Lecture 3: Monte Carlo methods and Markov Chain Monte Carlo: a CMB example

Licia Verde

1 Catching up

[real world effects...]

2 Introduction (or probabilities, again)

Let's go back to the issue of parameter estimation and error calculation. Here is the conceptual interpretation of what it means that en experiment measures some parameters (say cosmological parameters). There is some underlying true set of parameters $\vec{\alpha}_{true}$ that are only known to Mother Nature but not to the experimenter. There true parameters are statistically realized in the observable universe and random measurement errors are then included when the observable universe gets measured. This "realization" gives the measured data $\mathcal{D}_0$. Only $\mathcal{D}_0$ is accessible to the observer (you). Then you go and do what you have to do to estimate the parameters and their errors (chi-square, likelihood, etc...) and get $\vec{\alpha}_0$. Note that $\mathcal{D}_0$ is not a unique realization of the true model given by $\vec{\alpha}_{true}$: there could be infinitely many other realizations as *hypothetical data sets*, which could have been the measured one: $\mathcal{D}_1, \mathcal{D}_2, ...$ each of them with a slightly different fitted parameters $\vec{\alpha}_1, \vec{\alpha}_2....$ $\vec{\alpha}_0$ is one parameter set drawn from this distribution. The hypotetical ensamble of universes described by $\vec{\alpha}_i$ is called ensamble, and one expects that the expectation value $\langle\vec{\alpha}_i\rangle = \vec{\alpha}_{true}$. If we knew the distribution of $\vec{\alpha}_i - \vec{\alpha}_{true}$ we would know everything we need about the uncertainties in our measurement $\vec{\alpha}_0$. The goal is to infer the distribution of $\vec{\alpha}_i - \vec{\alpha}_{true}$ without knowing $\vec{\alpha}_{true}$.

3 Monte Carlo simulation of (synthetic) data sets

Here's what we do: we say that hopefully $\vec{\alpha}_0$ is not too wrong and we consider a fictitious world where $\vec{\alpha}_0$ was the true one. So it would not be such a big mistake to take the probability distribution of $\vec{\alpha}_0 - \vec{\alpha}_i$ to be that of $\vec{\alpha}_{true} - \vec{\alpha}_i$. In many cases we know how to simulate $\vec{\alpha}_0 - \vec{\alpha}_i$ and so we can simulate many synthetic realization of "worlds" where $\vec{\alpha}_0$ is the true underlying model. Then mimic the observation process of these fictitious Universes replicating all the observational errors and effects and from each of these fictitious universe estimate the parameters. Simulate enough of them and from $\vec{\alpha}_i^S - \vec{\alpha}_0$ you will be able to map the desired multi-dimensional probability distribution.

Example: analysis of large scale structure data.

4 Likelihoods, again

While the CMB temperature distribution is gaussian (or very close to Gaussian) the C_ℓ distrIbution is not. At high ℓ the Central limit theorem will ensure that the likelihood is well approximated by a Gaussian but at low ℓ this is not the case.

$$\mathcal{L}(T|C_\ell^{th}) \propto \frac{\exp[-(TS^{-1}T)/2]}{\sqrt{det(S)}} \tag{1}$$

where T denotes a vector of the temperature map and S_{ij} is the signal covariance:

$$S_{ij} = \sum_\ell \frac{(2\ell+1)}{4\pi} C_\ell^{th} P_\ell(\hat{n}_i \cdot \hat{n}_j) \tag{2}$$

and P_ℓ denote the Legendre polynomials.

If we then expans in spherical harmonics T we obtain:

$$\mathcal{L}(T|C_\ell^{th}) \propto \frac{\exp[-1/2|a_{\ell m}|^2/C_\ell^{th}]}{\sqrt{C_\ell^{th}}} \tag{3}$$

Isotropy means that we can sum over $m's$ thus:

$$-2\ln\mathcal{L} = \sum_\ell (2\ell+1)\left[\ln\left(\frac{C_\ell^{th}}{C_{ell}^{data}}\right) + \left(\frac{C_\ell^{data}}{C_\ell^{th}}\right) - 1\right] \tag{4}$$

where $C_\ell^{data} = \sum_m |a_{\ell m}|^2/(2\ell+1)$.

