Markov Chain Monte Carlo

David A. van Dyk

Statistics Section, Imperial College London

Smithsonian Astrophysical Observatory, March 2014
Outline

1. Background
   - Bayesian Statistics
   - Monte Carlo Integration
   - Markov Chains

2. Basic MCMC Jumping Rules
   - Metropolis Sampler
   - Metropolis Hastings Sampler

3. Practical Challenges and Advice
   - Diagnosing Convergence
   - Choosing a Jumping Rule
   - Transformations and Multiple Modes

4. Overview of Recommended Strategy
Outline

1. **Background**
   - Bayesian Statistics
   - Monte Carlo Integration
   - Markov Chains

2. **Basic MCMC Jumping Rules**
   - Metropolis Sampler
   - Metropolis Hastings Sampler

3. **Practical Challenges and Advice**
   - Diagnosing Convergence
   - Choosing a Jumping Rule
   - Transformations and Multiple Modes

4. **Overview of Recommended Strategy**
Bayesian Statistical Analyses: Likelihood

**Likelihood Functions:** The distribution of the data given the model parameters. E.g., \( Y \overset{\text{dist}}{\sim} \text{Poisson}(\lambda_S) \):

\[
\text{likelihood}(\lambda_S) = e^{-\lambda_S} \frac{\lambda_S^Y}{Y!}
\]

**Maximum Likelihood Estimation:** Suppose \( Y = 3 \)

The likelihood and its normal approximation.

Can estimate \( \lambda_S \) and its error bars.
Bayesian Analyses: Prior and Posterior Dist’ns

Prior Distribution: Knowledge obtained prior to current data.

Bayes Theorem and Posterior Distribution:

\[ \text{posterior}(\lambda) \propto \text{likelihood}(\lambda)\text{prior}(\lambda) \]

Combine past and current information:

Bayesian analyses rely on probability theory
Why be Bayesian?

- Avoid Gaussian assumptions
  - Methods like $\chi^2$ fitting implicitly assume a Gaussian model.
  - Many other methods rely on asymptotic Gaussian properties (e.g., stemming from central limit theorem).

- Bayesian methods rely directly on probability calculus.

- Designed to combine multiple sources of information and/or external sources of information.

- Modern computational methods allow us to work with specially-tailored models and methods.
  - Selection effects, contaminated data, observational biases, complex physics-based models, data distortion, calibration uncertainty, measurement errors, etc.
Simulating from the Posterior Distribution

- We can *simulate* or *sample* from a distribution to learn about its contours.
- With the sample alone, we can learn about the posterior.
- Here, $Y \overset{\text{dist}}{\sim} \text{Poisson}(\lambda_S + \lambda_B)$ and $Y_B \overset{\text{dist}}{\sim} \text{Poisson}(c\lambda_B)$.
Highly non-linear relationship among stellar parameters.
Model Fitting: Complex Posterior Distributions

Highly non-linear relationships among stellar parameters.
The classification of certain stars as field or cluster stars can cause multiple modes in the distributions of other parameters.
Complex Posterior Distributions

Standard Algorithm
one degree of freedom

Marginal Augmentation
one degree of freedom

2 log(σ)

μ/σ

log(σ)

log(q₁)

2 log(σ)

log(q₁)

log(q₁q₂)

2 log(σ)

log(q₁)

log(q₁q₂)
Complex Posterior Distributions

![Graph of posterior density vs. energy (keV)](image)

- Posterior density vs. energy (keV)
- Energy range: 0 to 10 keV
- Posterior density range: 0 to 0.5

David A. van Dyk
MCMC
Complex Posterior Distributions

- Energy of photon 1 (keV)
- Energy of photon 2 (keV)
Using Simulation to Evaluate Integrals

Suppose we want to compute

\[ I = \int g(\theta) f(\theta) d\theta, \]

where \( f(\theta) \) is a probability density function. If we have a sample

\[ \theta^{(1)}, \ldots, \theta^{(n)} \overset{\text{dist}}{\sim} f(\theta), \]

we can estimate \( I \) with

\[ \hat{I}_n = \frac{1}{n} \sum_{i=1}^{n} g(\theta^{(t)}). \]

In this way we can compute means, variances, and the probabilities of intervals.
We Need to Obtain a Sample

Our primary goal:

Develop methods to obtain a sample from a distribution

- The sample may be independent or dependent.
- Markov chains can be used to obtain a dependent sample.
- In a Bayesian context, we typically aim to sample the posterior distribution.

