4th IMPRS Astronomy Summer School

Drawing Astrophysical Inferences from Data Sets

William H. Press
The University of Texas at Austin

Lecture 5

Markov Chain Monte Carlo (MCMC)

Data set $\mathbf{D}$

Parameters $\mathbf{x}$ (sorry, we've changed notation!)

We want to go beyond simply maximizing $P(\mathbf{D}|\mathbf{x})$
and get the whole Bayesian posterior distribution of $\mathbf{x}$

Bayes says this is proportional to $\pi(\mathbf{x}) \equiv P(\mathbf{D}|\mathbf{x})P(\mathbf{x})$
but with an unknown proportionality constant (the Bayes denominator). It seems as if we need this denominator to find confidence regions, e.g., containing 95% of the posterior probability.

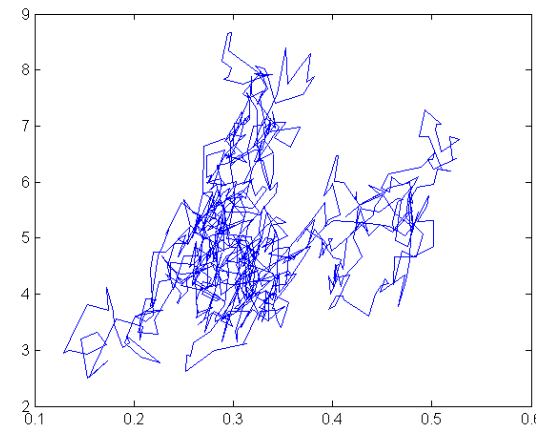
But no! MCMC is a way of drawing samples $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots$
from the distribution $\pi(\mathbf{x})$
without having to know its normalization!

With such a sample, we can compute any quantity of interest about the distribution of $\mathbf{x}$, e.g., confidence regions, means, standard deviations, covariances, etc.

Two ideas due to Metropolis and colleagues make this possible:

1. Instead of sampling unrelated points, sample a Markov chain $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots$ where each point is (stochastically) determined by the previous one by some chosen distribution $p(\mathbf{x}_i|\mathbf{x}_{i-1})$

Although locally correlated, it is possible to make this sequence ergodic, meaning that it visits every **x** in proportion to π(**x**).



2. Any distribution $p(\mathbf{x}_i|\mathbf{x}_{i-1})$ that satisfies

$$\pi(\mathbf{x}_1)\,p(\mathbf{x}_2|\mathbf{x}_1) = \pi(\mathbf{x}_2)\,p(\mathbf{x}_1|\mathbf{x}_2)$$

(“detailed balance”) will be such an ergodic sequence!

Deceptively simple proof: Compute distribution of x_1’s successor point

$$\int p(\mathbf{x}_2|\mathbf{x}_1)\pi(\mathbf{x}_1)\,d\mathbf{x}_1 = \pi(\mathbf{x}_2)\int p(\mathbf{x}_1|\mathbf{x}_2)\,d\mathbf{x}_1 = \pi(\mathbf{x}_2)$$

So how do we find such a $p(\mathbf{x}_i|\mathbf{x}_{i-1})$?

Metropolis-Hastings algorithm:

Pick more or less any "proposal distribution" $q(\mathbf{x}_2|\mathbf{x}_1)$
(A multivariate normal centered on $\mathbf{x}_1$ is a typical example.)

Then the algorithm is:

1. Generate a candidate point $\mathbf{x}_{2c}$ by drawing from the proposal distribution around $\mathbf{x}_1$

2. Calculate an "acceptance probability" by

Notice that the q's cancel out if symmetric on arguments, as is a multivariate Gaussian

$$\alpha(\mathbf{x}_1, \mathbf{x}_{2c}) = \min\left(1, \frac{\pi(\mathbf{x}_{2c})\, q(\mathbf{x}_1|\mathbf{x}_{2c})}{\pi(\mathbf{x}_1)\, q(\mathbf{x}_{2c}|\mathbf{x}_1)}\right)$$

3. Choose $\mathbf{x}_2 = \mathbf{x}_{2c}$ with probability α, $\mathbf{x}_2 = \mathbf{x}_1$ with probability $(1-\alpha)$

So, $$p(\mathbf{x}_2|\mathbf{x}_1) = q(\mathbf{x}_2|\mathbf{x}_1)\, \alpha(\mathbf{x}_1, \mathbf{x}_2), \qquad (\mathbf{x}_2 \neq \mathbf{x}_1)$$

It's something like: always accept a proposal that increases the probability, and sometimes accept one that doesn't. (Not exactly this because of ratio of q's.)

