Cosmic Microwave Background Radiation

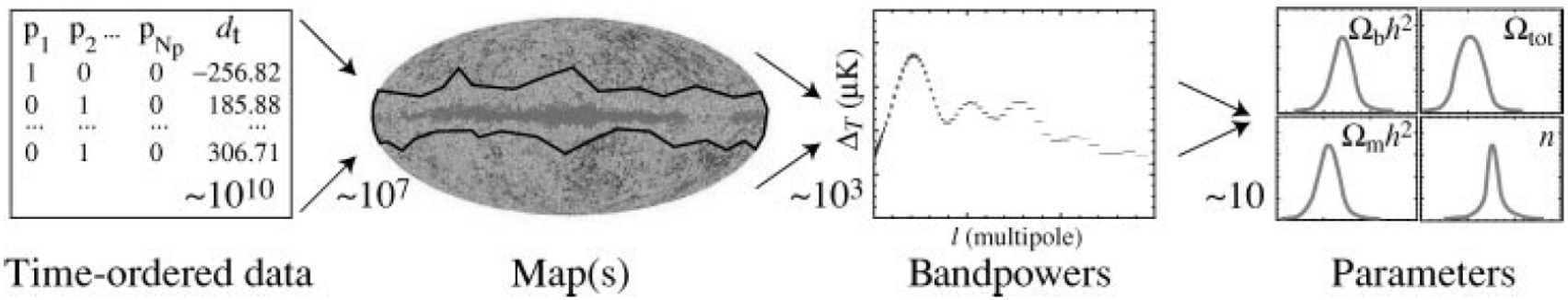
Lecture 3 : Parameter Estimation

Jayanti Prasad

Inter-University Centre for Astronomy & Astrophysics (IUCAA)
Pune, India (411007)

Autumn School on Cosmology (5 - 15th Nov 2013)
BITS PILANI

CMB Data Analysis pipeline



Data pipeline and radical compression. Maps are constructed for each frequency channel from the data timestreams, combined, and cleaned of foreground contamination by spatial (represented here by excising the galaxy) and frequency information. Bandpowers are extracted from the maps and cosmological parameters from the bandpowers. Each step involves a substantial reduction in the number of parameters needed to describe the data, from potentially $10^{10} \rightarrow 10$ for the Planck satellite.

In every step of CMB data analysis the aim is to reduce the volume of data without losing information.

Plan of the Talk

- Data Analysis Techniques
  - Inverse Problems
  - Chi-Square minimization
  - Maximum-Likelihood Estimation

Plan of the Talk

- Data Analysis Techniques
  - Inverse Problems
  - Chi-Square minimization
  - Maximum-Likelihood Estimation
- CMB data and Map making

Plan of the Talk

- Data Analysis Techniques
  - Inverse Problems
  - Chi-Square minimization
  - Maximum-Likelihood Estimation
- CMB data and Map making
- Cosmological Parameter Estimation
  - CMB Likelihood
  - Markov Chain Monte Carlo Methods

Plan of the Talk

- Data Analysis Techniques
  - Inverse Problems
  - Chi-Square minimization
  - Maximum-Likelihood Estimation
- CMB data and Map making
- Cosmological Parameter Estimation
  - CMB Likelihood
  - Markov Chain Monte Carlo Methods
- WMAP and Planck

Plan of the Talk

- Data Analysis Techniques
  - Inverse Problems
  - Chi-Square minimization
  - Maximum-Likelihood Estimation
- CMB data and Map making
- Cosmological Parameter Estimation
  - CMB Likelihood
  - Markov Chain Monte Carlo Methods
- WMAP and Planck
- Summary and Conclusions

Data Analysis Techniques

- Scientific observations can be represented by a data vector **d** which can be a time series $\{t\}$ or temperature map of the sky $\{T\}$ or something else.
- Data have information about some physical process for which we have a theoretical model represented by a set of parameters i.e., parameter vector Θ.
- One of its example is a Gaussian process represented by two parameters i.e., the mean μ and the variance σ^2. The probability of obtaining data d given a theoretical (Gaussian model (μ, σ^2)) is given by:

$$P(d|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(d-\mu)^2}{\sigma^2}\right] \qquad (1)$$

Data Analysis Techniques

- Note that finding out which of the theoretical model is better than others e.g., we should be fitting a parabola rather than a line, is different from finding the parameters of a model.
- In the present discussion we will not discuss model comparison, we will always assume a model is true and will try to find its parameters i.e., what are the values of the parameters of ΛCDM cosmological model.
- Values of parameters do not make sense unless we specify a model.
- Once we have a model (with its parameters) we can easily create data by simulating that model and this is called a forward problem.
- Finding parameters Θ of a model from the data $\mathbf{d}$ is called the inverse problem which is much harder to solve than the forward problem.

Linear Problem

- A problem is said to be a linear problem when the data **d** depends on model parameters linearly, for example fitting a polynomial is a linear problem:

$$d_i = \sum_{j=1}^{M} c_j \theta^j \quad \text{for} \quad i = 1, N \tag{2}$$

  where c_i are the model parameters.
- By definition there is no linear relationship between the data and model parameter for a non-linear problem, for example, fitting a Gaussian is a non-linear problem.
- Non-linear problems are harder to solve than linear problems.
- Dependency of cosmological data i.e., CMB map, on cosmological parameters is a non-linear problem.

Solvable problems

- Not all the inverse problems are solvable and in some cases we can easily find out why that is so.
- On the basis of whether the size N of the data vector $\mathbf{d}$ is larger, smaller or equal to the size M of the parameter vector Θ, there are three possibilities.
  - $N = M$: Unique solution is possible
  - $N > M$: Over constrained problem, χ^2 minimization, Unique solution
  - $N < M$: Under-constrained problem, ill posed problem, priors, regularization

  Note that in the above consideration we have assumed that all the data points are independent.
- One of the common methods to solve an inverse is to minimize a measure of misfit between the data and the theoretical model.

