Graduate Lecture
Dark Energy - Observational Evidence and Theoretical Modeling
Lecture III

Jochen Weller
Department of Physics & Astronomy
University College London

January 17, 2006

Contents

2 Observational Evidence for Accelerated Expansion from Supernovae 2

2.3 Parameter Estimation 2

2.3.3 Monte Carlo Markov Chain Sampling 2

Chapter 2

Observational Evidence for Accelerated Expansion from Supernovae

2.3 Parameter Estimation

2.3.3 Monte Carlo Markov Chain Sampling

[1] We have seen in the previous section that it takes longer and longer to calculate the posterior likelihood, the more parameters we have. In fact the computational time scaling of the grid based method goes like the power of N, where N is the number of parameters. A further problem we discussed for the grid based method is that we would actually use a finer grid in the peak of the distribution and a coarser one in the tails, where there is not much likelihood.

We are interested in the posterior distribution, given by Bayes' theorem

$$p(\boldsymbol{\theta}|D) = \frac{p(\boldsymbol{\theta})p(D|\boldsymbol{\theta})}{\int p(\boldsymbol{\theta})p(D|\boldsymbol{\theta})\, d\boldsymbol{\theta}} \tag{2.1}$$

In order to analyse the posterior distribution we require to calculate quantities, like moments, quantiles, highest posterior density etc., which in general is the expectation of a function of the parameters:

$$E\left[f(\boldsymbol{\theta})|D\right] = \frac{\int f(\boldsymbol{\theta})p(\boldsymbol{\theta})p(D|\boldsymbol{\theta})\, d\boldsymbol{\theta}}{\int p(\boldsymbol{\theta})p(D|\boldsymbol{\theta})\, d\boldsymbol{\theta}}$$

[1] Bibliography: Markov Chain Monte Carlo in practice, *Eds.* Gilks, Richardson, Spiegelhalter, CHAPMAN & HALL/CRC.

However this can be a tricky task, particularly for a larger number of parameters ($N > 3$). Further notice, that for most applications it is only necessary to calculate

$$E\left[f(\boldsymbol{\theta})|D\right] \propto \int f(\boldsymbol{\theta})p(\boldsymbol{\theta})p(D|\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

The problem is hence to calculate an integral. This can be done efficiently by a *Monte Carlo* integration. In order to simplify our notation we consider the following problem:

$$E\left[f(X)\right] = \int f(x)\pi(x)\, dx \ . \tag{2.2}$$

where X comprise of N continuous real variables and $\pi(x)$ is the distribution. The Monte Carlo method works by drawing samples $\{X_t, t = 1, ..., N_s\}$ from $\pi(\cdot)$ and then approximating

$$E\left[f(X)\right] \approx \frac{1}{N_s}\sum_{t=1}^{N_s} f(X_t) \ .$$

Note that if N_s is chosen large enough this is an adequate approximation. In general it is not possible to draw the X_t independently and directly from $\pi(\cdot)$, since $\pi(\cdot)$ can be any distribution. However, the X_t need not be independent, as long as they are generated in the correct proportions according to $\pi(\cdot)$.

This is given if one can create a Markov chain which has $\pi(\cdot)$ as stationary (limiting) distribution. Suppose we generate a sequence of random variables, $\{X_0, X_1, X_2, ...\}$, such that the next state is sampled from a distribution $p(X_{t+1}|X_t)$, i.e. the next state only depends on the current state of the chain and not on the entire history. This sequence is called a *Markov chain* with transition kernel (probability) $p(\cdot|\cdot)$. If the probability is well behaved (regular), the chain will gradually forget about the initial state X_0 and approach a stationary (or invariant) distribution after a sufficient sequence size. In Figure 2.16 we see an example of a sequence which approaches a stationary distribution. That means as t increases the sample points X_t, will look more and more like they are drawn from a stationary distribution $\phi(\cdot)$. After a sufficient long *burn-in* the points $\{X_t; t = m + 1, ..., n\}$ will be dependent samples approximately from $\phi(\cdot)$. We can now estimate $E[f(x)]$, where X has the distribution $\phi(\cdot)$

