Markov chain Monte Carlo

**Goal:** We have a target distribution $\pi(\theta)$ with unknown normalizing constant; $\pi(\theta) = K f(\theta)$ where $f$ is known but $K$ is not.

We wish to sample from $\pi$ in order to estimate

$$E[g(\theta)] = \int g(\theta) \pi(\theta) \, d\theta$$

**Why?**

1. We’re being Bayesians, and $\pi(\theta) = \pi(\theta|\text{data})$ is the posterior distribution of $\theta$.

2. We wish to explore the likelihood surface of $\theta$ (in the case that the posterior is proper when using a totally flat prior).

3. We wish to optimize the likelihood of $\theta$

**MCMC:**

1. Form a Markov chain (MC) with transition matrix $P(\theta^{(t+1)}|\theta^{(t)})$ having stationary distribution $\pi(\theta)$.

2. Generate $\theta^{(0)}, \theta^{(1)}, \theta^{(2)}, \ldots$ from the MC

3. $\text{ave}\{g(\theta^{(t)})\} \to E[g(\theta)]$ provided the MC is ergodic.

**Ref:** Gelman et al. (1995) Bayesian data analysis. Chapman & Hall.
Metropolis-Hastings algorithm

The first problem is to find a transition density, $P$, whose stationary distribution is the target, $\pi$.

There are several recipes to follow. The most general is the *Metropolis-Hastings algorithm*:

1. Given the current point $\theta^{(t)}$, pick $\theta^*$ from a jump density $J_t(\theta^*|\theta^{(t)})$.

2. Evaluate the ratio

$$r = \frac{\pi(\theta^*) / J_t(\theta^*|\theta^{(t)})}{\pi(\theta^{(t)}) / J_t(\theta^{(t)}|\theta^*)} = \frac{\pi(\theta^*) J_t(\theta^{(t)}|\theta^*)}{\pi(\theta^{(t)}) J_t(\theta^*|\theta^{(t)})}$$

3. Pick $U \sim \text{uniform}(0, 1)$.

4. If $U < r$, take $\theta^{(t+1)} = \theta^*$; otherwise $\theta^{(t+1)} = \theta^{(t)}$.

**Special case:** *Metropolis algorithm*

Let the jump distribution be symmetric, so that

$$J_t(\theta^*|\theta^{(t)}) = J_t(\theta^{(t)}|\theta^*)$$

In this case, we have $r = \pi(\theta^*) / \pi(\theta^{(t)})$.

Note that in both cases, we only need to know $\pi$ up to a constant.
Why does this work?

Assume (without loss of generality) that \( r \leq 1 \). Then
\[
\pi(\theta^{(t)}) J_t(\theta^* | \theta^{(t)}) r = \pi(\theta^*) J_t(\theta^{(t)} | \theta^*).
\]

It follows that \( \pi \) is a stationary distribution of the Markov chain.

**Trickier:** to show that \( \pi \) is the unique stationary dist’n, we need to show that the MC is irreducible, aperiodic and not transient.

**Aperiodic and not transient:** These properties follow provided that we’re dealing with proper distributions.

**Irreducible:** we need to be able to get from any state to any other state (eventually). This depends on the choice of the jump distribution \( J_t \). We need to at least be sure that there are no absorbing states.

In my experience (we’ll see below), the Gibbs sampler (described below) may be especially susceptible to lack of irreducibility.
Simple example

[Taken from Gelman et al (1995).]

Consider a standard bivariate normal density.

Let $J_t(\theta^*|\theta^{(t)})$ be draws from independent normal densities centered at $\theta^{(t)}$ with SDs = 0.2.

In the left panel below, we show 100 steps of such Markov chains started at five dispersed points.

In the right panel, we show the first 1000 steps.
Gibbs sampler

Suppose we can divide $\theta$ up into $d$ subvectors:
$\theta = (\theta_1, \ldots, \theta_d)$

Define $\theta_{-i} = (\theta_1, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d)$
(i.e., $\theta$ without its $i$th subvector).

In the Gibbs sampler, we draw $\theta_i^*$ from $\pi(\theta_i|\theta_{-i}^{(t)})$ and step
from $\theta^{(t)}$ to $\theta^{(t+1)} = (\theta_1^{(t)}, \ldots, \theta_{i-1}^{(t)}, \theta_i^*, \theta_{i+1}^{(t)}, \ldots, \theta_d^{(t)})$.

where $\pi(\theta_i|\theta_{-i}) = \pi(\theta)/\int \pi(\theta) \, d\theta_i$

It is easy to see that this is a special case of the Metropolis-Hastings algorithm, and that $r = 1$.