Map Making in CMB

- In CMB experiments like WMAP and Planck the time order data or TOD **d** depends on the sky temperature **T** in the following way:

$$D_i = A_{ij} T_j + N_j \tag{3}$$

  the index i and j are over time and pixels respectively.

- In order to make a map from the TOD we have to estimate T from D for which we can minimize the following function:

$$f(T) = (D - AT)'(D - AT) \tag{4}$$

  which gives:

$$\hat{T} = (A'A)^{-1} A'D \tag{5}$$

CMB Map making

- In place of $f(T)$ if we can also minimize χ^2 :

$$\chi^2(T) = (D - AT)' C_N^{-1} (D - AT) \tag{6}$$

where $C_N =< NN' >$ is the noise covariance matrix, then we get:

$$\hat{T} = (A' C_N^{-1} A)^{-1} A' C_N^{-1} D \tag{7}$$

which is called Maximum-Likelihood (ML) solution.

- The covariance matrix of the maximum likelihood solution is given by :

$$\mathcal{N} = (A' C_N^{-1} A)^{-1} \tag{8}$$

[Tegmark (1997)]

Chi-Square Distribution

- If we have a variable $y = \sum_{i=1}^{N} X_i^2$ and X is drwan from Gaussian random distribution then Y follows χ^2 distribution of degree ν:

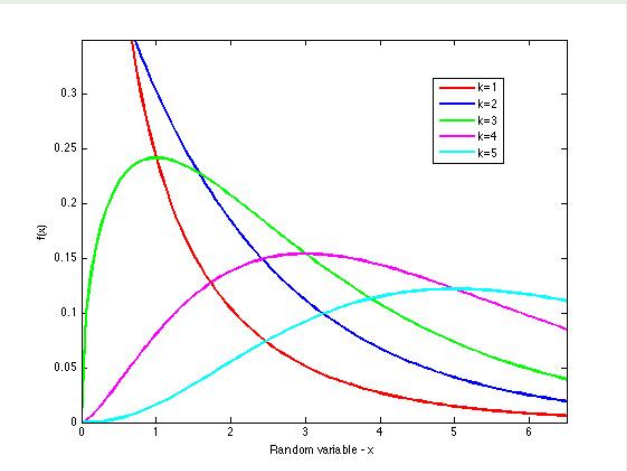
$$f_{\chi^2}(x) = \frac{x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}}{\Gamma(\frac{\nu}{2}) 2^{\frac{\nu}{2}}} \tag{9}$$

Chi-Square Distribution

- If we have a variable $y = \sum_{i=1}^{N} X_i^2$ and X is drwan from Gaussian random distribution then Y follows χ^2 distribution of degree ν:

$$f_{\chi^2}(x) = \frac{x^{\frac{\nu}{2}-1} e^{-\frac{x}{2}}}{\Gamma(\frac{\nu}{2}) 2^{\frac{\nu}{2}}} \tag{9}$$

- For large ν, χ^2 distribution approaches Gaussian.



CMB anisotropies

- CMB temperature anisotropies are expressed in terms of multipoles:

$$\frac{\Delta T(\hat{n})}{T} = \sum_{l=0}^{\infty} \sum_{m=-l}^{m=l} a_{lm} Y_{lm}(\hat{n}) \tag{10}$$

where

$$a_{lm} = \int \frac{\Delta T(\hat{n})}{T} Y_{lm}(\hat{n}) d\hat{n} \tag{11}$$

- Where a_{lm} follow the Gaussian distribution with zero mean and variance given by C_l:

$$< a_{lm} >= 0 \tag{12}$$

and

$$\langle a_{lm} a^*_{l'm'} \rangle = \delta_{ll'} \delta_{mm'} C_l \tag{13}$$

- An unbiased estimator of C_l is defined as:

$$\hat{C}_l = \frac{1}{2l+1} \sum_{m=-l}^{m=l} a_{lm} a^*_{lm} \tag{14}$$

- Note that the angular power spectrum C_l follows χ^2 distribution.

Maximum Likelihood Estimation

- The probability distribution $P(\Theta|\mathbf{d})$ (posterior) for model parameters Θ given data $\mathbf{d}$ can be related to the probability $P(\mathbf{d}|\Theta)$ (likelihood) of an experiment giving data $\mathbf{d}$ for model parameters Θ using the Bayes' theorem:

$$P(\Theta|\mathbf{d}) = \frac{P(\mathbf{d}|\Theta)P(\Theta)}{P(\mathbf{d})} \tag{15}$$

