MLSS 2009 practical

Markov chain Monte Carlo — version 2

Iain Murray

1 Introduction

To provide a demonstration of what MCMC can actually be used for, and to add a bit of interest, we will be solving a full probabilistic modelling task. This document fully describes the model, although most people won't have time to fully understand and implement it during the practical. Instead, code for the model's joint probability distribution is provided.

I have also provided code for two standard MCMC samplers. Don't feel you should implement these yourself. Note that Matlab's stats toolbox also provides (different) implementations of both samplers. If you really want to write a sampler, a suggestion is given later.

Your main tasks will be: 1) writing code to drive a standard sampler given a joint probability distribution and 2) investigating to what extent the samplers are working.

Download the code and data from:
`http://www.cs.toronto.edu/~murray/teaching/09mlss/`

2 Problem introduction

We will be solving a data-analysis problem inspired by Bovy et al. (2009).

Motivation: Some galaxies and stars have dynamics following very large scale orbits. Over the course of an experiment they will only move a very small way around their orbit: we effectively just see their current position and velocity. Given this 'snapshot' we would like to infer the physical properties of the system.

The simplest version of this problem involves seeing a 1-dimensional projection of a system in which objects are oscillating around a very massive 'black hole'. We will assume that each object is undergoing 'simple harmonic motion' (SHM). Each object has an unknown amplitude A_n and phase ϕ_n associated with it. What we can measure is each object's position x_n and a velocity v_n. From these we would like to infer the position m and resonant frequency ω associated with the black hole.

The system is described by:

$$\begin{aligned} x_n &= m + A_n \sin(\omega t + \phi_n) \\ v_n &= A_n\,\omega \cos(\omega t + \phi_n) \end{aligned} \tag{1}$$

Observed: $\{x_n, v_n\}_{n=1}^N$

Unobserved: $m, \omega, \{\phi_n, A_n\}_{n=1}^N$

We assume that our experiment is not conducted at any special time, so without loss of generality we can assert that our snapshot of the dynamics was taken at time $t\!=\!0$.

Notice that if we knew the properties of the black hole, $\{m, \omega\}$, then we could solve Equation (1) for all remaining unknowns (at $t\!=\!0$):

$$\begin{aligned} A_n &= \sqrt{(x_n - m)^2 + (v_n/\omega)^2} \\ \phi &= \mathrm{atan2}\big(x_n - m,\, v_n/\omega\big). \end{aligned} \tag{2}$$

The website provides position and velocity observations of 50 objects. Again, the task is to infer the postion, m, and resonant frequency, ω, of the black hole.

3 The model and its parameter posterior

This section defines a hierarchical Bayesian model. If you take my final model specification on trust, you should be able to build a sampler without following all the details. If you have time, please do explore your own versions of the model.

The way to solve most inference problems starts with writing down a joint probability containing at least our observations and variables of interest. Assuming the position and resonant frequency of the black hole were chosen independently:

$$p(\log\omega, m, \{x_n, v_n\}) \;=\; p(\{x_n, v_n\} \,|\, \log\omega, m)\, p(\log\omega)\, p(m).$$

Unless you have specific knowledge of my fictional black hole, I suggest you put wide uniform priors on m and $\log\omega$.

We can rewrite the likelihood term based on the fact that each (x_n, v_n) pair is a deterministic transformation of $(\log A_n, \phi_n)$. Computing A_n and ϕ_n from the data using Equation (2),

$$\begin{aligned} p(x_n, v_n \,|\, \omega, m) &= \begin{vmatrix} \frac{\partial x_n}{\partial \log A_n} & \frac{\partial x_n}{\partial \phi_n} \\ \frac{\partial v_n}{\partial \log A_n} & \frac{\partial v_n}{\partial \phi_n} \end{vmatrix}^{-1} p(\log A_n, \phi_n) \\ &= \left| -A_n^2\,\omega\sin^2(\phi_n) - A_n^2\,\omega\cos^2(\phi_n) \right|^{-1} p(\log A_n, \phi_n) \\ &= \frac{1}{A_n^2\,\omega}\, p(\log A_n, \phi_n). \end{aligned}$$

This makes it easier to specify what the likelihood is. It isn't obvious (to me) what sorts of positions and velocities would result from taking a snapshot from a randomly oscillating system. However, I can express the belief that if the bodies formed independently and aren't interacting much then all the phases are independent and uniform:

$$p(\phi_n) = \frac{1}{2\pi}. \tag{3}$$

The prior on amplitudes is harder. For a single body I would pick a uniform prior on $\log A_n$ covering a large range. However, my joint prior for all the amplitudes is *not*

$$p\big(\{\log A_n\}_{n=1}^N\big) = \prod_n \frac{1}{\log A_{\max} - \log A_{\min}}. \tag{4}$$