Exercise: show that for an experiment with (gaussian) noise the expression is the same but with the substitution $C_\ell^{th} \longrightarrow C_\ell^{th} + \mathcal{N}_\ell$ with $\mathcal{N}$ denoting the power spectrum of the noise.

Exercise: show that for a partial sky experiment (that covers a fraction of sky f_{sky} you can approximately write:

$$\ln\mathcal{L} \longrightarrow f_{sky}\ln\mathcal{L} \tag{5}$$

NB remember how the number of independent modes scales with the sky area.

As an aside... "But what do I do with polarization data?". Well... if the $a_{\ell m}^T$ are Gaussianly distributed also the $a_{\ell m}^E$ and $a_{\ell m}^B$ will be. So we can generalize the approach above using a vector $(a_{\ell m}^T a_{\ell m}^E a_{\ell m}^B)$. Let us consider a full sky, ideal experiment. Start by writing down the covariance, follow the same steps as above and show that:

$$\begin{aligned} -2\ln\mathcal{L} &= \sum_\ell (2\ell+1)\left\{\ln\left(\frac{C_\ell^{BB}}{\hat{C}_\ell^{BB}}\right) + \ln\left(\frac{C_\ell^{TT}C_\ell^{EE} - (C_\ell^{TE})^2}{\hat{C}_\ell^{TT}\hat{C}_\ell^{EE} - (\hat{C}_\ell^{TE})^2}\right)\right. \\ &+ \left.\frac{\hat{C}_\ell^{TT}C_\ell^{EE} + C_\ell^{TT}\hat{C}_\ell^{EE} - 2\hat{C}_\ell^{TE}C_\ell^{TE}}{C_\ell^{TT}C_\ell^{EE} - (C_\ell^{TE})^2} + \frac{\hat{C}_\ell^{BB}}{C_\ell^{BB}} - 3\right\}, \end{aligned} \tag{6}$$

where C_ℓ denotes C_ℓ^{theory} and $\hat{C}_\ell$ denotes C_ℓ^{data}.

It is easy to show that for a noisy experiment then $C_\ell^{XY} \longrightarrow C_\ell^{XY} + \mathcal{N}_\ell^{\mathcal{X}\mathcal{Y}}$ where $\mathcal{N}_\ell$ denotes the noise power spectrum and $X, Y = \{T, E, B\}$.

Exercise: generalize the above to partial sky coverage: for added complication take $f_{sky}^{TT} \neq f_{sky}^{EE} \neq f_{sky}^{bb}$. (this is often the case as the sky cut for polarization may be different from that of temperature(the foregrounds are different) and in general the cut (or the weighting) for B may need to be larger than that for E.

Let us now expand in Taylor series Equation (4) around its maximum by writing $C_\ell^d = c_\ell^{th}(1+\epsilon)$. For a single multipole ℓ,

$$-2\ln\mathcal{L}_\ell = (2\ell+1)[\epsilon - \ln(1+\epsilon)] \simeq (2\ell+1)\left(\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3} + \mathcal{O}(\epsilon^4)\right) . \tag{7}$$

We note that the Gaussian likelihood approximation is equivalent to the above expression truncated at ϵ^2: $-2\ln\mathcal{L}_{\mathrm{Gauss},\ell} \propto (2\ell+1)/2[(\widehat{\mathcal{C}}_\ell - \mathcal{C}_\ell^{\mathrm{th}})/\mathcal{C}_\ell^{\mathrm{th}}]^2 \simeq (2\ell+1)\epsilon^2/2$.