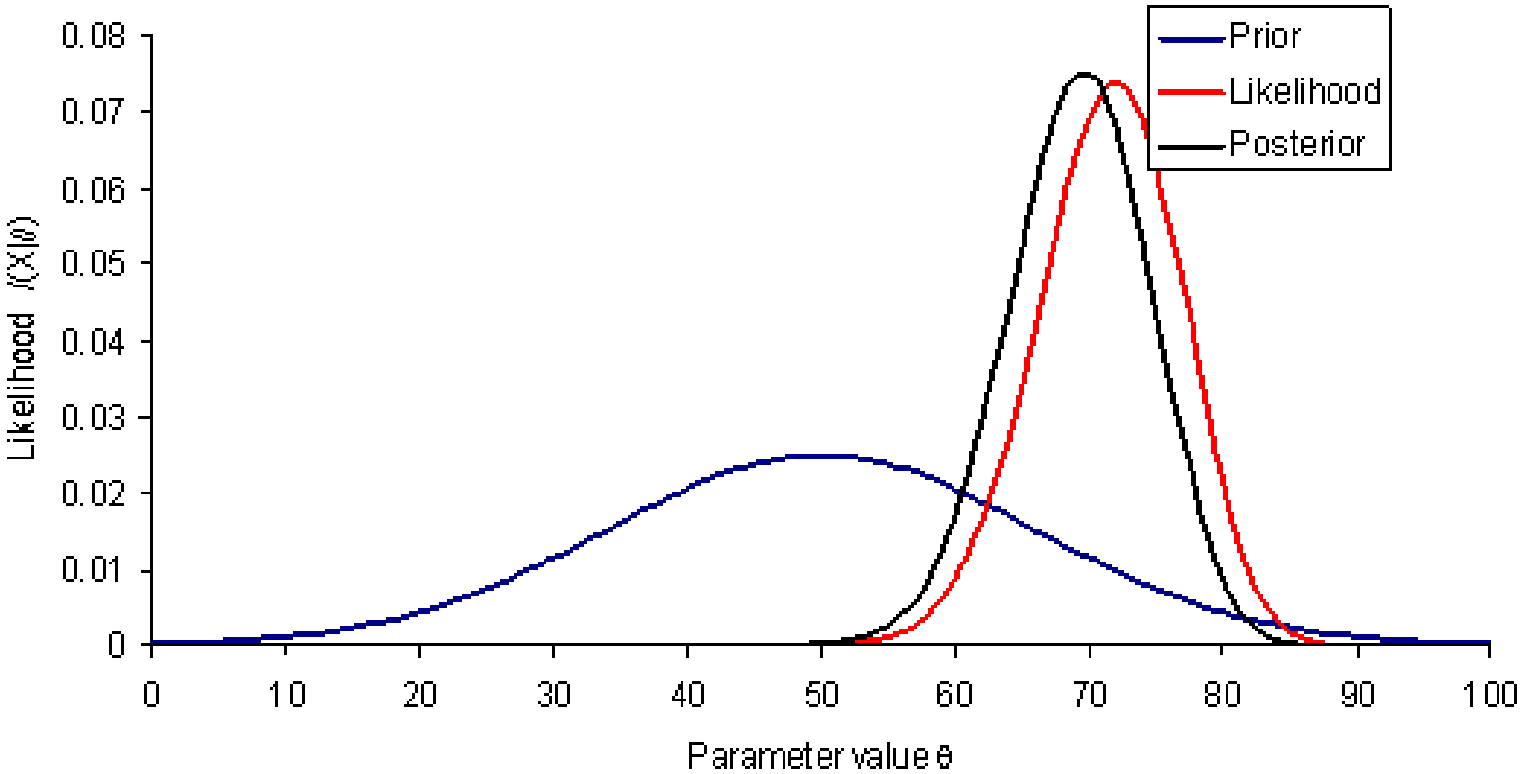
  where $P(\Theta)$ is called the prior and $P(\mathbf{d}) = \sum P(\mathbf{d}|\Theta)P(\Theta)$ is used for the normalization purpose.

- For the case of flat prior, posterior and likelihood are proportional:

$$P(\Theta|\mathbf{d}) = P(\mathbf{d}|\Theta) = L(\mathbf{d}|\Theta) \tag{16}$$

- In Bayesian formalism we can easily incorporate new data in analysis by considering the posterior of the old data as prior.

Bayesian Analysis



Note that when likelihood/posterior is not Gaussian then the average value of the parameter $< \theta >$ may not coincide with the value of θ_0 at which the likelihood/posterior is maximum.

Errors in maximum Likelihood estimation

- We are not only interested in finding the point Θ_0 or $< \Theta >$, we are also also interested errors.
- The spread of the likelihood function $P(\mathbf{d}|\Theta)$ around the maximum likelihood point Θ_0 can be used to find the error bars:
- Expanding the likelihood function $\mathcal{L} = -2 \log \log P(\mathbf{d}|\Theta)$ around the maximum likelihood point Θ_0:

$$\mathcal{L}(\Theta) = \mathcal{L}(\Theta_0) + (\theta^i - \theta_0^i) \frac{\partial \mathcal{L}(\Theta)}{\partial \theta^i}\bigg|_{\theta^i=\theta_0^i} + \frac{1}{2}(\theta^i - \theta_0^i)(\theta^j - \theta_0^j) \frac{\partial^2 \mathcal{L}(\Theta)}{\partial \theta^i \partial \theta^j}\bigg|_{\theta^i=\theta_0^i} + ... \tag{17}$$

Second derivative of the Likelihood function is called the Hessian :

$$H_{ij} = \frac{\partial^2 \mathcal{L}(\Theta)}{\partial \theta^i \partial \theta^j} \tag{18}$$

- There is an in-equality called the Cramer-Rao lower bound which says that the error bars (variance) on any estimator cannot be smaller than the inverse of information matrix I

$$C_{ij} \geq \frac{1}{I_{ij}} \tag{19}$$

where I is defined as:

$$I_{ij} = \left\langle \frac{\partial^2 \mathcal{L}(\Theta)}{\partial \theta^i \partial \theta^j} \right\rangle \tag{20}$$

- How accurately a parameter can be estimated from the likelihood depends upon how sensitive the likelihood function is on that parameter, which is quantified by the Hessian matrix.

