

Metropolis-Hastings Sampling

PUBH 8442: Bayes Decision Theory and Data Analysis

Eric F. Lock
UMN Division of Biostatistics, SPH
elock@umn.edu

03/31/2021

Overview of posterior simulation methods

- ▶ Direct sampling
- ▶ Non-iterative indirect sampling:
 - ▶ Importance sampling
 - ▶ Rejection sampling
- ▶ Markov chain Monte Carlo sampling:
 - ▶ Metropolis-Hastings algorithm
 - ▶ Gibbs sampling
- ▶ And many more!

Markov Chain Monte Carlo (MCMC)

- ▶ “Monte Carlo” refers to any method that uses random sampling to obtain results
- ▶ A *Markov chain* is a sequence of random variables $\theta^{(1)}, \theta^{(2)}, \dots$, satisfying the *Markov property*:

$$P(\theta^{(t+1)} | \theta^{(1)}, \dots, \theta^{(t)}) = P(\theta^{(t+1)} | \theta^{(t)}).$$

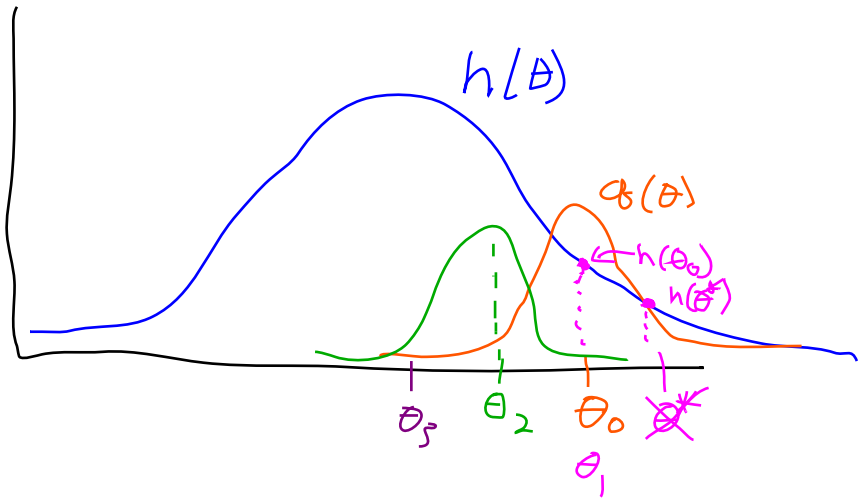
- ▶ Current *state* $t + 1$ can depend only on previous state t
- ▶ MCMC methods “adaptively” simulate from posterior $p(\theta | \mathbf{y})$
 - ▶ Current draw depends on previous draw
 - ▶ Draws converge to approximate dependent samples from $p(\theta | \mathbf{y})$

Metropolis-Hastings sampling

- ▶ Wish to draw $\theta^{(1)}, \theta^{(2)}, \dots$ from (potentially unnormalized) distribution h
 - ▶ e.g. $h(\theta) = p(\mathbf{y} | \theta)p(\theta)$
- ▶ Define a *proposal* density that depends on previous draw $\theta^{(t-1)}$: $q(\cdot | \theta^{(t-1)})$
- ▶ New draw is taken from $q(\cdot | \theta^{(t-1)})$, with a rejection step to encourage new draw has high density under h
- ▶ The *Metropolis* algorithm applies to symmetric q :

$$q(\theta^* | \theta^{(t-1)}) = q(\theta^{(t-1)} | \theta^*)$$

- ▶ *Metropolis-Hastings* algorithm extends to non-symmetric q .



The Metropolis Algorithm

- ▶ Specify an initial value $\theta^{(0)}$
- ▶ For $t = 1, \dots, T$, repeat:
 - ▶ Draw θ^* from $q(\cdot | \theta^{(t-1)})$
 - ▶ Compute $r = \frac{h(\theta^*)}{h(\theta^{(t-1)})}$
 - ▶ If $r \geq 1$, set $\theta^{(t)} = \theta^*$;
if $r < 1$, set $\theta^{(t)} = \begin{cases} \theta^* & \text{with probability } r \\ \theta^{(t-1)} & \text{with probability } 1 - r \end{cases}$.
- ▶ Often work with log-densities for computational reasons:

$$r = \exp\{\log(h(\theta^*)) - \log(h(\theta^{(t-1)}))\}$$

The Metropolis-Hastings Algorithm

- ▶ Specify an initial value $\theta^{(0)}$
- ▶ For $t = 1, \dots, T$, repeat:
 - ▶ Draw θ^* from $q(\cdot | \theta^{(t-1)})$
 - ▶ Compute $r = \frac{h(\theta^*)q(\theta^{(t-1)} | \theta^*)}{h(\theta^{(t-1)})q(\theta^* | \theta^{(t-1)})}$.
 - ▶ If $r \geq 1$, set $\theta^{(t)} = \theta^*$;
if $r < 1$, set $\theta^{(t)} = \begin{cases} \theta^* & \text{with probability } r \\ \theta^{(t-1)} & \text{with probability } 1 - r \end{cases}$.