$$\bar{f} = \frac{1}{n-m}\sum_{t=m+1}^{n} f(X_t) \ . \tag{2.3}$$

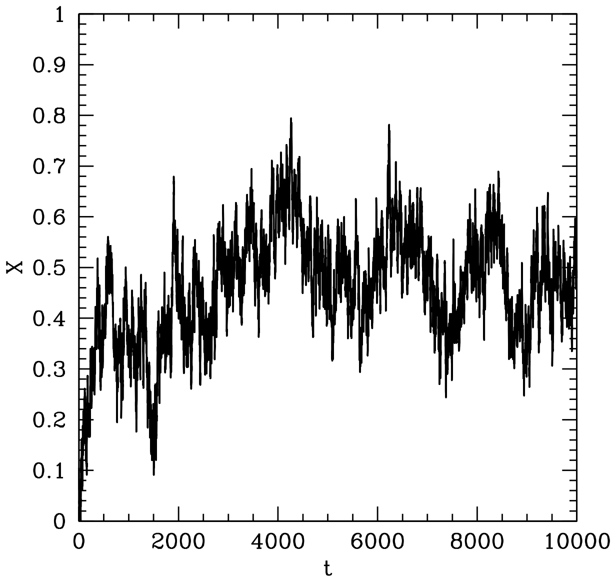

Figure 2.16: Markov chain sequence approaching stationary distribution.

The next step is to construct a Markov chain where $\phi(\cdot)$ is $\pi(\cdot)$. One possibility is the *Metropolis-Hastings* algorithm. At each time step a *candidate* point Y is chosen from a proposal distribution $q(\cdot|X_t)$, for example a multivariate Gaussian, with mean X_t and fixed covariance. The candidate point is then accepted with probability $\alpha(X_t, Y)$, where

$$\alpha(X,Y) = \min\left(1, \frac{\pi(Y)q(X|Y)}{\pi(X)q(Y|X)}\right) \ .$$

If the candidate point is accepted the next state becomes $X_{t+1} = Y$, if the candidate point is rejected the chain does not move, i.e. $X_{t+1} = X_t$. The Metropolis-Hastings algorithm is then:

1. Initialize to a random X_0.

2. Sample point from Y from $q(\cdot|X_t)$.

3. Sample a Uniform $(0,1)$ variable U.

4. If $U \leq \alpha(X_t, Y)$ set $X_{t+1} = Y$, otherwise set $X_{t+1} = X_t$.

5. increment t and start again at 2.

Interestingly the proposal distribution $q(\cdot|\cdot)$ can have any form and the stationary distribution of the chain will be $\pi(\cdot)$. The Metropolis algorithm itself (which we will exploit later) considers only symmetric proposals $q(Y|X) = q(X|Y)$. For examples $q(\cdot|X)$ a multivariate Gaussian with mean X and covariance matrix Σ. Hence we obtain

$$\alpha(X,Y) = \min\left(1, \frac{\pi(Y)}{\pi(X)}\right) .$$

One has to be careful how to choose Σ. If Σ is too small there will be a high acceptance rate and slow mixing, while a wide distribution will result in low acceptance and no movement of the chain, hence resulting in slow mixing as well. We will later discuss how to improve this.

Let us try and apply this now to the fitting of the Supernovae data, with all three parameters $\boldsymbol{\theta} = (\mathcal{M}, \Omega_{\mathrm{m},0}, \Omega_{\Lambda,0})^2$, where $\pi(\boldsymbol{\theta}) \propto \exp[-0.5\chi^2(\boldsymbol{\theta})]$.

```
SUBROUTINE CHAIN(FILENAME)

   INTEGER ::  I,NP,COUNT
   REAL ::  CHI2GET,CHIOLD
   REAL, ALLOCATABLE ::  PAROLD(:),RSHIFT(:)
   REAL ::  ACCEPT,ATEST
   CHARACTER(80) ::  FILENAME

   ALLOCATE(PAROLD(NPAR),RSHIFT(NPAR))
!  SET UP INITIAL SAMPLING POINT
   PAROLD=PAR
   CALL GAUSSRANDOM(RSHIFT)
   PAR = PAROLD+DPAR*RSHIFT
   CALL CHECK(PAR)
!  Calculate intial chi2 value
   CALL MATCHCHI2(PAR,CHI2GET)
   PAROLD=PAR
   CHIOLD=CHI2GET
   NP = 1
   COUNT=1
   OPEN(UNIT=11,FILE=FILENAME,STATUS='UNKNOWN',FORM='FORMATTED')
   DO I=1,NSAMPLE
!  get multivariate Gaussian with width RSHIFT
```