Problems

1. Show that for a case when the noise is Gaussian, maximizing likelihood is equivalent to minimization Chi-square.
2. Show that the maximum likelihood estimator is the minimum variance estimator.
3. Show that finding the solution of the linear problem $\mathbf{d} = A\Theta$ is equivalent to finding of the maximum of the following function:

$$f(\Theta) = \frac{1}{2}\Theta' A\Theta - \mathbf{d}'\Theta + c \tag{21}$$

4. Show that the minimum chi-square solution of the model $y = c_0 + c_1 x$ with data (x_i, y_i, σ_i) is given by:

$$c_0 = \frac{S_{01}S_{00} - S_{00}S_{01}}{SS_{00} - S_{00}^2} \quad \text{and} \quad c_1 = \frac{SS_{01} - S_{01}S_{10}}{SS_{00} - S_{00}^2} \tag{22}$$

where $S = \sum_{i=1}^{N} 1/\sigma_i^2$ and $S_{ij} = \sum_{k=1}^{N} x^i y^j/\sigma_k^2$. Also find the error.

- Sometime likelihood is not much sensitive to individual parameters (may not be Gaussian) but is highly sensitive to some combinations of those.
- For example, CMB Likelihood is almost Gaussian with respect to the combination $(\Omega_b h^2, \Omega_c h^2, \theta, A^*, t_z)$ of cosmological parameters [Chu et al. (2003)] where

$$A^* = \frac{A}{76,000}\left(\frac{0.05 Mpc^{-1}}{k_{\text{pivot}}}\right)^{1-n_s} e^{-2\tau}, \tag{23}$$

and

$$t = \frac{1}{\sqrt{\Omega_b h^2}} 2^{n_s - 1} \tag{24}$$

- Once we have the probability distribution $P(\Theta|\mathbf{d})$ for model parameters Θ we can statistics of the parameters:

$$< \Theta >= \int \Theta d\Theta P(\Theta|\mathbf{d}) \tag{25}$$

- In practice, before carrying out the above integral we find out the one dimensional probability distribution by marginalization over other parameters:

$$P(\theta_r) = \int d\theta_1 d\theta_2 .... d\theta_{r-1} d\theta_{r+1} ... d\theta_M P(\theta_1, \theta_2, ... s\theta_M) \tag{26}$$

and

$$< \theta_r >= \int \theta_r d\theta_r P(\theta_r) \tag{27}$$

- Carrying out multi-dimensional integration is very expansive i.e., computational cost grows as $O(n^M)$ where n is the number of grid points along one direction and M is the dimensionality of the parameter space.
- If we can replace the multi-dimensional integration by summation over a finite number of points which represent the probability distribution function then computational cost becomes manageable.

$$< \Theta >= \int \Theta d\Theta P(\Theta|\mathbf{d}) = \frac{1}{N} \sum_{i=1}^{N} \Theta_i P(\Theta_i|\mathbf{d}) \qquad (28)$$

- Markov-Chain Monte Carlo sampling samples the likelihood function in such a way that there are more point in the region where the likelihood function has the large values and less where it has small values.

Markov-Chain Monte Carlo

- When we toss a coin n times then the outcome of the n^{th} toss does not depend on the outcome of any of the previous outcomes.

Markov-Chain Monte Carlo

- When we toss a coin n times then the outcome of the n^{th} toss does not depend on the outcome of any of the previous outcomes.
- In a Markov-Chain the probability of a random variable X_n to have value x_n at step n depends on the probability of the variable X_{n-1} to have the value x_{n-1} at step $n-1$.

$$P(X_N) = P(X_n, X_{n-1})P(X_{n-1}) \tag{29}$$

where $P(X_n, X_{n-1})$ is called the transition probability, transition kernel or proposal density.

Markov-Chain Monte Carlo

- When we toss a coin n times then the outcome of the n^{th} toss does not depend on the outcome of any of the previous outcomes.
- In a Markov-Chain the probability of a random variable X_n to have value x_n at step n depends on the probability of the variable X_{n-1} to have the value x_{n-1} at step $n-1$.

$$P(X_N) = P(X_n, X_{n-1})P(X_{n-1}) \tag{29}$$

  where $P(X_n, X_{n-1})$ is called the transition probability, transition kernel or proposal density.
- In most cases transition kernel is symmetric:

$$P(X_n, X_{n-1}) = P(X_{n-1}, X_n) \tag{30}$$

Markov-Chain Monte Carlo

- When we toss a coin n times then the outcome of the n^{th} toss does not depend on the outcome of any of the previous outcomes.
- In a Markov-Chain the probability of a random variable X_n to have value x_n at step n depends on the probability of the variable X_{n-1} to have the value x_{n-1} at step $n-1$.

$$P(X_N) = P(X_n, X_{n-1})P(X_{n-1}) \tag{29}$$

  where $P(X_n, X_{n-1})$ is called the transition probability, transition kernel or proposal density.
- In most cases transition kernel is symmetric:

$$P(X_n, X_{n-1}) = P(X_{n-1}, X_n) \tag{30}$$

- The transition probability $P(X_n, X_{n-1})$ has the remarkable property that after an initial burn-in period it generates a sample which has the probability distribution $P(X)$.

Example

We consider the following Gaussian transition probability:

$$P(x_n|X_{n-1}) \propto \exp\left[-\frac{(X_n - X_{n-1})^2}{2\sigma^2}\right] \tag{31}$$

where σ is generally a fixed parameter called the step size. We can go from $(n-1)^{th}$ step to n^{th} step using the above transition probability using the Metropolis-Hasting algorithm.

- The choice of proposal density can affect the way the sampling algorithm works so it is advisable to use a proposal density which is of the similar shape of the distribution we are aiming to sample.