Also widely used for CMB studies is the lognormal likelihood for the equal variance approximation (Bond et al 1998): approximation is

$$-2\ln\mathcal{L}'_{\mathrm{LN}} = \frac{(2\ell+1)}{2}\left[\ln\left(\frac{\widehat{\mathcal{C}}_\ell}{\mathcal{C}_\ell^{\mathrm{th}}}\right)\right]^2 \simeq (2\ell+1)\left(\frac{\epsilon^2}{2} - \frac{\epsilon^3}{2}\right) . \tag{8}$$

Thus our approximation of likelihood function is given by the form,

$$\ln\mathcal{L} = \frac{1}{3}\ln\mathcal{L}_{\mathrm{Gauss}} + \frac{2}{3}\ln\mathcal{L}'_{\mathrm{LN}} , \tag{9}$$

where

$$\ln\mathcal{L}_{\mathrm{Gauss}} \propto -\frac{1}{2}\sum_{\ell\ell'}(\mathcal{C}_\ell^{\mathrm{th}} - \widehat{\mathcal{C}}_\ell)Q_{\ell\ell'}(\mathcal{C}_{\ell'}^{\mathrm{th}} - \widehat{\mathcal{C}}_{\ell'}) , \tag{10}$$

and

$$\ln\mathcal{L}_{\mathrm{LN}} = -1/2\sum_{\ell\ell'}(z_\ell^{\mathrm{th}} - \widehat{z}_\ell)\mathcal{Q}_{\ell\ell'}(z_{\ell'}^{\mathrm{th}} - \widehat{z}_{\ell'}), \tag{11}$$

where $z_\ell^{\mathrm{th}} = \ln(\mathcal{C}_\ell^{\mathrm{th}} + \mathcal{N}_\ell)$, $\widehat{z}_\ell = \ln(\widehat{\mathcal{C}}_\ell + \mathcal{N}_\ell)$ and $\mathcal{Q}_{\ell\ell'}$ is the local transformation of the curvature matrix Q to the lognormal variables z_ℓ,

$$\mathcal{Q}_{\ell\ell'} = (\mathcal{C}_\ell^{th} + \mathcal{N}_\ell)Q_{\ell\ell'}(\widehat{\mathcal{C}}_{\ell'}^{th} + \mathcal{N}_{\ell'}). \tag{12}$$

[.....]

Note that for the latest WMAP release at low ℓ the likelihood is computed directly from the maps $\vec{m}$. The standard likelihood is given by

$$L(\vec{m}|S)d\vec{m} = \frac{\exp\left[-\frac{1}{2}\vec{m}^t(S+N)^{-1}\vec{m}\right]}{|S+N|^{1/2}}\frac{d\vec{m}}{(2\pi)^{3n_p/2}}, \tag{13}$$

where $\vec{m}$ is the data vector containing the temperature map, $\vec{T}$, as well as the polarization maps, $\vec{Q}$, and $\vec{U}$, n_p is the number of pixels of each map, and S and N are the signal and noise covariance matrix ($3n_p \times 3n_p$), respectively. As the temperature data are completely dominated by the signal at such low multipoles, noise in temperature may be ignored. This simplifies the form of likelihood as

$$L(\vec{m}|S)d\vec{m} = \frac{\exp\left[-\frac{1}{2}\vec{\tilde{m}}^t(\tilde{S}_P + N_P)^{-1}\vec{\tilde{m}}\right]}{|\tilde{S}_P + N_P|^{1/2}} \frac{d\vec{\tilde{m}}}{(2\pi)^{n_p}} \frac{\exp\left(-\frac{1}{2}\vec{T}^t S_T^{-1}\vec{T}\right)}{|S_T|^{1/2}} \frac{d\vec{T}}{(2\pi)^{n_p/2}}, \quad (14)$$

where S_T is the temperature signal matrix ($n_p \times n_p$), the new polarization data vector, $\vec{\tilde{m}} = (\tilde{Q}_p,\ \tilde{U}_p)$ and $\tilde{S}_P$ is the signal matrix for the new polarization vector with the size of $2n_p \times 2n_p$.

