Markov Chain Monte Carlo

Andrew Siegel

June 5, 2017

Introduction

- Consider a collection of random variables

$$X_1, X_2, ..., X_n$$

- Interpret X_n as "the state of a system at time n"
- $X_n \in \{1...N\}$, i.e. denote the state as an integer even though it can represent anything.
- Define P_{ij}, $i, j = 1, ..., N$ as the probability that the system in state i will transition to state j. The P_{ij} are called *transition probabilities*.
- For a Markov chain P_{ij} is independent of the past state of the system, i.e. is the same for all times.

Example

- Let's start with a very simple example
- Imagine two positions, call them X_1 and X_2.
- A person starts at one position and each, say, hour, transitions to a new position (or stays in current one) with probabilities given by

$$\begin{aligned}
P_{11} &= .9: \text{ probability of moving from } X_1 \text{ to } X_1 \\
P_{12} &= .1: \text{ probability of moving from } X_1 \text{ to } X_2 \\
P_{21} &= .2: \text{ probability of moving from } X_2 \text{ to } X_1 \\
P_{22} &= .8: \text{ probability of moving from } X_2 \text{ to } X_2
\end{aligned}$$

Simple Example, cont

- Note that we can write this as a matrix

$$P = \begin{bmatrix} .9 & .1 \\ .2 & .8 \end{bmatrix}$$

  where the rows sum to one, indicating that the system has to be in some state at each time step.
- A key question is, given an initial position, what are the the probabilities after 2,3,...,k steps?
- And, as $k \to \infty$, is a steady state reached where the fractions of time spent in each state are constant?
- Turns out that a unique steady state (*stationary probabilities*) are guaranteed if the Markov chain is *irreducible* and *aperiodic*. More on those soon. First let's gain some further intuition.

```
function [ pi ] = MarkovChain( X, pdf, nsteps)

[N  ~]= size(pdf);      %number of states in process
cdf = (cumsum(pdf'))';  %cdf
pi = zeros(N,1);        %# of time state j visited

for i=1:nsteps          %iterate over chain
    rn = rand();        %uniform random number
    tmp = cdf(X,:);     %pdf for state transitions
                        %from current state
    j = 1;
    cutoff = tmp(j);
    while (rn > cutoff)
        j  = j + 1;
        cutoff = tmp(j);
    end
    X=j;
    pi(j) = pi(j) + 1;

end
pi=pi./nsteps;
```

Multi-step probabilities

- Let's create a 3x3 example to calculate multi-step probabilities
- Here is a transition probability table that expresses the probability of one of three weather conditions on: rainy, snowy, or nice.

$$\begin{array}{cc} & \begin{array}{ccc} R & N & S \end{array} \\ P = & \begin{pmatrix} .5 & .25 & .25 \\ .5 & 0 & .5 \\ .25 & .25 & .5 \end{pmatrix} \begin{array}{c} R \\ N \\ S \end{array} \end{array}$$

- Let's calculate e.g. the probability of snow in 2 days given rain today:

$$p_{13}^{(2)} = p_{11}p_{13} + p_{12}p_{23} + p_{13}p_{33}$$

Multi-step probabilities, cont.

- This can be read as a sum of conditional probabilities, i.e
  - Probability of transitioning from 1 to 1 times the probability of transitioning from one to 3 PLUS
  - Probability of transitioning from 1 to 2 times the probability of transitioning for 2 to 3 PLUS
  - Probability of transitioning from 1 to 3 times the probability of transitioning from 3 to 3.
- This then easily generalizes to

$$p_{ij}^{(2)} = \sum_{l=1}^{r} p_{il} p_{lj} = P^2$$

and furthermore

$$p_{ij}^{(k)} = P^k$$

Stationary Probabilities for Weather Example

- Taking our matrix P to a large power (20 by trial and error) yields:

$$P^{(20)} = \begin{bmatrix} .4 & .2 & .4 \\ .4 & .2 & .4 \\ .4 & .2 & .4 \end{bmatrix}$$

- If we consider an initial state of $x = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$ (or equivalently $\begin{bmatrix} 0 & 1 & 0 \end{bmatrix}$ or $\begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$) we get the stationary probabilities as:

$$\pi = xP = \begin{bmatrix} .4 \\ .2 \\ .4 \end{bmatrix}$$

Properties of Markov Chains

- Some event must occur, so P_{ij} sum to one:

$$\sum_{j=1}^{N} P_{ij} = 1, \qquad i = 1, ..., N$$

- A Markov Chain is *irreducible* if every state can eventually be reached by every other state:
- Denote π_j as the long-run proportion of time that the system spends in state j. The π_j satisfies

$$\begin{aligned} \pi_j &= \sum_{i=1}^{N} \pi_i P_{ij}, \qquad j = 1, ..., N \\ \sum_{j=1}^{N} \pi_j &= 1 \end{aligned}$$