[Lewis & Bridle (2002)]

Metropolis-Hesting Algorithm

- The first step of the algorithm is to set the initial value of the random variable i.e., $X(n = 0) = X_0$ which should not be very far from the best fit value (maximum likelihood value) of the parameter.
- Once the initial step is set, we can find a proposed value Y for the $(n + 1)^{th}$ step using the proposal density $P(Y|X_n)$.
- In order to decide whether we should accept Y as X_{n+1} we compute the Metropolis ratio r

$$r = \frac{P(Y)P(X_n|Y)}{P(X_n)P(Y|X_n)} = \frac{P(Y)}{P(X_n)} \quad \text{for symmetric proposal density}$$

- If $r \geq 1$ then we set $X_{n+1} = Y$ otherwise we accept Y with probability r i.e., we draw a uniform random number U and set $X_{n+1} = Y$ only when $U > r$.

[Gregory (2005)]

CMB Likelihood

- CMB temperature and polarization observations can constrain cosmological parameters if the likelihood function can be computed exactly.
- Computing the likelihood function exactly in a brute force way is computationally challenging since it involves inversion of the covariance matrix i.e., $O(N^3)$ computation.
- In Cosmological parameter estimation a theoretical model is represented by its angular power spectrum C_l.
- For a set cosmological parameters we can compute the angular power spectrum C_l using publicly available Boltzmann codes like CMBFAST and CAMB and try to fit that with observed C_l.

- From Bayes theorem the posterior for the parameter C_l with data T is given by:

$$P(C_l|T) = \frac{P(T|C_l)P(C_l)}{P(T)} \tag{32}$$

where C_l is the theoretical C_l and T is the observed sky map of CMB anisotropies.

$$T(\hat{n}) = \sum_{lm} a_{lm} Y_{lm}(\hat{n}) \tag{33}$$

- Since computing the exact likelihood function is challenging, approximations are generally made (Gaussian Likelihood, Gibbs sampling etc).

$$L(T|C_l) \propto \frac{1}{\sqrt{|S|}} \exp[-(TS^{-1}T)/2] \tag{34}$$

where the covariance matrix S is related to angular power spectrum:

$$< T(\hat{n}_i)T(\hat{n}_j) >= S_{ij} = \sum_l \frac{2l+1}{4\pi} C_l P_l(\hat{n}_i.\hat{n}_j) \tag{35}$$

[Verde et al. (2003); Hamimeche & Lewis (2008)]

- In terms of C_l the likelihood function can be written as:

$$L(T|C_l) = \prod_{lm} \frac{1}{\sqrt{C_l}} \exp[-|a_{lm}|^2/(2C_l)] \tag{36}$$

- Since we observe only one sky so we cannot measure the power spectra directly, but instead form the rotationally invariant estimators, C_l, for full-sky CMB maps given by

$$\hat{C}_l = \frac{1}{2l+1} \sum_{m=-l}^{m=l} |a_{lm}|^2 \tag{37}$$

Problem 5

Show that $\hat{C}_l$ as given by equation (37) is an unbiased estimator i.e., $< \hat{C}_l >= C_l$ (true power spectrum).

- Note that the likelihood has a maximum when $C_l = \hat{C}_l$ so $\hat{C}_l$ is the MLE.

Problem 6

From equation (36) show that:

$$\chi^2 = -2 \log L(\hat{C}_l | C_l) = \sum_l (2l+1) \left[\log \left(\frac{C_l}{\hat{C}_l} \right) + \frac{\hat{C}_l}{C_l} - 1 \right] \quad (38)$$

This expression for likelihood does not consider:

- Finite resolution of the detector - window function
- Detector noise
- Cut-sky $f_{\rm sky}$ to avoid foreground etc., which leads correlations among different multi-poles.

When all these factors are taken into account computing the likelihood function becomes challenging.

CosmoMC

- In general, cosmological parameters from a CMB experiment like WMAP or Planck are estimated using a publicly available Markov Chain Monte Carlo code called CosmoMC [Lewis & Bridle (2002)].
- CosmoMC has been successfully used to estimate cosmological parameters from WMAP data and a detail discussion and working of the code is also discussed in a WMAP first year paper [Verde et al. (2003)]
- CosmoMC used publicly available code CAMB [Lewis et al. (2000)] for computing theoretical C_ls.
- WMAP team has provided a code for computing the likelihood from the temperature and polarization data.

Cosmological Parameters

S. No	Parameter	Description
1	$\Omega_b h^2$	physical baryon density
2	$\Omega_{\rm DM} h^2$	physical dark matter density (CDM+massive neutrinos)
3	θ	100 $\times$ ratio of the angular diameter distance to the LSS sound horizon
4	τ	reionization optical depth
5	Ω_K	spatial curvature
6	f_ν	neutrino energy density as fraction of $\Omega_{\rm DM} h^2$
7	w	constant equation of state parameter for scalar field dark energy
8	n_s	spectral index for scalar power spectrum
9	n_t	spectral index for tensor power spectrum
10	n_{run}	running for the index for scalar power spectrum
11	$log[10^{10} A_s]$	Amplitude of the scalar power spectrum
12	r	ratio of tensor to scalar primordial amplitudes at pivot scale
13	A_{SZ}	SZ template amplitude, as in WMAP
14	Ω_Λ	energy density (parameter) Cosmological constant
15	Age/Gyr	Age of the universe
16	Ω_m	dark matter density
17	σ_8	Mass variance at 8 Mpc
18	z_{re}	Redshift of the reionization
19	r_{10}	tensor-scalar C_l amplitude at l=10
20	H_0	Hubble parameter is H0 km/s/Mpc

CosmoMC has 20 parameters which can be estimated from a CMB data set. Note that not all the parameters are independent, in fact we can change just seven parameters (shown in red) to fit a CMB data set (WMAP).