5 Markov Chain Monte Carlo (MCMC)

When dealing with high dimensional likelihoods (i.e. many parameters) the process of mapping the likelihood (or the posterior) surface can become very expensive. For example for CMB studies the modles considered have from 6 to 11+ parameters. Every model evaluation even with a fact code such as CAMB can take up to minutes per iteration. A grid-based likelihood analysis would require prohibitive amounts of CPU time. For example, a coarse grid (~ 20 grid points per dimension) with six parameters requires $\sim 6.4 \times 10^7$ evaluations of the power spectra. At 1.6 seconds per evaluation, the calculation would take $\sim$ 1200 days. Christensen & Meyer (2000) proposed using Markov Chain Monte Carlo (MCMC) to investigate the likelihood space. This approach has become the standard tool for CMB analyses. MCMC is a method to simulate posterior distributions. In particular one simulates sampling the posterior distribution $\mathcal{P}(\alpha|x)$, of a set of parameters α given event x, obtained via Bayes' Theorem

$$\mathcal{P}(\alpha|x) = \frac{\mathcal{P}(x|\alpha)\mathcal{P}(\alpha)}{\int \mathcal{P}(x|\alpha)\mathcal{P}(\alpha)d\alpha}, \quad (15)$$

where $\mathcal{P}(x|\alpha)$ is the likelihood of event x given the model parameters α and $\mathcal{P}(\alpha)$ is the prior probability density; α denotes a set of cosmological parameters (e.g., for the standard, flat ΛCDM model these could be, the cold-dark matter density parameter Ω_c, the baryon density parameter Ω_b, the spectral slope n_s, the Hubble constant –in units of 100 km s^{-1} Mpc^{-1})– h, the optical depth τ and the power spectrum amplitude A), and event x will be the set of observed $\widehat{\mathcal{C}}_\ell$. The MCMC generates random draws (i.e. simulations) from the posterior distribution that are a "fair" sample of the likelihood surface. From this sample, we can estimate all of the quantities of interest about the posterior distribution (mean, variance, confidence levels). The MCMC method scales approximately linearly with the number of parameters, thus allowing us to perform likelihood analysis in a reasonable amount of time.

A properly derived and implemented MCMC draws from the joint posterior density $\mathcal{P}(\alpha|x)$ once it has converged to the stationary distribution. The primary consideration in implementing MCMC is determining when the chain has *converged*. After an initial *"burn-in"* period, all further samples can be thought of as coming from the stationary distribution. In other words the chain has no dependence on the starting location.

Another fundamental problem of inference from Markov chains is that there are always areas of the target distribution that have not been covered by a finite chain. If the MCMC is run for a very long time, the ergodicity of the Markov chain guarantees that eventually the chain will cover all the target distribution, but in the short term the simulations cannot tell us about areas where they have not been. It is thus crucial that the chain achieves good *"mixing"*. If the Markov chain does not move rapidly throughout the support of the target distribution because of poor *mixing*, it might take a prohibitive amount of time for the chain to fully explore the likelihood surface. Thus it is important to have a convergence criterion and a mixing diagnostic. Plots of the sampled MCMC parameters or likelihood values versus iteration number are commonly used to provide such criteria (left panel of Figure **??**). However, samples from a chain are typically serially correlated; very high auto-correlation leads to little movement of the chain and thus makes the chain to "appear" to have converged. For a more detailed discussion see Gilks (). Using a MCMC that has not fully explored the likelihood surface for determining cosmological parameters will yield *wrong* results.

5.1 Markov Chains in Practice

here are the necessary steps to run a simple MCMC for the CMB temperature power spectrum. It is straightforward to generalize these instructions to include the temperature-polarization power spectrum and other datasets. The MCMC is essentially a random walk in parameter space, where the probability of being at any position in the space is proportional to the posterior probability.

1) Start with a set of cosmological parameters $\{\alpha_1\}$, compute the $\mathcal{C}_\ell^1$ and the likelihood $\mathcal{L}_1 = \mathcal{L}(\mathcal{C}_\ell^{1\rm th}|\widehat{\mathcal{C}}_\ell)$.

2) Take a random step in parameter space to obtain a new set of cosmological parameters $\{\alpha_2\}$. The probability distribution of the step is taken to be Gaussian in each direction i with r.m.s given by σ_i. We will refer below to σ_i as the "step size". The choice of the step size is important to optimize the chain efficiency (see §**??**)

3) Compute the $\mathcal{C}_\ell^{2\rm th}$ for the new set of cosmological parameters and their likelihood $\mathcal{L}_2$.

