Technical Computing for the Earth Sciences, Lecture 8:

Bayes Theorem, MCMC beyond plain Metropolis

EARS 80.03

“Bayesian Statistics”?

“Bayesian Statistics”?

Can refer to two related things

1. Anything that makes use of *Bayes’ Theorem*
   (also sometimes called *Bayes Rule*)
   - $\mathrm{P(A|B) = P(B|A)*P(A)/P(B)}$
     - or: $\mathrm{P(A|B)\ P(B) = P(B|A)\ P(A)}$
     - May call $\mathrm{P(A)}$ “*Prior*” and $\mathrm{P(A|B)}$ “*Posterior*”

“Bayesian Statistics”?

Can refer to two related things

1. Anything that makes use of *Bayes’ Theorem*
   (also sometimes called *Bayes Rule*)
   - $\mathrm{P(A|B) = P(B|A)*P(A)/P(B)}$
     - or: $\mathrm{P(A|B)\ P(B) = P(B|A)\ P(A)}$
     - May call $\mathrm{P(A)}$ “*Prior*” and $\mathrm{P(A|B)}$ “*Posterior*”
2. A *philosophy* of statistics - “Bayesian” vs “Frequentist”
   - Frequentist: Assigns probabilities to observations only
     “*If I repeated this experiment many times, how often I would get result X*”
   - Bayesian: Probabilities as degrees of belief / (un)certainty
     “*What is the probability that X is true*”

“Bayesian Statistics”?

Philosophy of statistics - [“Bayesian” vs “Frequentist”]

- Differences in terminology:
  - Frequentists talk about “*confidence intervals*”;
    Bayesians talk about “*credible intervals*”.
  - If you see “priors” and “posteriors” think Bayesian.
    If you see hypothesis tests, think Frequentist.
- Bayesians put assumptions in their “prior” distributions; Frequentists may encode similar assumptions in their choice of which technique(s) to apply to which problem(s)
- Any practical problem that can be solved with one approach can generally be solved with the other as well, but one approach may be more natural than the other

The Metropolis Algorithm

The Metropolis algorithm

The Metropolis algorithm is an adaptation of a random walk with an acceptance/rejection rule to converge to the specified target distribution. The algorithm proceeds as follows.

1. Draw a starting point θ^0, for which $p(\theta^0|y) > 0$, from a *starting distribution* $p_0(\theta)$. The starting distribution might, for example, be based on an approximation as described in Section 13.3. Or we may simply choose starting values dispersed around a crude approximate estimate of the sort discussed in Chapter 10.
2. For $t = 1, 2, \ldots$:
   (a) Sample a *proposal* θ^* from a *jumping distribution* (or *proposal distribution*) at time t, $J_t(\theta^*|\theta^{t-1})$. For the Metropolis algorithm (but not the Metropolis-Hastings algorithm, as discussed later in this section), the jumping distribution must be *symmetric*, satisfying the condition $J_t(\theta_a|\theta_b) = J_t(\theta_b|\theta_a)$ for all θ_a, θ_b, and t.
   (b) Calculate the ratio of the densities,
   $$r = \frac{p(\theta^*|y)}{p(\theta^{t-1}|y)}. \tag{11.1}$$
   (c) Set
   $$\theta^t = \begin{cases} \theta^* & \text{with probability } \min(r, 1) \\ \theta^{t-1} & \text{otherwise.} \end{cases}$$

Gelman et al., p 278.

The Metropolis-Hastings Algorithm

The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm generalizes the basic Metropolis algorithm presented above in two ways. First, the jumping rules J_t need no longer be symmetric; that is, there is no requirement that $J_t(\theta_a|\theta_b) \equiv J_t(\theta_b|\theta_a)$. Second, to correct for the asymmetry in the jumping rule, the ratio r in (11.1) is replaced by a ratio of ratios:

$$r = \frac{p(\theta^*|y)/J_t(\theta^*|\theta^{t-1})}{p(\theta^{t-1}|y)/J_t(\theta^{t-1}|\theta^*)}. \tag{11.2}$$

(The ratio r is always defined, because a jump from θ^{t-1} to θ^* can only occur if both $p(\theta^{t-1}|y)$ and $J_t(\theta^*|\theta^{t-1})$ are nonzero.)

Allowing asymmetric jumping rules can be useful in increasing the speed of the random walk. Convergence to the target distribution is proved in the same way as for the Metropolis algorithm. The proof of convergence to a unique stationary distribution is identical. To prove that the stationary distribution is the target distribution, $p(\theta|y)$, consider any two points θ_a and θ_b with posterior densities labeled so that $p(\theta_b|y)J_t(\theta_a|\theta_b) \geq p(\theta_a|y)J_t(\theta_b|\theta_a)$. If θ^{t-1} follows the target distribution, then it is easy to show that the unconditional probability density of a transition from θ_a to θ_b is the same as the reverse transition.

Gelman et al., p 279.

The Metropolis-Hastings Algorithm

Relation between the jumping rule and efficiency of simulations

The ideal Metropolis-Hastings jumping rule is simply to sample the proposal, θ^*, from the target distribution; that is, $J(\theta^*|\theta) \equiv p(\theta^*|y)$ for all θ. Then the ratio r in (11.2) is always exactly 1, and the iterates θ^t are a sequence of independent draws from $p(\theta|y)$. In general, however, iterative simulation is applied to problems for which direct sampling is not possible.

A good jumping distribution has the following properties:

- For any θ, it is easy to sample from $J(\theta^*|\theta)$.
- It is easy to compute the ratio r.
- Each jump goes a reasonable distance in the parameter space (otherwise the random walk moves too slowly).
- The jumps are not rejected too frequently (otherwise the random walk wastes too much time standing still).

We return to the topic of constructing efficient simulation algorithms in the next chapter.

Gelman et al., p 280.

Other MCMC variants

- *Hamiltonian Monte Carlo* (HMC)
    - A variant of Metropolis-Hastings where walkers have momentum
- *No-U-Turn Sampling* (NUTS)
    - A variant of HMC where you can’t turn around
- MCMC with simulated annealing
- MCMC with genetic algorithms
- etc…