# An introduction to Markov Chain Monte Carlo techniques

G. J. A. Harker

University of Colorado

ASTR5550, 19th March 2012

# Outline

Introduction
- Bayesian inference: recap
- MCMC: when to use it and why
- A simple MCMC algorithm

Example: extracting a 21-cm signal from the Dark Ages
- Simulating the data
- Building the likelihood function for *DARE*
- Results

Subtleties and extensions
- Initial values and burn-in
- Convergence and termination
- Dealing with shortcomings of MCMC

# Bayes' theorem in suitable notation

$$\Pr(\boldsymbol{\Theta}|\boldsymbol{D},H) = \frac{\Pr(\boldsymbol{D}|\boldsymbol{\Theta},H)\Pr(\boldsymbol{\Theta}|H)}{\Pr(\boldsymbol{D}|H)}$$

- $\boldsymbol{\Theta}$ is a vector of $N$ parameters defined within some model (or hypothesis) $H$.
- $\boldsymbol{D}$ is a vector storing the data.
- $\Pr(\boldsymbol{\Theta}|\boldsymbol{D},H)$ is the *posterior* probability distribution.
- $\Pr(\boldsymbol{D}|\boldsymbol{\Theta},H)$ is the *likelihood*, $\mathcal{L}(\boldsymbol{\Theta})$.
- $\Pr(\boldsymbol{\Theta}|H)$ is the *prior* probability distribution, $\pi(\boldsymbol{\Theta})$.
- $\Pr(\boldsymbol{D}|H)$ is known as the *evidence*, $\mathcal{Z}$.
- $\mathcal{Z} = \int \pi(\boldsymbol{\Theta})\mathcal{L}(\boldsymbol{\Theta})\mathrm{d}^N\boldsymbol{\Theta}$

# Monte Carlo methods: when to use them and why

- Monte Carlo methods can be used to map out the posterior, which often can't be computed analytically.
- With this probability distribution in hand we can compute a number of useful quantities, such as mean values and errors on arbitrary functions of the parameters.
- Monte Carlo methods can help overcome the 'curse of dimensionality': to compute the likelihood on a grid with $k$ points on a side, we need $k^N$ computations of the likelihood.
  - Even small $k$ or $N$ can become problematic if the likelihood is costly to compute.
  - Unfortunately this is often the case, e.g. need to run CMBFast to compute a power spectrum in CMB experiments.
- Ideally we want to build up our picture of the posterior from independent, random draws.

# When and why to use MCMC specifically?

- Independent, random draws often can't be achieved in practice, or would be very inefficient (e.g. rejection sampling) for the current problem.
- This often occurs with complicated distributions or large numbers of dimensions.
- MCMC can be used to generate dependent draws which are approximately from the desired distribution.
- MCMC is actually a big family of methods and at least one will likely be suitable for your problem, but often the simple (and popular) Metropolis-Hastings algorithm will be fine.
- The posterior then only needs to be evaluated up to some normalizing constant.

# General approach: a random walk through parameter space

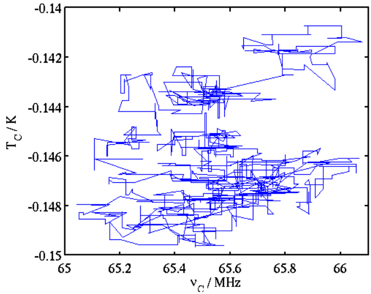


Figure: A two-dimensional slice through parameter space.