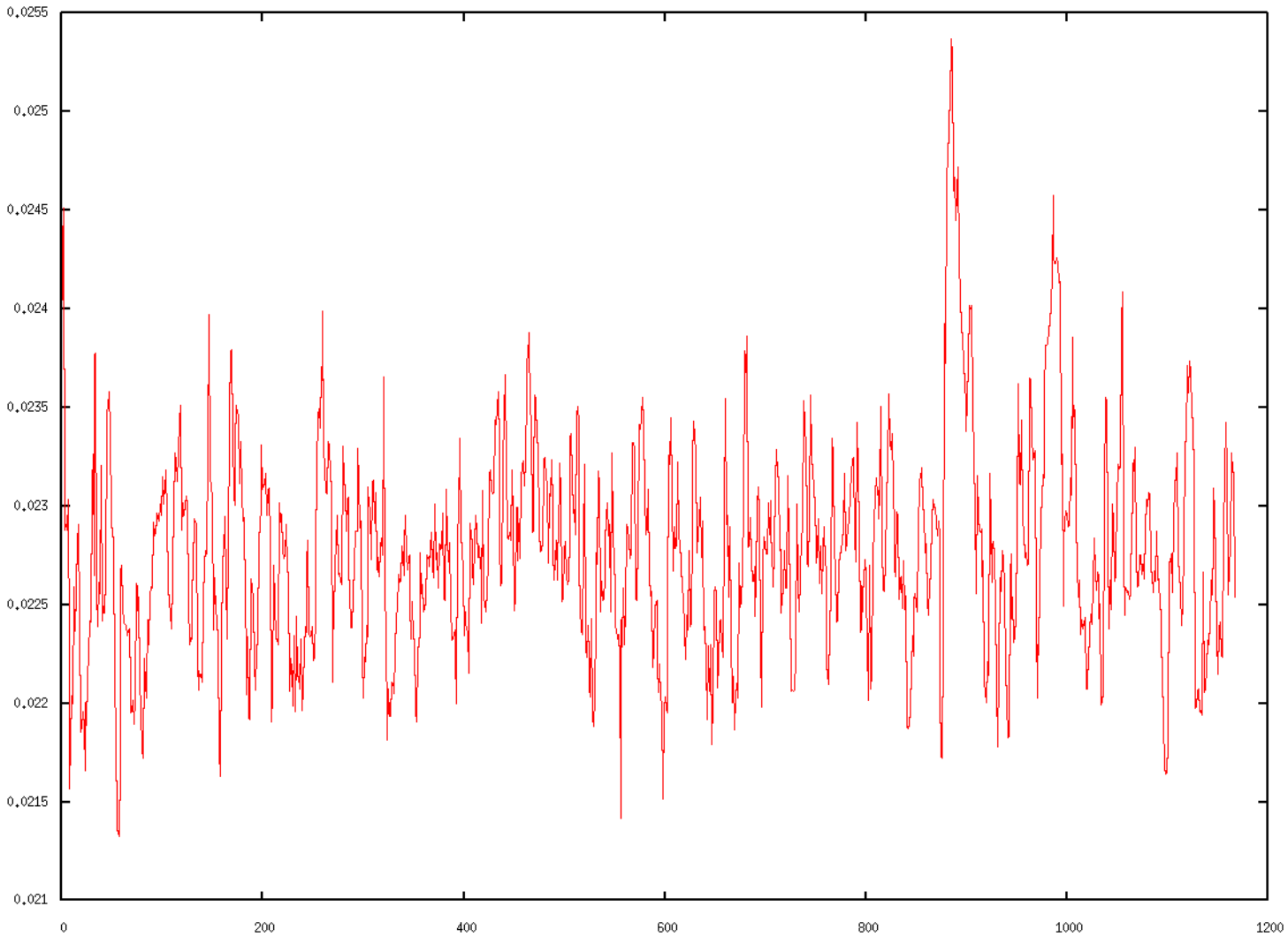
How CosmoMC works?

- Corresponding to every cosmological parameter θ_i we want to estimate we give the (1) search range (2) an starting point and guess (2) mean value and (3) standard deviation around the mean.
- At every step we compute the likelihood at current location Θ in the parameter space and move to new location Θ_1 in the parameter space by taking a random step and compute:

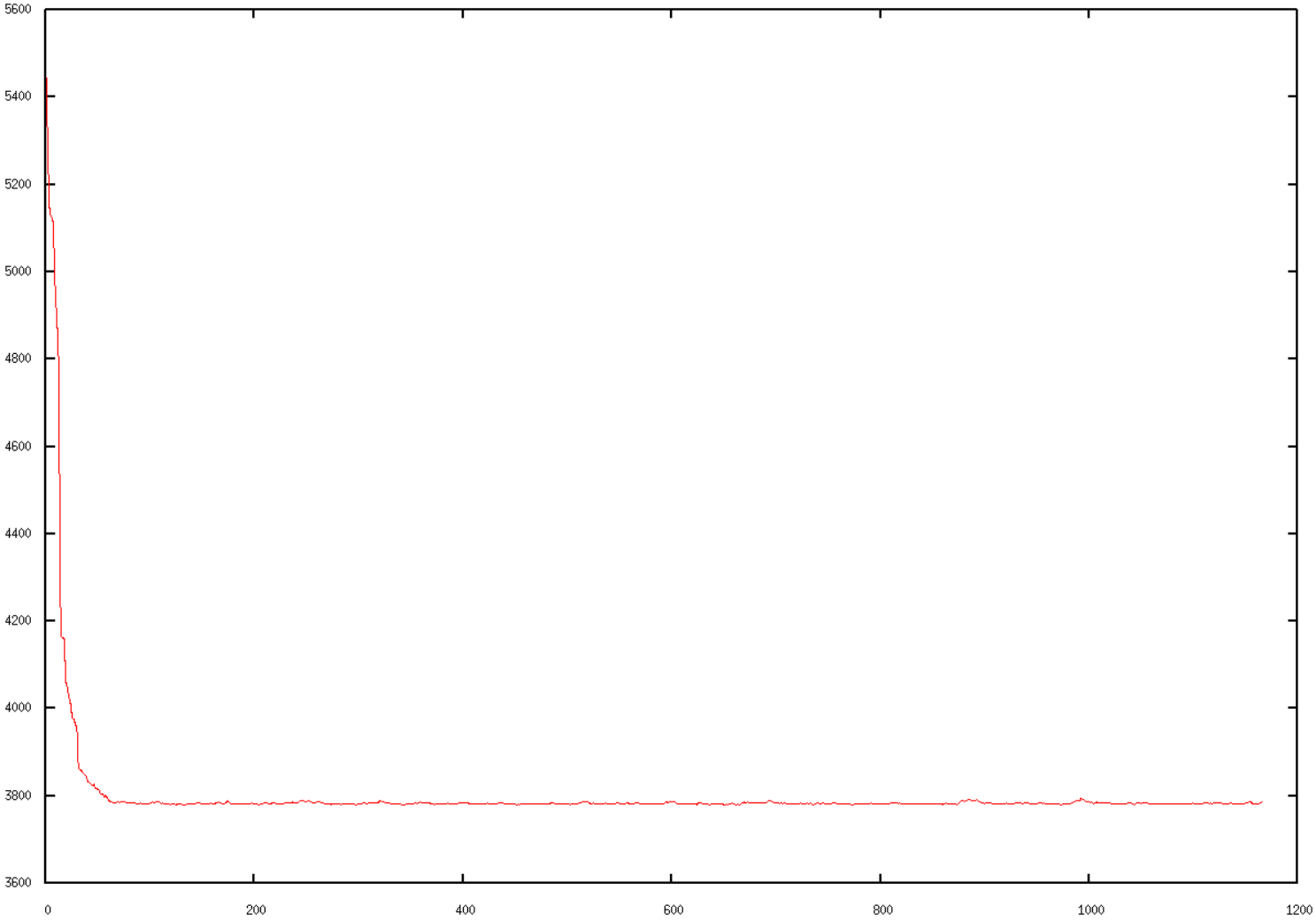
$$r = \frac{\mathcal{L}(\Theta_1)}{\mathcal{L}(\Theta_1)} \tag{39}$$

  if $r > 1$ then we accept Θ_1 as the new point in the chain otherwise compare r with a uniform random number u and accept Θ_1 only when $r > u$.
- We start multiple chains (random walks) in parallel and test for a convergence criteria at every step.
- Once we reach convergence, we compute the various statistical measures from the chains.