4.a) If $\mathcal{L}_2/\mathcal{L}_1 \geq 1$, "take the step" i.e. save the new set of cosmological parameters $\{\alpha_2\}$ as part of the chain, then go to step 2 after the substitution $\{\alpha_1\} \longrightarrow \{\alpha_2\}$.

4.b) If $\mathcal{L}_2/\mathcal{L}_1 < 1$, draw a random number x from a uniform distribution from 0 to 1. If $x \geq \mathcal{L}_2/\mathcal{L}_1$ "do not take the step", i.e. save the parameter set $\{\alpha_1\}$ as part of the chain and return to step 2. If $x < \mathcal{L}_2/\mathcal{L}_1$, " take the step", i.e. do as in 4.a).

5) For each cosmological model run four chains starting at randomly chosen, well-separated points in parameter space. When the convergence criterion is satisfied and the chains have enough points to provide reasonable samples from the a posteriori distributions (i.e. enough points to be able to reconstruct the 1- and 2-σ levels of the marginalized likelihood for all the parameters) stop the chains.

It is clear that the MCMC approach is easily generalized to compute the joint likelihood of *WMAP* data with other datasets.

5.2 Improving MCMC Efficiency

MCMC efficiency can be seriously compromised if there are degeneracies among parameters.

[figure]

The Markov chain efficiency can be improved in different ways. Here we report the simplest way....

Reparameterization

We describe below the method we use to ensure convergence and good mixing. Degeneracies and poor parameter choices slow the rate of convergence and mixing of the Markov Chain. There is one near-exact degeneracy (the geometric degeneracy) and several approximate degeneracies in the parameters describing the CMB power spectrum Bond et al (1994), Efstathiou&Bond(1999). The numerical effects of these degeneracies are reduced by finding a combination of cosmological parameters (e.g., Ω_c, Ω_b, h, etc.) that have essentially orthogonal effects on the angular power spectrum. The use of such parameter combinations removes or reduces degeneracies in the MCMC and hence speeds up convergence and improves mixing, because the chain does not have to spend time exploring degeneracy directions. Kosowsky, milosavljevic &Jimenez (2002) introduced a set of reparameterizations to do just this. In addition, these new parameters reflect the underlying physical effects determining the form of the CMB power spectrum (we will refer to these as physical parameters). This leads to particularly intuitive and transparent parameter dependencies of the CMB power spectrum.

For the 6 parameters LCDM model these "normal" or "physical" parameters are: the physical energy densities of cold dark matter, $\omega_c \equiv \Omega_c h^2$, and baryons, $\omega_b \equiv \Omega_b h^2$, the characteristic angular scale of the acoustic peaks,

$$\theta_A = \frac{r_s(a_{dec})}{d_A(a_{dec})}, \tag{16}$$

where a_{dec} is the scale factor at decoupling,

$$r_s(a_{dec}) = \frac{c}{H_0\sqrt{3}} \int_0^{a_{dec}} \left[\left(1 + \frac{3\Omega_b}{4\Omega_\gamma}\right) \left((1-\Omega)x^2 + \Omega_\Lambda x^{1-3w} + \Omega_m x + \Omega_{rad}\right) \right]^{-1/2} dx \tag{17}$$

is the sound horizon at decoupling, and

$$d_A(a_{dec}) = \frac{c}{H_0} \int_{a_{dec}}^1 \left[(1-\Omega)x^2 + \Omega_\Lambda x^{1-3w} + \Omega_m x + \Omega_{rad}\right]^{-1/2} dx \tag{18}$$

is the angular diameter distance at decoupling, where H_0 denotes the Hubble constant and c is the speed of light. Here $\Omega_m = \Omega_c + \Omega_b$, Ω_Λ denotes the vacuum energy density parameters, w is the equation of state of the dark energy component, $\Omega = \Omega_m + \Omega_\Lambda$ and the radiation density parameter $\Omega_{\rm rad} = \Omega_\gamma + \Omega_\nu$, Ω_γ, Ω_ν are the the photon and neutrino density parameters respectively. For reionization sometmes the parameter $\mathcal{Z} \equiv \exp(-2\tau)$ is used, where τ denotes the optical depth to the last scattering surface (not the decoupling surface).