- Build up a chain of positions in parameter space, $\mathbf{\Theta}_1$, $\mathbf{\Theta}_2$,. . .
- Markov chain: probability density that the $n^{\text{th}}$ point is at $\mathbf{\Theta}_n$ depends only on $\mathbf{\Theta}_{n-1}$ and not $\mathbf{\Theta}_{n-2}$, $\mathbf{\Theta}_{n-3}$, . . .
- This transition probability, $T(\mathbf{\Theta}_i|\mathbf{\Theta}_{i-1})$, is chosen so that the points in the chain are samples from the approximate distribution we wish to find.
- There are a number of different algorithms which

# Properties of the chain

- If the chain is:
  - Irreducible (all possible states can be reached from any point in the chain);
  - Aperiodic (doesn't get trapped in cycles)

  then it converges to some invariant distribution.
- The aim is to make that equilibrium distribution, $p(\boldsymbol{\Theta})$, the one we want to find.
- In equilibrium, detailed balance is satisfied:

$$p(\boldsymbol{\Theta}_{n+1})T(\boldsymbol{\Theta}_{n+1}|\boldsymbol{\Theta}_n) = p(\boldsymbol{\Theta}_n)T(\boldsymbol{\Theta}_n|\boldsymbol{\Theta}_{n+1})$$

# Metropolis-Hastings Algorithm

- Choose an arbitrary *proposal density*, $q(\mathbf{\Theta}_n, \mathbf{\Theta}_{n+1})$, from which are drawn suggestions for new positions in the chain.
- A proposal for the next position in the chain is accepted with probability

$$\alpha(\mathbf{\Theta}_n, \mathbf{\Theta}_{n+1}) = \min\left\{1, \frac{\Pr(\mathbf{\Theta}_{n+1}|\boldsymbol{D}, H)q(\mathbf{\Theta}_{n+1}, \mathbf{\Theta}_n)}{\Pr(\mathbf{\Theta}_n|\boldsymbol{D}, H)q(\mathbf{\Theta}_n, \mathbf{\Theta}_{n+1})}\right\}$$

- Otherwise, the chain stays where it is.
- This ensures that the equilibrium distribution is our posterior probability distribution.
- If $\mathbf{\Theta}_n$ is a sample from the the equilibrium distribution, then detailed balance holds.
- A symmetric $q$ is often used, simplifying the expression for $\alpha$.

# Metropolis-Hastings Algorithm

- Detailed balance only holds when the initial point is already a sample from the equilibrium distribution. The chain can take a while to reach equilibrium, so early points are discarded (the ‘burn-in’ phase).
- Successive links in the chain are correlated: to obtain approximately independent samples, often only one in every $n_{\mathrm{thin}}$ samples is retained (‘thinning’ the chain).
- A simple Metropolis-Hastings algorithm (Metropolis et al. 1953; Hastings 1970) can be run in an embarrassingly parallel way, since several chains can be run at the same time and they don’t have to communicate with each other.
- There are a variety of other samplers available: Gibbs (need to be able to compute conditional PDFs), slice sampling, Hamiltonian Monte Carlo (more complicated), and many more, but Metropolis-Hastings is generally applicable and usually OK.

# Choosing a proposal distribution

- One diagnostic for how well the chain is exploring parameter space is the *acceptance ratio*, *a*, the fraction of proposed moves which are accepted.
  - This should be of order a few tens of percent.
  - Too low: have to perform many likelihood computations for each new link.
  - Too high: adjacent links will be highly correlated; more thinning needed; may find it hard to move to other peaks (tough in general for vanilla MCMC).

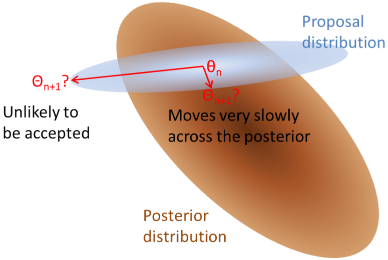


Figure: Ideally, the proposal distribution should be matched in shape and size to the posterior distribution.

- The main factor influencing *a* is the choice of *q*.

# Choosing a proposal distribution

- Common to choose a multidimensional Gaussian centred on the current $\mathbf{\Theta}$ as the proposal distribution, so that setting $q$ just involves setting the covariance matrix of this Gaussian.
- For problems in small numbers of dimensions, it may be good enough to set the width of the Gaussian in different directions by hand, either heuristically or using some knowledge about the likely shape of the posterior.
- Otherwise, it's sensible to try to find the covariance matrix of $q$ more automatically, e.g. from the covariance of posterior samples from an early part of the run. This also means that a general move is not along parameter axes (otherwise, normally just change some random subset of the parameters in each step).
- Cannot update $q$ too often as the algorithm is running, as this breaks the 'Markov' property of the Markov chain.

# Example: extracting a 21-cm signal from the Dark Ages

- A number of experiments are attempting to detect highly redshifted 21-cm radiation from the epoch of reionization and the cosmic dark ages.
- This will allow the effect of the first stars, galaxies and black holes on the surrounding IGM to be studied.
- Need to look at low radio frequencies, $\lesssim 200\,\text{MHz}$.
- The *Dark Ages Radio Explorer (DARE)* looks at 40–120 MHz ($z \sim$ 11–34) from lunar orbit.
- Many other sources of radiation are present at these frequencies.
- These large foregrounds must be modelled and removed accurately; this also puts stringent requirements on the calibration of the instrument.

# A model for the sky-averaged 21-cm signal

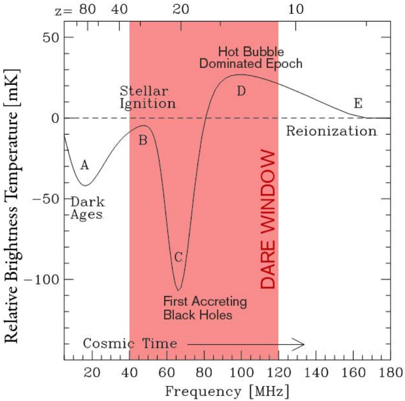


Figure: Reference model for the sky-averaged 21-cm signal

- The sky-averaged 21-cm signal depends on the Lyman-$\alpha$ and X-ray radiation fields.
  - Ly-$\alpha$ couples 21-cm 'spin temperature' to the gas temperature.
  - X-rays heat the gas.
  - Energetic photons ionize hydrogen
- The spectrum has a few key 'turning points' (A,B,C,D,E): positions of these are used to parametrize the different possible histories.

# Foregrounds for sky-averaged 21-cm experiments

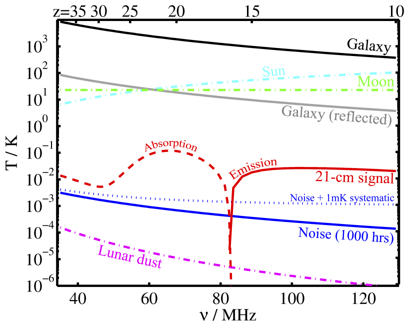


Figure: Foregrounds and noise for a *DARE*-like experiment, compared to the expected signal

- Several foregrounds dominate the signal and must be carefully modelled.
  - Synchrotron, free-free etc. from our Galaxy.
  - A sea of unresolved extragalactic sources.
  - The Sun and the thermal emission of the Moon.
  - Reflections of other foregrounds from the Moon.
- Some other foregrounds are neglected here: Jupiter, plasma from impacts of dust on the antenna, . . .

# A synthetic dataset from the *Dark Ages Radio Explorer*

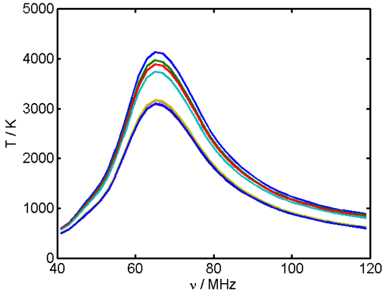


Figure: Eight noisy spectra, each assuming 375 hours of integration

- The spectra turn over at low frequencies: not physical, but owing to the instrument response.
- This must also be modelled! Tiny errors in assumed instrument response drown the signal when convolved with huge foregrounds.
- To achieve required accuracy, must fit the instrument model along with the signal and the foregrounds from the data.

# Building the likelihood function for *DARE*

- Likelihood function depends on a large number of parameters, and must be computed numerically.
- We want good estimates of the science parameters and the errors on them, and to study correlations between parameters.
- Ideally suited to MCMC!
- For computational reasons, in fact we always work with $\log \mathcal{L}$, $\log \pi$, etc. Ranging over the prior space, the un-logged posterior probability varies by many orders of magnitude.

# Building the likelihood function for *DARE*

- We assume that *DARE* collects spectra between 40 and 120 MHz in eight independent sky regions (with different foregrounds, but the same 21-cm signal and instrument).
- Signal:
  - Six parameters, $\{\nu_{\{B,C,D\}}, T_{\{B,C,D\}}\}$, the position in frequency and temperature of the three turning points in the DARE band.
  - Interpolate between these with a cubic spline, giving $T_{\rm sky}(\nu)$.
- Foregrounds:
  - Diffuse foregrounds: log $T$ is a third-order polynomial in log $\nu$ in each sky region.

    $$\log T^i_{\rm FG} = \log T^i_0 + a^i_1 \log(\nu/\nu_0) + a^i_2 [\log(\nu/\nu_0)]^2 + a^i_3 [\log(\nu/\nu_0)]^3 \,,$$

    $i = 1, 2, \ldots, 8.$

# Building the likelihood function for *DARE*

- Foregrounds, contd.:
  - Sun: different normalization but same shape in each sky region:

    $$\log T^{i}_{\text{Sun}} = \log T_0^{i,\text{Sun}} + a_1^{\text{Sun}} \log(\nu/\nu_0) + a_2^{\text{Sun}} [\log(\nu/\nu_0)]^2 + a_3^{\text{Sun}} [\log(\nu/\nu$$

    $i = 1, 2, \ldots, 8.$
  - Moon: $T_{\text{Moon,eff}}(\nu) = $ const., around 230 K attenuated by the backlobe of the *DARE* beam to an effective $\sim 23$ K.
- Total sky spectrum

$$T^{i}_{\text{sky}}(\nu) = \left[ T_{\text{sig}}(\nu) + T^{i}_{\text{FG}}(\nu) \right] [1 + r_{\text{Moon}}] + T_{\text{Moon,eff}} \ ,$$

$i = 1, 2, \ldots, 8$, where $r$ is the reflectivity of the Moon.

# Effect of the instrument on the spectrum

- Expected antenna temperature

$$T_{\mathrm{ant}}(\nu) = \left[1 - |\Gamma(\nu)|^2\right] T_{\mathrm{sky}}(\nu) \\ + \left[1 + 2\epsilon|\Gamma(\nu)| \cos(\beta(\nu)) + |\Gamma(\nu)|^2\right] T_{\mathrm{rcv}}\ .$$

- $\Gamma(\nu)$ is the complex reflection coefficient, having amplitude $|\Gamma(\nu)|$ and phase $\beta(\nu)$.
- $|\Gamma(\nu)|$ and $\beta(\nu)$ are parametrized using the first ten coefficients of their discrete cosine transform. Higher-frequency components are assumed not present as the response has been assumed to be smooth.
- $T_{\mathrm{rcv}}$ is the receiver temperature, assumed constant.
- $\epsilon$ is a correlation coefficient.

# Modelled *DARE* instrument response, and the parameters of the model

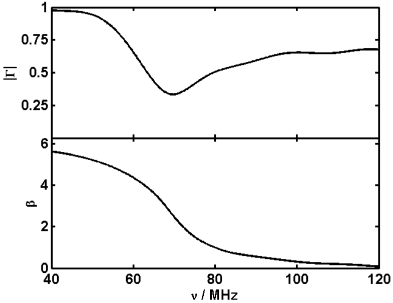


Figure: Amplitude and phase of the reflection coefficient as a function of frequency

Table: Summary of the parameter space

| Parameter group | No. of params. |
|---|---|
| 21-cm signal | $3 \times 2 = 6$ |
| Diffuse foregrounds | $4 \times 8 = 32$ |
| Sun | $8 + 3 = 11$ |
| Moon | 2 |
| Instrument | 22 |
| Total | 73 |

# Putting it all together: the *DARE* likelihood function

- The thermal noise on the spectrum is Gaussian, and its RMS is predicted with the radiometer equation:

$$\sigma(\nu) = \frac{T_{\mathrm{ant}}(\nu)}{\sqrt{2Bt}}$$

- $B$ is the width of a frequency channel (effective width, if some parts are discarded because of radio recombination lines, etc.), typically $\sim$ 1 MHz.
- $t$ is the integration time (typically a few hundred hours for these global signal experiments).
- Factor of $\sqrt{2}$ comes because we assume a crossed pair of dipoles.

# Putting it all together: the *DARE* likelihood function

- Probability density of measuring the temperature $T^i_{\rm meas}(\nu_j)$ if the noise-free antenna temperature is $T^i_{\rm ant}(\nu_j|\mathbf{\Theta})$ is

$$p_{ij} = \frac{1}{\sqrt{2\pi\sigma_i^2(\nu_j|\mathbf{\Theta})}} e^{-[T^i_{\rm meas}(\nu_j) - T^i_{\rm ant}(\nu_j|\mathbf{\Theta})]^2/2\sigma_i^2(\nu_j|\mathbf{\Theta})} .$$

- Assuming each sky area and frequency channel is independent, then

$$\mathcal{L}(\boldsymbol{T}_{\rm meas}|\mathbf{\Theta}) = \prod_{i=1}^{n_{\rm areas}} \prod_{j=1}^{n_{\rm freq}} p_{ij} .$$

- $\boldsymbol{T}_{\rm meas}$ is the noisy spectrum generated using the ‘true’ parameters (our synthetic, ‘measured’ dataset), and $T^i_{\rm ant}(\nu_j|\mathbf{\Theta})$ can be computed as before assuming any $\mathbf{\Theta}$.

# Results: constraints on the position of the turning points and the shape of the signal

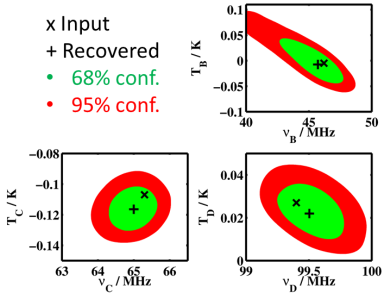


Figure: Marginalizing can give constraints on individual parameters or subsets of parameters.

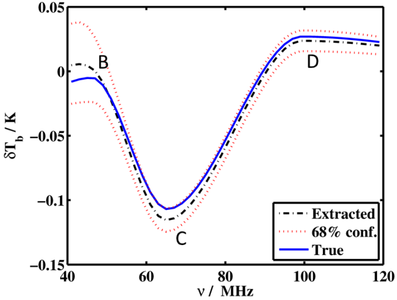
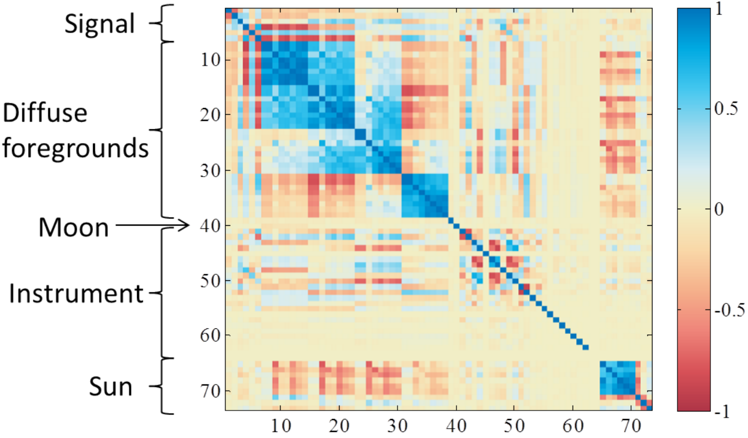


Figure: Rigorous constraints can also be found on arbitrary functions of the parameters; don't just have to do 'error propagation'.

# Scaled covariance matrix of the parameters



# Choosing initial values and burn-in

- One strategy for deciding where to start the chains is just to pick random points within the prior volume. This can be problematic as they might take a while to find the region of high likelihood.
- A standard minimization routine (see earlier in this course) can usually find the high-likelihood region more efficiently, then you can start the chains off from there. This might fail for multimodal functions.
- To work out how many 'burn-in' points to discard (often $\sim$ a few thousand), easiest is just to look at the chains: it's usually good strategy to plot things as you go along anyway, just to make sure everything's going OK.
- Might need to discard fewer points if the chains start out in the high-density region.

# Deciding how many points is enough

- If the chain is mixing well, then you might not need that many samples of the equilibrium distribution (i.e. no more than a few hundred, if $N$ is small) to compute the quantities you want with reasonable accuracy. So how do you know if the chain is well mixed?
- Visual inspection of diagnostic plots:
    - Iteration number vs. parameter value: shouldn't get stuck in one area for a while and then jump around.
    - Running mean of a parameter: should be converging nicely to some value.
    - Autocorrelation of the parameter value along the chain: this should not remain high at large lag.
    - This might also help you choose the amount of thinning, which otherwise is usually chosen heuristically (keeping somewhere between one draw in every few draws, and one draw in every 100).

# Deciding how many points is enough: a more quantitative method

- A variety of diagnostics are available to check if your chains are converged, but none is perfect.
- The Gelman-Rubin test (Gelman & Rubin 1992) is popular and straightforward, but requires you to run multiple chains (you'll usually want to be doing this anyway).
- The idea is to compare the variance within each chain to the variance between chains (after discarding the first half of the chain). The between-chain variance shouldn't be too large if all the chains are well converged.

# The Gelman-Rubin test in more detail

- Start with $m$ chains of length $2n$, discard the first $n$ points from each, and then for each parameter:
  - Compute $B/n$, the variance between the $m$ sequence means;
  - Compute $W$, the average of the $m$ within-chain sample variances;
  - Estimate the target variance as $\hat{\sigma}^2 = \frac{n-1}{n} W + \frac{B}{n}$;
  - Find the 'potential scale reduction factor' $\hat{R} = \sqrt{\hat{\sigma}^2/W}$.
- We need $\hat{R} \lesssim 1.1$ for each parameter (preferably less).
- Finally, this lets us combine all the chains together to estimate the posterior.

# Complicated or pathological likelihoods

- It's possible to come across problems where it seems very difficult to get MCMC to work efficiently.
- Multimodal posteriors are especially awkward.
  - A method such as simulated annealing allows the chain to take bigger steps early on, and hopefully find all the peaks.
  - There are other families of related methods which might work better, such as 'nested sampling'
- Likelihood contours might be narrow and strongly curved: unless you can reformulate the problem, the chains might explore them very slowly. Techniques like nested sampling can work here too.

# Model selection and Bayesian evidence

- Sometimes we may want to compare how different models, with different parametrizations, perform in fitting some data set.
- Models with more parameters in general have more freedom, so one can penalize the more flexible models to reflect this (e.g. the Bayesian Information Criterion, BIC).
- A more rigorous way to compare models is to use the evidence, $\mathcal{Z}$, but unfortunately MCMC only computes the posterior up to an arbitrary normalization.
- It's possible to compute the evidence using results of MCMC, but in general this is very expensive (much more so than running MCMC in the first place).
- If this is the aim from the start, it's worth considering alternative methods.

# Further Reading

- W. H. Press, S. A. Teukolsky, W. T. Vetterling and B. P. Flannery
  *Numerical Recipes*
  Cambridge University Press, 2007
  The usual practical introduction to the topic can be found in Chapter 15.

- A. Lewis and S. Bridle
  Cosmological parameters from CMB and other data: A Monte Carlo approach
  *Phys. Rev. D*, **66**, 103511 (2002).
  A nice astronomical example of the use of MCMC, with some good references, and useful tips for writing your own MCMC code.