STA 365: Applied Bayesian Statistics

Boris Babic
Assistant Professor, University of Toronto

Week 7B: Markov Chain Monte Carlo

Intro

- So far we have looked at cases where the posterior has a convenient functional form, making the distributions easy to summarize.
- For example, with a beta posterior we can use pbeta() and qbeta() in R to obtain posterior estimates/intervals and construct hypothesis tests.
- But this is not always the case.

Intro

- In general, y comes from $f(y|\theta)$ and $\theta \sim \pi(\theta)$. After $Y = y$ has been observed, $L(\theta) = f(y|\theta)$.
- The posterior is, as always:

$$\pi(\theta \mid y) = \frac{\pi(\theta)L(\theta)}{\int \pi(\theta)L(\theta)d\theta}. \tag{1}$$

- If the prior and likelihood do not combine in a convenient way, $\int \pi(\theta)L(\theta)d\theta$ can be difficult to evaluate analytically.
- Likewise, summaries of the posterior distribution can be difficult to evaluate analytically when the posterior is not expressed in a recognized closed form. For example, the posterior mean of θ is in general given by:

$$\mathrm{E}(\theta \mid y) = \frac{\int \theta\pi(\theta)L(\theta)d\theta}{\int \pi(\theta)L(\theta)d\theta}. \tag{2}$$

Intro

- With large enough n an option is to use normal approximations. For example, let $\pi(\theta) \propto 1$ and $f(y|\theta) \overset{\text{iid}}{\sim} \text{Bin}(n, \theta)$. Then posterior is $\pi(\theta|y) \sim \text{Beta}(y+1, n-y+1)$ which can be approximated by $\text{N}(\frac{y}{n}, \frac{y(n-y)}{n^3})$.
- But this is not an advisable general practice. For example, $0 < \theta < 1$ in inference for proportions, so at a minimum we would have to truncate the distribution.
- Markov chain Monte Carlo (MCMC) algorithms can be used to simulate the posterior from general Bayesian models and draw samples from it.
- Before we get to the Markov Chain part, we will motivate Monte Carlo sampling.

The Monte Carlo Method

- Let θ be a parameter of interest and let $y_1, ...y_n$ be the numerical values of a sample from a distribution $\pi(y_1, ...y_n|\theta)$. Suppose we could sample some number S of independent random θ values from the posterior distribution:

$$\theta^{(1)}, ..., \theta^{(S)} \overset{\text{iid}}{\sim} \pi(\theta|y).$$

- Then the empirical distribution of the samples $\theta^{(1)}, ..., \theta^{(S)}$ would approximate $\pi(\theta|y)$, and the approximation would improve with increasing S.
- The empirical distribution of $\theta^{(1)}, ..., \theta^{(S)}$ is known as a Monte Carlo approximation to $\pi(\theta|y)$.
- Many computer languages and computing environments have procedures for simulating this sampling process. We call this simulation process Monte Carlo.
- Basically a fancy way of saying we can take quantities of interest of a distribution from simulated draws from the distribution.
- R has built-in functions to simulate iid samples from most of the distributions we will use. For example, you can generate monte carlo samples from a standard normal distribution using rnorm(n, mean $= 0$, sd $= 1$).

The Monte Carlo Method

- Let $g(\theta)$ be (almost any) function of θ.
- The law of large numbers says that if $\theta^{(1)}, ..., \theta^{(S)}$ are iid samples from $\pi(\theta|y)$ then

$$\frac{1}{S}\sum_{s=1}^{S} g(\theta^{(s)}) \rightarrow \mathrm{E}[g(\theta)|y] = \int g(\theta)\pi(\theta|y d\theta \text{ as } S \rightarrow \infty.$$

- This implies that as $S \rightarrow \infty$
  - $\overline{\theta} = \frac{1}{S}\sum_{s=1}^{S} \theta^{(s)} \rightarrow \mathrm{E}[\theta|y]$
  - $\frac{1}{S-1}\sum_{s=1}^{S} (\theta^{(s)} - \overline{\theta})^2 \rightarrow \mathrm{Var}(\theta|y)$
  - $\frac{1}{S}\ \#(\theta^{(s)} \leq c) \rightarrow \Pr(\theta \leq c|y)$.
  - The empirical distribution of $\theta^{(1)}, ..., \theta^{(S)} \rightarrow \pi(\theta|y)$.
  - The α-percentile of $\theta^{(1)}, ..., \theta^{(S)} \rightarrow \theta_\alpha$.
- Just about any aspect of a posterior distribution we may be interested in can be approximated arbitrarily exactly with a large enough Monte Carlo sample.

Introduction to Markov Chains

- Often we're going to want to draw samples from any posterior distribution, not just ones that have a recognized form.
- Indeed, that's where the big benefit from computation comes in – the distributions with a nice recognized form are ones which typically have a convenient closed form expression in the form of a conjugate model.
- In the more general case, we will sample using a class of algorithms known as Markov Chain Monte Carlo.
- And we will be able to sample from just about any posterior distribution.

