PROCEEDINGS OF SCIENCE

Methods and tools for statistical analyses of CMB data

Anthony Lasenby and Michael Hobson*
Astrophysics Group, Cavendish Laboratory, Madingley Road, Cambridge CB3 0HE, UK

E-mail: a.n.lasenby@mrao.cam.ac.uk, mph@mrao.cam.ac.uk

Bayesian inference provides a unified framework for data analysis. Bayesian methods for parameter estimation are now used extensively in cosmology, but Bayesian model selection is still relatively new to the CMB community. We begin by introducing the concept of the Bayesian evidence and explain its use in model selection. After illustrating the approach with some simple examples, methods for computing the evidence are presented, including Gaussian approximation, the Savage–Dickey density ratio, thermodynamic integration and nested sampling. We also present some recent methods for performing efficient nested sampling. The use of the evidence in a cosmological context is then illustrated through some brief case studies. In particular, we outline how the evidence can be useful in general cosmological model selection, primordial power spectrum modelling, rotating universe modelling, checking the consistency of datasets, component separation and object detection.

CMB and Physics of the Early Universe
20-22 April 2006
Ischia, Italy

*Talk delivered by both authors.

PoS(CMB2006)014

1. Introduction

The general topic of statistical analysis of CMB data is a very wide area. Rather than present a very broad overview, we instead concentrate on a single unifying topic, namely the use of the evidence in Bayesian inference. Bayesian methods for parameter estimation are now very widely accepted within the CMB community, but the use of the evidence to select between different models for the data is a relatively recent development. We begin by introducing the concept of the evidence, then explain how it can be computed and conclude by illustrating its use in a number of cosmological examples.

2. Model selection and Bayesian evidence

Let us begin by defining the evidence in completely general terms. Suppose we collect a set of N data points D_i $(i = 1,2,\ldots,N)$, which we denote collectively as the data vector $\boldsymbol{D}$. Suppose further that we propose some model (or hypothesis) H for the data that depends on a set of M parameters θ_j $(j = 1,\ldots,M)$, that we denote by the parameter vector $\boldsymbol{\theta}$.

Bayes' theorem states that

$$\Pr(\boldsymbol{\theta}|\boldsymbol{D},H) = \frac{\Pr(\boldsymbol{D}|\boldsymbol{\theta},H)\Pr(\boldsymbol{\theta}|H)}{\Pr(\boldsymbol{D}|H)}, \tag{2.1}$$

where the meaning of each term is as follows. The prior $\Pr(\boldsymbol{\theta}|H)$ represents our state of knowledge (or prejudices) about the parameter values before analysing the data. This is modulated by the likelihood, $\Pr(\boldsymbol{D}|\boldsymbol{\theta},H)$, of the data given a particular set of parameter values. This product gives (to within a constant factor) the posterior $\Pr(\boldsymbol{\theta}|\boldsymbol{D},H)$, which encodes all the inferences regarding the parameters $\boldsymbol{\theta}$. The normalisation of the posterior is given by the evidence $\Pr(\boldsymbol{D}|H)$, and it is this quantity that may be used to decide which of a set of alternative models best describes the data.

Suppose, for example, that we have two alternative models H_0 and H_1 for describing a dataset $\boldsymbol{D}$, where H_0 depends on the parameter set $\boldsymbol{\theta}_0$, and H_1 on the set $\boldsymbol{\theta}_1$. For H_i $(i = 0,1)$, the probability density associated with the observed data $\boldsymbol{D}$ is

$$\Pr(\boldsymbol{D}|H_i) = \int \Pr(\boldsymbol{D}|\boldsymbol{\theta}_i,H_i)\Pr(\boldsymbol{\theta}_i|H_i)\,d\boldsymbol{\theta}_i. \tag{2.2}$$

In either case H_0 or H_1, the evidence is the average of the likelihood with respect to the prior. Thus a model has a large evidence if more of its allowed parameter space is likely, given the data. Conversely, a model has a small evidence if there are large areas of its allowed parameter space with low likelihood values. Hence evidence naturally incorporates Occam's razor: a simpler theory is preferred to a more complicated one, unless the latter is significantly better at describing the data. In performing model selection, one then merely needs to consider the ratio

$$\frac{\Pr(H_1|\boldsymbol{D})}{\Pr(H_0|\boldsymbol{D})} = \frac{\Pr(\boldsymbol{D}|H_1)}{\Pr(\boldsymbol{D}|H_0)}\frac{\Pr(H_1)}{\Pr(H_0)}, \tag{2.3}$$

in which the prior probabilities of the hypotheses also appear. It is often true that $\Pr(H_0) = \Pr(H_1)$, in which case the preferred model is simply that with the largest evidence. In some cases, however, the priors are not equal and the proper form (2.3) should be used.

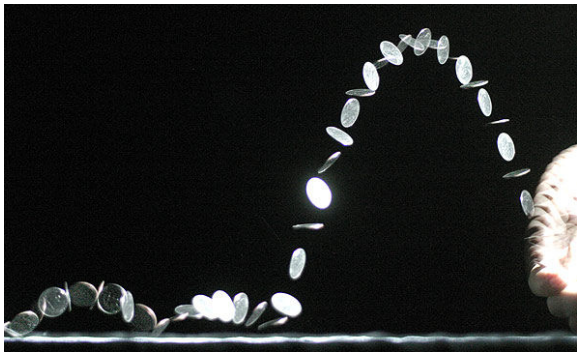


Figure 1: A coin tossing experiment.

2.1 A simple example

A real case (reported in the newspapers) in which Bayesian model selection has been applied concerned a Belgian one Euro coin (see [16]). In a coin tossing experiment the coin came down heads 140 times, and tails 110. What is the evidence ratio for H_1 'it is biased' versus H_0 'it is fair'?

Clearly one can only answer this with a definite form for H_1, so let us assume H_1 corresponds to a uniform prior over $[0,1]$ for the probability p of heads: so $\Pr(p|H_1) = 1$. Then,

$$\Pr(D|H_1) = \int \Pr(D|p,H_1)\Pr(p|H_1)\,dp = \int_0^1 p^{n_H}(1-p)^{n_T}\,dp = \frac{n_H!n_T!}{(n_H+n_T+1)!}$$

Meanwhile, if the coin is fair then $\Pr(D|H_0) = (1/2)^{n_H+n_T}$. Thus, for the numbers given, the ratio of evidences is

$$\frac{\Pr(D|H_1)}{\Pr(D|H_0)} = \frac{2^{250}\,140!\,110!}{251!} = 0.48.$$

If the two hypothesis are equally likely a priori, so that $\Pr(H_0) = \Pr(H_1)$, then (2.3) shows the hypothesis H_0 that the coin is fair is favoured by 2 to 1 relative to our alternative hypothesis H_1. As discussed in [16], by different choice of priors on p, tailored to be more favourable to the outcome actually observed, it is possible to reverse the sense of this comparison. However, even the most extreme choice of prior is unable to match the type of 'probablity' in favour of H_1 that standard frequentist significance methods yield.

2.2 Another simple example (more relevant to astronomy)

Suppose we have data at known sample points and want to know if there is a 'trend' present (see Fig. 2). Thus, the two alternative models for the data are: H_1: $y_i = a_0 + a_1 x_i + \varepsilon_i$ and H_0: $y_i = a_0 + \varepsilon_i$, where ε is a noise vector belonging to $N(0,\sigma^2)$ (say). To perform a model selection, we need to specify priors on a_0 and a_1. If we let these be uniform (and uncorrelated) over $(-\infty,\infty)$ then we can perform the integrals analytically (the actual form of priors is not too important if data is definitive). In this case, one finds that

$$\frac{\Pr(D|H_1)}{\Pr(D|H_0)} = \sqrt{\frac{2\pi\sigma^2}{\sum(x_i-\bar{x})^2}}\exp\left\{\frac{[\sum(x_i-\bar{x})(y_i-\bar{y})]^2}{2\sigma^2\sum(x_i-\bar{x})^2}\right\}, \tag{2.4}$$

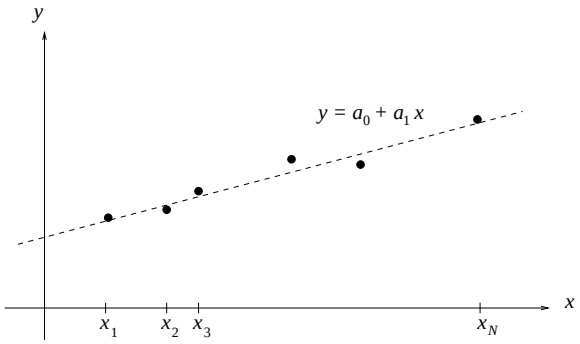


Figure 2: Linear regression.

which depends on the (positive) exponential of the correlation coefficient squared. Note that one aspect we have glossed over here is that the use of an infinite range has led to improper priors on a_0 and a_1, and there has been no attempt in (2.4) to deal with the infinite normalisation factors which arise from these. Despite this, the emergence in this approach of the correlation coefficient as the important statistic, is clearly satisfying.

3. Evaluation of the evidence

In general, evaluation of the evidence integral (2.2) has to be performed numerically, and is computationally challenging. For ease of notation, let us first rewrite Bayes' theorem (2.1) as

$$\Pr(\theta|\boldsymbol{d},H_i) = \frac{\Pr(\boldsymbol{d}|\theta,H_i)\Pr(\theta|H_i)}{\Pr(\boldsymbol{d}|H_i)} \quad \rightarrow \quad P_i(\theta) = \frac{L_i(\theta)\pi_i(\theta)}{E_i},$$

so that the evidence integral becomes

$$E_i = \int L_i(\theta)\pi_i(\theta)\,d\theta.$$

If the dimension M of the parameter space is small ($M \lesssim$ few), one may calculate the unnormalised posterior $\overline{P}(\theta) = L(\theta)\pi(\theta)$ over a grid in parameter space and perform simple quadrature to obtain the evidence trivially. For higher-dimensional problems, this approach rapidly becomes impossible and one needs to find alternative methods.

3.1 Gaussian approximation to the posterior

The simplest approach is to use a multivariate Gaussian approximation to the 'unnormalised' posterior about its peak (see e.g. [7])

$$\overline{P}_i(\theta) \approx \overline{P}_i(\hat{\theta}) \exp\left[-\tfrac{1}{2}(\theta-\hat{\theta})^{\mathrm{t}}\,\boldsymbol{V}^{-1}(\theta-\hat{\theta})\right],$$

where $\boldsymbol{V}^{-1} = -\boldsymbol{H} = -\nabla\nabla \ln \overline{P}_i(\theta)\big|_{\theta=\hat{\theta}}$ is the inverse covariance matrix. In this approximation, the evidence integral is analytic and given by $E_i \approx (2\pi)^{M_i/2}|\boldsymbol{V}_i|^{1/2} L_i(\hat{\theta})\pi_i(\hat{\theta})$. Hence the log evidence ratio is

$$\ln\left(\frac{E_0}{E_1}\right) = \ln\left(\frac{\hat{L}_0}{\hat{L}_1}\right) + \tfrac{1}{2}\left[(M_0-M_1)\ln(2\pi) + \ln\left(\frac{|\boldsymbol{V}_0|}{|\boldsymbol{V}_1|}\right)\right] + \ln\left(\frac{\hat{\pi}_0}{\hat{\pi}_1}\right),$$

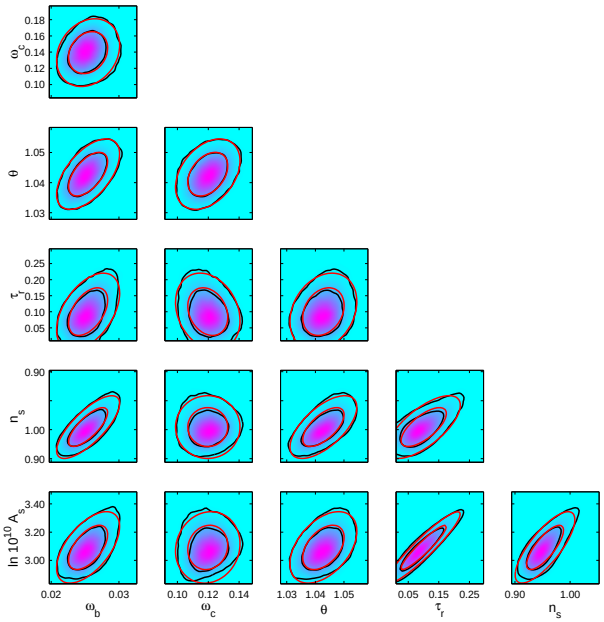


Figure 3: Gaussian approximation to the posterior for the standard ΛCDM model with parameters $(\omega_b, \omega_c, \theta, \tau, \ln A, n_s)$, where θ is the ratio of the angular diameter distance to the sound horizon at decoupling; the data sets used are WMAP1, ACBAR, CBI, VSA, SDSS and HST. The black contours are derived from 10^4 MCMC samples and the red contours are the Gaussian approximation (from [27]).

where the hats on variables denote their values at the peak of the posterior. In particular, we note that $\ln(\hat{L}_0/\hat{L}_1) = -\frac{1}{2}\Delta\hat{\chi}^2$ if the likelihoods are Gaussian, and $\ln(\hat{\pi}_0/\hat{\pi}_1) = \ln(\Delta\theta_1/\Delta\theta_0)$ if the priors are uniform with widths $\Delta\theta_i$.

In using this approach, it can be useful to choose 'normal' parameters to improve the accuracy of the approximation. In cosmological parameter estimation in particular, the Gaussian approximation can be made very accurate by using the 'physical' variables proposed by [13] (see Fig. 3).

3.2 Savage–Dickey density ratio

The Gaussian approximation is poor for complicated or multimodal posteriors, especially in the wings of the distribution and at any abrupt cut-offs resulting from priors. One can, however, calculate an exact evidence for an arbitrary posterior using the Savage–Dickey density ratio [6], provided: (i) H_0 and H_1 are nested hypothesis, which implies $L_0(\theta) = L_1(\theta, \psi = \psi_0)$; and (ii) the prior on the parameters is separable, which imples $\pi_1(\theta, \psi) = \pi_0(\theta)\pi_1(\psi)$. In this case, the evidence ratio becomes

$$\frac{E_0}{E_1} = \frac{P_1(\psi_0)}{\pi_1(\psi_0)},$$

where $P_1(\psi_0) = \int P_1(\theta, \psi_0)\, d\theta$ is the properly normalised marginalised posterior for the model H_1, evaluated at $\psi = \psi_0$. The main problem with this method, however, is that the estimation of the marginalised posterior needs MCMC sampling, often requiring some annealing to probe the wings of the distribution. Hence the resulting evidence value is stochastic and is often just as diffcult to evaluate in practice as the more general methods we discuss below.

3.3 MCMC sampling and thermodynamic integration

Calculating the evidence using MCMC sampling (with annealing) from the full posterior requires no assumptions regarding hypotheses or priors. The basic method is thermodynamic integration (see e.g. [9]). One begins by defining

$$E(\lambda) = \int L^{\lambda}(\theta)\pi(\theta)\,d\theta, \tag{3.1}$$

so the required evidence value is $E(1)$. One then performs MCMC sampling from $L^{\lambda}(\theta)\pi(\theta)$, starting with $\lambda = 0$ and slowly raising its value according to some annealing schedule until $\lambda = 1$. The N_s samples corresponding to any particular value of λ may be used to obtain an estimate of the quantity

$$\langle \ln L \rangle_{\lambda} \equiv \frac{\int (\ln L) L^{\lambda} \pi \, d\theta}{\int L^{\lambda} \pi \, d\theta} \approx \frac{1}{N_s} \sum_{i=1}^{N_s} \ln L(\theta_i),$$

From (3.1), this quantity can also be written as

$$\langle \ln L \rangle_{\lambda} = \frac{1}{E} \frac{dE}{d\lambda} = \frac{d \ln E}{d\lambda},$$

and so the (log of) the evidence is given by

$$\ln E(1) = \ln E(0) + \int_0^1 \langle \ln L \rangle_{\lambda} \, d\lambda \approx \sum_{j=1}^{N_{\lambda}} \langle \ln L \rangle_{\lambda_j} \Delta\lambda_j,$$

where we use the fact that $E(0) = 1$, and where N_{λ} and $\Delta\lambda_j$ are the number of λ values and the corresponding stepsizes in the annealing schedule.

Although entirely general in its applicability, thermodynamic integration clearly produces evidence values that are stochastic. The major problem, however, is that accurate evidence values require slow annealing. Moreover, common schedules, such as linear or geometric ones, can get stuck in local maxima. Nonetheless, [21] proposes a 'selective annealing' method in which 'bad' regions die without tunneling and no 'good' sample is ever destroyed. This method is not so troubled by the existence of local optima, but the annealing still slows at phase transitions of the system. It is also worth noting that, independent of its use in thermodynamic integration, annealing can greatly improve MCMC chain mobility during burn-in by applying the likelihood gradually.