- For any function h on the state space:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} h(X_i) = \sum_{j=1}^{N} \pi_j h(j)$$

- A Markov Chain is said to be *aperiodic* if for some $n > 0$ and some state j:

$$P\{X_n = j | X_0 = j\} > 0 \text{ and } P\{X_{n+1} = j | X_0 = j\} > 0$$

- That is, if any single state can possibly be repeated between successive time intervals.
- We will not go into details here, but without this restriction analysis of the Markov chain is more complicated.

Stationary Probabilities

- In the typical case where we can assume a stationary, non-reducible Markov process, then it can be show that

$$\pi_j = \lim_{n \to \infty} P\{X_n = j\},\ j = 1, ...N$$

- That is, the π_j give the long term probabilities of the system being in a given state.
- Another way to think of this is after many iterations of the state, the π_j represent the fractions of time the systems was in the j'th state.

Markov Matrices

It is useful to take a linear algebra perspective. A Markov Matrix has the following properties

- All entries ≥ 0.
- Columns sum to 1
- Example

$$A = \begin{bmatrix} .1 & .01 & .3 \\ .2 & .99 & .3 \\ .7 & 0 & .4 \end{bmatrix}$$

- Key properties
  - $\lambda = 1$ is a eigenvalue
  - all other $|\lambda_i| < 1$

Eigenvalue 1

- How does linear algebra show us this?
- Notice that for a Markov matrix the columns of $A - I$ add to zero. In our example
$$A - I = \begin{bmatrix} -.9 & .01 & .3 \\ .2 & -.01 & .3 \\ .7 & 0 & -.6 \end{bmatrix}$$
- Also use that the eigenvalues of A^T are equal to the eigenvalues of A.

Simple Example

- Imagine that we know the probabilities each year that a person will move from, say, Mass. to California, and vice versa.
- Enter these probabilities as the columns of our Markov matrix

$$P = \begin{bmatrix} .9 & .2 \\ .1 & .8 \end{bmatrix}$$

- The matrix P states the following:
  - In a given year the probability of a person living in Cal staying there .9, and of moving to Mass is .1.
  - In a given year, the probability of a person living in Mass and staying there is .8, and of moving to Cal. is .2;

Simple Example,cont

- Question: What are the steady state probabilities?

Simulating Dependent Samples

- Suppose we want to generate the value of a random variable X distributed according to $P\{X = j\} = p_j, j = 1, ..., N$
- Can we generate aperiodic Markov chain with limiting probabilities $p_j, j = 1, ..., N$ and run chain for n steps to obtain the value of X_n?
- Want to estimate typically

$$E[h(x)] = \sum_{j=1}^{N} h(j) p_j$$

- Can be done using estimator

$$\frac{1}{n} \sum_{i=1}^{n} h(X_i)$$

Simulating Dependent Samples, cont.

- Or to eliminate dependence on initial state, start with future state

$$\hat{\theta} = \frac{1}{n-k} \sum_{i=k+1}^{n} h(X_i)$$

- Question arises of how to estimate MSE of estimator, i.e.

$$MSE = E\left[\left(\hat{\theta} - \sum_{j-1}^{N} h(j)p_j\right)^2\right]$$

since adjacent samples are no longer uncorrelated, and variance of the mean depends on the variance of the process as well as the covariance between samples.

Batching

- One approach is to use *batching*, where samples are clustered into roughly independent subgroups.
- Define the j'th batch average Y_j as

$$Y_j = \frac{1}{r} \sum_{i=k+(j-1)r+1}^{k+jr} h(X_i), j = 1, ..., s$$

- Treat the Y_j as if they were independent and identically distributed variance σ^2.
- Then the estimate of MSE is just $\hat{\sigma}^2/s$

Definition of Ergodicity

- A Markov chain is said to be ergodic if there exists a positive integer T_0 such that for all pairs of states i, j in the Markov chain, if it is started at time 0 in state i then for all $t > T_0$, the probability of being in state j at time t is greater than 0.
- For a Markov chain to be ergodic, two technical conditions are required of its states and the non-zero transition probabilities; these conditions are known as irreducibility and aperiodicity. Informally, the first ensures that there is a sequence of transitions of non-zero probability from any state to any other, while the latter ensures that the states are not partitioned into sets such that all state transitions occur cyclically from one set to another.

Irreducibility

- A Markov chain is said to be *irreducible* if, $\forall a, b \in \mathcal{X}$, where $\mathcal{X}$ denotes the state space,

$$P(x_t = b | x_o = a) > 0.$$

- Conceptually, a chain is irreducible if it is possible to reach any state regardless of the starting point.
- Note: This does not say that you can reach any state from any other state.
- Irreducubility is a necessary condition for many of the the key Markov theorems (subsuquent slides).

Aperiodicity

- An irreducible Markov chain is *aperiodic* if the following condition holds for any starting point *a*:

$$gcd\{t : P(x_t = a|x_o = a) > 0\} = 1$$

- This is read: "The greatest common denominator of all times for which a return is possible to node *a* most be one."
- Conceptually this indicates that it is possible for any node to be visited at any time, e.g. certain nodes node restricted to odd times, etc.

Aperiodicity, cont.

- More generally, an aperiodic graph is a directed graph whose cycles have lcd of 1.
- Consider any given node and ask how many paths get us back to that node. If answer is, say, 4,5,8,12 ... we say node is aperiodic becuase paths are only divisible by 1.
- Note that this focuses just a specific node while aperiodicity deals with cycles to all nodes. But "to all nodes" invovles creating a global list of all paths to each node and look for lcd of this list, i.e.
  4,5,8,12
  2,4,6,8,10
  4,6,8,10
  6,8,10,12
- Above list still has lcd 1, so it suffices to find only a single node with cycle path lengths not divisible by 1. but notice that you could have 3,5,7

Ergodic Theorem for Markov Chains

- Intuition on Markov chains is that they *fill the event space* if run for enough iterations
- This is formalized with the following theorem
- If $\{x_0, x_1, \cdots, x_n\}$ is an irreducible discrete Markov chain with stationary distribution π, then as $n \to \infty$

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) \to E[f(x)]$$

- Note that the x_i are not i.i.d but rather are typically correlated. The correlations can be both a benefit and a drawback depending on the application.

Long-time probability

- If furthermore the Markov chain is *aperiodic* then for all x, x_o in the event space, as $n \rightarrow \infty$

$$P(x_n = X | x_o = X_o) \rightarrow \pi(x)$$

- Note that without aperidocity the Markov chain may still have stationary probabilities.
- However apoeriodicity is required to ensure that the next draw from the chain represents a draw from the underlying pdf. Subtle difference.
- This make sense since aperiodicity essentially forces an ordering at each step.

Detailed Balance

- A probability mass function π on $\mathcal{X}$ satisfies *detailed balance* if for all a and b

$$\pi_a P_{ab} = \pi_b P_{ba}.$$

- Note: if π satisfies detailed balance then π is a stationary distribution

$$\pi_b = \sum_a \pi_a P_{ab} = \sum_a \pi_b P_{ba} = \pi_b \sum_a P_{ba} = \pi_b$$

- Note that satisfying detailed balance is a very special case of a stationary distribution – in most cases this will not be the case.

Metropolis Algorithm

- Given a probability mass function π on $\mathcal{X}$ and $f : \mathcal{X} \rightarrow \mathcal{R}$, goal is to sample from π or approximate $E[f(x)]$.
- This is of great interest when π and or f are extremely complicated.
- Metropolis: Construct a Markov chain with stationary probabilities π, then invoke the ergodic theorem.
- Metropolis Hastings: Slightly more general, trial matrix does not need to be symmetric.

Metropolis Algorithm

- Begin with *proposal matrix* Q, $Q_{ab} \in \mathcal{X}$. Can be any stochastic matrix but choice affects performance. Must be symmetric for Metropolis, not Metropolis Hastings
- Choose Q
- Choose arbitrary x_o in $\mathcal{X}$
- for i=0,1,...,n-1
    - Sample x from $Q(x_i, x)$
    - Sample $u = U(0, 1)$
    - if $u < \frac{\pi(x)}{\pi(x_i)}$ then $x_{i+1} = x$, else $x_{i+1} = x_i$
- endfor

Metropolis-Hastings Algorithm

- Begin with *proposal matrix* Q, $Q_{ab} \in \mathcal{X}$. Can be any stochastic matrix but choice affects performance.
- Choose Q
- Choose arbitrary x_o in $\mathcal{X}$
- for i=0,1,...,n-1
    - Sample x from $Q(x_i, x)$
    - Sample $u = U(0,1)$
    - if $u < \frac{\pi(x)q(x,x_i)}{\pi(x_i)q(x_i,x)}$ then $x_{i+1} = x$, else $x_{i+1} = x_i$
- endfor