A quick introduction to Bayesian MCMC

Based on pages 188-9, and 216-8 of Yang (2014)
Suppose we want to estimate parameter θ when the data x are given. We use Bayes's theorem

$$f(\theta|x) = \frac{f(\theta)f(x|\theta)}{f(x)} = \frac{f(\theta)f(x|\theta)}{\int f(\theta)f(x|\theta)\,\mathrm{d}\theta}, \quad (1)$$

where $f(\theta)$ is the *prior distribution*, $f(x|\theta)$ is the likelihood (the probability of data x given parameter θ), and $f(\theta|x)$ is the *posterior distribution*. The *marginal probability* of the data, $f(x)$, is a normalizing constant, to make $f(\theta|x)$ integrate to one. Equation (1) thus says that the posterior is proportional to the prior times the likelihood.

JC distance between two sequences (Example 7.1 p.216, Yang 2014 Molecular Evolution: A Statistical Approach). Consider the use of the JC69 model to estimate the distance θ between the human and orangutan 12S rRNA genes from the mitochondrial genome. The data are summarized as $x = 90$ differences out of $n = 948$ sites. The MLE was found to be $\hat{\theta} = 0.1015$, with the 95% confidence interval to be (0.0817, 0.1245). To apply the Bayesian approach, we specify an exponential prior with mean $\mu = 0.2$, with density

$$f(\theta) = \tfrac{1}{\mu}\mathrm{e}^{-\theta/\mu}. \quad (2)$$

The probability of the data (that is, the likelihood) is given by the binomial probability as

$$f(x|\theta) = p^x(1-p)^{n-x} = (\tfrac{3}{4} - \tfrac{3}{4}\mathrm{e}^{-4\theta/3})^x(\tfrac{1}{4} + \tfrac{3}{4}\mathrm{e}^{-4\theta/3})^{n-x}, \quad (3)$$

where $p = (\tfrac{3}{4} - \tfrac{3}{4}\mathrm{e}^{-4\theta/3})$ is the probability of difference at the site according to the JC69 model.

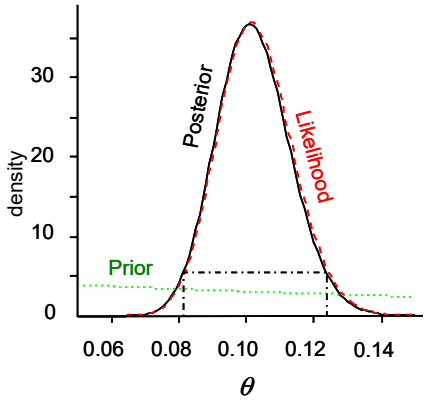


Figure 6.3 from Yang (2014). Prior and posterior densities for sequence distance θ under the JC69 model. The likelihood is shown as well, rescaled to match up the posterior density. Note that the posterior density is the prior density times the likelihood, followed by a change of scale to make the area under the posterior density curve equal to one. The data analyzed here are the human and orangutan mitochondrial 12S rRNA genes, with $x = 90$ differences at $n = 948$ sites. The 95% HPD interval, (0.08116, 0.12377), is indicated on the graph.

To apply equation (**1**), we calculate the integral in the denominator numerically, to give $f(x) = 5.16776 \times 10^{-131}$. Then equation (**1**) is straightforward to use, with the prior and likelihood given in (2) and (3). The figure below shows the posterior density $f(\theta|x)$, plotted together with the prior and scaled likelihood. The posterior looks very similar to the likelihood, and the prior is nearly flat at the neighborhood of the peak of the likelihood. This is because the dataset is large. The posterior mean is found by numerical integration to be 0.10213, with standard deviation 0.01091. The mode is at $\theta =$ 0.10092, and the 95% equal-tail credibility interval (CI) is (0.08191, 0.12463).

Bayesian computation - Markov chain Monte Carlo

For most problems, the prior $f(\theta)$ and the likelihood $f(\theta|x)$ are easy to calculate, but the marginal probability of the data $f(x)$, *i.e.*, the normalizing constant, is hard to calculate, as it involves integrals often of high dimensions. The breakthrough is the development of Markov chain Monte Carlo (MCMC) algorithms, which have made modern Bayesian inference computationally feasible.

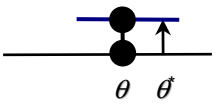
Instead of calculating the posterior density $f(\theta|x)$ analytically, as we did in the JC distance example above, an MCMC algorithm generates a sample from this distribution. Suppose the sample

is $\theta_1, \theta_2, \ldots, \theta_N$. One can then use it to calculate the posterior mean, standard deviation, etc. and construct the 95% interval (called the 95% credibility interval). Also one construct a histogram and smoothen it to generate an estimate of the whole posterior density $f(\theta|x)$.

Here we construct an MCMC algorithm to estimate the sequence distance under the JC69 model. The data are again $x = 90$ differences out of $n = 948$ sites. The prior is $f(\theta) = \frac{1}{\mu} e^{-\theta/\mu}$, with $\mu = 0.2$. As the posterior $f(\theta|x)$ is a function of θ, with the data x given, it is convenient to write $\pi(\theta) = f(\theta|x)$. We use a sliding window of size w around the current value θ to propose a new value θ^*.

MCMC algorithm (Example 7.1 p.216, Yang 2014):

1. Initialize: set $\theta = 0.05$, say.
2. (Proposal) Propose a new state from a sliding window: $\theta^* \sim U(\theta - w/2, \theta + w/2)$. That is, generate a $U(0, 1)$ random number r, and set $\theta^* = \theta - w/2 + wr$. If $\theta^* < 0$, set $\theta^* = -\theta^*$.