We first discuss an independent method:

Rejection Sampling
Suppose we cannot sample \( f(\theta) \) directly, but can find \( g(\theta) \) with

\[
f(\theta) \leq Mg(\theta)
\]

for some \( M \).

1. Sample \( \tilde{\theta} \overset{\text{dist}}{\sim} g(\theta) \).
2. Sample \( u \overset{\text{dist}}{\sim} \text{Unif}(0, 1) \).
3. If

\[
u \leq \frac{f(\tilde{\theta})}{Mg(\tilde{\theta})}, \text{ i.e., if } uMg(\tilde{\theta}) \leq f(\tilde{\theta})
\]

accept \( \tilde{\theta} \): \( \theta^{(t)} = \tilde{\theta} \).

Otherwise reject \( \tilde{\theta} \) and return to step 1.

How do we compute \( M \)?
Consider the distribution:

We must bound $f(\theta)$ with some unnormalized density, $Mg(\theta)$. 
Imagine that we sample uniformly in the red rectangle:

\[ \theta \sim g(\theta) \text{ and } y = uMg(\theta) \]

Accept samples that fall below the dashed density function.

*How can we reduce the wait for acceptance??*
How can we reduce the wait for acceptance??

**Improve \( g(\theta) \) as an approximation to \( f(\theta) \)!**
What is a Markov Chain

A Markov chain is a sequence of random variables,

$$\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots$$

such that

$$p(\theta^{(t)}|\theta^{(t-1)}, \theta^{(t-2)}, \ldots, \theta^{(0)}) = p(\theta^{(t)}|\theta^{(t-1)}).$$

A Markov chain is generally constructed via

$$\theta^{(t)} = \varphi(\theta^{(t-1)}, U^{(t-1)})$$

with $U^{(1)}, U^{(2)}, \ldots$ independent.
What is a Stationary Distribution?

A stationary distribution is any distribution $f(x)$ such that

$$f(\theta^{(t)}) = \int p(\theta^{(t)}|\theta^{(t-1)})f(\theta^{(t-1)})d\theta^{(t-1)}$$

If we have a sample from the stationary dist’n and update the Markov chain, the next iterate also follows the stationary dist’n.

What does a Markov Chain at Stationarity Deliver?

Under regularity conditions, the density at iteration $t$,

$$f^{(t)}(\theta|\theta^{(0)}) \rightarrow f(\theta) \quad \text{and} \quad \frac{1}{n} \sum_{t=1}^{n} h(\theta^{(t)}) \rightarrow E_f[h(\theta)]$$

We can treat $\{\theta^{(t)}, t = N_0, \ldots N\}$ as an approximate correlated sample from the stationary distribution.

**GOAL:** Markov Chain with Stationary Dist’n = Target Dist’n.
Outline

1. Background
   - Bayesian Statistics
   - Monte Carlo Integration
   - Markov Chains

2. Basic MCMC Jumping Rules
   - Metropolis Sampler
   - Metropolis Hastings Sampler

3. Practical Challenges and Advice
   - Diagnosing Convergence
   - Choosing a Jumping Rule
   - Transformations and Multiple Modes

4. Overview of Recommended Strategy
The Metropolis Sampler

Draw $\theta^{(0)}$ from some starting distribution.

For $t = 1, 2, 3, \ldots$

Sample: $\theta^*$ from $J_t(\theta^*|\theta^{(t-1)})$

Compute: $r = \frac{p(\theta^*|y)}{p(\theta^{(t-1)}|y)}$

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

Note
- $J_t$ must be symmetric: $J_t(\theta^*|\theta^{(t-1)}) = J_t(\theta^{(t-1)}|\theta^*)$.
- If $p(\theta^*|y) > p(\theta^{(t-1)}|y)$, jump!
The Random Walk Jumping Rule

Typical choices of $J_t(\theta^*|\theta(t-1))$ include
- Unif $(\theta(t-1) - k, \theta(t-1) + k)$
- Normal $(\theta(t-1), kl)$
- $t_{df}(\theta(t-1), kl)$

$J_t$ may change, but may not depend on the history of the chain.

How should we choose $k$? Replace $I$ with $M$? How?
An Example

A simplified model for high-energy spectral analysis.

