11: Basics of Markov chain Monte Carlo Simulation

Taylor

University of Virginia

Introduction

In this section, we discuss **Markov chain Monte Carlo** techniques, which all produce correlated draws from the posterior of interest.

Overview

Even though we might not even have a time series model, we construct a Markov chain

$$\theta^1, \theta^2, \ldots, \theta^N$$

Even though these draws are correlated, they are all **marginally** distributed according to the posterior

$$p(\theta \mid y).$$

We estimate expectations with sample means:

$$E[h(\theta) \mid y] \approx \frac{1}{N} \sum_{i=1}^{N} h(\theta^i).$$

Typical MCMC Output

A 2-parameter model:

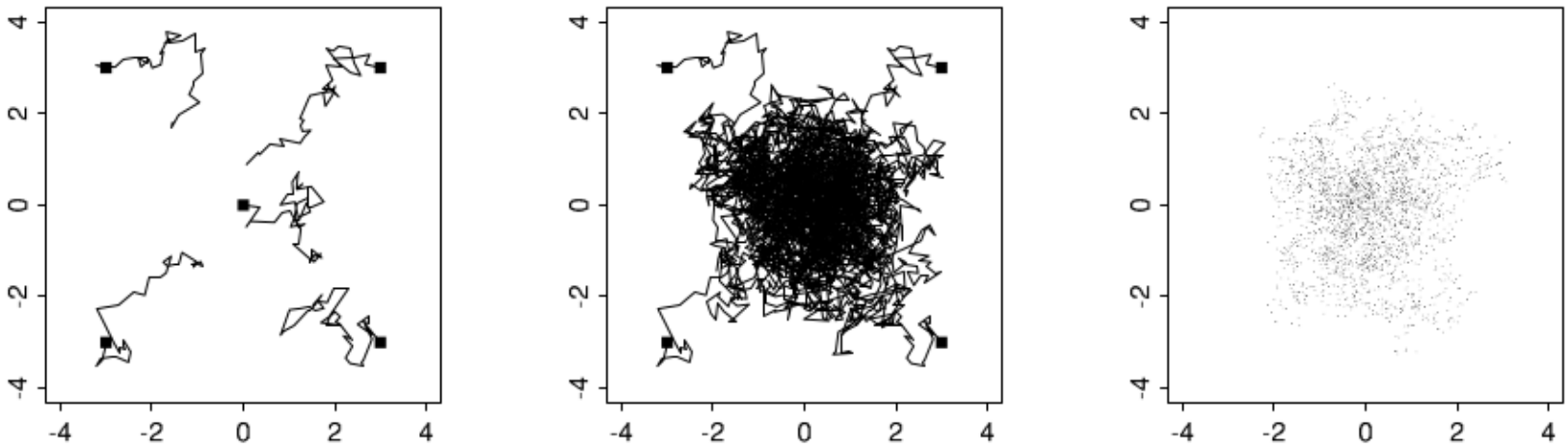


Figure 11.1 *Five independent sequences of a Markov chain simulation for the bivariate unit normal distribution, with overdispersed starting points indicated by solid squares. (a) After 50 iterations, the sequences are still far from convergence. (b) After 1000 iterations, the sequences are nearer to convergence. Figure (c) shows the iterates from the second halves of the sequences; these represent a set of (correlated) draws from the target distribution. The points in Figure (c) have been jittered so that steps in which the random walks stood still are not hidden. The simulation is a Metropolis algorithm described in the example on page 278, with a jumping rule that has purposely been chosen to be inefficient so that the chains will move slowly and their random-walk-like aspect will be apparent.*

Typical MCMC Output

Another 2-parameter model:

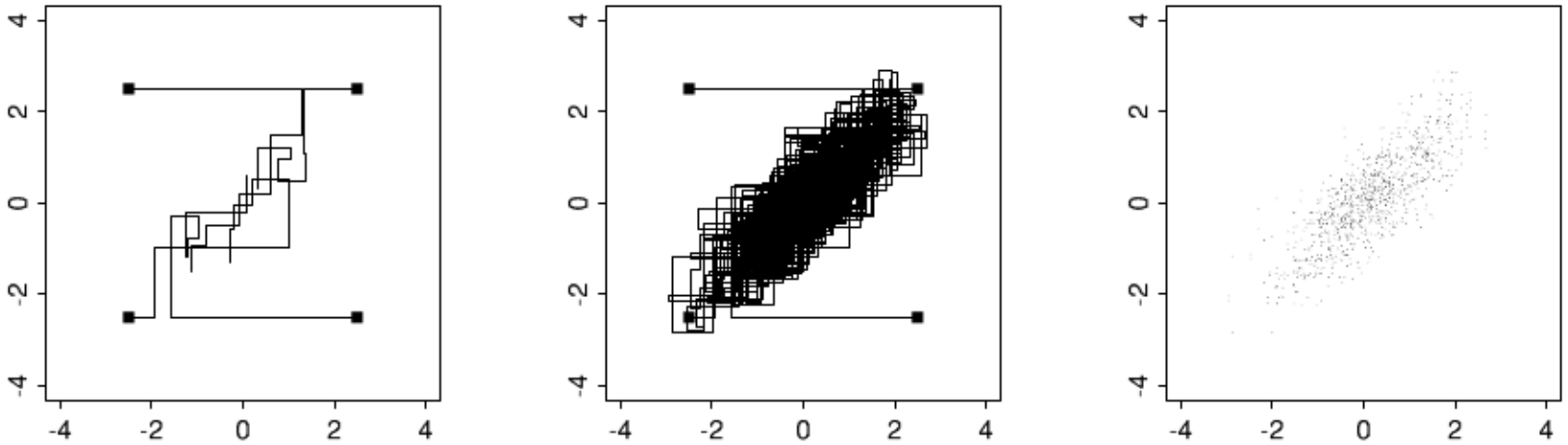
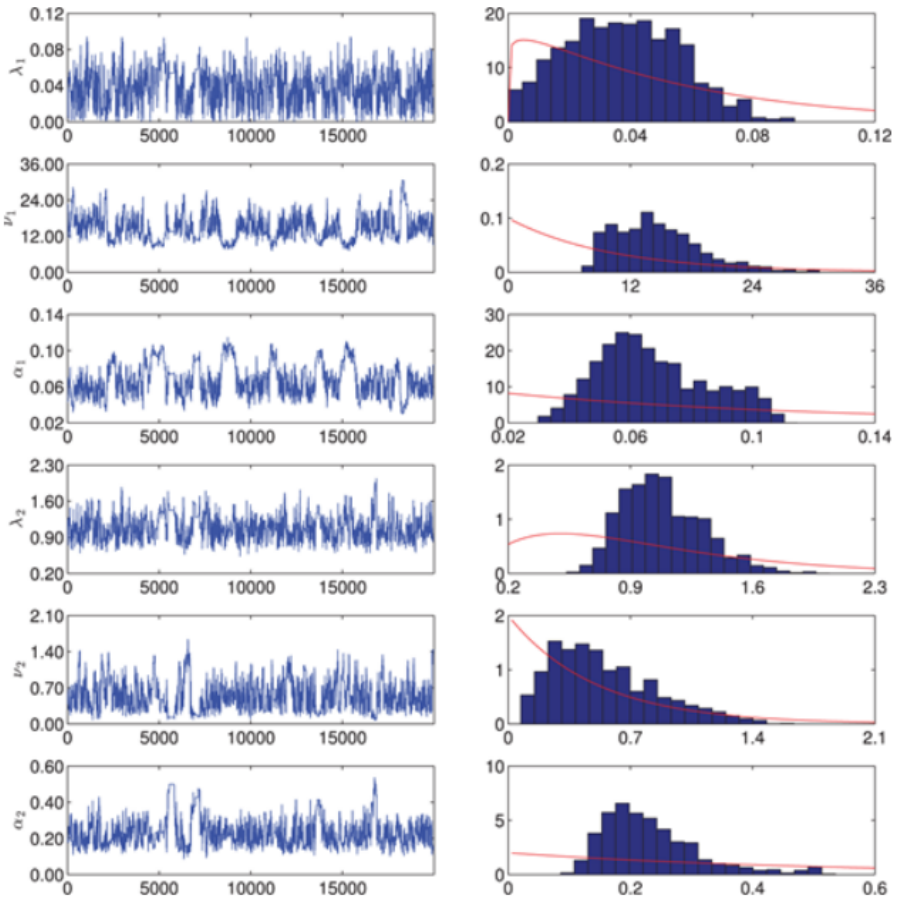


Figure 11.2 *Four independent sequences of the Gibbs sampler for a bivariate normal distribution with correlation $\rho = 0.8$, with overdispersed starting points indicated by solid squares. (a) First 10 iterations, showing the componentwise updating of the Gibbs iterations. (b) After 500 iterations, the sequences have reached approximate convergence. Figure (c) shows the points from the second halves of the sequences, representing a set of correlated draws from the target distribution.*

Typical MCMC Output

A 6-parameter model from:
`https://academic.oup.com/jfec/article/14/2/278/1751519`

Figure 2.