3. (Accept or reject) If $\pi(\theta^*) > \pi(\theta)$, accept the proposal; otherwise accept it with probability $\frac{\pi(\theta^*)}{\pi(\theta)}$. If the proposal is accepted, set $\theta \leftarrow \theta^*$; otherwise θ is unchanged. Another way of describing this step is to accept the proposal with probability

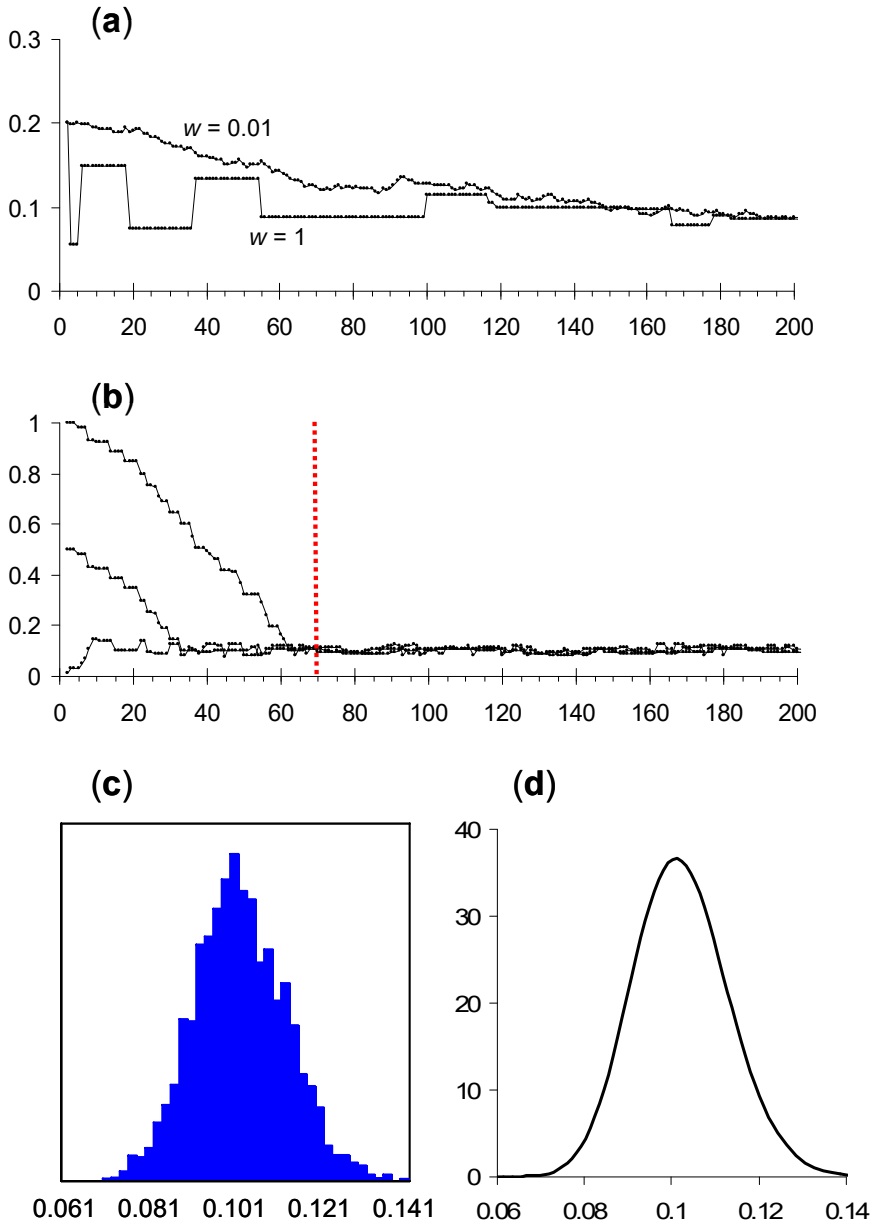
$$\alpha = \min\left(1, \frac{\pi(\theta^*)}{\pi(\theta)}\right). \quad (4)$$

4. Print out θ.
5. Go to step 2.

Note about the R program. The above algorithm is implemented in a small R program, listed at the end of this document. Please look at the code. Install R and RStudio on your computer and copy the code into R studio. Experiment with different starting values and window sizes. Uncomment the last line of the code to plot a histogram.

Further reading

Chapters 6 & 7 in Yang 2014 *Molecular Evolution: A Statistical Approach*, OUP.



MCMC runs for estimating sequence distance θ under the JC69 model. The data consists of $x = 90$ differences between two sequences out of $n = 948$ sites. (**a**) Two chains with the window size either too small ($w = 0.01$) or too large ($w = 1$). Both chains started at $\theta = 0.2$. The chain with $w = 0.01$ has an acceptance proportion of 91%, so that almost every proposal is accepted. However, this chain takes tiny baby steps and does not explore the space well. The chain with $w = 1$ has the acceptance proportion 7%, so that most proposals are rejected. The chain often stays at the same state for many iterations without a move. Further experiment shows that the window size $w = 0.1$ leads to an acceptance rate of 35%, and is near optimum. (**b**) Three chains started from $\theta = 0.01$, 0.5, and 1, with window size 0.1. It appears that after about 70 iterations, the three chains become indistinguishable and have reached stationarity, so that a *burn-in* of 100 iterations appears sufficient for those chains. (**c**) Histogram estimation constructed from 10,000 iterations. (**d**) Posterior density obtained from a long chain of 10,000,000 iterations, sampling every 10 iterations, estimated using a kernel density smoothing algorithm. (This is figure 7.2 in Yang 2014 *Molecular Evolution: A Statistical Approach*, OUP)