Markov chain Monte Carlo

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

We wish to sample from π in order to estimate

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

Why?

1. We're being Bayesians, and $\pi(\theta) = \pi(\theta|\text{data})$ is the posterior distribution of θ.
2. We wish to explore the likelihood surface of θ (in the case that the posterior is proper when using a totally flat prior).
3. We wish to optimize the likelihood of θ

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. 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 distibution is the target, π.

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

1. Given the current point $\theta^{(t)}$, pick $\theta^\star$ from a jump density $J_t(\theta^\star|\theta^{(t)})$.
2. Evaluate the ratio
$$r = \frac{\pi(\theta^\star)/J_t(\theta^\star|\theta^{(t)})}{\pi(\theta^{(t)})/J_t(\theta^{(t)}|\theta^\star)} = \frac{\pi(\theta^\star)J_t(\theta^{(t)}|\theta^\star)}{\pi(\theta^{(t)})J_t(\theta^\star|\theta^{(t)})}$$
3. Pick $U \sim \text{uniform}(0,1)$.
4. If $U < r$, take $\theta^{(t+1)} = \theta^\star$; otherwise $\theta^{(t+1)} = \theta^{(t)}$.

Special case: *Metropolis algorithm*

Let the jump distribution be symmetric, so that
$$J_t(\theta^\star|\theta^{(t)}) = J_t(\theta^{(t)}|\theta^\star)$$

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

Note that in both cases, we only need to know π up to a constant.

Why does this work?

Assume (without loss of generality) that $r \leq 1$. Then $\pi(\theta^{(t)}) J_t(\theta^\star|\theta^{(t)})\, r = \pi(\theta^\star) J_t(\theta^{(t)}|\theta^\star)$.

It follows that π is a stationary distribution of the Markov chain.

Trickier: to show that π 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

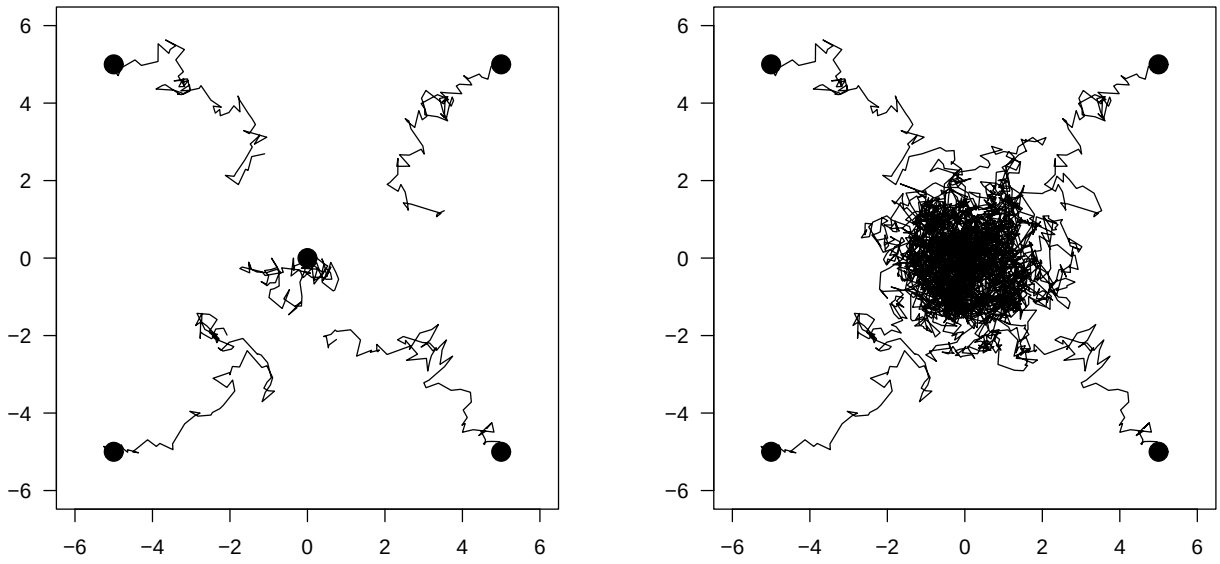
[Taken from Gelman et al (1995).]

Consider a standard bivariate normal density.