**Model:**

Consider a perfect detector:

1. 1000 energy bins, equally spaced from 0.3keV to 7.0keV,
2. \( Y_i \sim \text{Poisson} \left( \alpha E_i^{-\beta} \right) \), with \( \theta = (\alpha, \beta) \),
3. \( E_i \) is the energy, and
4. \( (\alpha, \beta) \sim \text{Unif}(0, 100) \).

**The Sampler:**

We use a Gaussian Jumping Rule,

- centered at the current sample, \( \theta(t) \)
- with standard deviations equal 0.08 and correlation zero.
Simulated Data

2288 counts were simulated with $\alpha = 5.0$ and $\beta = 1.69$. 

![Graph showing simulated data with red curve indicating expected counts]
Chains “stick” at a particular draw when proposals are rejected.
The Joint Posterior Distribution

Scatter Plot of Posterior Distribution

α
β

David A. van Dyk  MCMC
\[ E(\alpha|Y) \approx 5.13, \quad SD(\alpha|Y) \approx 0.11, \quad \text{and a 95\% CI is } (4.92, 5.41) \]
Marginal Posterior Dist’n of Power Law Param

\[ \mathbb{E}(\beta|Y) \approx 1.71, \quad \text{SD}(\beta|Y) \approx 0.03, \quad \text{and a 95\% CI is (1.65, 1.76)} \]
The Metropolis-Hastings Sampler

A more general Jumping rule:

Draw $\theta^{(0)}$ from some starting distribution.

For $t = 1, 2, 3, \ldots$

Sample: $\theta^*$ from $J_t(\theta^* | \theta^{(t-1)})$

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

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

Note

- $J_t$ may be any jumping rule, it needn’t be symmetric.
- The updated $r$ corrects for bias in the jumping rule.
The Independence Sampler

Use an approximation to the posterior as the jumping rule:

\[ J_t = \text{Normal}_d(\text{MAP estimate, Curvature-based Variance Matrix}). \]

MAP estimate = argmax_\theta p(\theta | y)

Variance \approx \left[ - \frac{\partial^2}{\partial \theta \cdot \partial \theta} \log p(\theta | Y) \right]^{-1}

Note: \( J_t(\theta^*|\theta^{(t-1)}) \) does not depend on \( \theta^{(t-1)} \).
The Independence Sampler

The Normal Approximation may not be adequate.

- We can inflate the variance.
- We can use a heavy tailed distribution, e.g., lorentzian or $t$. 

\[\theta \quad f(\theta)\]

\[0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7\]

\[0.0 \quad 0.1 \quad 0.2 \quad 0.3 \quad 0.4\]

\[0.00 \quad 0.10 \quad 0.20 \quad 0.30 \quad 0.40\]

\[\theta \quad f(\theta)\]

\[0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7\]

\[0.0 \quad 0.1 \quad 0.2 \quad 0.3 \quad 0.4\]
A simplified model for high-energy spectral analysis.

- We can fit \( (\alpha, \beta) \) with a general mode finder (e.g., Levenberg-Marquardt).
- Requires coding likelihood (e.g. Cash statistic), specifying starting values, etc.
- Base choice of parameter on quality of normal approx.
  - MLE is invariant to transformations.
  - Variance matrix of transform is computed via \textit{delta method}.
- Can use the jumping rule:
  \[ J_t = \text{Normal}_2(\text{MAP est}, \text{Curvature-based Variance Matrix}). \]
Very little “sticking” here: acceptance rate is 98.8%.
Autocorrelation is essentially zero: nearly independent sample!!
This result depends critically on access to a very good approximation to the posterior distribution.
Outline

1. Background
   - Bayesian Statistics
   - Monte Carlo Integration
   - Markov Chains

2. Basic MCMC Jumping Rules
   - Metropolis Sampler
   - Metropolis Hastings Sampler

3. Practical Challenges and Advice
   - Diagnosing Convergence
   - Choosing a Jumping Rule
   - Transformations and Multiple Modes

4. Overview of Recommended Strategy
Has this Chain Converged?


David A. van Dyk
MCMC
Has this Chain Converged?

Comparing multiple chains can be informative!