View large Download slide

MCMC output for the two-factor model. Left panels are trace plots of parameters. In the right

Variance of our Estimator

Let $\sigma^2 = \text{Var}[h(\theta^i)]$, $\rho(h) = \text{Corr}\left(h(\theta^i), h(\theta^{h+i})\right)$

$$N \,\text{Var}\left[\frac{1}{N}\sum_{i=1}^{N} h(\theta^i)\right] = N \,\text{Cov}\left(\frac{1}{N}\sum_{i=1}^{N} h(\theta^i), \frac{1}{N}\sum_{j=1}^{N} h(\theta^j)\right)$$

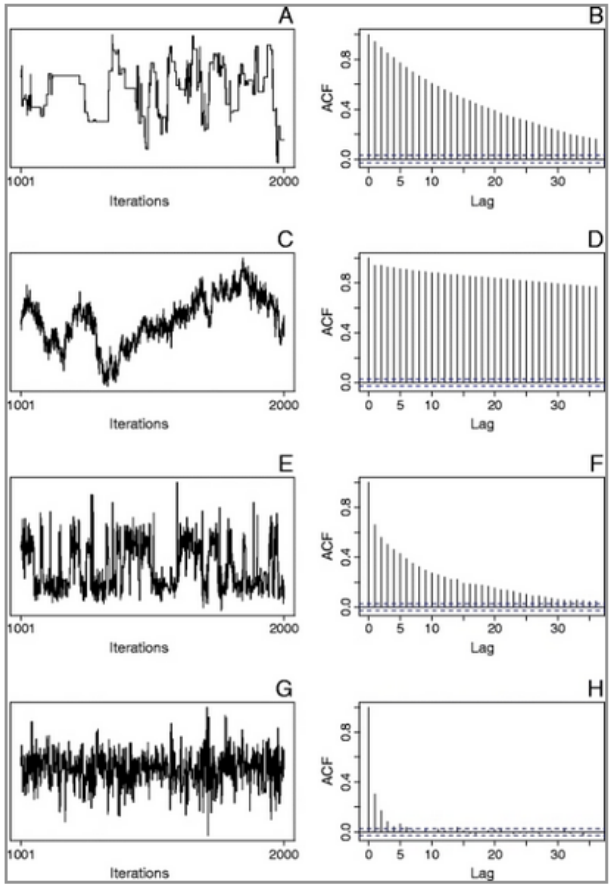
$$= \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} \text{Cov}\left(h(\theta^i), h(\theta^i)\right) \qquad \text{(bilinearity)}$$

$$= \sigma^2\left\{1 + 2\sum_{h=1}^{N}\frac{N-h}{N}\rho(h)\right\} \qquad \text{(count diagonally)}$$

$$\to \sigma^2 \underbrace{\left\{1 + 2\sum_{h=1}^{\infty}\rho(h)\right\}}_{\text{correlation is bad}}$$

Typical MCMC Output

Assessing the integrated autocorrelation with acf plots:



bad, bad, less bad, good.

https://openi.nlm.nih.gov/detailedresult?img=PMC3218285_13428_2011_114_Fig10_HTML&req=4

Another issue

The previous problem assumes each draw is distributed according to the posterior.