In each step of the Markov chain, we generally update each of the subvectors of $\theta$ (a set of $d$ Gibbs steps). This can be done either in a fixed order or in a random order.
Another simple example

[Again, taken from Gelman et al (1995).]

Consider a single observation \((y_1, y_2)\) from a bivariate normal distribution with unknown means \((\theta_1, \theta_2)\) and known covariance matrix \(\Sigma\) with variances 1 and covariance \(\rho\).

With a uniform prior on \(\theta\), we have \(\theta | y \sim \text{normal}(y, \Sigma)\).

While we can easily sample from this directly, that’s look at what Gibbs sampling would give:

\[
\begin{align*}
\theta_1 | \theta_2, y & \sim \text{normal}(y_1 + \rho(\theta_2 - y_2), 1 - \rho^2) \\
\theta_2 | \theta_1, y & \sim \text{normal}(y_2 + \rho(\theta_1 - y_1), 1 - \rho^2)
\end{align*}
\]

In implementing the Gibbs sampler, we do the following:

1. Pick a starting point \((\theta_1^{(0)}, \theta_2^{(0)})\)
2. Draw \(\theta_1^{(t+1)}\) from \(\pi(\theta_1 | y, \theta_2^{(t)})\)
3. Draw \(\theta_2^{(t+1)}\) from \(\pi(\theta_2 | y, \theta_1^{(t+1)})\).
4. Repeat steps 2–3.
Simple example (continued)

This figure shows the results for $y = (0, 0)$ and $\rho = 0.8$, starting at four dispersed points.

The left panel contains the first 100 steps; the right panel contains the first 1000 steps.

Note: If it is difficult to sample from $\pi(\theta_i|\theta_{-i})$, we can use an approximation $g(\theta_i|\theta_{-i})$ and then use the general form of the ratio in the Metropolis-Hastings algorithm to fix things up.
Major issues

1. Pick a good sampler
   - Move through the parameter space quickly
   - Don’t reject too often

2. How many samples? (assessment of convergence)

3. Estimate $E[g(\theta)]$ by $\frac{1}{N} \sum_{t=1}^{N} g(\theta^{(t)})$
   - “Burn in”: drop the first $M$ steps
     [Gelman et al: drop first half]
   - Keep only every $k$th sample
     [Gelman et al: keep every sample]
Picking a good sampler

1. Easy to sample from $J(\theta^* | \theta)$
2. Easy to calculate the ratio $r$
3. In the Gibbs sampler, you want the components to be approximately independent (reparameterize?)
4. In Metropolis-Hastings:
   - Want steps to be big
   - Don’t want to reject too often
5. Data augmentation (like in EM) can be helpful
6. MVN(\(\theta, \Sigma\)) where \(\theta\) is a \(p\)-vector:
   - Jump using N(\(\theta^{(t)}, c^2\Sigma\))
   - Optimal \(c \approx 2.4/\sqrt{p}\)
   - Reject $\sim 44\%$ when \(p = 1\); $\sim 23\%$ when \(p > 5\).
   - Adaptive approach? (use initial samples to adjust the jump density)
Assessing convergence

1. Run multiple chains with dispersed starting points
2. For each scalar of interest (and many that are not of interest), $\psi$, consider

$$B = \frac{n}{J-1} \sum_{j=1}^{J} (\psi_{.j} - \psi_{..})^2, \quad W = \frac{1}{J} \sum_{j=1}^{J} s_{j}^2$$

where $s_{j}^2 = \sum_{i=1}^{n} (\psi_{ij} - \psi_{.j})^2 / (n - 1)$

Gelman et al suggest monitoring $R = [\frac{n-1}{n} W + \frac{1}{n} B] / W$, looking for $\sqrt{R} < 1.2$.