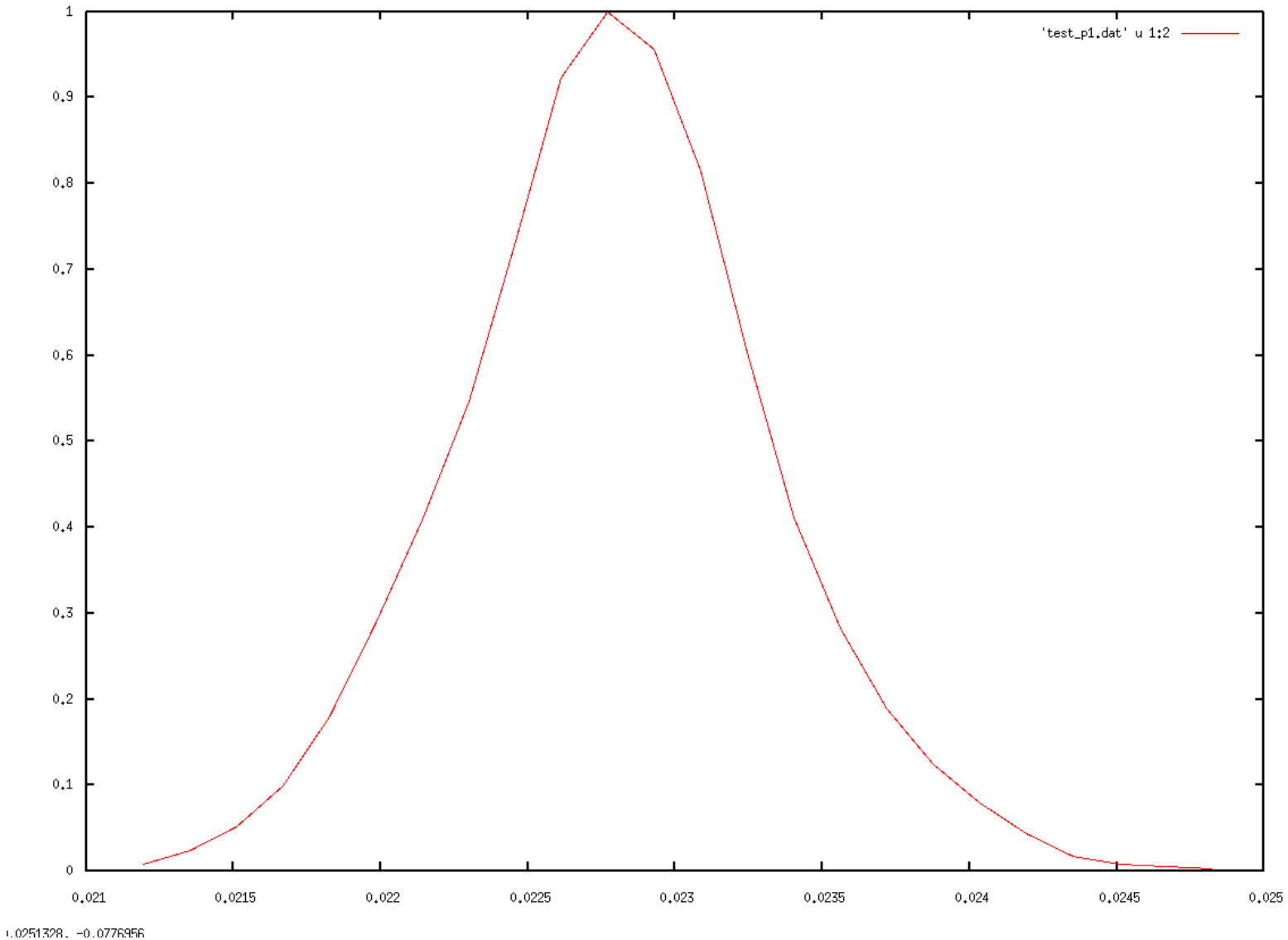
CosmoMC



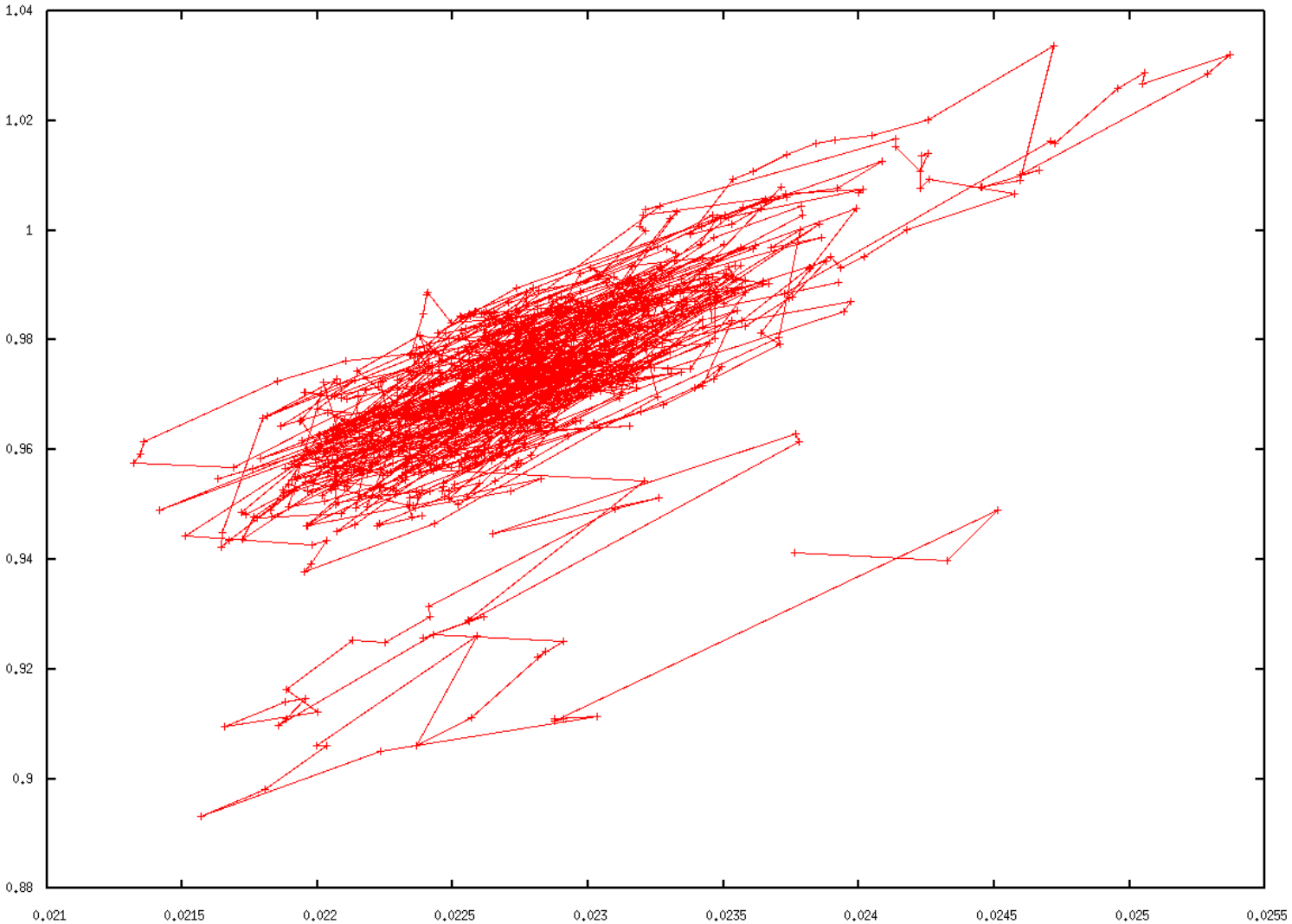
A typical Markov chain in CosmoMC.



As a chain progresses $\chi^2 = -2 \log L$ decreases



From the chains we can plot the probability distribution for parameters.



We can also plot scatter and contour plots from the chains.

Summary and Conclusions

References

Chu, M., Kaplinghat, M., & Knox, L. 2003, Astrophys. J. , 596, 725

Gregory, P. C. 2005, Bayesian Logical Data Analysis for the Physical Sciences: A Comparative Approach with 'Mathematica' Support (Cambridge University Press)

Hamimeche, S., & Lewis, A. 2008, Phys. Rev. D , 77, 103013

Lewis, A., & Bridle, S. 2002, Phys. Rev. D , 66, 103511

Lewis, A., Challinor, A., & Lasenby, A. 2000, Astrophys. J. , 538, 473

Tegmark, M. 1997, Astrophys. J. Lett. , 480, L87

Verde, L., et al. 2003, Astrophys. J. Suppl. Ser. , 148, 195