Metropolis-Hastings Algorithm

Yingbo Li

Clemson University

MATH 9810

Metropolis-Hastings: Motivation

- Sometimes drawing from symmetric proposal distribution $J(\theta \mid \theta^{(s)})$ not efficient, i.e., takes long time for chain to converge
- Example of such inefficiency:
  - Suppose $p(\theta \mid Y)$ has long tail like a Gamma distribution.
  - Normal proposal with small variance: takes long time to traverse distribution repeatedly.
  - Normal proposal with large variance: many proposed θ with small posterior density, so too small rate of acceptance

Metropolis-Hastings: Motivation

- In cases like last slide, ideal to propose values in tail roughly in same proportion as they appear in $p(\theta \mid Y)$.
- For example, $J \sim$ Gamma might be a closer approximation to $p(\theta \mid Y)$ than $J \sim$ Normal.
- But, Gamma distribution is not symmetric proposal distribution
- Have to correct the acceptance ratio r for this fact; otherwise, we might inaccurately favor values with high density in J that may not be high density in $p(\theta \mid Y)$
- This leads to the Metropolis-Hastings (M-H) algorithm

Metropolis-Hastings Algorithm

Suppose we want to estimate $p(\theta|Y)$ using M-H

- Propose a new $\theta^* \sim J(\theta \mid \theta^{(s)})$ where J is an arbitrary distribution (certain restrictions apply)
- Compute Metropolis-Hastings ratio

$$\alpha = \min\left\{1, \frac{p(\theta^* \mid Y)J(\theta^{(s)} \mid \theta^*)}{p(\theta^{(s)} \mid Y)J(\theta^* \mid \theta^{(s)})}\right\}$$

- Set

$$\theta^{(s+1)} = \begin{cases} \theta^* & \text{with probability } \alpha \\ \theta^{(s)} & \text{with probability } 1-\alpha \end{cases}$$

Features of M-H Jumping Distribution

- It is easy to sample from $J(\theta \mid \theta^{(s)})$ and to compute α.
- $J(\theta \mid \theta^{(s)})$ must depend only on $\theta^{(s)}$ and not previous values of θ in the chain
- $J(\theta|\theta^{(s)})$ must be such that you can get to any value of the parameter space for θ eventually from any $\theta^{(s)}$
- $J(\theta|\theta^{(s)})$ must be such that you don't return periodically to any particular value of θ
- You get to specify $J(\theta|\theta^{(s)})$. Use tuning to select one that leads to roughly 35% of new proposed θ^* accepted.
- Can use different jumping distributions in different iterations, i.e., J is allowed to depend on s.
  But J cannot dependent on the draws, i.e., $\theta^{(s)}$.

Special Cases of M-H Algorithm

- Metropolis algorithm: symmetric jump $J(\theta^* \mid \theta^{(s)}) = J(\theta^{(s)} \mid \theta^*)$

$$\alpha = \min\left\{1, \frac{p(\theta^* \mid Y)}{p(\theta^{(s)} \mid Y)}\right\}$$

- Gibbs sampler: jumping distribution equals the target distribution, i.e., $J(\theta^* \mid \theta^{(s)}) = p(\theta^* \mid Y)$, hence

$$\alpha = \min\left\{1, \frac{p(\theta^* \mid Y)p(\theta^{(s)} \mid Y)}{p(\theta^{(s)} \mid Y)p(\theta^* \mid Y)}\right\} = 1$$

- Since J can be different in different iterations, we can update each dimension of the parameter vector one at a time, using either Gibbs, Metropolis, or M-H update.

Examples of Jumping Distributions

- Continuous $\theta \in \mathbb{R}$: symmetric random walk:

$$J(\theta \mid \theta^{(s)}) = \mathsf{N}(\theta^{(s)}, c^2) \text{ or } \mathsf{Unif}(\theta^{(s)} - c, \theta^{(s)} + c)$$

- Continuous $\theta \in [a, b]$: reflecting random walk:

$$\theta^* \sim \mathsf{Unif}(\theta^{(s)} - c, \theta^{(s)} + c)$$

  If $\theta^* < a$, use $a + (\theta^* - a)$; if $\theta^* > b$, use $b - (\theta^* - b)$;

- Continuous $\theta \in \mathbb{R}^+$: in addition to reflecting random walk, can also use symmetric random walk on $\log(\theta)$:

$$\log \theta^* \sim \mathsf{N}(\log \theta^{(s)}, c^2) \text{ or } \mathsf{Unif}(\log \theta^{(s)} - c, \log \theta^{(s)} + c)$$

  Here the jumping distribution is no longer symmetric.

- Continuous $\theta \in (0, 1)$: symmetric random walk on $\log(\theta/(1 - \theta))$

Examples of Jumping Distributions

- Discrete $\theta \in \mathbb{Z}$: symmetric random walk:

$$\theta^* = \begin{cases} \theta^{(s)} + 1 & \text{with probability } 1/2 \\ \theta^{(s)} - 1 & \text{with probability } 1/2 \end{cases}$$

- Discrete $\theta \in \mathbb{N} \cup \{0\}$: reflecting random walk
  If $\theta^{(s)} \geq 1$, the same as above; if $\theta^{(s)} = 0$, then set $\theta^* = 1$.
- Discrete $\theta \in \{0, 1\}$:

$$\theta^* = \begin{cases} 1 & \text{if } \theta^{(s)} = 0 \\ 0 & \text{if } \theta^{(s)} = 1 \end{cases}$$

Ergodic Theorem

Theorem

If $\{x^{(1)}, x^{(2)}, ...\}$ is an irreducible, aperiodic and recurrent Markov chain, then there is a unique probability distribution π such that as $s \to \infty$,

- $P(x^{(s)} \in A) \longrightarrow \pi(A)$ *for any set* A*;*
- $\frac{1}{S} \sum_s g(x^{(s)}) \longrightarrow \int g(x)\pi(x)dx.$

The distribution π is the *stationary distribution* of the Markov chain. If

1. $x^{(s)} \sim \pi$, and
2. $x^{(s+1)}$ is generated from the Markov chain starting at $x^{(s)}$,

then $x^{(s+1)} \sim \pi$.

“Proof” that $\pi(\theta) = p(\theta \mid Y)$ for M-H algorithm

Suppose $\theta^{(s)} \sim p(\theta \mid Y)$, we need to show $\theta^{(s+1)} \sim p(\theta \mid Y)$, too.

Suppose θ_a and θ_b are two values of θ such that $p(\theta_a \mid Y)J(\theta_b \mid \theta_a) \geq p(\theta_b \mid Y)J(\theta_a \mid \theta_b)$, then

$$\begin{aligned}
p(\theta^{(s)} = \theta_a, \theta^{(s+1)} = \theta_b) &= p(\theta_a \mid Y) \cdot J(\theta_b \mid \theta_a) \cdot \frac{p(\theta_b \mid Y)J(\theta_a \mid \theta_b)}{p(\theta_a \mid Y)J(\theta_b \mid \theta_a)} \\
&= p(\theta_b \mid Y)J(\theta_a \mid \theta_b) \\
&= p(\theta^{(s)} = \theta_b, \theta^{(s+1)} = \theta_a)
\end{aligned}$$

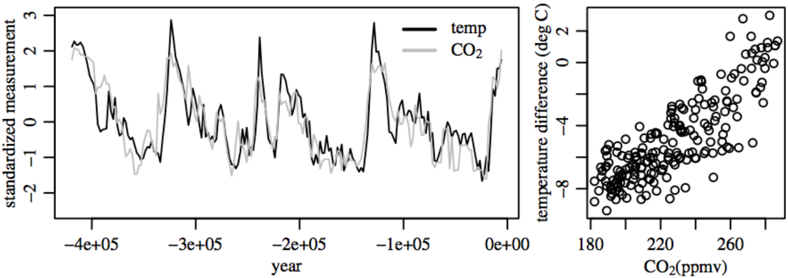
Hence marginal distribution of $\theta^{(s+1)}$ is

$$\begin{aligned}
p(\theta^{(s+1)} = \theta) &= \int p(\theta^{(s+1)} = \theta, \theta^{(s)} = \theta')d\theta' \\
&= \int p(\theta^{(s+1)} = \theta', \theta^{(s)} = \theta)d\theta' \\
&= p(\theta^{(s)} = \theta)
\end{aligned}$$