Proof: $$p(\mathbf{x}_2|\mathbf{x}_1) = q(\mathbf{x}_2|\mathbf{x}_1)\,\alpha(\mathbf{x}_1,\mathbf{x}_2), \qquad (\mathbf{x}_2 \neq \mathbf{x}_1)$$

$$\begin{aligned}
\pi(\mathbf{x}_1)\,q(\mathbf{x}_2|\mathbf{x}_1)\,\alpha(\mathbf{x}_1,\mathbf{x}_2) &= \min[\pi(\mathbf{x}_1)\,q(\mathbf{x}_2|\mathbf{x}_1),\ \pi(\mathbf{x}_2)\,q(\mathbf{x}_1|\mathbf{x}_2)] \\
&= \min[\pi(\mathbf{x}_2)\,q(\mathbf{x}_1|\mathbf{x}_2),\ \pi(\mathbf{x}_1)\,q(\mathbf{x}_2|\mathbf{x}_1)] \\
&= \pi(\mathbf{x}_2)\,q(\mathbf{x}_1|\mathbf{x}_2)\,\alpha(\mathbf{x}_2,\mathbf{x}_1)
\end{aligned}$$

which is just detailed balance!

("Gibbs sampler", beyond our scope, is a special case of Metropolis-Hastings. See, e.g., NR3.)

Let's do an MCMC example to show how it can be used with models that might be analytically intractable (e.g., discontinuous or non-analytic).
[This is the example worked in NR3.]

The lazy birdwatcher problem

- You hire someone to sit in the forest and look for mockingbirds.
- They are supposed to report the time of each sighting t_i
  - But they are lazy and only write down (exactly) every k_1 sightings (e.g., k_1= every 3rd)
- Even worse, at some time t_c they get a young child to do the counting for them
  - He doesn't recognize mockingbirds and counts grackles instead
  - And, he writes down only every k_2 sightings, which may be different from k_1
- You want to salvage something from this data
  - E.g., average rate of sightings of mockingbirds *and* grackles
  - Given only the list of times
  - That is, k_1, k_2, and t_c are all unknown nuisance parameters
- This all hinges on the fact that every second (say) event in a Poisson process is statistically distinguishable from every event in a Poisson process at half the mean rate
  - same mean rates
  - but different fluctuations
  - We are hoping that the difference in fluctuations is enough to recover useful information
- Perfect problem for MCMC

Waiting time to the kth event in a Poisson process with rate λ is distributed as Gamma(k,λ)

$$\tau = t_{i+k} - t_i$$

$$p(\tau|k,\lambda) = \frac{\lambda^k}{(k-1)!}\tau^{k-1}e^{-\lambda\tau}$$

And non-overlapping intervals are independent: $t_{i+k} - t_i$

$t_{i+2k} - t_{i+k}$

Proof:

$$\begin{aligned} p(\tau)d\tau &= P(k-1 \text{ counts in } \tau) \times P(\text{last } d\tau \text{ has a count}) \\ &= \text{Poisson}(k-1, \lambda\tau) \times (\lambda\, d\tau) \\ &= \frac{(\lambda\tau)^{k-1}}{(k-1)!}e^{-\lambda\tau}\lambda\, d\tau \end{aligned}$$

So

$$P(\mathbf{D}|\mathbf{x}) = \prod_{t_i \le t_c} p(t_{i+1} - t_i \,|\, k_1, \lambda_1) \times \prod_{t_i > t_c} p(t_{i+1} - t_i \,|\, k_2, \lambda_2)$$

What shall we take as our proposal generator?
This is often the creative part of getting MCMC to work well!