3. May wish to transform to approximate normality before doing (2).

4. Definitely monitor $\pi(\theta^{(t)})$ as well

5. Can be very hard to look at plots of everything (and they may mislead you).
Example: T-cell assay

Data

<table>
<thead>
<tr>
<th>cells alone</th>
<th>gD2</th>
<th>gB2</th>
<th>Tetox</th>
</tr>
</thead>
<tbody>
<tr>
<td>179 249 460</td>
<td>2133 2528 2700</td>
<td>2171 1663 6200</td>
<td>761 9864 —</td>
</tr>
<tr>
<td>346 1540 306</td>
<td>8299 1886 3245</td>
<td>1699 2042 3374</td>
<td>183 7748 —</td>
</tr>
<tr>
<td>117 249 1568</td>
<td>1174 4293 979</td>
<td>1222 1536 2406</td>
<td>6497 2492 6188</td>
</tr>
<tr>
<td>184 414 308</td>
<td>2801 2437 1776</td>
<td>2193 3211 1936</td>
<td>2492 5134 927</td>
</tr>
<tr>
<td>797 233 461</td>
<td>1076 1527 2866</td>
<td>2205 2278 2215</td>
<td>3725 3706 4051</td>
</tr>
<tr>
<td>305 348 480</td>
<td>3475 902 3654</td>
<td>2046 1285 1187</td>
<td>9899 5891 3646</td>
</tr>
<tr>
<td>1090 159 89</td>
<td>1472 90 3639</td>
<td>657 2393 1814</td>
<td>3330 4174 2389</td>
</tr>
<tr>
<td>280 571 329</td>
<td>4449 3643 881</td>
<td>3462 2118 1013</td>
<td>8793 4313 672</td>
</tr>
</tbody>
</table>

Data:

\[ y_{ij} = \text{scintillation count for well } i, j. \]

\[ x_{ij} = \sqrt{y_{ij}} \]

\[ k_{ij} = (\text{unobserved}) \text{ no. responding cells in well } i, j. \]

Model:

\[ k_{ij} \sim \text{Poisson}(\lambda_i) \]

\[ x_{ij} | k_{ij} \sim \text{normal}(a + bk_{ij}, \sigma^2) \]

\( (x_{ij}, k_{ij}) \) are mutually independent
Example (continued)

Modes in the likelihood surface
(found using EM with 10,000 starting points)

<table>
<thead>
<tr>
<th>(\lambda_0)</th>
<th>(\lambda_1)</th>
<th>(\lambda_2)</th>
<th>(\lambda_3)</th>
<th>(a)</th>
<th>(b)</th>
<th>(\sigma)</th>
<th>log lik</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.32</td>
<td>3.03</td>
<td>2.82</td>
<td>4.37</td>
<td>16.73</td>
<td>10.34</td>
<td>3.52</td>
<td>0.00</td>
</tr>
<tr>
<td>1.18</td>
<td>5.40</td>
<td>4.95</td>
<td>7.49</td>
<td>12.16</td>
<td>6.69</td>
<td>2.15</td>
<td>-0.07</td>
</tr>
<tr>
<td>0.17</td>
<td>2.10</td>
<td>1.95</td>
<td>3.07</td>
<td>17.44</td>
<td>14.56</td>
<td>4.18</td>
<td>-0.77</td>
</tr>
<tr>
<td>0.51</td>
<td>3.89</td>
<td>3.56</td>
<td>5.58</td>
<td>15.72</td>
<td>8.35</td>
<td>3.58</td>
<td>-0.97</td>
</tr>
<tr>
<td>0.73</td>
<td>4.62</td>
<td>4.25</td>
<td>6.58</td>
<td>14.58</td>
<td>7.27</td>
<td>3.43</td>
<td>-1.35</td>
</tr>
<tr>
<td>1.64</td>
<td>6.79</td>
<td>6.29</td>
<td>9.35</td>
<td>10.81</td>
<td>5.51</td>
<td>1.89</td>
<td>-1.67</td>
</tr>
<tr>
<td>1.57</td>
<td>6.22</td>
<td>5.80</td>
<td>8.61</td>
<td>10.60</td>
<td>6.02</td>
<td>2.13</td>
<td>-1.86</td>
</tr>
<tr>
<td>2.59</td>
<td>7.76</td>
<td>7.25</td>
<td>10.34</td>
<td>5.75</td>
<td>5.47</td>
<td>1.88</td>
<td>-2.55</td>
</tr>
<tr>
<td>2.71</td>
<td>8.52</td>
<td>7.97</td>
<td>11.41</td>
<td>6.65</td>
<td>4.88</td>
<td>1.80</td>
<td>-3.70</td>
</tr>
</tbody>
</table>

Note that at the modes, \(\lambda_1\), \(\lambda_2\), and \(\lambda_3\) are essentially collinear.
Example (continued)

Why use MCMC here?

1. It’s a nice example to play with
2. We may want to be Bayesian
3. Explore the likelihood surface
4. Random optimizer (see bit in Lange on simulated annealing)

Prior on $\theta = (\lambda_0, \lambda_1, \lambda_2, \lambda_3, a, b, \sigma^2)$: $p(\theta) \propto 1/\sigma^2 \prod_i e^{-\lambda_i}$

Flat improper priors on $a, b, \sigma$; $\lambda$'s are indep gamma$(1, 1)$

We’ll use a Gibbs sampler, including the augmenting the data with the missing $k$'s.
Example: Gibbs sampler

1. Sample $k$'s
   
   \[ \Pr(k_{ij} = s|\theta, x_{ij}) \propto \frac{e^{-\lambda_i} \lambda_i^s}{s!} \phi\left(\frac{x_{ij} - a - bs}{\sigma}\right) \]

2. Sample $\lambda$'s:
   
   Given $k$ and $x$, $\lambda_i \sim \text{gamma}(\sum_j k_{ij} + 1, n + 1)$

3. Regress $x$ on $k$ (with intercept) to get $\hat{\beta}$ and residual variance $s^2$.

4. Sample $\sigma$:
   
   Given $k$ and $x$, $\sigma^2$ is a scaled inverse-$\chi^2(n - 2, s^2)$.
   
   [Sample $X \sim \chi^2(n - 2)$ and let $\sigma = s \sqrt{(n - 2)/X}$]

5. Sample $a$ and $b$:
   
   Given $k$, $x$ and $\sigma$, $(ab)'$ is normal with mean $\hat{\beta}$ and variance matrix $\sigma^2(Z'Z)^{-1}$ where $Z = (1_k)$
   
   [See Gelman et al, ch 8]
Results

I started a chain running at each of my 9 previously-discovered modes.

To my surprise, the chain seemed to mix quite well! The following table shows Gelman's statistic comparing the within- and between-chain variation.

<table>
<thead>
<tr>
<th>n</th>
<th>$\lambda_0$</th>
<th>$\lambda_1$</th>
<th>$\lambda_2$</th>
<th>$\lambda_3$</th>
<th>a</th>
<th>b</th>
<th>$\sigma$</th>
<th>log lik</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>1.637</td>
<td>2.067</td>
<td>2.023</td>
<td>2.182</td>
<td>1.478</td>
<td>3.747</td>
<td>1.130</td>
<td>1.092</td>
</tr>
<tr>
<td>1000</td>
<td>1.201</td>
<td>1.252</td>
<td>1.251</td>
<td>1.268</td>
<td>1.159</td>
<td>1.427</td>
<td>1.046</td>
<td>1.024</td>
</tr>
<tr>
<td>2500</td>
<td>1.065</td>
<td>1.107</td>
<td>1.106</td>
<td>1.120</td>
<td>1.048</td>
<td>1.241</td>
<td>1.017</td>
<td>1.009</td>
</tr>
<tr>
<td>5000</td>
<td>1.021</td>
<td>1.025</td>
<td>1.026</td>
<td>1.028</td>
<td>1.021</td>
<td>1.048</td>
<td>1.010</td>
<td>1.004</td>
</tr>
<tr>
<td>10000</td>
<td>1.008</td>
<td>1.009</td>
<td>1.009</td>
<td>1.010</td>
<td>1.007</td>
<td>1.022</td>
<td>1.004</td>
<td>1.002</td>
</tr>
</tbody>
</table>

Taking the second halves of each chain, we get the following:

<table>
<thead>
<tr>
<th></th>
<th>$\lambda_0$</th>
<th>$\lambda_1$</th>
<th>$\lambda_2$</th>
<th>$\lambda_3$</th>
<th>a</th>
<th>b</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ave</td>
<td>0.21</td>
<td>2.14</td>
<td>1.98</td>
<td>3.11</td>
<td>17.57</td>
<td>14.07</td>
<td>4.51</td>
</tr>
<tr>
<td>sd</td>
<td>0.11</td>
<td>0.41</td>
<td>0.38</td>
<td>0.55</td>
<td>0.87</td>
<td>1.40</td>
<td>0.73</td>
</tr>
<tr>
<td>2.5%ile</td>
<td>0.06</td>
<td>1.54</td>
<td>1.40</td>
<td>2.31</td>
<td>16.09</td>
<td>10.07</td>
<td>3.56</td>
</tr>
<tr>
<td>97.5%ile</td>
<td>0.47</td>
<td>3.17</td>
<td>2.91</td>
<td>4.54</td>
<td>19.50</td>
<td>15.36</td>
<td>6.36</td>
</tr>
</tbody>
</table>
Results

Here are scatter plots of $b$ vs $\sigma$ for two of the chains. (Left—100 steps; center—1000 steps; right—all samples)

Here are histograms from the second halves of the chains.