Lecture 7: Markov Chain Monte Carlo

Professor Alexander Franks

2020-11-16

Announcements

Homework 4 out.

- 3 Q's

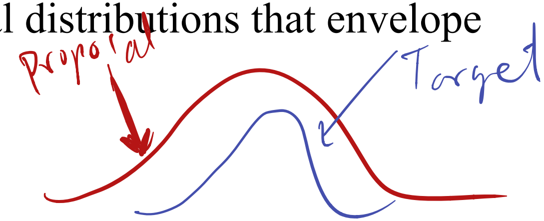
Monte Carlo estimation

$= \int \theta\, p(\theta|y)\, d\theta$

- $\bar{\theta} = \sum_{s=1}^{S} \theta^{(s)}/S \to \mathrm{E}[\theta|y_1, \ldots, y_n]$ $= \int (\theta - E(\theta|y))^2\, p(\theta|y)$
- $\sum_{s=1}^{S} \left(\theta^{(s)} - \bar{\theta}\right)^2 / (S-1) \to \mathrm{Var}[\theta|y_1, \ldots, y_n]$
- $\#\left(\theta^{(s)} \leq c\right)/S \to \Pr(\theta \leq c|y_1, \ldots, y_n)$ $= \int I[\theta \leq c]\, p(\theta|y)\, dy$
- the α-percentile of $\left\{\theta^{(1)}, \ldots, \theta^{(S)}\right\} \to \theta_\alpha$

Sampling from the posterior distributions

- The Monte Carlo methods we discussed previously assumed we could easily get samples from the posterior, e.g. with `rnorm`
- In general, sampling from a general probability distribution is hard
- Want to call `rcomplicatedistribution()` but don't have it
  - Inversion and rejection can work well for low dimensional posteriors
- In high dimensions, these approaches aren't sufficient
  - Near impossible to find good proposal distributions that envelope target
  - Or rejection rate is extremely high



Markov Chain Monte Carlo

- We want independent random samples, $\theta^{(s)}$ from $p(\theta \mid y_1, \ldots y_n)$
- But there is no good way to get independent samples
- Alternative, create a sequence of correlated samples with the correct **limiting** distribution

time series

- Markov Chain Monte Carlo gives us a way to generate correlated samples from a distribution

$\theta^{(s)} \sim MCMC$ (correlated samples)

As long $E[\theta^{(s)}] = E[\theta|y]$

$$E\left[\frac{1}{n}\sum \theta^{(s)}\right] \xrightarrow{\lim} E[\theta|y]$$

(linearity of expectation)

Correlation doesn't hurt us in this particular regard.

Monte Carlo Error

- Reminder: $\bar{\theta} = \sum_{s=1}^{S} \theta^{(s)} / S$ and S is the number of samples.
- If the samples are independent,

Monte Carlo Error

$$\mathrm{Var}(\bar{\theta}) = \frac{1}{S^2} \sum_{s=1}^{S} \mathrm{Var}(\theta^{(s)}) = \frac{\mathrm{Var}(\theta \mid y_1, \ldots y_n)}{S} = O\left(\frac{1}{S}\right)$$

- If the samples are *positively correlated*,

$$\frac{1}{S^2}\left(\sum_{i=1}^{S} \mathrm{Var}(\theta^{s}) + \sum_{i \neq j} \mathrm{cov}(\theta^{i}, \theta^{j}) \right)$$

0 when indep.

$$\mathrm{Var}(\bar{\theta}) = \frac{1}{S^2} \sum_{s,t} \mathrm{Cov}(\theta^{(s)}, \theta^{(t)}) > \frac{\mathrm{Var}(\theta \mid y_1, \ldots y_n)}{S}$$

- MCMC methods have higher Monte Carlo error due to positive dependence between samples.
- Hope to minimize dependendence and hence MC error

Basics of Markov Chains

Markov Chains: Big Picture

- For standard Monte Carlo, we make use of the law of large number to approximate posterior quantities

- The law of large numbers can still apply to random variables that are not independent

- We have a sequence of random variables indexed in time, θ_t

- We'll be using a *discrete-time* Markov Chain: $t \in 0, 1, \ldots T$

- The observations, $\theta^{(t)}$ can be discrete or continous ("discrete-state" or "continuous-state" Markov Chain)

Discrete-state Markov Chains

- Let $\theta^{(t)} \in 1, 2, \ldots M$ be the state space for the Markov Chain
- A sequence is called a markov chain if

$$Pr(\theta^{(t+1)} \mid \theta^{(t)}, \theta^{(t-1)} \ldots \theta^{(1)}) = Pr(\theta^{(t+1)} \mid \theta^{(t)})$$