For t_c, step by small additive changes (e.g., normal)

For λ_1 and λ_2, step by small multiplicative changes (e.g., lognormal)

In the acceptance probability the ratio of the q's in

$$\alpha(\mathbf{x}_1, \mathbf{x}_{2c}) = \min\left(1, \frac{\pi(\mathbf{x}_{2c})\, q(\mathbf{x}_1|\mathbf{x}_{2c})}{\pi(\mathbf{x}_1)\, q(\mathbf{x}_{2c}|\mathbf{x}_1)}\right)$$

is just x_{2c}/x_1, because

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{1}{2}\left[\frac{\log(x) - \mu}{\sigma}\right]^2\right)$$

Bad idea: For $k_{1,2}$ step by 0 or ±1

This is bad because, if the λ's have converged to about the right rate, then a change in *k* will throw them way off, and therefore nearly always be rejected. Even though this appears to be a "small" step of a discrete variable, it is not a small step in the model!

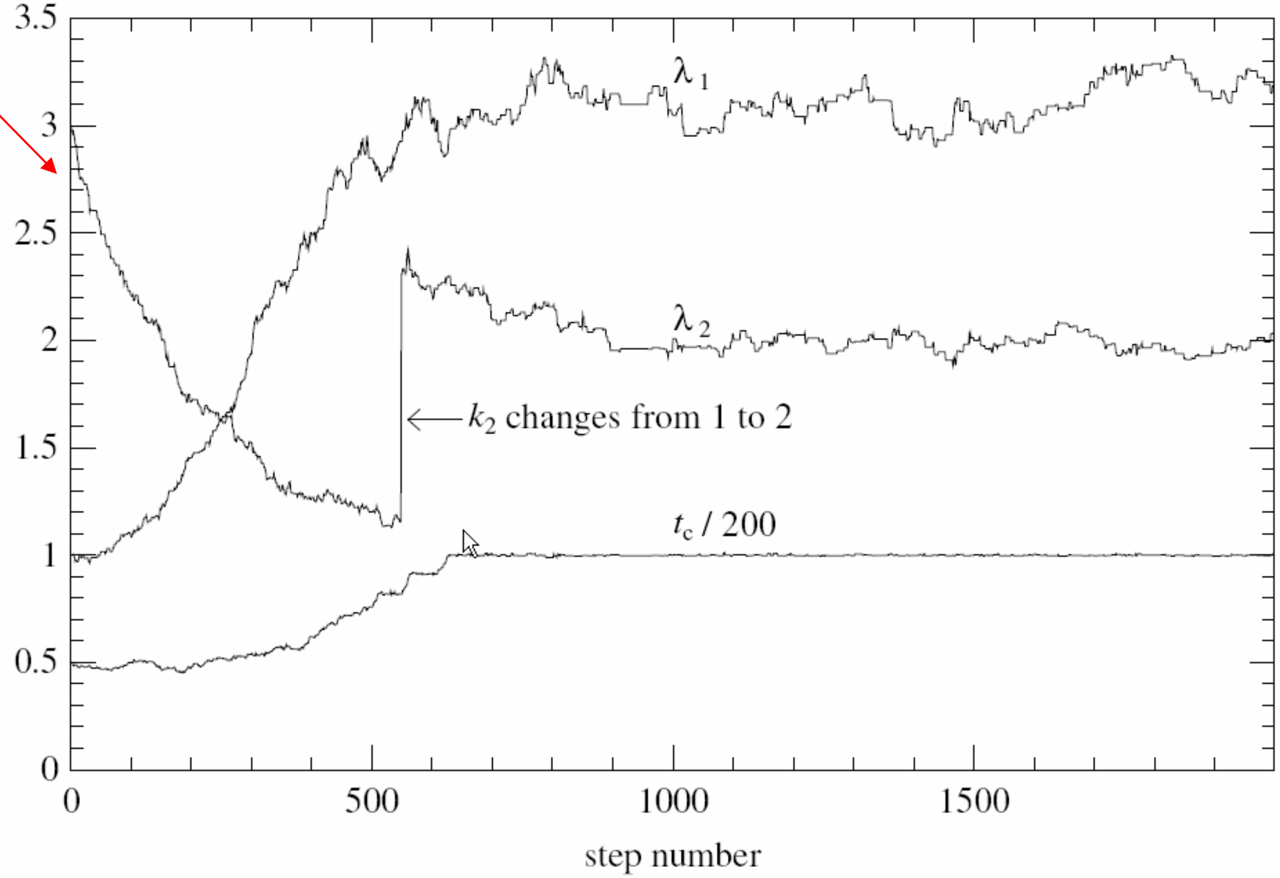
Good idea: For $k_{1,2}$ step by 0 or ±1, also changing $\lambda_{1,2}$ so as to keep λ/k constant in the step

This is genuinely a small step, since it changes only the clumping statistics, by the smallest allowed amount.