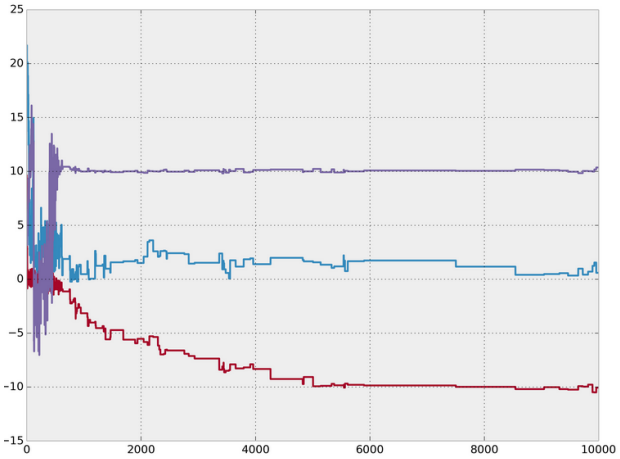
If we start the chain far away from the mode, how long until it converges? How can we be sure it has converged?

Another issue

The previous problem assumes each draw is distributed according to the posterior.

If we start the chain far away from the mode, how long until it converges? How can we be sure it has converged?

Trace plots help. We also have convergence diagnostics (more on this later).



You could throw away 6000 iterations as a **burn in** or **warm-up**.

Metropolis-Hastings algorithm

At iteration $t-1$ you have θ^{t-1}. Propose $\theta^* \sim q(\theta \mid \theta^{t-1})$, and accept this draw with probability

$$a(\theta^{t-1}, \theta^*) = \min\left\{1, \frac{p(\theta^* \mid y)q(\theta^{t-1} \mid \theta^*)}{p(\theta^{t-1} \mid y)q(\theta^* \mid \theta^{t-1})}\right\}.$$

If you accept, $\theta^t = \theta^*$. Otherwise, $\theta^t = \theta^{t-1}$.

Many algorithms are a special case of this one.

Metropolis-Hastings algorithm

Why it's widely-applicable:

$$
\begin{aligned}
a(\theta^{t-1}, \theta^*) &= \min\left\{1, \frac{p(\theta^* \mid y)q(\theta^{t-1} \mid \theta^*)}{p(\theta^{t-1} \mid y)q(\theta^* \mid \theta^{t-1})}\right\} \\
&= \min\left\{1, \frac{p(\theta^*, y)/p(y)q(\theta^{t-1} \mid \theta^*)}{p(\theta^{t-1}, y)/p(y)q(\theta^* \mid \theta^{t-1})}\right\} \\
&= \min\left\{1, \frac{p(y \mid \theta^*)p(\theta^*)q(\theta^{t-1} \mid \theta^*)}{p(y \mid \theta^{t-1})p(\theta^{t-1})q(\theta^* \mid \theta^{t-1})}\right\}.
\end{aligned}
$$

Don't need to know the normalizing constant/marginal likelihood/evidence!

Metropolis-Hastings algorithm

$$a(\theta^{t-1}, \theta^*) = \min\left\{1, \frac{p(\theta^* \mid y)q(\theta^{t-1} \mid \theta^*)}{p(\theta^{t-1} \mid y)q(\theta^* \mid \theta^{t-1})}\right\}.$$

1. When is $a(\theta^{t-1}, \theta^*)$ big?
2. When is $a(\theta^{t-1}, \theta^*)$ small?
3. How should we pick $q(\theta^* \mid \theta^{t-1})$?
4. What is the ideal $q(\theta^* \mid \theta^{t-1})$?
5. What if $q(\theta^* \mid \theta^{t-1})$ is too peaked?
6. What if $q(\theta^* \mid \theta^{t-1})$ is too diffuse?

Metropolis-Hastings algorithm

Let's pick $q(\theta \mid \theta^{t-1}) = \text{Normal}(\theta^{t-1}, \sigma^2 \mathbf{I})$

`https://chi-feng.github.io/mcmc-demo/app.html#RandomWalkMH,banana`

Independent Metropolis Hastings

If $q(\theta \mid \theta^{t-1}) = q(\theta)$ (i.e. propose independently of past values), then you have the **Independent Metropolis Hastings** algorithm, which has acceptance probabilities

$$a(\theta^{t-1}, \theta^*) = \min \left\{ 1, \frac{p(\theta^* \mid y) q(\theta^{t-1})}{p(\theta^{t-1} \mid y) q(\theta^*)} \right\}.$$

The proposals are iid, but the chain is Markovian!

Metropolis algorithm

If $q(\theta^* \mid \theta^{t-1}) = q(\theta^{t-1} \mid \theta^*)$ (i.e. q is **symmetric**), the acceptance probability

$$a(\theta^{t-1}, \theta^*) = \min\left\{1, \frac{p(\theta^* \mid y)q(\theta^{t-1} \mid \theta^*)}{p(\theta^{t-1} \mid y)q(\theta^* \mid \theta^{t-1})}\right\}.$$

becomes

$$a(\theta^{t-1}, \theta^*) = \min\left\{1, \frac{p(\theta^* \mid y)}{p(\theta^{t-1} \mid y)}\right\}.$$

and they call this the **Metropolis** algorithm (drop the "Hastings").

The Gibbs Sampler

Say there are two parameters: $\theta = (\theta_1, \theta_2)$. The **Gibbs sampler** alternates between

1. $\theta_1^t \sim p(\theta_1 \mid \theta_2^{t-1}, y)$
2. $\theta_2^t \sim p(\theta_2 \mid \theta_1^t, y)$

If there are more parameters: $\theta = (\theta_1, \theta_2, \ldots, \theta_d)$. The Gibbs sampler alternates between

1. $\theta_1^t \sim p(\theta_1 \mid \theta_{2:d}^{t-1}, y)$
2. $\theta_2^t \sim p(\theta_2 \mid \theta_1^t, \theta_{3:d}^{t-1}, y)$
3. $\vdots$
4. $\theta_d^t \sim p(\theta_d \mid \theta_{1:d-1}^t, y)$

This is only possible if you can sample from the **conditional posteriors** (i.e. need conditional conjugacy).

The Gibbs Sampler

Example on page 277:

1. $\theta_1 \mid \theta_2, y \sim \text{Normal}(y_1 + \rho(\theta_2 - y_2), 1 - \rho^2)$
2. $\theta_2 \mid \theta_1, y \sim \text{Normal}(y_2 + \rho(\theta_1 - y_1), 1 - \rho^2)$

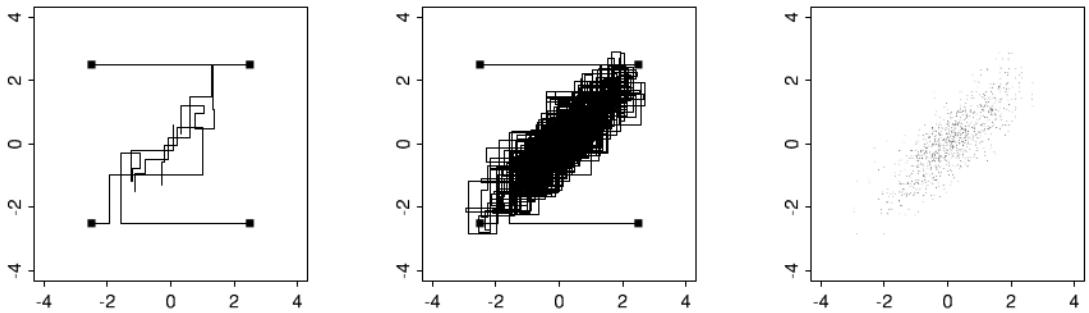


Figure 11.2 *Four independent sequences of the Gibbs sampler for a bivariate normal distribution with correlation $\rho = 0.8$, with overdispersed starting points indicated by solid squares. (a) First 10 iterations, showing the componentwise updating of the Gibbs iterations. (b) After 500 iterations, the sequences have reached approximate convergence. Figure (c) shows the points from the second halves of the sequences, representing a set of correlated draws from the target distribution.*

Back to Assessing convergence for scalars $\psi_{i,j}$

Run m chains for n iterations, $i = 1, \ldots, n$ and $j = 1 \ldots, m$.

$$\bar{\psi}_{\cdot\cdot} = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} \psi_{ij} \qquad \text{(overall average)}$$

$$\bar{\psi}_{\cdot j} = \frac{1}{n} \sum_{i=1}^{n} \psi_{ij} \qquad \text{(chain average)}$$

$$s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\psi_{ij} - \bar{\psi}_{\cdot j})^2 \qquad \text{(chain sd)}$$

$$W = \frac{1}{m} s_j^2 \qquad \text{(within-sequence variance)}$$

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\bar{\psi}_{\cdot j} - \bar{\psi}_{\cdot\cdot})^2$$

$$\hat{\text{var}}^{+}(\psi \mid y) = \frac{n-1}{n} W + \frac{1}{n} B.$$