STA250 Notes

Matt Meisner

16 October 2013

1 Notes on Markov Chains

1. Burn-in.
   (a) We typically discard the first portion of our samples that we acquire through Gibbs Sampling or the Metropolis Hastings algorithm; this is called the "burn-in."
   (b) Since the chain has not yet reached the stationary distribution that we know corresponds to the target distribution we wish to approximate, these initial samples don't accurately decribe the posterior of interest.
   (c) There is not an accepted, theoretically sound way for determining how many samples the burn-in should be. So, we try to be conservative, and throw out more than we need to. One way of choosing the burn-in size is to inspect the trace plot, which is a plot of the parameter estimates against iteration number; one can choose the number of iterations for the burn-in as the approximate number of iterations until the parameter estimates seem to be converging to a single value, and after wild fluctuations have subsided.
   (d) A larger burn-in may be needed if we start simulations at very wrong starting values. Fortunately, regardless of where we start, the chain, in theory, converges to the stationary distribution (which is our target distribution).
2. Sample dependence.
   (a) Since the subsequent state of a Markov chain depends on its current state, the samples we get through Gibbs Sampling or the MH algorithm are dependent upon one another.
   (b) Autocorrelation is one way of quantifying this; one can see how samples, taken different numbers of iterations apart, are correlated with each other other.
   (c) The R function acf() computes autocorrelations and produces beautiful plots.
   (d) While we would prefer completely independent samples from the posterior, we don't have a way of doing this (unless we know the posterior and can sample from it, but then there's no need to try to approximate it!).
   (e) Fortunately, unless the dependencies are very large, this doesn't cause major problems. But if the dependencies are very large, then our exploration of parameter space can be slow.
   (f) As long as we have a Markov chain that is ergodic, then our earlier result about Monte Carlo integration (which assumed iid samples) generalizes to dependent samples: we know we converge to the stationary distribution as the number of samples goes to infinity.
3. Effective sample size.
   (a) Our samples aren't independent, and $10,000$ dependent samples provide less information than $10,000$ independent samples.
   (b) Effective sample size can be formally defined, but the general idea is that it tries to quantify how many independent samples a certain number of dependent samples is equivalent to.

2 Metropolis Hastings Algorithm

1. We need a method of obtaining an ergodic Markov chain. Gibbs sampling is one such method, but not the only one. The MH algorithm is another.

2. The MH algorithm satisfies the "detailed balance condition:"

$$\pi(x)p(x,y) = \pi(y)p(y,x), \qquad \forall\ x, y.$$

   Where π is the stationary distribution and p is the transition density, i.e. $p(x,y)$ is the probability of transitioning to state y conditional on being in state x, and $p(y,x)$ is the probability of transitioning to state x conditional on being in state y.

3. When the detailed balance condition holds, existence of a stationary distribution π is guaranteed.

4. Algorithm:

   (a) Set $t = 0$ and select the starting value $\theta^{(0)}$.

   (b) Given the current state $\theta^{(t)}$, propose a new value θ^* from a proposal density $p(\theta^{(t)}, \cdot)$, that depends on the current state. Note that this p is NOT the same as the transition density, also called p, described above!

   (c) Accept the proposed value with probability α, where:

$$\alpha = \min\left\{1, \frac{\pi(\theta^*)p(\theta^*, \theta^{(t)})}{\pi(\theta^{(t)})p(\theta^{(t)}, \theta^*)}\right\}$$

   If accepted, set $\theta^{(t+1)} = \theta^*$, else set $\theta^{(t+1)} = \theta^{(t)}$.

   Note that if we use a normal proposal distribution, the above simplies to:

$$\alpha = \min\left\{1, \frac{\pi(\theta^*)}{\pi(\theta^{(t)})}\right\}$$

   Since the normal distribution is symmetric, implying: $p(\theta^*, \theta^{(t)}) = p(\theta^{(t)}, \theta^*)$. This special case is the Metropolis algorithm.

   (d) Increment $t \mapsto t + 1$ and return to step b.

5. A critical feature of the MH algorithm is that we only need to know the posterior density, $p(\theta|x)$, up to proportionality. That's because the ugly marginal $p(x)$ cancels out when we calculate $\frac{\pi(\theta^*)}{\pi(\theta^{(t)})}$. Any algorithm requiring $p(x)$ would be essentially useless, since it's not typically computationally feasible to quickly calculate it.

6. The proposal distribution $p(\theta^{(t)}, \theta^*)$, which describes the probability of proposed values of θ given that we are at $\theta^{(t)}$, is often chosen to be $N(\theta^{(t)}, v^2)$.

   (a) v^2 is a tuning parameter that we control.

   (b) Make it too large, and ridiculous values will be prosposed, very few proposals will be accepted, and exploring the interesting part of the parameter space will be slow. A signature of this problem in a trace plot is that the parameter estimates will stay at one value for several steps, only occasionally jumping to new values.

   (c) On the other hand, make v^2 too small, and the acceptance rate will be extremely high, but the changes in parameter values at each time point will be tiny. Again, exploring the posterior density will be slow. In this case, the trace plot shows very tiny changes at each subsequent iteration.

   (d) In practice, v^2 is typically chosen by trial and error; we ideally want an acceptance rate in the 20-60% range, but this may not always be possible, especially for multivariate parameters.

(e) We may tweak v^2 during the burn-in to try to get an acceptance rate in that range, but we may not change it afterwards, unless we want to venture into the challenging world of "adaptive MCMC," which is a world fraught with danger.

7. We can combine MH and Gibbs sampling in an MCMC implementation, and this can often increase efficiency. Gibbs sampling requires that we can analytically write out the full conditional distribution of a parameter conditioned on all of the other parameters. For some parameters in some models, this might be impossible or tedious. If that is the case, then we can substitute MH sampling for these parameters (while still using Gibbs sampling when these conditionals are known).