Let's try it.

We simulate 1000 t_i's with the secretly known λ_1=3.0, λ_2=2.0, t_c=200, k_1=1, k_2=2

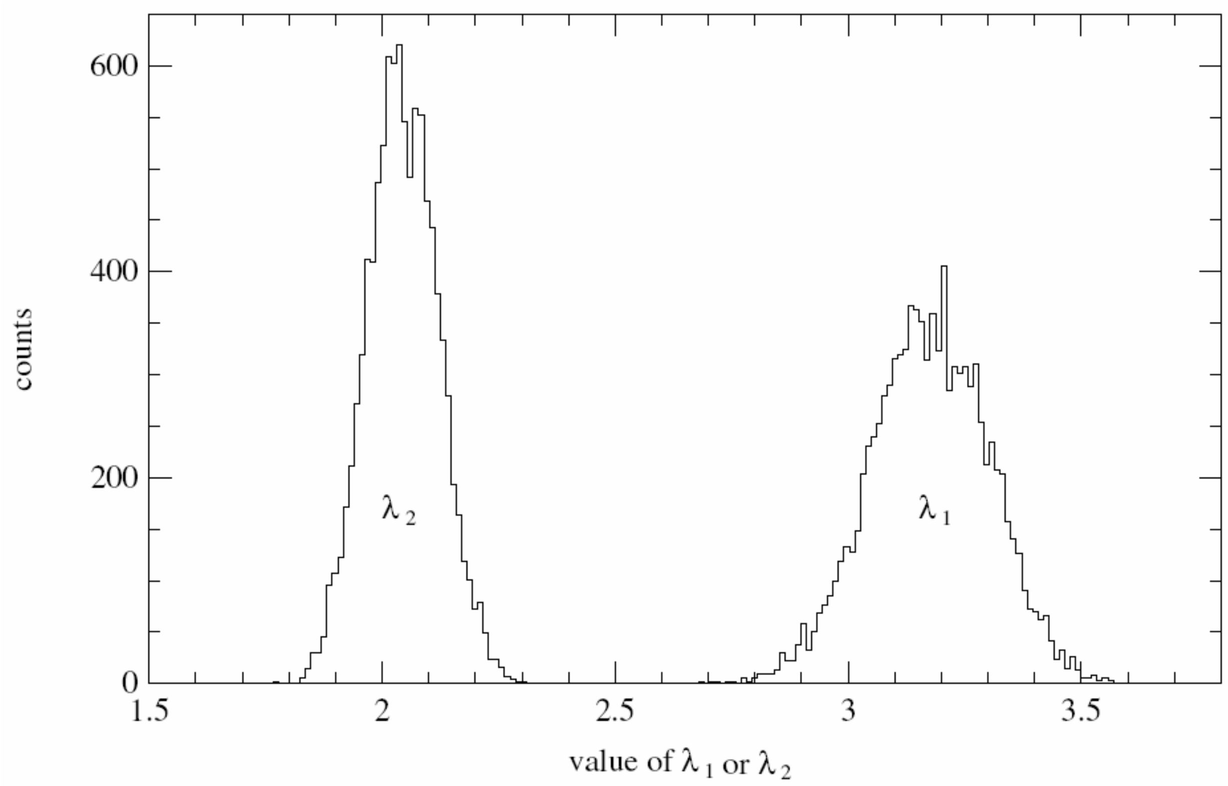
Start with wrong values λ_1=1.0, λ_2=3.0, t_c=100, k_1=1, k_2=1



"burn-in" period while it locates the Bayes maximum

ergodic period during which we record data for plotting, averages, etc.

Histogram of quantities during a long-enough ergodic time



These are the actual Bayesian posteriors of the model!

Could as easily do joint probabilities, covariances, etc., etc.

Notice does not converge to being centered on the true values, because the (finite available) data is held fixed. Convergence is to the Bayesian posterior for that data.