- ▶ Under mild conditions, $\theta^{(t)}$ converges in distribution to a draw from posterior as $t \rightarrow \infty$
 - ▶ See, e.g., <http://web1.sph.emory.edu/users/hwu30/teaching/statcomp/papers/chibGreenbergMH.pdf>
- ▶ The Metropolis-Hastings algorithm is identical to the Metropolis if q is symmetric
- ▶ In practice, a good initial value $\theta^{(0)}$ will have high posterior density
 - ▶ Could initialize by posterior mode, if possible: $\theta^{(0)} = \hat{\theta}$
 - ▶ Alternatively, could make a guess or generate $\theta^{(0)}$ from prior

Choice of proposal density

- ▶ A common choice for q is a normal distribution centered at previous draw:

$$q(\theta^* | \theta^{(t-1)}) = \text{Normal}(\theta^{(t-1)}, \sigma^2)$$

If θ is multivariate, replace σ^2 with Σ

- ▶ Higher σ^2 often leads to low acceptance ratio
 - ▶ Proposals θ^* may be far away from areas in which p concentrates (“big jumps”)
- ▶ Lower σ^2 often leads to high acceptance ratio
 - ▶ Proposals θ^* are close to $\theta^{(t-1)}$. Many iterations needed to cover larger areas of parameter space.
- ▶ Would like to compromise between these two extremes

Choice of proposal density

- ▶ As a rule of thumb, accepting about 20% – 70% of proposals is reasonable
- ▶ Can vary q to give the desired rejection rate
- ▶ Some algorithms adjust q adaptively during sampling
- ▶ Alternatively, for $q = \text{Normal}(\theta^{(t-1)}, \sigma^2)$, let σ^2 be an approximation to posterior variance.
 - ▶ Recall Bayesian CLT: $\text{Var}_{\theta | \mathbf{y}} \theta \approx (I_{\theta}^P(\mathbf{y}))^{-1}$

Other Considerations

- ▶ Beginning iterations are dependent on initial value
 - ▶ Especially if initial value is far from concentration of posterior.
- ▶ Typical to ignore M beginning iterations as *burn in*
 - ▶ Burn in can vary: $M = 1,000$, $M = 5,000$ or even $M = 100,000$ iterations
 - ▶ May adjust proposal distribution during burn-in
- ▶ Aim for *stationarity* after burn-in:
The probability distribution of θ_t does not depend on t
 - ▶ Initial iterations not stationary because of dependence on $\theta^{(0)}$
 - ▶ Eventually iterations will be approximately stationary.
 - ▶ The stationary distribution is the posterior:

$$p(\theta^{(t)}) \approx p(\theta | \mathbf{y}) \text{ for } t > M$$

Other Considerations

- ▶ $\theta^{(t)}$ for $t \geq M$ are kept as draws from posterior
- ▶ To validate burn-in, can run from different initializations
 - ▶ See if they converge to similar distributions after burn-in
- ▶ In general, want low dependence between MCMC samples
 - ▶ Low *autocorrelation*: $\text{cor}(\theta^{(t)}, \theta^{(t-1)})$.
 - ▶ Leads to better convergence toward stationary posterior
 - ▶ Leads to lower uncertainty in results from posterior draws

Example: Basketball shooting (cont.)

- ▶ Consider the shooting percentage for a basketball team over n games: $\mathbf{y} = (y_1, \dots, y_n)$
- ▶ Model $y_i \stackrel{iid}{\sim} \text{Beta}(\theta, 2)$ for $\theta > 0$

$$p(y_i | \theta) = \theta(1 + \theta)y_i^{\theta-1}(1 - y_i)$$

- ▶ Use a Gamma(a, b) prior for θ
- ▶ Then,

$$p(\theta | \mathbf{y}) \propto \theta^{n+a-1}(\theta + 1)^n e^{-b\theta} \left(\prod_{i=1}^n y_i \right)^\theta$$
$$:= h(\theta)$$

Example: Basketball shooting (cont.)