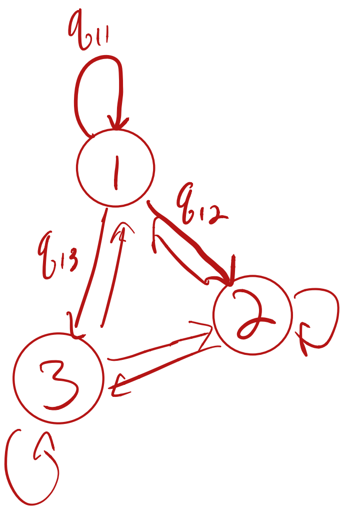
for all $t \geq 0$

- The **Markov property**: given the entire past history, $\theta^{(1)}, \ldots \theta^{(t)}$, the most recent $\theta^{(t+1)}$ depends only on the immediate past, $\theta^{(t)}$

Memory of 1 time period

The Transition Matrix

- Define $q_{ij} = Pr(\theta^{(t+1)} \mid \theta^{(t)})$ is the transition probability from state i to state j
- The $M \times M$ matrix $Q = (q_{ij})$ is called the *transition matrix* of the Markov Chain



3-state example

$$Q = \begin{array}{cc} & \text{To} \\ \text{From} & \begin{bmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{bmatrix} \end{array}$$

The Transition Matrix

3-state example

$$Q = \begin{bmatrix} q_{11} & q_{12} & q_{13} \\ q_{21} & q_{22} & q_{23} \\ q_{31} & q_{32} & q_{33} \end{bmatrix}$$

$q_{11} + q_{12} + q_{13} = 1$

- The rows of the transition matrix sum to 1
- Note: $Q^n = (q_{ij}^{(n)})$ is is the probability of transitionining from i to j in n steps

$(Q \times Q)_{ij}$ = Prob. of going to j in 2 steps, given start in i.

- A Markov Chain is **regular** if Q^n has strictly positive entries for some value of $n > 1$
- A Markov Chain is **irreducible** if for any two states i and j it is possible to go from i to j in a finite number of steps (with positive probability)

"Graph is connected"

Reducible

The limiting distribution

- A regular, irreducible Markov chain has a **limiting probability distribution**
- Describes the long-run fraction of time does the Markov Chain spend in each state (in the long run)
    - *Does not* depend on where the chain starts

long-run
island
probabilities.

- Let $\pi = (\pi_1, \dots \pi_M)$ be a row vector of probabilities associated with each state, such that $\sum_{i=1}^{M} = \pi_i = 1$
    - The limiting distribution converges to π, which is said to be **stationary** because $\cancel{\pi Q = \pi}$ $Q^T \pi = \pi$
    - If you sample from the limiting distribution and then transtion, the result is still distributed according to the limiting distribution

e.g. $\pi = (1/4, 1/4, 1/2)$

Markov Chain Example

- Sociologists often study social mobility using a Markov chain.
- In this example, the state space is `{low income, middle income, and high income}` of families
- Let $\mathbf{Q}$ be the transition matrix from parents income to childrens income

Child

Parents $$\mathbf{Q} = \begin{array}{c|ccc|} & \text{Lower} & \text{Middle} & \text{Upper} \\ \hline \text{Lower} & 0.40 & 0.50 & 0.10 \\ \text{Middle} & 0.05 & 0.70 & 0.25 \\ \text{Upper} & 0.05 & 0.50 & 0.45 \end{array}$$

Multi-step Transition Probabilities

Grandkids

2-step transition probabilities

$$\mathbf{Q}^2 = \mathbf{Q} \times \mathbf{Q} = \begin{array}{ccc} L & M & U \\ \multicolumn{3}{c}{\left|\begin{matrix} 0.1900 & 0.6000 & 0.2100 \\ 0.0675 & 0.6400 & 0.2925 \\ 0.0675 & 0.6000 & 0.3325 \end{matrix}\right|} \end{array}$$

parents

4-step transition probabilities

4th Generation

$$\mathbf{Q}^4 = \mathbf{Q}^2 \times \mathbf{Q}^2 = \left|\begin{matrix} 0.0908 & 0.6240 & 0.2852 \\ 0.0758 & 0.6256 & 0.2986 \\ 0.0758 & 0.6240 & 0.3002 \end{matrix}\right|$$

Multi-step Transition Probabilities

4-step transition probabilities

$$\mathbf{Q}^4 = \mathbf{Q}^2 \times \mathbf{Q}^2 = \begin{vmatrix} 0.0908 & 0.6240 & 0.2852 \\ 0.0758 & 0.6256 & 0.2986 \\ 0.0758 & 0.6240 & 0.3002 \end{vmatrix}$$

8-step transition probabilities

$$\mathbf{Q}^8 = \mathbf{Q}^4 \times \mathbf{Q}^4 = \begin{vmatrix} 0.0772 & 0.6250 & 0.2978 \\ 0.0769 & 0.6250 & 0.2981 \\ 0.0769 & 0.6250 & 0.2981 \end{vmatrix}$$

Converging to limiting distribution.