Using Multiple Chains

- Compare results of multiple chains to check convergence.
- Start the chains from distant points in parameter space.
- Run until they appear to give similar results
  - ... or they find different solutions (multiple modes).
Consider $M$ chains of length $N$: \{\psi_{nm}, n = 1, \ldots, N\}.

\[
B = \frac{N}{M - 1} \sum_{m=1}^{M} (\bar{\psi}_m - \bar{\psi})^2
\]

\[
W = \frac{1}{M} \sum_{m=1}^{M} s_m^2 \quad \text{where} \quad s_m^2 = \frac{1}{N - 1} \sum_{n=1}^{N} (\psi_{nm} - \bar{\psi}_m)^2
\]

Two estimates of $\text{Var}(\psi | Y)$:

1. $W$: underestimate of $\text{Var}(\psi | Y)$ for any finite $N$.
2. $\hat{\text{var}}^+ (\psi | Y) = \frac{N-1}{N} W + \frac{1}{N} B$: overestimate of $\text{Var}(\psi | Y)$.

\[
\hat{R} = \sqrt{\frac{\hat{\text{var}}^+ (\psi | Y)}{W}} \downarrow 1 \quad \text{as the chains converge.}
\]
Spectral Analysis: effect on burn in of power law parameter

- **Sample 1** with **\( \sigma = 0.005 \)**:
  - Acceptance rate: 87.5%
  - Lag one autocorrelation: 0.98

- **Sample 2** with **\( \sigma = 0.08 \)**:
  - Acceptance rate: 31.6%
  - Lag one autocorrelation: 0.66

- **Sample 3** with **\( \sigma = 0.4 \)**:
  - Acceptance rate: 3.1%
  - Lag one autocorrelation: 0.96

---

**Choice of Jumping Rule with Random Walk Metropolis**

David A. van Dyk

MCMC
Higher Acceptance Rate is not Always Better!

Aim for 20% (vectors) - 40% (scalars) acceptance rate
Statistical Inference and Effective Sample Size

- **Point Estimate:** \( \bar{h}_n = \frac{1}{n} \sum h(\theta(t)) \) (estimate of \( E(h(\theta) | x) \))

- **Variance Estimate:** \( \text{Var}(\bar{h}_n) \approx \frac{\sigma^2}{n} \frac{1+\rho}{1-\rho} \) with (not \( \text{var}(\theta) \))

  \[
  \sigma^2 = \text{Var}(h(\theta)) \text{ estimated by } \hat{\sigma}^2 = \frac{1}{n-1} \sum_{t=1}^{n} [h(\theta(t)) - \bar{h}_n]^2,
  \]

  \[
  \rho = \text{corr} \left[ h(\theta(t)), h(\theta(t-1)) \right] \text{ estimated by }
  \]

  \[
  \hat{\rho} = \frac{1}{n-1} \frac{\sum_{t=2}^{n} [h(\theta(t)) - \bar{h}_n][h(\theta(t-1)) - \bar{h}_n]}{\sqrt{\sum_{t=1}^{n-1} [h(\theta(t)) - \bar{h}_n]^2 \sum_{t=2}^{n} [h(\theta(t)) - \bar{h}_n]^2}}
  \]

- **Interval Estimate:** \( \bar{h}_n \pm t_d \sqrt{\text{Var}(\bar{h}_n)} \) with \( d = n \frac{1-\rho}{1+\rho} - 1 \)

The effective sample size is \( n \frac{1-\rho}{1+\rho} \).
Sample from $N(0, 1)$ with random walk Metropolis with $J_t = N(\theta(t), \sigma)$.

What is the Effective Sample Size here? and $\sigma$?
Illustration of the Effective Sample Size

What is the Effective Sample Size here? and $\sigma$?
Illustration of the Effective Sample Size

What is the Effective Sample Size here? and σ?
Illustration of the Effective Sample Size

What is the Effective Sample Size here? and $\sigma$?
Lag One Autocorrelation

Small Jumps versus Low Acceptance Rates

![Graph showing lag 1 autocorrelation against log(sigma)]
Effective Sample Size

Balancing the Trade-Off

![Graph showing effective sample size vs. log(sigma)]
Acceptance Rate

Bigger is not always Better!!

High acceptance rates only come with small steps!!
Finding the Optimal Acceptance Rate

- Effective sample size vs. log(sigma)
- Acceptance rate vs. log(sigma)
A whole new set of issues arise in higher dimensions...

Tradeoff between high autocorrelation and high rejection rate:
- more acute with high posterior correlations
- more acute with high dimensional parameter
Random Walk Metropolis with High Correlation

In principle we can use a correlated jumping rule, but

- the desired correlation may vary, and
- is often difficult to compute in advance.
Random Walk Metropolis with High Correlation

What random walk jumping rule would you use here?

Remember: you don’t get to see the distribution in advance!
Random Walk Metropolis for Spectral Analysis:

Scatter Plot of Posterior Distribution

Autocorrelation for alpha

Why is the Mixing SO Poor?!??
Consider the Scales of $\alpha$ and $\beta$:

A new jumping rule: std dev for $\alpha = 0.110$, for $\beta = 0.026$, and corr $= -0.216$. 
Improved Convergence

Original Jumping Rule:

- Autocorrelation for alpha
- Hist of 500 Draws excluding Burn-in
- Posterior Density
Improved Convergence

Improved Jumping Rule:

Autocorrelation for alpha

Hist of 500 Draws excluding Burn-in

Original Eff Sample Size = 19, Improved Eff Sample Size = 75, with $n = 500$. 
Strategy: When using

- Normal \((\theta^{(t-1)}, kM)\) or better yet
- \(t_{df}(\theta^{(t-1)}, kM)\)

try using the variance-covariance matrix from a standard fitted model for \(M\)

... at least when there is standard mode-based model-fitting software available.
Parameter transformations can greatly improve MCMC.

Recall the Independence Sampler:

The normal approximation is not as good as we might hope...
But if we use the square root of $\theta$: 

![Graphs showing transformation to normality](image-url)
Transforming to Normality

And...

The normal approximation is much improved!
Transforming to Normality

Working with Gaussian or symmetric distributions leads to more efficient Metropolis and Metropolis Hastings Samplers.

General Strategy:

- Transform to the Real Line.
- Take the log of positive parameters.
- If the log is “too strong”, try square root.
- Probabilities can be transformed via the logit transform:

\[
\log\left(\frac{p}{1 - p}\right)
\]

- More complex transformations for other quantities.
- Try out various transformations using an initial MCMC run.
- Statistical advantages to using normalizing transforms.
Removing Linear Correlations

Linear transformations can remove linear correlations
Removing Linear Correlations

... and can help with non-linear correlations.
Multiple Modes

- Scientific meaning of multiple modes.
- Do not focus only on the major mode!
- “Important” modes.
- Challenging for Bayesian and Frequentist methods.
- Consider Metropolis & Metropolis Hastings.
- Value of excess dispersion.
Multiple Modes

1. Use a mode finder to “map out” the posterior distribution.
   1. Design a jumping rule that accounts for all of the modes.
   2. Run separate chains for each mode.
2. Use one of several sophisticated methods tailored for multiple modes.
   2. Parallel Tempering.
   3. Many other specialized methods.
Outline

1. Background
   - Bayesian Statistics
   - Monte Carlo Integration
   - Markov Chains

2. Basic MCMC Jumping Rules
   - Metropolis Sampler
   - Metropolis Hastings Sampler

3. Practical Challenges and Advice
   - Diagnosing Convergence
   - Choosing a Jumping Rule
   - Transformations and Multiple Modes

4. Overview of Recommended Strategy
Overview of Recommended Strategy

(Adopted from *Bayesian Data Analysis*, Section 11.10, Gelman et al. (2005), Second Edition)

1. Start with a crude approximation to the posterior distribution, perhaps using a mode finder.
2. Simulate directly, avoiding MCMC, if possible.
3. If necessary use MCMC with one parameter at a time updating or updating parameters in batches:

   **Two-Step Gibbs Sampler:**
   
   Step 1: Sample $\theta^{(t)} \overset{\text{dist}}{\sim} p(\theta \mid \phi^{(t-1)}, Y)$
   
   Step 2: Sample $\phi^{(t)} \overset{\text{dist}}{\sim} p(\phi \mid \theta^{(t)}, Y)$

4. Use Gibbs draws for closed form complete conditionals.
Overview of Recommended Strategy- Con’t

5 Use metropolis jumps if complete conditional is not in closed form. Tune variance of jumping distribution so that acceptance rates are near 20% (for vector updates) or 40% (for single parameter updates).

6 To improve convergence, use transformations so that parameters are approximately independent.

7 Check for convergence using multiple chains.

8 Compare inference based on crude approximation and MCMC. If they are not similar, check for errors before believing the results of the MCMC.