- ▶ Observe $n = 20$ games with $\sum_{i=1}^{20} \log y_i = -9.89$
- ▶ Prior $a = b = 1$
- ▶ Previously approximated posterior using Bayesian CLT:

$$p(\theta | \mathbf{y}) \approx \text{Normal}(3.24, 0.33)$$

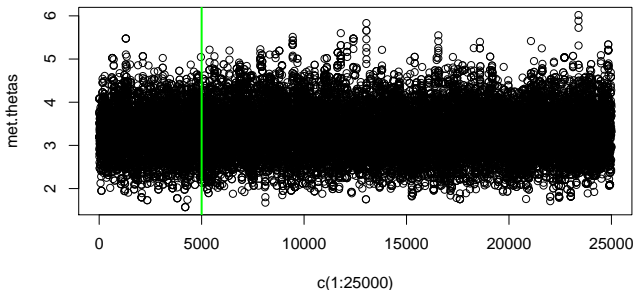
- ▶ Now, use Metropolis sampling to draw from $p(\theta | \mathbf{y})$.
 - ▶ Use asymptotic approximation to motivate $\theta^{(0)}$ and q

Example: Basketball shooting (cont.)

- ▶ Apply Metropolis algorithm, with
 - ▶ Initial value $\theta^{(0)} = 3.24$
 - ▶ Proposal density $p(\cdot | \theta^{(t-1)}) = \text{Normal}(\theta^{(t-1)}, 0.33)$
 - ▶ Unnormalized posterior $h(\theta)$
- ▶ Run for $T = 25,000$ iterations
- ▶ Treat the first $M = 5,000$ iterations as burn-in
- ▶ Remaining $N = 20,000$ as draws from $p(\theta | \mathbf{y})$
http://www.ericfrazierlock.com/Metropolis-Hastings_Sampling_Rcode1.r

Example: Basketball shooting (cont.)

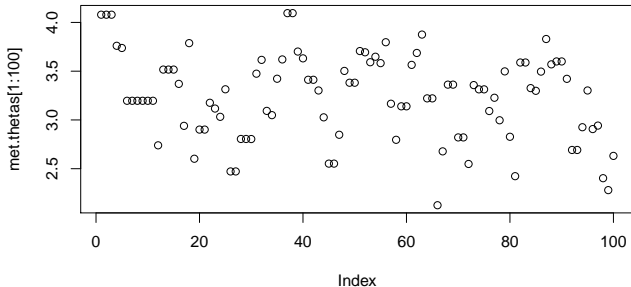
- Simulated iterations $\theta^{(1)}, \theta^{(2)}, \dots$:



- Proposal acceptance rate = 70%
- Autocorrelation of draws $r = 0.778$

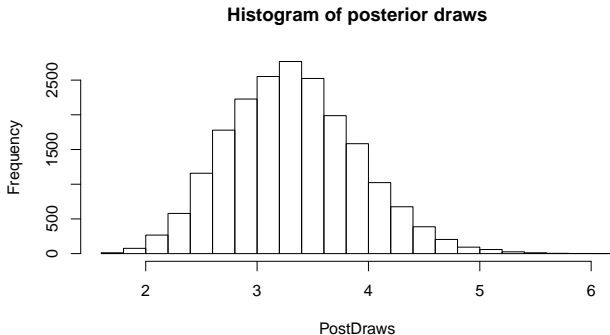
Example: Basketball shooting (cont.)

- First 100 iterations:



Example: Basketball shooting (cont.)

- Estimated posterior density:



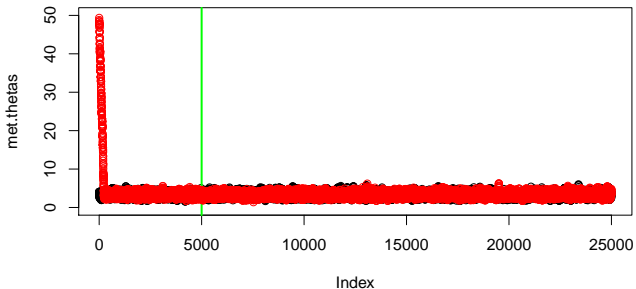
Example: Basketball shooting (cont.)