Idea of MCMC

~~The limiting distribution~~

Design Q so $\pi = P(\theta | y)$

$$\mathbf{Q}^{\infty} = \mathbf{1}\pi = \begin{vmatrix} \pi_1 & \pi_2 & \pi_3 \\ \pi_1 & \pi_2 & \pi_3 \\ \pi_1 & \pi_2 & \pi_3 \end{vmatrix}$$

π is limiting distribution.

```r
Q <- matrix(c(0.4, 0.05, 0.05,
              0.5, 0.7, 0.5,
              0.1, 0.25, 0.45),
            ncol=3)

p <- eigen(t(Q))$vectors[, 1]
stationary_probs <- p/sum(p)
stationary_probs
```

$Q^T \pi = \pi$ (defn of stationary)

limiting distn.

```
## [1] 0.07692308 0.62500000 0.29807692
```

```r
stationary_probs %*% Q
```

Still in stationary distn.

```
##               [,1]  [,2]      [,3]
## [1,] 0.07692308 0.625 0.2980769
```

Markov Chain Monte Carlo

- Incredible idea: create a Markov Chain with the desired limiting distribution
    - Want the limiting distribution to be the posterior distribution
- Unlike the previous examples, we will mostly work with *infinite* state space
- Instead of a transition matrix we have a transition kernel which is a conditional probability, $p(\theta^{(t+1)} \mid \theta^{(t)})$

Design

- Want $p(\theta^{(t+1)} \mid \theta^{(t)})$ to have limiting distribution $p(\theta \mid y)$
    - If we run the random walk for long enough, $\theta^{(t)}$ will be distributed approximately according to $p(\theta \mid y)$
- The Metroplis algorithm tells us how to construct such a transition ~~matrix~~ kernel

Generalizing the rejection sampler

- Make a small tweak to the rejection sampler
- Sample from a proposal, $q(\theta)$, doesn't have to envelope $p(\theta \mid y)$!
- If $p(\theta \mid y) > 0$ then we need $q(\theta) > 0$ (same support)
- Unlike the rejection sampler, we never "throw out" samples
- Instead, at each iteration we have a choice:
    - Accept the new proposed sample

  or

    - Or **repeat** the previous sample again

Generalizing the rejection sampler

1. Initialize θ_0 to be the starting point for you Markov Chain
2. Choose a *proposal distribution*, $J(\theta^*)$ "Jump" density
   - Propose a candidate value for the next sample
   - Best performance if density is very similar to target
3. Generate the candidate θ^* from the proposal distribution, J
4. Compute $r = \min(1, \frac{p(\theta^*|y)}{p(\theta_t|y)})$

   Posterior density at proposed value

   Posterior density at proposed value at current value
5. Set $\theta_{t+1} \leftarrow \theta^*$ with probability r
   - Generate a uniform random number $u \sim Unif(0, 1)$
   - If $u < r$ we accept θ^* as our next sample
   - $\boxed{\text{Else } \theta_{t+1} \leftarrow \theta_t}$ (we do not update the sample this time)

Intuition

- If $p(\theta^* \mid y) > p(\theta_t \mid y)$ accept with probability 1
    - The proposed sample has higher posterior density than the previous sample
    - Always accept if we increase the posterior probability density
- If $p(\theta^* \mid y) < p(\theta_t \mid y)$ accept with probability $r < 1$
    - Accept with probability less than 1 if probability density would decrease
    - Relative frequency of θ^* vs θ_t in our samples should be $\frac{p(\theta^*|y)}{p(\theta_t|y)}$

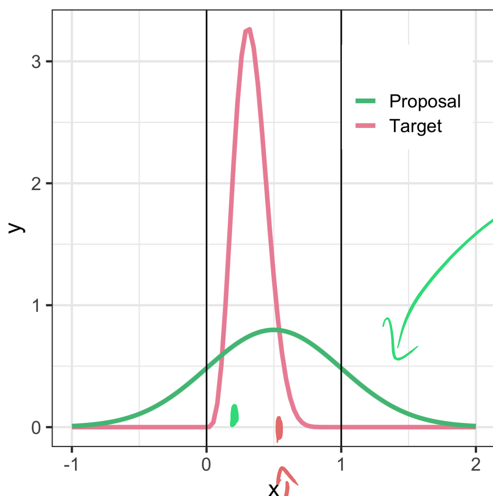
Independence Sampler

- The previous algorithm is known as an "Independence Sampler"
- Let $P(\theta \mid y)$ be a Beta(5, 10) posterior distribution
- Propose from a distribution $J(\theta^*) \sim N(0.5, 1)$

Confusing Name!

Samples are <u>not</u> independent.

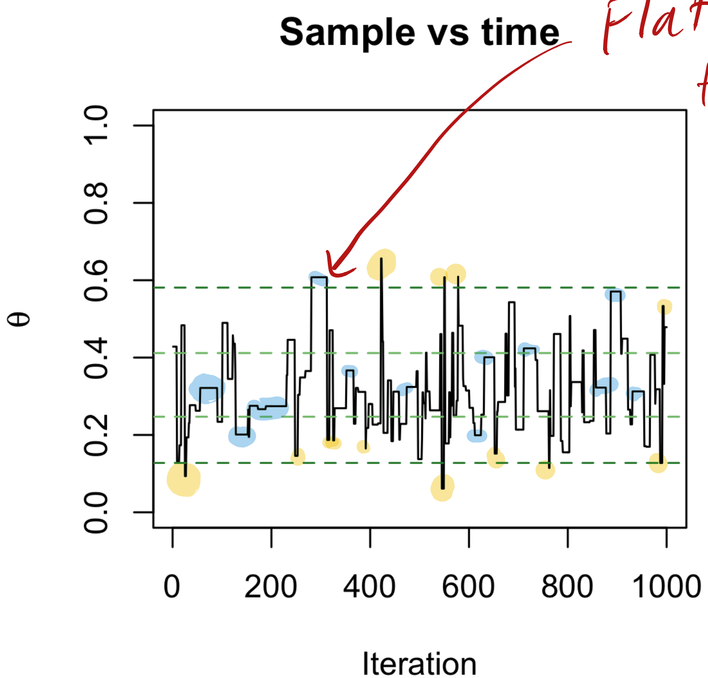
$$\frac{P(\theta^* \mid y)}{P(\theta_0 \mid y)}$$



θ_0

Proposal (Doesn't envelope target)

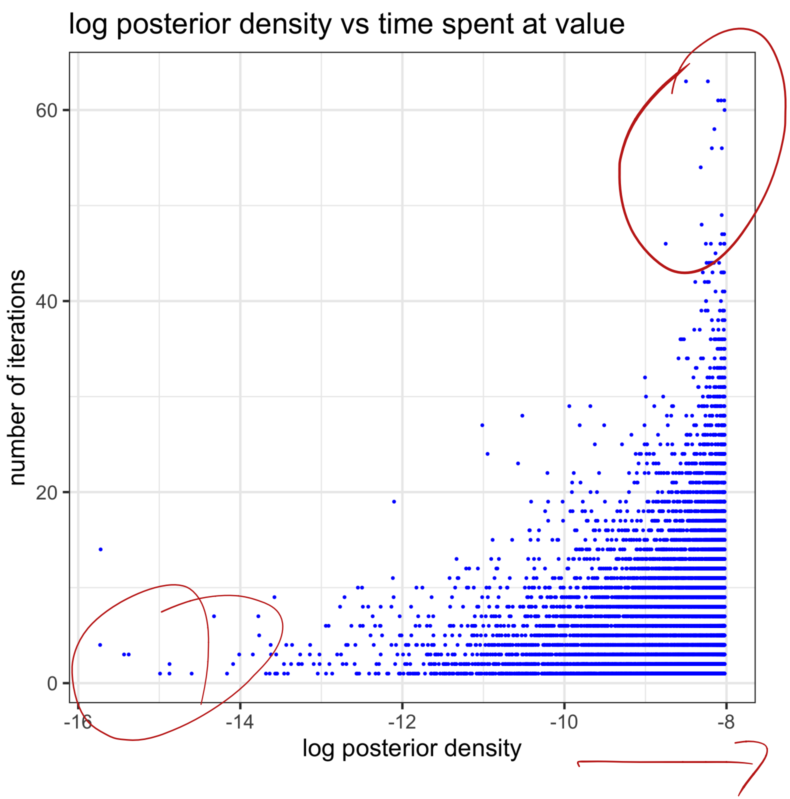
Independence Sampler



Flat parts show that we're staying in a state.

Note and source of confusion: samples are correlated over time for the "independence sampler".

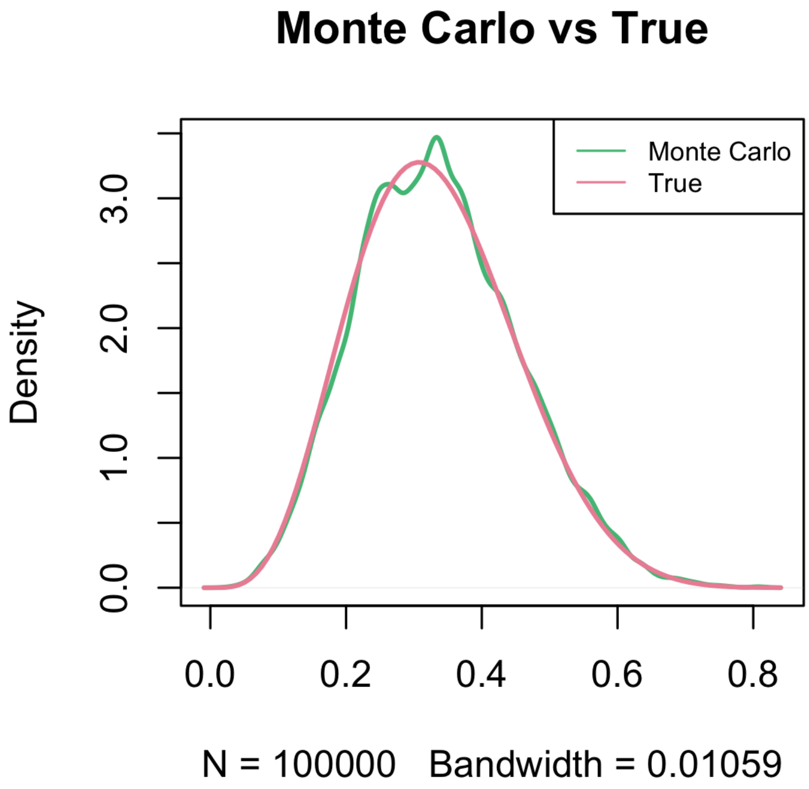
Weighting by waiting



time at a given state.

Where did the sampler get stuck? Where does it quickly leave?

Independence Sampler



The Metropolis Algorithm

- The Metropolis Algorithm generalizes the independence sampler
- Allow the proposal distribution to depend on the most recent sample
  - Independence: $J(\theta^*)$, e.g. $\theta^* \sim N(0.5, 1)$ Does not depend on θ_t
  - Metropolis: $J(\theta^* \mid \theta_t)$, e.g. $\theta^* \sim N(\theta_t, 1)$
- Independence sampler: "Independence" refers to the proposal being fixed (**not** independence samples)!
- Metropolis sampler: a "moving" proposal distribution

Proposal density changes in time.

The Metropolis Algorithm

1. Initialize θ_0 to be the starting point for you Markov Chain
2. Choose a proposal distribution, $J(\theta^* \mid \theta_t)$
   - Propose a candidate value for the next sample
   - Must have symmetry: $J(\theta^* \mid \theta_t) = J(\theta_t \mid \theta^*)$
3. Generate the candidate θ^* from the proposal distribution, J
4. Compute $r = \min(1, \frac{p(\theta^*|y)}{p(\theta_t|y)})$
5. Set $\theta_{t+1} \leftarrow \theta^*$ with probability r
   - Generate a uniform random number $u \sim Unif(0, 1)$
   - If $u < r$ we accept θ^* as our next sample
   - Else $\theta_{t+1} \leftarrow \theta_t$ (we do not update the sample this time)

Metropolis Algorithm

- Let $P(\theta \mid y)$ be a Beta(5, 10) posterior distribution
- 1-d sampling: lets try sampling from the Beta using the Metropolis algorithm
- Initialize θ_0 to 0.9
  - Note that the probability of drawing a value larger than 0.9 from a Beta(5, 10) is smaller than 1e-8
  - Our initial value is far from the high posterior density
  - In the long run this won't matter
- Define transition kernel $J(\theta_{t+1} \mid \theta_t)$ as $\theta^* \sim N(\theta_t, \tau^2)$
  - How does choice of τ^2 effect performance of MC sampler?