MVE550 2023 Lecture 6
Compendium chapter 2
Inference for Markov chains
Hidden Markov Models (HMMs)

Petter Mostad

Chalmers University

November 15, 2023

Inference for Markov chains

- We have looked at (discrete-time, homogeneous) Markov Chains $X_0, X_1, \ldots,$ with discrete state spaces.
- The can be described by describing the distribution of X_0 and the transition matrix P.
- In many applications, these *parameters* of the chain will be unknown, and must be *inferred* from data.
- We will limit ourselves to looking at cases where
  - the distribution of X_0 is known,
  - the state space is finite,
  - the data is an *observed* sequence $x_0, x_1, \ldots, x_t$ from the chain,
  - we use the data *and contextual knowledge* to make inference about the transition matrix P.
- Following the *Bayesian paradigm* we do not make an *estimate* for P, but instead we find a *posterior distribution* for P, and use this to make predictions.

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 $\text{E}(p) = \frac{\alpha}{\sum_{j=1}^k \alpha_j}$.

The Multinomial Dirchlet conjugacy, predictions

If $p \sim \text{Dirichlet}(\alpha_1, \dots, \alpha_k)$ and $x \sim \text{Multinomial}(n, p)$, then

- The predictive distribution is given by

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

- For example, if e_i is the vector with 1 at place i and zeros elsewhere, then $\pi(x = e_i) = \frac{\alpha_i}{\sum_{j=1}^k \alpha_j}$.
- For example, if x_{new} is a vector of new counts, then, as $p \mid x \sim \text{Dirichlet}(x + \alpha)$, we get

$$\pi(x_{new} = e_i \mid x) = \frac{x_i + \alpha_i}{n + \sum_{j=1}^k \alpha_j}.$$

- The α_i in the prior can be called *pseudo-counts*.
- For x_{new} with more than one count, prediction probabilities can be computed with the full formula above.

Inference for P: Summary

- Represent the rows $P_1, \ldots, P_k$ of P as random variables: Decide on *priors* for each, representing contextual knowledge. (One may also use a joint prior!)
- Find the posteriors $P_i \mid \text{data}$, where the data consists of counts of observed transitions from state i. (With a joint prior one gets a joint posterior!)
- To predict the continuation of a chain: Either first simulate $\tilde{P}$ from the posteriors and predict using this $\tilde{P}$, or predict one step at a time, adding prediction to data each time.
- In practice, one can use Dirichlet priors. The parameters of the Dirichlet priors are called pseudo-counts. The posteriors are then also Dirichlet distributions.
- If the chain is at state i and one uses the prior $P_i \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_k)$ for row i of P, the probabilities of the next state are given by the vector

$$\frac{x + \alpha}{n + \sum_{j=1}^{k} \alpha_j}.$$

Example: Not quite a Markov chain

Exercise 2.20 from Dobrow:

- Let $X_0, X_1, \ldots$ 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 $Y_0, Y_1, \ldots$ 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

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

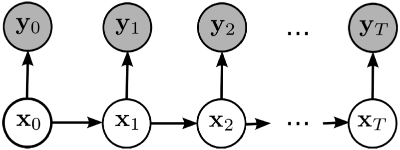
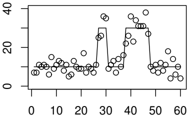


Figure: A hidden Markov model.

- 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:

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

- 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 $X_0, \dots, X_n$ given the data?
  - What is the probability distribution for a single X_i given the data?

  We do not focus on these questions here.
- 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. (See below).
  - Inference may even be done based only on observations of the Y_i and some assumptions on the X_i (we do not consider this).

Inference for HMMs: Summary

- Just as for inference for Markov chains: Consider the transition matrix P and the emission matrix Q (containing probabilities $\Pr(Y_s = j \mid X_s = i)$) as random variables.
- Decide on priors (a standard choice uses Dirichlet distributions).
- To predict: Either: Simulate from the posterior (Dirichlet distributions) for P and Q, and then simulate values for the hidden chain and observable Y's. Or: Simulate one step at a time, and add simulated values to data.