- ▶ Repeat algorithm with different initializations and proposal densities
 - ▶ $\theta^{(0)} = 50$
 - ▶ $p(\cdot | \theta^{(t-1)}) = \text{Normal}(\theta^{(t-1)}, 0.01)$
 - ▶ $p(\cdot | \theta^{(t-1)}) = \text{Normal}(\theta^{(t-1)}, 50)$

- ▶ Explore effect on Markov chain, sensitivity of results
http://www.ericfrazerlock.com/Metropolis-Hastings_Sampling_Rcode1.r

Example: Basketball shooting (cont.)

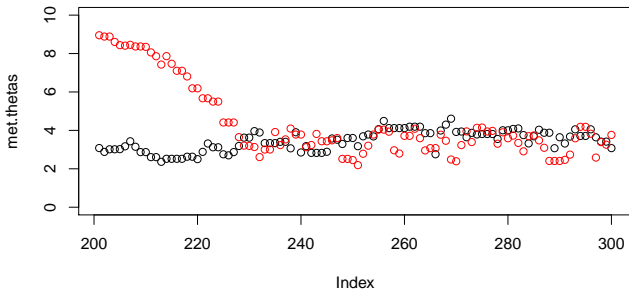
- Simulated iterations with $\theta^{(0)} = 50$ (red)



- Draws are indistinguishable after burn-in

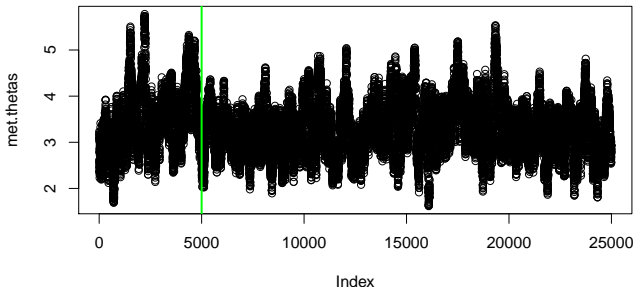
Example: Basketball shooting (cont.)

- Iterations 200-300:



Example: Basketball shooting (cont.)

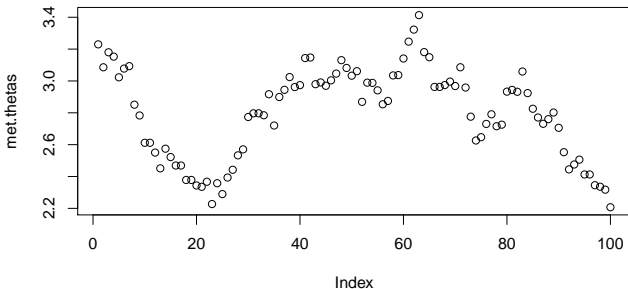
- Simulated iterations with $p(\cdot | \theta^{(t-1)}) = \text{Normal}(\theta^{(t-1)}, 0.01)$



- Proposal acceptance rate = 94%
- Autocorrelation of draws $r = 0.987$

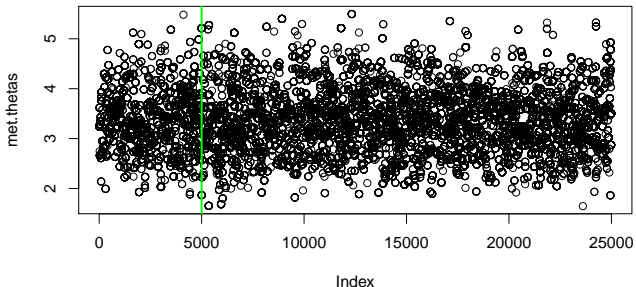
Example: Basketball shooting (cont.)

- First 100 draws:



Example: Basketball shooting (cont.)

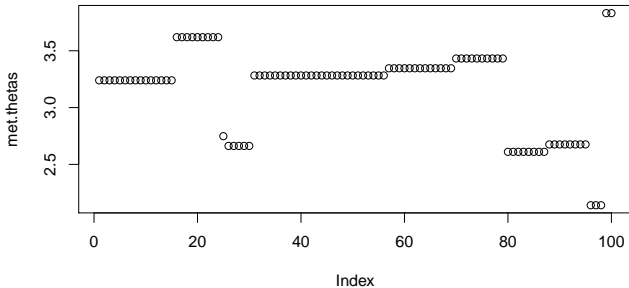
- Simulated iterations with $p(\cdot | \theta^{(t-1)}) = \text{Normal}(\theta^{(t-1)}, 50)$



- Proposal acceptance rate = 10%
- Autocorrelation of draws $r = 0.870$

Example: Basketball shooting (cont.)

- First 100 draws:



Example: Basketball shooting (cont.)

- Comparison of posterior density estimates:

