SAMSI Astrostatistics Tutorial

More Markov chain Monte Carlo & Demo of Mathematica software

Phil Gregory

University of British Columbia

2006
Bayesian Logical Data Analysis for the Physical Sciences

Contents:

1. Role of probability theory in science
2. Probability theory as extended logic
3. The how-to of Bayesian inference
4. Assigning probabilities
5. Frequentist statistical inference
6. What is a statistic?
7. Frequentist hypothesis testing
8. Maximum entropy probabilities
9. Bayesian inference (Gaussian errors)
10. Linear model fitting (Gaussian errors)
11. Nonlinear model fitting
12. Markov chain Monte Carlo
13. Bayesian spectral analysis
14. Bayesian inference (Poisson sampling)

Appendix A. Singular value decomposition
Appendix B. Discrete Fourier Transform
Appendix C. Difference in two samples
Appendix D. Poisson ON/OFF details
Appendix E. Multivariate Gaussian from maximum entropy.
Outline

1. Introduce MCMC and parallel tempering
2. Technical difficulties
3. Return to simple spectral line problem
4. Tests for convergence
5. Exoplanet examples
6. Model comparison (global likelihoods)
In a Bayesian analysis we need to perform a lot of integration.

Example 1: to find the marginal posterior probability density for the orbital period, $P$, we need to integrate the joint posterior over all the other parameters.

$$p(P \mid D, M_1, I) = \int dK dV d\chi d\varepsilon d\omega \ p(P, K, V, \chi, \varepsilon, \omega \mid D, M_1, I)$$

Example 2: to find the posterior probability of a model we need to evaluate the global likelihood of the model $p(D \mid M_1, I)$. This requires an integral over all the model parameters.

$$p(D \mid M_1, I) = \int dP dK dV d\chi d\varepsilon d\omega \ p(P, K, V, \chi, \varepsilon, \omega \mid M_1, I) \ p(D \mid M_1, P, K, V, \chi, \varepsilon, \omega, I)$$

Markov chain Monte Carlo (MCMC) algorithms provide a powerful means for efficiently computing integrals in many dimensions.
MCMC to the rescue

In straight Monte Carlo integration independent samples are randomly drawn from the volume of the parameter space. The price that is paid for independent samples is that too much time is wasted sampling regions where posterior probability density is very small.

Suppose in a one-parameter problem the fraction of the time spent sampling regions of high probability is $10^{-1}$. Then in an M-parameter problem, this fraction could easily fall to $10^{-M}$.

MCMC algorithms avoid the requirement for completely independent samples, by constructing a kind of random walk in the model parameter space such that the number of samples in a particular region of this space is proportional to the posterior density for that region.

The random walk is accomplished using a Markov chain, whereby the new sample, $X_{t+1}$, depends on previous sample $X_t$ according to an entity called the transition probability or transition kernel, $p(X_{t+1} | X_t)$. The transition kernel is assumed to be time independent.

The remarkable property of $p(X_{t+1} | X_t)$ is that after an initial burn-in period (which is discarded) it generates samples of $X$ with a probability density equal to the desired posterior $p(X|D,I)$.
A simple Metropolis-Hastings MCMC algorithm

\[ P(X|D,M,I) = \text{target posterior probability distribution} \]

(X represents the set of model parameters)

1. Choose \( X_0 \) an initial location in the parameter space. Set \( t = 0 \).

2. Repeat \{ \}

   \begin{itemize}
   \item Obtain a new sample \( Y \) from a proposal distribution \( q(Y|X_t) \) that is easy to evaluate. \( q(Y|X_t) \) can have almost any form.
   \item Sample a Uniform \((0, 1)\) random variable \( U \).
   \item If \( U \leq \frac{p(Y|D,I)}{p(X_t|D,I)} \times \frac{q(X_t|Y)}{q(Y|X_t)} \), then set \( X_{t+1} = Y \)
   \item Otherwise set \( X_{t+1} = X_t \)
   \item Increment \( t \)
   \end{itemize}

This factor = 1 for a symmetric proposal distribution like a Gaussian
Conditions for convergence

Remarkably, for a wide range of proposal distributions $q(Y \mid X)$, the Metropolis-Hastings algorithm generates samples of $X$ with a probability density which converges on the desired target posterior $p(X \mid D, I)$, called the stationary distribution of the Markov chain.

For the distribution of to converge to a stationary distribution, the Markov chain must have three properties (Roberts, 1996).

1. It must be irreducible. That is, from all starting points, the Markov chain must be able (eventually) to jump to all states in the target distribution with positive probability.

2. It must be aperiodic. This stops the chain from oscillating between different states in a regular periodic movement.

3. It must be reversible.

$$
p(X' \mid D, I) \ p(X \mid X') = p(X' \mid D, I) \ p(X' \mid X)$$
A comparison of the samples from three Markov Chain Monte Carlo runs using Gaussian proposal distributions with differing values of the standard deviation.

In this example the posterior probability distribution consists of two 2 dimensional Gaussians.
A comparison of the autocorrelation functions for three Markov Chain Monte Carlo runs using Gaussian proposal distributions with differing values of the standard deviation.

Bayesian Logical Data Analysis for the Physical Sciences © Cambridge University Press 2005
The simple Metropolis-Hastings MCMC algorithm can run into difficulties if the probability distribution is multi-modal with widely separated peaks. It can fail to fully explore all peaks which contain significant probability, especially if some of the peaks are very narrow.

One solution is to run multiple Metropolis-Hastings simulations in parallel, employing probability distributions of the kind

$$
\pi (X | D, M, \beta, I) = p (X | M, I) \ p (D | X, M, I)^{\beta} \quad (0 < \beta \leq 1)
$$

$\beta = 1$ corresponds to our desired target distribution. The others correspond to progressively flatter probability distributions.

I learned about parallel tempering from this book.

At intervals, a pair of adjacent simulations are chosen at random and a proposal made to swap their parameter states. The update can be accepted/rejected with a Metropolis-Hastings criterion. At time $t$, simulation $\beta_i$ is in state $X_{t,i}$ and simulation $\beta_{i+1}$ is in state $X_{t,i+1}$. If the swap is accepted by the test given below then these states are interchanged. Accept the swap with probability

$$r = \min \left\{ 1, \frac{\pi(X_{t,i+1}|D, \beta_i, I) \pi(X_{t,i}|D, \beta_{i+1}, I)}{\pi(X_{t,i}|D, \beta_i, I) \pi(X_{t,i+1}|D, \beta_{i+1}, I)} \right\}$$

The swap allows for an exchange of information across the ladder of simulations.

In the low $\beta$ simulations, radically different configurations can arise, whereas at higher $\beta$, a configuration is given the chance to refine itself.

Final results are based on samples from the $\beta = 1$ simulation. Samples from the other simulations can be used to evaluate the Bayes Factor in model selection problems.
Technical Difficulties

1. Deciding on the burn-in period.

2. Choosing a good choice for the characteristic width of each proposal distribution, one for each model parameter.

   For Gaussian proposal distributions this means picking a set of proposal \( \sigma \)'s. This can be very time consuming for a large number of different parameters.

3. Deciding how many iterations are sufficient.

4. Deciding on a good choice of tempering levels (\( \beta \) values).
Schematic of a Bayesian Markov Chain Monte Carlo program for nonlinear model fitting. The program incorporates a control system that automates the selection of Gaussian proposal distribution σ's.
Return to simple spectral line problem

Now assume 4 unknowns: line center frequency, $T$, line width, and an extra noise term with unknown standard deviation

Parameters $f$, $T$, $lw$, $s$

All channels have Gaussian noise characterized by $\sigma = 1$ mK. The noise in separate channels is independent. The line center frequency $\nu_0 = 37$. 

Bayesian Logical Data Analysis for the Physical Sciences © Cambridge University Press 2005
Tests for convergence

1. Examine plots of the MCMC iterations for each parameter.

2. Divide the post burn-in iterations into two halves and plot one on top of the other using different colors.

3. Compute the Gelman-rubin statistic for each parameter from the repeated runs.
Tests for convergence

Compare iterations from multiple chains

For each parameter compute Gelman-Rubin statistic

GR should be close to 1.0

Measured GR values < 1.02
Gelman-Rubin Statistic

Let $\theta$ represent one of the model parameters.

Let $\theta^i_j$ represent the $i^{th}$ iteration of the $j^{th}$ chain.

Extract the last $\eta$ post burn-in iterations for each simulation.

Mean within chain variance $W = \frac{1}{m(\eta - 1)} \sum_{j=1}^{m} \sum_{i=1}^{\eta} (\theta^i_j - \bar{\theta}_j)^2$

Between chain variance $B = \frac{\eta}{m - 1} \sum_{j=1}^{m} (\bar{\theta}_j - \bar{\theta})^2$

Estimated variance $\hat{V}(\theta) = \left(1 - \frac{1}{\eta}\right)W + \frac{1}{\eta}B$

Gelman – Rubin statistic $= \sqrt{\frac{\hat{V}(\theta)}{W}}$

The Gelman – Rubin statistic should be close to 1.0 (e.g. < 1.05) for all parameters for convergence

HD 73526 MCMC orbital parameter iterations

Evidence for 3 periods
THREE LOW-MASS PLANETS FROM THE ANGLO-AUSTRALIAN PLANET SEARCH

C. G. Tinney,² R. Paul Butler,³ Geoffrey W. Marcy,⁴,⁵ Hugh R. A. Jones,⁶,⁷ Alan J. Penny,⁸,⁹ Chris McCarthy,³,⁵ Brad D. Carter,¹⁰ and Debra A. Fischer⁴,⁵

Data

Evidence for a second planet in HD 208487

Gregory, P. C. (2005), AIP Conference Proceeding 803, p. 139, 2005

Best fit $P_1 = 129.517$; $K_1 = 15.9778$; $V = 7.92451$; $\chi_1 = 0.184585$; $e_1 = 0.224581$; $\omega_1 = 2.20511$;
$P_2 = 998.085$; $K_2 = 9.81886$; $\chi_2 = 0.723248$; $e_2 = 0.189448$; $\omega_2 = 2.74554$

RMS residual = 4.2 m s$^{-1}$

$\chi^2_v = 0.83$
HD 208487

$P_1 = 129.52 \text{ d}$
$e_1 = 0.22$

$P_2 = 998 \text{ d}$
$e_2 = 0.19$

Model compare
Model selection

To answer the model selection question, we compute the odds ratio (abbreviated simply by the odds) of model $M_1$ to model $M_2$.

Expand numerator and denominator with Bayes’ theorem

\[
O_{12} = \frac{p(M_1 | D, I)}{p(M_2 | D, I)} = \frac{\frac{p(M_1 | I) p(D | M_1, I)}{p(D | I)}}{\frac{p(M_2 | I) p(D | M_2, I)}{p(D | I)}} = \frac{p(M_1 | I)}{p(M_2 | I)} \cdot \frac{p(D | M_1, I)}{p(D | M_2, I)}
\]

posterior probability ratio

prior probability ratio

Bayes factor

$p(D | M_1, I)$, the called the global likelihood of $M_1$.

\[
p(D | M_1, I) = \int_T p(D, T | M_1, I) dT
\]

Expanded with product rule

\[
= \int_T p(T | M_1, I) p(D | M_1, T, I) dT
\]

The global likelihood of a model is equal to the weighted average likelihood for its parameters.
Model selection

One way to compute the global likelihood of the mode, $\log[p(D | M_1, I)]$.

$$\log[p(D | M_1, I)] = \int \langle \ln p(D | M, P, K, V, \chi, e, \omega, I) \rangle_{\beta} d\beta$$

$$\langle \ln p(D | M, P, K, V, \chi, e, \omega, I) \rangle_{\beta}$$

= expectation value of $\ln[p(D | M, P, K, V, \chi, e, \omega, I)]$ for given $\beta$

$$= \frac{1}{n} \sum_{t=1}^{n} \ln[p(D | M_1, P_{t,\beta}, K_{t,\beta}, V_{t,\beta}, \chi_{t,\beta}, e_{t,\beta}, \omega_{t,\beta}, I)]$$

where $n =$ number of MCMC iterations.

See my book for derivation.
One way to compute the global likelihood of the model, \( \log[p(D | M1, I)] \).

\[
\begin{align*}
\log[p(D | M1, I)] &= \int \langle \ln p(D | M_E, P, K, V, \chi, e, \omega, I) \rangle_\beta \, d\beta \\
\langle \ln p(D | M_E, P, K, V, \chi, e, \omega, I) \rangle_\beta &= \text{expectation value of } \ln [p(D | M_E, P, K, V, \chi, e, \omega, I)] \text{ for given } \beta \\
&= \frac{1}{n} \sum_{t=1}^{n} \ln[p(D | M1, P_t, K_t, V_t, \chi_t, e_t, \omega_t, I)]
\end{align*}
\]

where \( n = \text{number of MCMC iterations.} \)

30 tempering levels
p(M_{1j}|D,I) = 1.4 \times 10^{-55} \quad \text{(from tempering levels)}
\[\text{Blow-up}\]
p(M_{1j}|D,I) = 2.5 \times 10^{-55} \quad \text{(from restricted Monte Carlo integration)}
p(M_{0s}|D,I) = 1.5 \times 10^{-59} \quad \text{(from numerical integration)}