[2] Of course in a realistic situation one would choose the likelihood which is analytically marginalized over $\mathcal{M}$, but for illustrational purposes we choose the three parameter fits.

```
     CALL GAUSSRANDOM(RSHIFT)
!  Shift parameter values
     PAR = PAROLD+DPAR*RSHIFT
!  Check if parameter values are within given bounds,
!  otherwise shift to bound
     CALL CHECK(PAR)
!  calculate chi2 value
     CALL MATCHCHI2(PAR,CHI2GET)
!  calculate uniform random number
     CALL RANDOM_NUMBER(ACCEPT)
!  ratio of likelihoods
     ATEST = EXP(-0.5*(CHI2GET-CHIOLD))
!  Metropolis - Hastings criteria
     IF (MIN(1.0,ATEST)>=ACCEPT) THEN
       write(11,*) COUNT,NP,CHIOLD,PAROLD
       PAROLD=PAR
       CHIOLD=CHI2GET
       NP=1
       COUNT=COUNT+1
     ELSE
!  do not move and increase counter
       NP=NP+1
     END IF
  END DO
  CLOSE(11)
  DEALLOCATE(PAROLD,RSHIFT)

END SUBROUTINE CHAIN
```

Example code for fitting Supernovae data with an MCMC chain. We choose as the initial values for the chain: $\mathcal{M} = 16$, $\Omega_{\mathrm{m},0} = 0.1$ and $\Omega_{\Lambda,0} = 0.1$, the covariance of the proposal distribution are chosen $\Delta\mathcal{M} = 0.05$, and $\Delta\Omega = 0.02$ for the densities. The bounds are chosen to be $14 \leq \mathcal{M} \leq 17$, $0 \leq \Omega_{\mathrm{m},0} \leq 3.0$ and $-2 \leq \Omega_{\Lambda,0} \leq 3$. We stop the algorithm at an overall of 100.000 samples (accepted *and* unaccepted). Figure 2.16 shows the begining of the chain in the $\Omega_{\mathrm{m},0}$ variable. In Figure 2.17 we show the number of samples of the Markov chain sequence, within bins of the parameters.

In Figure 2.18 we show the two dimensional joint likelihoods.

The next question to address how we know for how long to run the chain? This is the question if the sequence has converged and is truly stationary. One of the simplest methods to address this question, is by running several

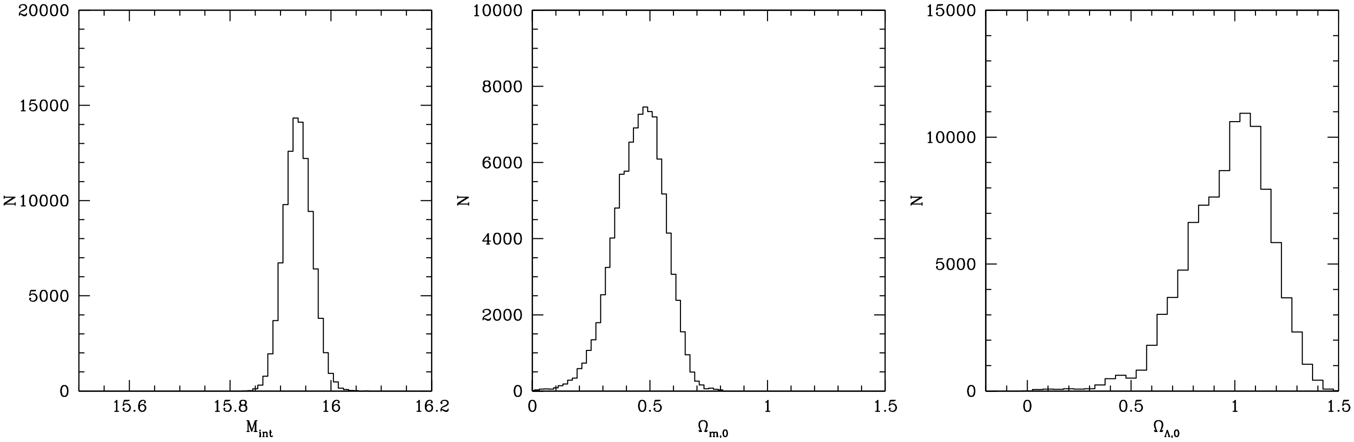

Figure 2.17: Histogram of the MCMC sampled distributions.

chains in parallel, with *over-dispersed* starting values and compare estimates $\bar{f}$.

The fundamental problem of inference from a Markov chain simulation is that there will always be areas of the target distribution that have not been covered by the finite chain. First we have to set up multiple chains (maybe run in parallel), with over-dispersed initial points. This is essential for a successful diagnostic. Over-dispersion can be achieved after running a single chain initially and get an idea about the distribution of this chain. One can then use the variance of this chain to achieve over-dispersion.

Figure 2.19 shows the results for three over-dispersed chains for our Supernovae fitting procedure with over-dispersed initial points. It is evident from the Figure that all three chains begin to converge already at 1000 (accepted) steps. Let us assume we are interested in a quantity ψ from the chain. These can be the parameters or any function of the parameters. Let us further assume that we run m parallel sequences of length n and label the quantities (ψ_{ij}), $j = 1, ..., n$ and $i = 1, ..., m$.

We hence compute two quantities: The between sequence variance B and the within-sequence variances W.

$$B = \frac{n}{m-1} \sum_{i=1}^{m} \left(\bar{\psi}_i - \bar{\psi}\right)^2 \ ,$$

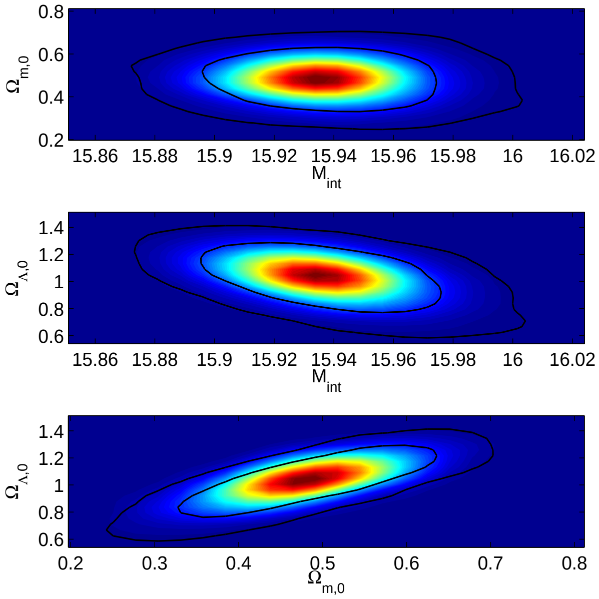

Figure 2.18: Joint likelihoods of the various parameter combinations. The color scheme correspond to the density of the sample over the parameter space, while the solid line are the 68% and 95% marginalized joint likelihoods. This figure has been produced from the chains with the `getdist` program provided with the `COSMOMC` package by Lewis and Bridle (2003). Note that this is for the entire sample compiled by Riess et al. 2004, not just the so called Gold Sample.

where

$$\bar{\psi}_i = \frac{1}{n}\sum_{j=1}^{n}\psi_{ij} \qquad \bar{\psi} = \frac{1}{m}\sum_{i=1}^{m}\bar{\psi}_i \ .$$

Further we define

$$W = \frac{1}{m}\sum_{i=1}^{m} s_i^2 \ ,$$

where

$$s_i^2 = \frac{1}{n-1}\sum_{j=1}^{n}\left(\psi_{ij} - \bar{\psi}_i\right)^2 \ .$$

So W is just the *average variance* of all the chains, while B measures the *variance of the averages* of the chains. Note that the between-sequence

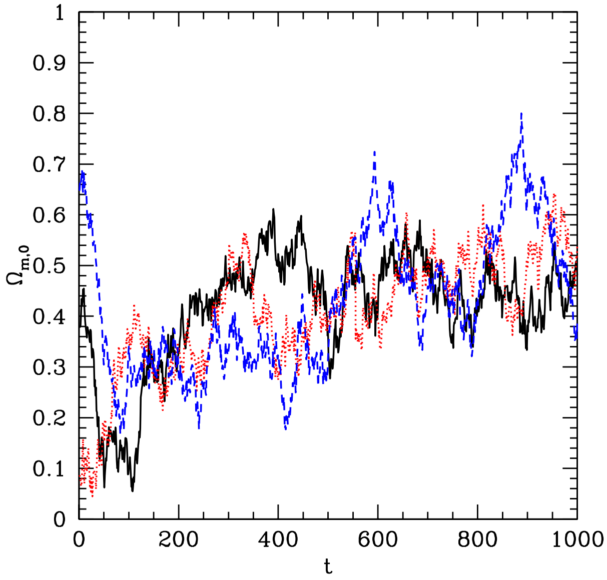

Figure 2.19: Start of three chains with over-dispersed initial conditions.