These reparameterizations are useful because the degeneracies are non-linear, that is they are not well described by ellipses in parameter space. For degeneracies that are well approximated by ellipses in parameter space it is possible to find the best reparameterization automatically. This is what the code **CosmoMC** (see tutorials) does.

[figure]

[mention slow/ fast parameters]

Step size optimization The choice of the step size in the Markov Chain is crucial to improve the chain efficiency and speed up convergence. If the step size is too big, the acceptance rate will be very small; if the step size is too small the acceptance rate will be high but the chain will exhibit poor mixing. Both situations will lead to slow convergence.

5.3 Convergence and Mixing

you should always use a convergence and mixing criterion when running MCMC's Let's illustrate here the method proposed by Gelman & Rubin 1992 as an example. They advocate comparing several sequences drawn from different starting points and checking to see that they are indistinguishable. This method not only tests convergence but can also diagnose poor mixing. Let us consider M chains starting at well-separated points in parameter space; each has $2N$ elements, of which we consider only the last N: $\{y_i^j\}$ where $i = 1, .., N$ and $j = 1, .., M$, i.e. y denotes a chain element (a point in parameter space) the index i runs over the elements in a chain the index j runs over the different chains. We define the mean of the chain

$$\bar{y}^j = \frac{1}{N}\sum_{i=1}^{N} y_i^j, \tag{19}$$

and the mean of the distribution

$$\bar{y} = \frac{1}{NM}\sum_{ij=1}^{NM} y_i^j\,. \tag{20}$$

We then define the variance between chains as

$$B_n = \frac{1}{M-1}\sum_{j=1}^{M}(\bar{y}^j - \bar{y})^2, \tag{21}$$

and the variance within a chain as

$$W = \frac{1}{M(N-1)}\sum_{ij}(y_i^j - \bar{y}^j)^2. \tag{22}$$

The quantity

$$\hat{R} = \frac{\frac{N-1}{N}W + B_n\left(1+\frac{1}{M}\right)}{W} \tag{23}$$

is the ratio of two estimates of the variance in the target distribution: the numerator is an estimate of the variance that is unbiased if the distribution is stationary, but is otherwise an overestimate. The denominator is an underestimate of the variance of the target distribution if the individual sequences did not have time to converge.

The convergence of the Markov chain is then monitored by recording the quantity $\hat{R}$ for all the parameters and running the simulations until the values for $\hat{R}$ are always < 1.03.

Question: how does the MCMC sample the prior if all one actually computes is the likelihood?

5.4 MCMC Output Analysis

Now that you have your multiple chains and the convergence criterium says they are converged whot do you do? First discard *burn in* and merge the chains. Since the MCMC passes objective tests for convergence and mixing, the density of points in parameter space is proportional to the posterior probability of the parameters. (Note that cosmomc saves repeated steps as the same entry in the file but with a weight equal to the repetitions: the MCMC gives to each point in parameter space a "weight" proportional to the number of steps the chain has spent at that particular location.). The marginalized distribution is obtained by projecting the MCMC points. This is a great advantage compared to the grid-based approach where multi-dimensional integrals would have to be performed. The MCMC basically performs a Monte Carlo integration. the density of points in the n-dimensional space is proportional to the posterior, and best fit parameters and multi-dimensional confidence levels can be found as illustrated in the last class.

Note that the global maximum likelihood value for the parameters does not necessarily coincide with the expectation value of their marginalized distribution if the likelihood surface is not a multi-variate Gaussian.

A virtue of the MCMC method is that the addition of extra data sets in the joint analysis can efficiently be done with minimal computational effort from the MCMC output if the inclusion of extra data set does not require the introduction of extra parameters or does not drive the parameters significantly away from the current best fit. If the likelihood surface for a subset of parameters from an external (independent) data set is known, or if a prior needs to be added *a posteriori*, the joint posterior surface can be obtained by multiplying the new probability distribution with the posterior distribution of the MCMC output.

[example COSMOMC]