Example: Regression with Correlated Errors

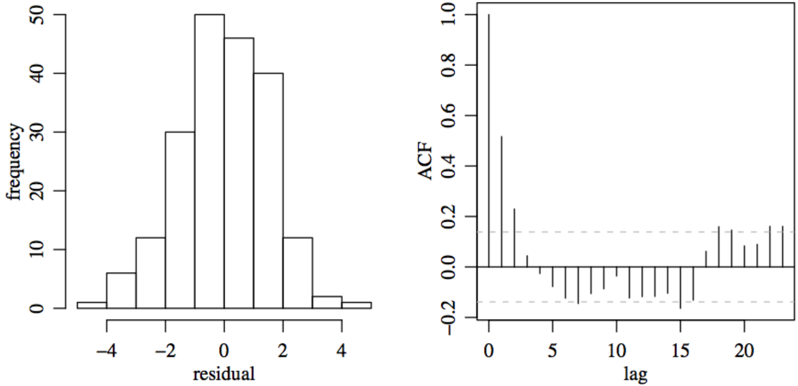
Historical CO_2 and temperature data

- Response: temperature
- Predictor: CO_2 concentration
- $n = 200$, roughly one observation for every 2000 years



Normal Linear Regression: Autocorrelation in Residuals

$$Y \sim \mathsf{N}(X\beta, \sigma^2 I_n)$$



Covariance of Y should not be a scalar matrix!

Normal Linear Regression with AR(1) Errors

$$Y \sim \mathsf{N}(X\beta, \Sigma), \quad \Sigma = \sigma^2 C_\rho,\ [C_\rho]_{i,j} = \rho^{|i-j|}$$

It seems reasonable to assume that ρ is positive; and to be stationary, it cannot be greater than 1.

Prior distributions:

$$\beta \sim \mathsf{N}(\beta_0, \Sigma_0),\ \sigma^2 \sim \mathsf{IG}(\nu 0/2, \nu_0\sigma_0^2/2),\ \rho \sim \mathsf{Unif}(0,1)$$

Conditional posteriors:

$$\beta \mid \sigma^2, \rho, Y \sim \mathsf{N}(\beta_n, \Sigma_n)$$

$$\Sigma_n = (X^T C_\rho^{-1} X/\sigma^2 + \sigma_0^{-1})^{-1}, \beta_n = \Sigma_n(X^T C_\rho^{-1} Y/\sigma^2 + \sigma_0^{-1}\beta_0)$$

$$\sigma^2 \mid \beta, \rho, Y \sim \mathsf{IG}((\nu 0 + n)/2, (\nu_0\sigma_0^2 + SSR_\rho)/2)$$

$$SSR_\rho = (Y - X\beta)^T C_\rho^{-1}(Y - X\beta)$$

$$p(\rho \mid \sigma^2, \beta, Y) \propto \cdots$$

Posterior Computation

Hybrid of Gibbs and Metropolis sampling: given $\beta^{(s)}, \sigma^{2(s)}, \rho^{(s)}$,

- Gibbs update of β:
  $\beta^{(s+1)} \sim \mathsf{N}(\cdot, \cdot)$, which depends on $\sigma^{2(s)}, \rho^{(s)}$.
- Gibbs update of σ:
  $\sigma^{2(s+1)} \sim \mathsf{IG}(\cdot, \cdot)$, which depends on $\beta^{(s+1)}, \rho^{(s)}$.
- Metropolis update of ρ:
  $\rho^{(s+1)} \propto p(\rho \mid \beta^{(s+1)}, \sigma^{2(s+1)}, Y)$.