Metropolis and Metropolis-Hastings algorithms

张世明 李芝茂

The Metropolis-Hastings algorithm is a general term for a family of Markov chain simulation methods that are useful for sampling from Bayesian posterior distributions.

The Metropolis algorithm

This algorithm proceeds as follows.

1. Draw a starting point θ_0, for which p(θ_0|y) > 0, from a starting distribution P_0 (θ).

2. For t = 1, 2, . . . :

(a) Sample a proposal θ^* from a jumping distribution at time t , $J_t(\theta^* | \theta_{t-1})$.

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

$$r = p(\theta^*|y)/p(\theta_{t-1}|y)$$

(c) Set

$$\theta_t = \begin{cases} \theta^* & \text{with probability } \min(r, 1) \\ \theta_{t-1} & \text{otherwise.} \end{cases}$$

The Metropolis algorithm

Given the current value θ_t, the transition distribution T_t $(\theta_t|\theta_{t-1})$ of the Markov chain is thus a mixture of a point mass at θ_t = θ_{t-1}, and a weighted version of the jumping distribution adjusts for the acceptance rate .

The algorithm requires the ability to calculate the ratio r for all (θ, θ^*), and to draw θ from the jumping distribution J_t $(\theta^*|\theta)$ for all θ and t. In addition, step (c) requires the generation of a uniform random number.

When θ_t = θ_{t-1} , this still counts as an iteration in the algorithm.

The Metropolis algorithm

Why does the Metropolis algorithm work?

The proof that the sequence of iterations $\theta_1, \theta_2, \ldots$ converges to the target distribution has two steps: first, it is shown that the simulated sequence is a Markov chain with a unique stationary distribution, and second, it is shown that the stationary distribution equals this target distribution.

The Metropolis algorithm

Why does the Metropolis algorithm work?

Now consider any two such points θ_a and θ_b , drawn from p(θ|y) and labeled so that p(θ_b|y) ≥ p(θ_a|y). The unconditional probability density of a transition from θ_a to θ_b is

$$p(\theta_{t-1} = \theta_a, \theta_t = \theta_b) = p(\theta_a | y)\ J_t(\theta_b | \theta_a),$$

The Metropolis algorithm

Why does the Metropolis algorithm work?

The unconditional probability density of a transition from θ_b to θ_a is:

$$p(\theta_{t-1} = \theta_b, \theta_t = \theta_a)$$

$$= p(\theta_b|y)\, Jt(\theta_a|\theta_b)[p(\theta_a|y) / p(\theta_b|y)]$$

$$= p(\theta_a|y)\, J_t(\theta_a|\theta_b)$$

which is the same as the probability of a transition from θ_a to θ_b. Since their joint distribution is symmetric, θt and $\theta t-1$ have the same marginal distributions, and so $p(\theta|y)$ is the stationary distribution of the Markov chain of θ.

The Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm generalizes the basic Metropolis algorithm presented above in two ways.

The Metropolis-Hastings algorithm

First, the jumping rules Jt need no longer be symmetric;

The Metropolis-Hastings algorithm

Second, to correct for the asymmetry in the jumping rule, the ratio r in (11.1) is replaced by a ratio of ratios:

$$r = \frac{p(\theta * \mid y) \,/\, Jt(\theta * \mid \theta t - 1)}{p(\theta t - 1 \mid y) \,/\, Jt(\theta t - 1 \mid \theta *)}$$

The Metropolis-Hastings algorithm

Allowing asymmetric jumping rules can be useful in increasing the speed of the random walk.

The Metropolis-Hastings algorithm

To prove that the stationary distribution is the target distribution, $p(\theta|y)$, consider any two points θa and θb with posterior densities labeled so that $p(\theta b|y)Jt(\theta a|\theta b) \geq p(\theta a|y)Jt(\theta b|\theta a)$.

The Metropolis-Hastings algorithm

If θt−1 follows the target distribution, then it is easy to show that the unconditional probability density of a transition from θa to θb is the same as the reverse transition.

Relation between the jumping rule and efficiency of simulations

The ideal Metropolis-Hastings jumping rule is simply to sample the proposal, $\theta*$, from the target distribution; that is, $J(\theta*|\theta) \equiv p(\theta*|y)$ for all θ.

Relation between the jumping rule and efficiency of simulations

Then the ratio r in (11.2) is always exactly 1, and the iterates θt 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.

Relation between the jumping rule and efficiency of simulations

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.

Relation between the jumping rule and efficiency of simulations

- 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).

Thank you