of all communication classes with transient states. Let remaining communication classes be $R_1, R_2, \ldots, R_m$.
- Each R_i must necessarily be *closed* in the sense that no states outside R_i are accessible from R_i.
- Ordering states according to T, R_1, ..., R_m, the transition matrix can be written

$$P = \begin{bmatrix} * & * & \cdots & * \\ 0 & P_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & P_m \end{bmatrix}.$$

- We get

$$P^n = \begin{bmatrix} * & * & \cdots & * \\ 0 & P_1^n & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & P_m^n \end{bmatrix}$$

and can take the limits of each P_i^n, if they exist.

Absorbing chains

- State i is *absorbing* if $P_{ii} = 1$.
- A Markov chain is *absorbing* if it has at least one absorbing state.
- By reordering the states, the transition matrix for an absorbing chain can be written in block form
$$P = \begin{bmatrix} Q & R \\ \mathbf{0} & I \end{bmatrix}.$$
where I is the identity matrix, $\mathbf{0}$ is a matrix of zeros, and Q corresponds to transient states.
- We can prove by induction that
$$P^n = \begin{bmatrix} Q^n & \left(I + Q + Q^2 + \cdots + Q^{n-1}\right) R \\ \mathbf{0} & I \end{bmatrix}.$$
- Taking the limit and using $\lim_{n\to\infty} Q^n = 0$ we get
$$\lim_{n\to\infty} P^n = \begin{bmatrix} \mathbf{0} & (I-Q)^{-1} R \\ \mathbf{0} & I \end{bmatrix} = \begin{bmatrix} \mathbf{0} & FR \\ \mathbf{0} & I \end{bmatrix}.$$
- $F = (I-Q)^{-1} = \lim_{n\to\infty} I + Q + \cdots + Q^n$ is called the *fundamental matrix*.

Absorbing chains, cont

- The probability to be absorbed in a particular absorbing state given a start in a transient state is given by the entries of FR.
- Further, the expected number of visits in transient state j for a chain that starts in the transient state i is given by F_{ij}. (See proof in Dobrow).
- Thus, the expected number of steps until absorbtion is given by the vector $F\mathbf{1}^t$.
- Note: Given an irreducible Markov chain. To compute the expected number of steps needed to go from state i to the first visit to state j, one can change the chain into one where state j is absorbing, and compute the expected number of steps until absorbtion using the theory above.

Example: First detection of a particular sequence

- Assume you want to find the expected number of steps until you detect HTTH in a sequence of fair coin flips.
- Build a Markov chain where the states indicate how far into the sequence you have read so far. Make the state HTTH absorbing.
- Find the transition matrix in canonical block form.

MVE550 2022 Lecture 5
Compendium chapters 2 and 3
Hidden Markov Models (HMM)
Inference for Markov chains and HMMs

Petter Mostad

Chalmers University

November 15, 2022

Overview

- Hidden Markov Models: Introduction and examples
- Inference questions for HMMs.
- The Multinomial-Dirichlet conjugacy.
- Some inference for Markov chains.
- Some inference for HMMs.

Example: Not quite a Markov chain

Exercise 2.20 from Dobrow:

- Let $X_0, X_1, \dots$ be a Markov chain with transition matrix

$$P = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ p & 1-p & 0 \end{bmatrix}$$

  for some $0 < p < 1$. Let g be the function defined by

$$g(x) = \begin{cases} 0, & \text{if } x = 1 \\ 1, & \text{if } x = 2,3 \end{cases}$$

  If we let $Y_n = g(X_n)$ for $n \geq 0$ is $Y_0, Y_1, \dots$ a Markov chain?
- Common phenomenon: The underlying process may reasonably be a Markov chain, but what we observe is not!

Hidden Markov Models

- A Hidden Markov Model (HMM) consists of
  - a Markov chain $X_0, \ldots, X_n, \ldots,$, and
  - another sequence $Y_0, \ldots, Y_n, \ldots,$ so that

$$\Pr(Y_k \mid Y_0, \ldots, Y_{k-1}, X_0, \ldots, X_k) = \Pr(Y_k \mid X_k)$$

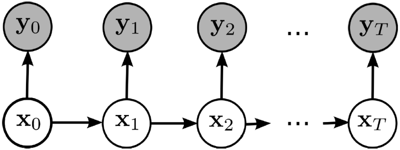
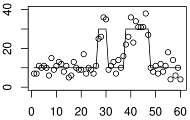


Figure: A hidden Markov model.

- In some models we instead have $\Pr(Y_k \mid Y_0, \ldots, Y_{k-1}, X_0, \ldots, X_k) = \Pr(Y_k \mid Y_{k-1}, X_k)$. There are then extra arrows from y_{k-1} to y_k in the figure above.
- Generally, $Y_0, \ldots, Y_k \ldots,$ are *observed*, while $X_0, \ldots, X_k \ldots,$ are *hidden*.
- In our applications, the X_k have a finite state space and the Y_k are discrete.

Example 1: Cough medicine

- Each day i a pharmacy sells Y_i bottles of cough medicine. We assume $Y_i \sim \text{Poisson}(X_i)$ where X_i is the "underlying demand", X_i has possible values 10 and 30, and is modelled by a Markov chain with transition matrix $P = \begin{bmatrix} 0.95 & 0.05 \\ 0.2 & 0.8 \end{bmatrix}$.
- A simulation from the flu model. The full line represents the underlying expected demand for cough-medicine, based on whether there is a flu-infection in the area or not. The dots represent the observed actual sales of the medicine.



- Can we learn about the presence of flu-infection from sales of cough-medicine?

Example 2: CpG islands

- DNA sequences may be modelled as Markov chains, with possible values A, C, G, T and the positions along the sequence as the steps in the chain.
- So-called "CpG islands" are sequences where the transition matrix (P_+) appears to be slightly different from the transition matrix (P_-) of of non-CpG islands:

$$P_+ = \begin{bmatrix} 0.180 & 0.274 & 0.426 & 0.120 \\ 0.171 & 0.368 & 0.274 & 0.188 \\ 0.161 & 0.339 & 0.375 & 0.125 \\ 0.079 & 0.355 & 0.384 & 0.182 \end{bmatrix}, \; P_- = \begin{bmatrix} 0.300 & 0.205 & 0.285 & 0.210 \\ 0.322 & 0.298 & 0.078 & 0.302 \\ 0.248 & 0.246 & 0.298 & 0.208 \\ 0.177 & 0.239 & 0.292 & 0.292 \end{bmatrix}.$$

- To detect CpG islands in a new DNA string, we set up a HMM where the underlying variable X_i has the two states: "CpG island" and "non-CpG island".

What questions do we want to ask?

- When the parameters of the HMM are known, we want to know about the values of the hidden variables X_i. For example:
  - What is the most likely sequence $X_0, \ldots, X_n$ given the data?
  - What is the probability distribution for a single X_i given the data?
- When the parameters of the HMM are not known, we need to infer these from some data.
  - If data with all X_i and Y_i known is available, inference for parameters is based on counts of transitions.
  - Inference may even be done based only on observations of the Y_i and some assumptions on the X_i (not done in this course).

The Multinomial Dirchlet conjugacy

- A vector $x = (x_1, \ldots, x_k)$ of non-negative integers has a Multinomial distribution with parameters n and p, where $n > 0$ is an integer and p is a probability vector of length k, if $\sum_{i=1}^k x_i = n$ and the probability mass function is given by

$$\pi(x \mid n, p) = \frac{n!}{x_1! x_2! \ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}.$$

- A vector $p = (p_1, \ldots, p_k)$ of non-negative real numbers satisfying $\sum_{i=1}^k p_i = 1$ has a Dirichlet distribution with parameter vector $\alpha = (\alpha_1, \ldots, \alpha_k)$, if it has probability density function

$$\pi(p \mid \alpha) = \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2) \cdot \Gamma(\alpha_k)} p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1} \ldots p_k^{\alpha_k - 1}.$$

- We have conjugacy in this case: $p \mid x \sim \text{Dirichlet}(\alpha + x)$.
- If $p \sim \text{Dirichlet}(\alpha)$ then $\mathrm{E}(p) = \frac{\alpha}{\sum_{j=1}^k \alpha_j}$.

The Multinomial Dirchlet conjugacy, predictions

- The (prior) predictive distribution is given by

$$\pi(x) = \frac{n!}{x_1! \ldots x_k!} \cdot \frac{\Gamma(\alpha_1 + x_1)}{\Gamma(\alpha_1)} \cdots \frac{\Gamma(\alpha_k + x_k)}{\Gamma(\alpha_k)} \cdot \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\Gamma(\sum_{i=1}^k \alpha_i + x_i)}.$$

- For example, if $p \sim \text{Dirichlet}(\alpha)$, the predicted probability that the next observation is of type i is

$$\pi(x = e_i = (0, ..., 1, \ldots, 0) \mid \alpha) = \frac{\alpha_i}{\sum_{j=1}^k \alpha_j}.$$

Inference for finite state space Markov chains

- Example: You have observed 0, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0 from a Markov chain with possible values 0 and 1. What is the transition matrix?
- First, make table with counts of transitions:

	0	1
0	3	3
1	3	1

.
- A reasonable guess for a transition matrix is then

$$P = \begin{bmatrix} 3/6 & 3/6 \\ 3/4 & 1/4 \end{bmatrix}.$$

- What should happen if we have never observed a transition $i \to j$ for two states i and j?
- What should happen if we have never observed any transition from a state i?

One solution: pseudo-counts

- Idea: If the count is zero, add some small positive number, a *pseudo-count*, so that the frequency becomes non-zero.
- The pseudo-count does not need to be an integer.
- To be "fair", we may add the same pseudo-count to all counts. We often use pseudo-counts equal to 1.
- In the example above, with pseudo-counts 1, the count table becomes

	0	1
0	4	4
1	4	2

and the transition matrix becomes

$$P = \begin{bmatrix} 4/8 & 4/8 \\ 4/6 & 2/6 \end{bmatrix}.$$

- Note how the influence of pseudo-counts approaches zero when the actual counts increase.
- What should happen if the state space is infinite?
- Generally, is there a theoretic framework to put this into?

Bayesian inference for Markov chains

- Write $P_1, \ldots, P_k$ for the k rows of P, and view each P_i as an independent random variable.
- Note that observed data (counts of transitions from each state i) is Multinomially distributed given P_i.
- If we assume $P_i \sim \text{Dirichlet}(\alpha_i)$ for some vector $\alpha_i = (\alpha_{i1}, \ldots, \alpha_{ik})$, and the counts for transitions out of i are given in the vector $c_i = (c_{i1}, \ldots, c_{ik})$, then the posterior for P_i becomes $\text{Dirichlet}(\alpha_i + c_i)$.
- Note that the expectected posterior becomes the vector

$$\mathrm{E}\left(P_i \mid \text{data}\right) = \frac{\alpha_i + c_i}{\alpha_{i1} + \cdots + \alpha_{ik} + c_{i1} + \cdots + c_{ik}}$$

  So the α_{ij} correspond exactly to pseudo-counts!
- The prior $\text{Dirichlet}(1, 1, \ldots, 1)$, with all pseudo-counts equal to 1 corresponds to a uniform distribution on the set of all probability vectors P_i that sum to 1.

More conclusions from the Bayesian framework

- We can show that, using any prior, if the sequence $X_0, X_1, \dots, X_n$ is observed as data, then the posterior probabilities for X_{n+1} are $\mathrm{E}\left(P_{x_n}\right)$.
- We can extend this to compute the probability of any sequence $X_{n+1}, \dots, X_{n+r}$ given data $X_0, \dots, X_n$.
- When the prior is Dirichlet as above, we can use the predictive distribution found above.
- If we know *a priori* that certain transitions are impossible, we can incorporate this into the prior: For example, using the prior $P_i \sim \text{Dirichlet}(1, 1, 0)$,means that transitions from state i to state 3 have probability zero.
- It is also possible to construct priors for the transition matrix P that represent other types of prior information, for example that the Markov chain must be time reversible.

Inference for the parameters of HMMs

Assume an HMM model where $X_i \in \{0, 1\}$, $Y_i \in \{1, 2, 3\}$, and we have observed both states in some stretch of data:

X	0	0	0	0	1	1	1	1	1	0
Y	1	2	1	1	2	3	2	3	3	1

- Counting transitions, we get

	0	1
0	3	1
1	1	4

and

	1	2	3
0	4	1	0
1	0	2	3

.

- In practice, we can use pseudocounts just as in the Markov chain case. In the example above, using all pseudocounts equal to 1, we get

$$P = \begin{bmatrix} 4/6 & 2/6 \\ 2/7 & 5/7 \end{bmatrix}, Q = \begin{bmatrix} 5/8 & 2/8 & 1/8 \\ 1/8 & 3/8 & 4/8 \end{bmatrix}$$

where P is the transition matrix of the Markov chain, and Q is the stochastic matrix of transition probabilities from X_i to Y_i.

- As for Markov chains, these results can be obtained by using priors for P and Q that are products of Dirichlet distributions.

More on inference of parameters for HMMs

- The Bayesian paradigm may be used to make predictions for later observations: In the example above, with $X_0, \dots X_9, Y_0, \dots Y_9$ observed, the probability vector with the three possible values of Y_{10} can be computed with the matrix product $\mathrm{E}\left(P_{x_9}\right)\mathrm{E}\left(Q\right)$.
- The priors can be adapted to incorporate actual prior information.
- For example, prior knowledge about the transitions from states of X_i to states of Y_i might lead you to model $Y_i \sim \text{Poisson}(\lambda_{X_i})$, so for each value of X_i the Y_i are Poisson distributed with parameter λ_{X_i}. Fixing a prior also on the λ_{X_i} parameters, we may then find the posteriors for these in similar ways as we have done before.

More inference questions for HMMs

- We focused above on the case where (some) parameters of the HMM are not fully known.
- If the HMM parameters are given and the Y_i are observed, the goal may instead be to learn about the values of the X_i (these methods are not part of the course):
  - Find the sequence $X_0, \ldots, X_k$ with the maximum probability given the observed $Y_0, \ldots, Y_k$ and the given model: The *Viterbi algorithm*.
  - Find the marginal distribution for each X_i given the observed $Y_0, \ldots, Y_k$ and the model: The Forward-Backward algorithm.
  - Find the *joint distribution* of $X_0, \ldots, X_k$ given the observed $Y_0, \ldots, Y_k$ and the model. In practice: Find a sequence $X_0, \ldots, X_k$ that is a *sample* from this joint distribution. This may also be done with a Forward-Backward algorithm.