If we were told some amplitudes we would change our beliefs about future amplitude. We wouldn't change our beliefs using the prior in Equation (4), which asserts that we know what distribution the amplitudes come from (we don't). Instead we need to build a hierarchical model and infer the distribution over amplitudes along with the properties of the black hole:

$$p\big(\{\log A_n\}_{n=1}^N\big) = \int \mathrm{d}\theta\; p(\theta) \prod_n p(\log A_n \,|\, \theta)$$

The family of possible densities $p(\log A_n \,|\, \theta)$ should be flexible to allow learning the true amplitude density given enough data. *For the purposes of this practical* I am asserting that you should use a mixture of two Gaussians.

$$\begin{aligned} p(\log A_n \,|\, \theta) \;&=\; \pi_1\,\mathcal{N}(\log A_n;\, \mu_1, \sigma_1^2) + (1-\pi_1)\,\mathcal{N}(\log A_n;\, \mu_2, \sigma_2^2) \\ \theta \;&\equiv\; \{\pi_1, \mu_1, \mu_2, \log\sigma_1, \log\sigma_2\}. \end{aligned} \tag{5}$$

Put broad uniform priors on all parameters in θ. Even though if you had more time you may want to add some more structure.

Having — finally — defined our model, the posterior distribution over all the unknowns is:

$$\begin{aligned} p(\log\omega, m, \theta \,|\, \{x_n, v_n\}) \;&\propto\; p(\log\omega, m, \theta, \{x_n, v_n\}) \\ &\propto\; p(\log\omega)\, p(m)\, p(\theta) \prod_{n=1}^{N} \left[\frac{1}{A_n^2\,\omega}\, p(\log A_n \,|\, \theta)\right]. \end{aligned} \tag{6}$$

where A_n is a function of the observations, see (2). Apart from $p(\log A_n \,|\, \theta)$, defined in (5), set all other priors to uniform distributions with sensible widths.

4 Markov chain Monte Carlo sampling

Write code to sample from the posterior distribution in Equation (6) using the provided MCMC code and `log_pstar.m`, which evaluates (6) up to a constant. This will require writing a function taking a single vector argument containing the parameters that you wish to sample. A handle to this function, `@function_name`, can then be passed to the samplers.

If you are keen to implement actual MCMC code yourself, here is an idea: implement a version of Metropolis that updates each parameter one at a time. The slice sampling code already updates one parameter at a time, while my Metropolis code adds a spherical perturbation to the whole parameter vector.

One thing you might want to think about is initialization. Depending on the sampler, you may have "burn in" problems if you set the first state of the Markov chain to be very far from the bulk of the posterior. Are there very simple ways of setting initial values by looking at the data or computing some simple statistics? Is good initialization necessary for this problem?

Use Matlab's `plot` to plot the time series of individual parameters and one parameter against another. You can also use `hist` to look at an approximation to the marginal posterior of a parameter. Create estimates with an expression of your uncertainty for ω and m.

5 Questions

MCMC sampling:

1. What is the effect of Metropolis's step-size parameter?
2. Is the way you initialize the chain critical for Metropolis and/or slice sampling?
3. What are the relative advantages of slice-sampling and Metropolis? Can you say good and bad things about both of them?
4. Why have I taken logs of quantities like ω and A? Need I have bothered? Does taking logs affect the model and/or the sampler?
5. The true values are $\omega = 875.2$ and $m = 31.79$. Are your posterior beliefs consistent with this? If not, what do you think went wrong with the sampling, modelling or both?

The model: for anyone who manages to get onto thinking about the model.

1. Do you believe the uniform independent priors for the phases, ϕ_n? How might you check whether this is reasonable for a given dataset, or evaluate alternatives?
2. Is the hierarchical prior necessary; is the prior of Equation (4) really so bad?
3. What effect does changing the details of the hierarchical model have? How might you check the assumptions made, or evaluate alternatives?
4. Various quantities have broad uniform priors. Is this reasonable? What would happen (if anything) in the limit where the widths of priors were made infinite?

References

J. Bovy, I. Murray, and D. W. Hogg. The gravitational force law in the solar system, 2009. arXiv:0903.5308.