Introduction to Markov Chains

- A bunch of draws of θ that are each slightly dependent on the previous one.
- The chain wanders around the parameter space, remembering only where it has been in the last period.
- How the chain jumps from one state to another at each period is governed by what we will call the transition matrix P.

Example of a Markov Chain

- Consider a simple example of a discrete Markov Chain (Albert and Hu, 2020).
- Suppose a person takes a random walk on a number line on the values 1, 2, 3, 4, 5, 6. If the person is currently at an interior value (2, 3, 4, or 5), in the next step she is equally likely to remain at that number or move to an adjacent number. If she does move, she is equally likely to move left or right. If the person is currently at one of the end values (1 or 6), in the next step she is equally likely to stay still or move to the adjacent location.
- A Markov chain describes such probabilistic movement between states.
- Here there are 6 possible states.
- The probability that the walker moves to another location depends only on her current location and not on previous locations visited.

Definition of a Markov Chain

- More formally, a Markov Chain is a stochastic process in which future states are independent of past states given the present state.
- Stochastic process: a consecutive set of random (not deterministic) quantities defined on some known state space. For example, our space is Ω, the parameter space. And the word consecutive implies a time component, usually indexed by t.
- Consider a draw $\theta^{(t)}$ to be a state at iteration t. The next draw $\theta^{(t+1)}$ is dependent only on the current state $\theta^{(t)}$ and not on any past states.
- This satisfies the Markov property:

$$\Pr(\theta^{(t+1)}|\theta^{(1)}, \theta^{(2)}, ..., \theta^{(t)}) = \Pr(\theta^{(t+1)}|\theta^{(t)})$$

- A stochastic process satisfying the Markov property is a Markov chain.

Working With Markov Chains

- We describe movement between states in terms of transition probabilities.
- The transition probabilities are summarized by means of a transition matrix P.
- For k discrete states P is a $k \times k$ matrix.
- In our example this is

$$P = \begin{bmatrix} .50 & .50 & 0 & 0 & 0 & 0 \\ .25 & .50 & .25 & 0 & 0 & 0 \\ 0 & .25 & .50 & .25 & 0 & 0 \\ 0 & 0 & .25 & .50 & .25 & 0 \\ 0 & 0 & 0 & .25 & .50 & .25 \\ 0 & 0 & 0 & 0 & .50 & .50 \end{bmatrix}$$

- The first row gives the probabilities of moving to all states 1 through 6 in a single step from location 1, the second row gives the transition probabilities in a single step from location 2, and so on.
- For continuous state space (infinite possible states), the transition kernel is a bunch of conditional probability density functions: $\pi(\theta^{(t+1)})|\theta^{(t)})$.

Working With Markov Chains

- We represent one's current location as a probability row vector

$$p = (p_1, p_2, p_3, p_4, p_5, p_6),$$

where p_i represents the probability that the person is currently in state i. For example $p = (1, 0, 0, 0, 0, 0)$ indicates the current location is state 1 without any uncertainty.

- If $p^{(j)}$ represents the location of the traveler at step j then the location of the traveler at the j + 1 step is given by the matrix product

$$p^{(j+1)} = p^{(j)} P.$$

.

- if $p^{(j)}$ represents the location at step j,then the location of the traveler after m additional steps, $p^{(j+m)}$ is given by the matrix product

$$p^{(j+m)} = p^{(j)} P^m,$$

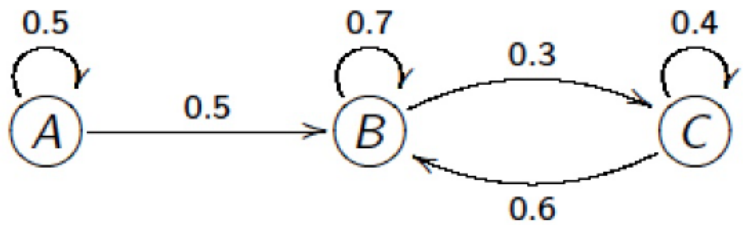
where P^m indicates the matrix P multiplied by itself m times.

Some intuition about convergence

- Suppose the person is currently in state 3: $p = (0, 0, 1, 0, 0, 0)$.
- After 4 steps, the probabilities of the states are given by $p \times P^4 = (0.10938, 0.25, 0.27734, 0.21875, 0.11328, 0.03125)$ and after 10 steps the probabilities are $p \times P^{10} = (0.12032, 0.23521, 0.21778, 0.18977, 0.1619, 0.075016)$.
- As one takes an infinite number of moves, the probability of landing at a particular state does not depend on the initial starting state.
- Instead, the probability of being in each state appears to converge to some distribution: i.e., a vector stating the probability of being in each state.
- In this case, that vector is exactly $p = (0.1, 0.2, 0.2, 0.2, 0.2, 0.1)$.
- This is known as the stationary distribution.
- In the Bayesian approach, we want to use a Markov Chain which converges to the posterior distribution as its stationary distribution, and then sample from that distribution.
- We need to understand when Markov Chains converge to a stationary distribution.

Properties of Markov Chains

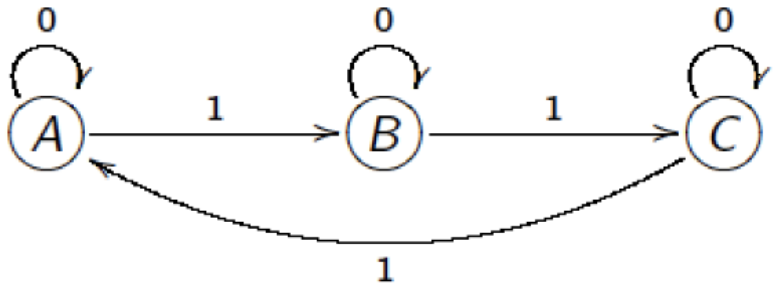
- There are several important properties of Markov chains.
- A Markov chain is irreducible if it is possible go from any state to any other state (not necessarily in one step). This is true in our example.
- The following chain is reducible, or not irreducible.



- The chain is not irreducible because we cannot get to A from B, and we cannot get to A from C, regardless of the number of steps we take.

Markov Chains

- A state in a Markov chain is periodic if the chain can return to the state only at multiples of some integer larger than 1. This example is aperiodic. Another way to put this: A Markov chain is aperiodic if the only length of time for which the chain repeats some cycle of values is the trivial case with cycle length equal to one. Hence, as long as the chain is not repeating an identical cycle, then the chain is aperiodic.
- Let A, B, and C denote the states (analogous to the possible values of θ) in a 3-state Markov chain. The following chain is periodic with period 3, where the period is the number of steps that it takes to return to a certain state.



Markov Chains

- A Markov chain is recurrent if for any given state i, if the chain starts at i, it will eventually return to i with probability 1.
- A Markov chain is positive recurrent if the expected return time to state i is finite; otherwise it is null recurrent.

Stationary Distribution

- Let w be a probability vector, a row vector with S components so that $s \sum_S w_i = 1$ and $w_i \geq 0$ for all $i \in S$. If $w = wP$, then this probability vector w is called the stationary distribution.
- That is,

$$w_i = \sum_S w_j p_{ji}$$

  for all i in S. In words, w_i is the dot product between w and the $i$the column of P.
- In our example, $w_1 = 0.1$ and indeed $(0.1, 0.2, 0.2, 0.2, 0.2, 0.2, .1)' \cdot (0.5, 0.25, 0, 0, 0, 0)' = 0.1$
- The typical MCMC algorithm constructs the Markov chain so that it converges to a stationary distribution regardless of our starting points.
- We can devise a Markov chain whose stationary distribution is our desired posterior distribution $\pi(\theta|y)$, then we can run this chain to get draws that are approximately from $\pi(\theta|y)$ once the chain has converged.

Stationary Distribution: Our example

- Suppose that the person begins at state 3 which is represented by the vector $p = (0, 0, 1, 0, 0, 0)$.
- If we multiply this vector by the matrix P, we obtain the probabilities of being in all six states after one move. If we multiply p by P n times we get the probabilities of being in the different states after n moves.
- In our example, if we multiply P 100 times in R, we obtain the constant vector w that is equal to $(0.1, 0.2, 0.2, 0.2, 0.2, 0.1)$. This is the stationary distribution.

Ergodic Theorem

- Let $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(M)}$ be M values from a Markov chain that is *aperiodic*, *irreducible* and *positive recurrent* (then the chain is ergodic), and $E[g(\theta)] < \infty$.
- Then with probability 1,

$$\frac{1}{M}\sum_{i=1}^{M} g(\theta_i) \to \int_{\Theta} g(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}, \text{ as } M \to \infty$$

where π is the stationary distribution

- This is the Markov chain analog to the strong law of large numbers (SLLN), and it allows us to ignore the dependence between draws of the Markov chain when we calculate quantities of interest from the draws.

Ergodicity

- If our Markov chain is aperiodic, irreducible, and positive recurrent, then it is ergodic and it has a unique stationary distribution.
- The ergodic theorem allows us to do Monte Carlo integration by calculating quantities of interest from our draws, ignoring the dependence between draws.

Burn-in

- Since convergence usually occurs regardless of our starting point, we can usually pick any feasible (for example, picking starting draws that are in the parameter space) starting point.
- However, the time it takes for the chain to converge varies depending on the starting point.
- As a matter of practice, most people throw out a certain number of the first draws, known as the burn-in. This is to make our draws closer to the stationary distribution and less dependent on the starting point.
- However, it is unclear how much we should burn-in since our draws are all slightly dependent and we do not know exactly when convergence occurs.