Let $J_t(\theta^\star|\theta^{(t)})$ 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 θ 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., θ without its ith subvector).

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

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 θ (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 (θ_1, θ_2) and known covariance matrix Σ with variances 1 and covariance ρ.

With a uniform prior on θ, 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:

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

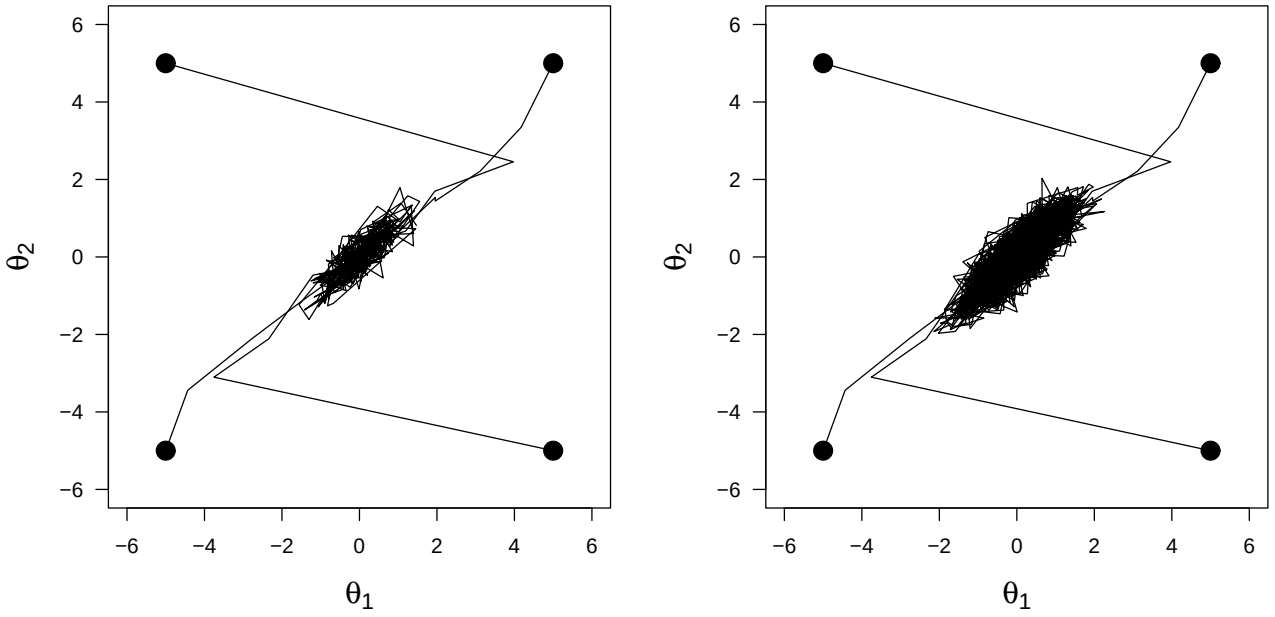
In implementing the Gibbs sampler, we do the following:

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 kth sample
     [Gelman et al: keep every sample]

Picking a good sampler

1. Easy to sample from $J(\theta^\star|\theta)$
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(θ, Σ) where θ 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), ψ, consider

$$B = \frac{n}{J-1}\sum_{j=1}^{J}(\psi_{\cdot j} - \psi_{\cdot\cdot})^2, \qquad W = \frac{1}{J}\sum_{j=1}^{J} s_j^2$$

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

   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

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

Data:

$y_{ij} =$ scintillation count for well i, j.

$x_{ij} = \sqrt{y_{ij}}$

$k_{ij} =$ (unobserved) 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 mututally independent

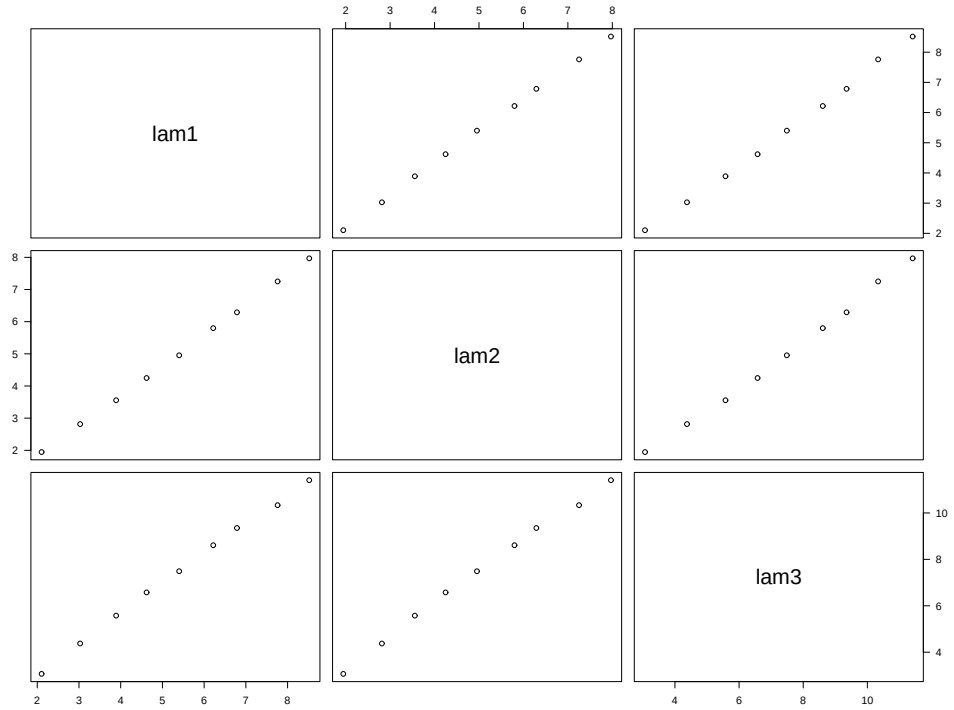
Example (continued)

Modes in the likelihood surface

(found using EM with 10,000 starting points)

λ_0	λ_1	λ_2	λ_3	a	b	σ	log lik
0.32	3.03	2.82	4.37	16.73	10.34	3.52	0.00
1.18	5.40	4.95	7.49	12.16	6.69	2.15	-0.07
0.17	2.10	1.95	3.07	17.44	14.56	4.18	-0.77
0.51	3.89	3.56	5.58	15.72	8.35	3.58	-0.97
0.73	4.62	4.25	6.58	14.58	7.27	3.43	-1.35
1.64	6.79	6.29	9.35	10.81	5.51	1.89	-1.67
1.57	6.22	5.80	8.61	10.60	6.02	2.13	-1.86
2.59	7.76	7.25	10.34	5.75	5.47	1.88	-2.55
2.71	8.52	7.97	11.41	6.65	4.88	1.80	-3.70

Note that at the modes, λ_1, λ_2, and λ_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, σ; λ'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

   $\Pr(k_{ij} = s|\theta, x_{ij}) \propto \frac{e^{-\lambda_i}\lambda_i^2}{s!}\phi(\frac{x_{ij}-a-bs}{\sigma})$

2. Sample λ'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 σ:

   Given k and x, σ^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 σ, $(ab)'$ is normal with mean $\hat{\beta}$ and variance matrix $\sigma^2(Z'Z)^{-1}$ where $Z = (1k)$

   [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.

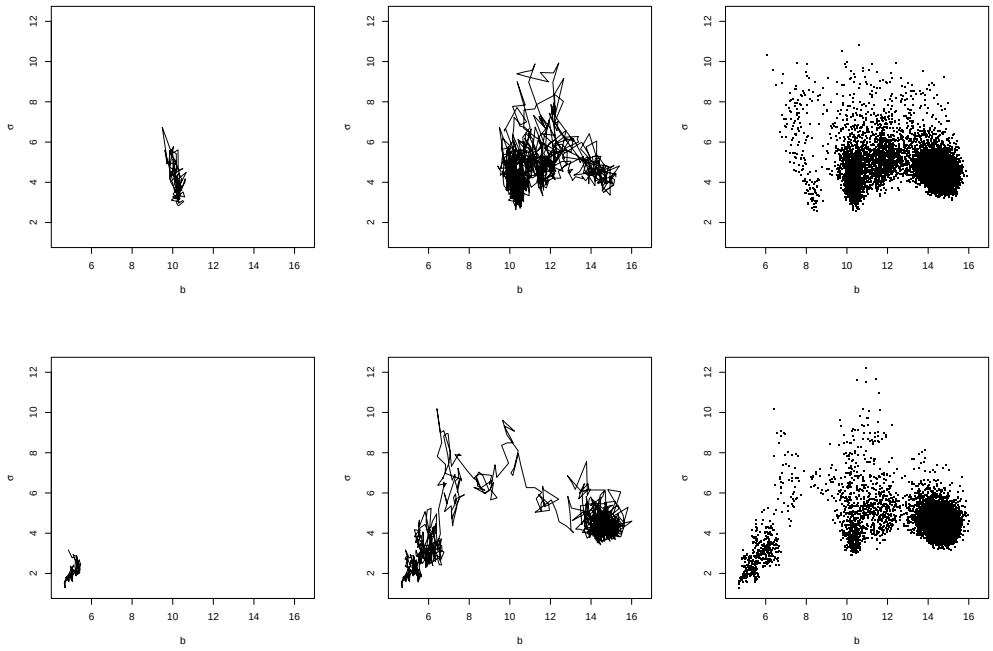
	$\sqrt{R}$							
n	λ_0	λ_1	λ_2	λ_3	a	b	σ	log lik
100	4.621	8.702	7.998	9.546	3.884	30.027	2.109	1.264
500	1.637	2.067	2.023	2.182	1.478	3.747	1.130	1.092
1000	1.201	1.252	1.251	1.268	1.159	1.427	1.046	1.024
2500	1.065	1.107	1.106	1.120	1.048	1.241	1.017	1.009
5000	1.021	1.025	1.026	1.028	1.021	1.048	1.010	1.004
10000	1.008	1.009	1.009	1.010	1.007	1.022	1.004	1.002

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

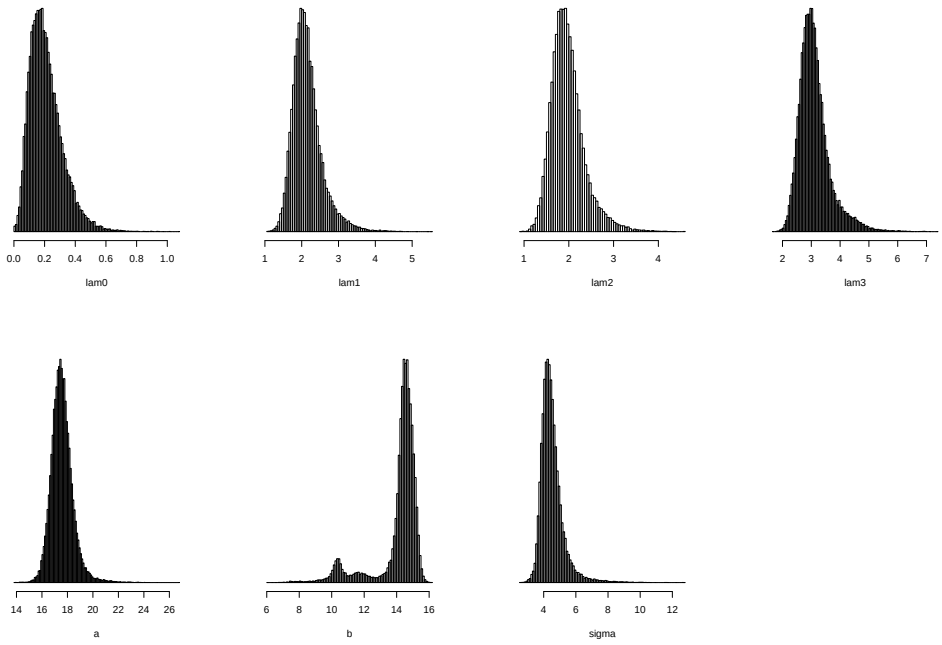
	λ_0	λ_1	λ_2	λ_3	a	b	σ
ave	0.21	2.14	1.98	3.11	17.57	14.07	4.51
sd	0.11	0.41	0.38	0.55	0.87	1.40	0.73
2.5%ile	0.06	1.54	1.40	2.31	16.09	10.07	3.56
97.5%ile	0.47	3.17	2.91	4.54	19.50	15.36	6.36

Results

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



Here are histograms from the second halves of the chains.



0
2
4
6
8
10
λ3
0
2
4
6
8
λ2