variance B contains a factor n because it is based on the variance of the within-sequence means, $\bar{\psi}_i$, each of which is an average of n values ψ_{ij}.

One estimate of the variance of ψ in the target distribution is

$$\widehat{\mathrm{var}}(\psi) = \frac{n-1}{n} W + \frac{1}{n} B \ ,$$

which is an overestimate. Further W is an underestimate of the target variance, because individual chains had not have time to cover the target distribution. For $n \to \infty$ both estimates approach the target variance $\mathrm{var}(\psi)$. Convergence can now be established by monitoring

$$\sqrt{\hat{R}} = \sqrt{\frac{\widehat{\mathrm{var}}(\psi)}{W}} \ , \tag{2.4}$$

which approaches 1 at convergence. Note that there are many other convergence criteria some of which are discussed in Gilks et al. (1996). For our example we ran 5 chains with over-dispersed initial conditions and obtained $\hat{R}^2(\mathcal{M}) - 1 = 0.0446$, $\hat{R}^2(\Omega_{\mathrm{m},0}) - 1 = 0.0118$, $\hat{R}^2(\Omega_{\Lambda,0}) - 1 = 0.0529$, while Gelman (1996) recommends values below 0.1. Finally the marginalized parameters are: $\mathcal{M} = 15.93 \pm 0.03$, $\Omega_{\mathrm{m},0} = 0.47 \pm 0.10$ and $\Omega_{\Lambda,0} = 1.02 \pm 0.17$.

Again we want to emphasize that this is for the entire sample, not just the Gold sample.

Finally we want briefly discuss, how to improve the efficiency of the sampling. This is essentially achieved by choosing appropriate parameters and sampling directions and step size. Gaussian distribution theory suggest that the most efficient proposal density is shaped like the target distribution scaled by a factor of about $2.4/\sqrt{d}$, where d is the number of parameters. The scale and shape of the target distribution can be estimated from early simulation draws and then the proposal density can be adaptively altered. In Figure 2.20 we show an example of a target distribution. In general we would calculate the covariance matrix for early samples, and calculate the eigenvalues and eigenvectors. We hence would choose a proposal Gaussian density according to the eigendirections and eigenvalues, which guarantees that most samples lie within the estimated target distribution. This step will then be updated at later stages of the sampling. However one has to be careful if there are large non-linear (or non-Gaussian) parameter degeneracies. In this case it is sometimes useful to analyse the covariance matrix of the logarithm of the parameters and work in this mapped parameter space.

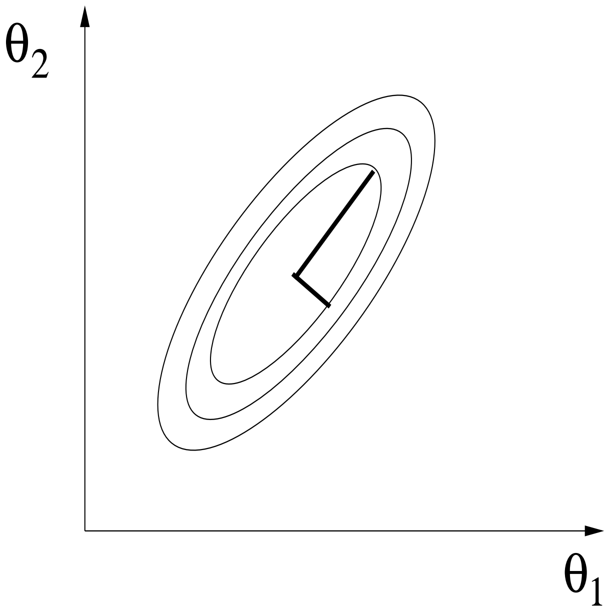

Figure 2.20: If we choose the proposal density as the target distribution, we can increase the efficiency immensely

.