3.4 Nested sampling

A new technique for efficient evidence evaluation (and the production of posterior samples) has recently been proposed by [22]. In this approach, one begins by defining the quantity

$$X(\lambda) = \int_{L(\theta) > \lambda} \pi(\theta) \, d\theta,$$

which is simply the prior mass contained within the region of parameter space over which the value of the likelihood exceeds λ. It is useful also to define the inverse function $L(X)$, such that $L(X(\lambda)) = \lambda$. One may now make a change of variable that converts the multi-dimensional evidence integral (2.2) into the one-dimensional integral (see Fig. 4).

$$E = \int L(\theta)\pi(\theta)\,d\theta = \int_0^1 L(X)\,dX.$$

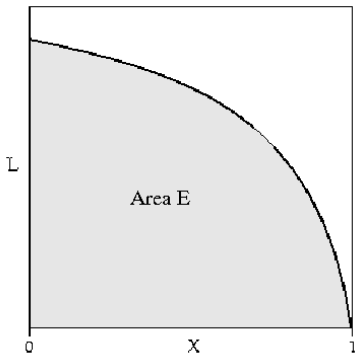
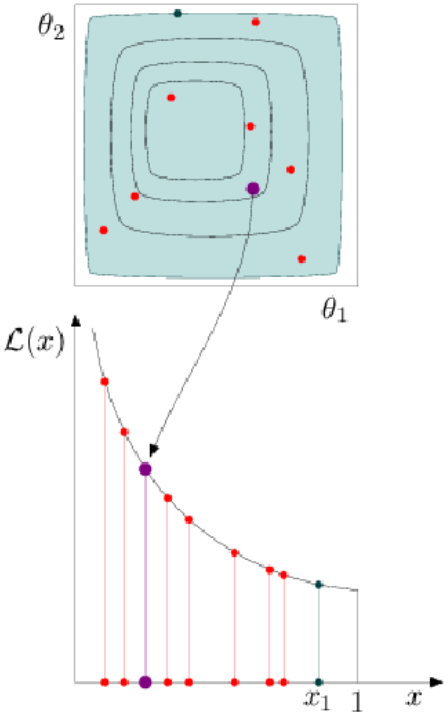


Figure 4: Geometrical interpretation of the variables in nested sampling.



1. Set $j=0$; initially $X_0=1$, $E=0$
2. Sample N points $\{\theta_i\}$ randomly from $\pi(\theta)$ and calculate their likelihoods
3. Set $j \to j+1$
4. Find point with lowest likelihood value (L_j)
5. Remaining prior volume $X_j = t_j X_{j-1}$ where $\Pr(t_j|N) = N t_j^{N-1}$; or just use $\langle t_j \rangle = N/(N+1)$
6. Increment evidence $E \to E + L_j w_j$
7. Remove lowest likelihood point from active set
8. Replace with new point sampled from $\pi(\theta)$ within hard-edged region $L(\theta) > L_j$
9. If $L_{\max} X_j < \alpha E$ (where $\alpha =$ some tolerance) $\Rightarrow E \to E + X_j \sum_{i=1}^{N} L(\theta_i)/N$; stop else goto 3

Figure 5: The nested sampling algorithm and its pictoral representation (the latter from [16]).

Let us now suppose one can evaluate $L_j = L(X_j)$ where $0 < X_m < \cdots < X_2 < X_1 < 1$. In this case, one can therefore estimate E by any numerical method:

$$E = \sum_{j=1}^{m} L_j w_j, \tag{3.2}$$

where w_j are an appropriate set of weights; for a simple trapezium rule $w_j = \frac{1}{2}(X_{j-1} - X_{j+1})$.

In nested sampling the summation (3.2) is performed as illustrated in Fig. 5. The key advantages are that: (i) in cosmological applications nested sampling typically requires ~ 100 times

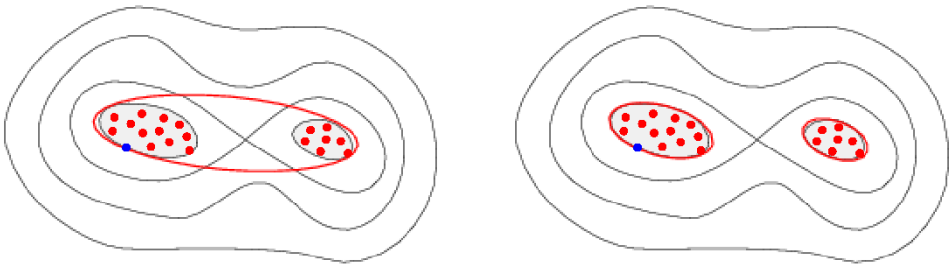


Figure 6: Practical nested sampling: single ellipsoidal method (left) and clustered ellipsoids method (right).

fewer samples than thermodynamic integration to calculate evidence to same accuracy; (ii) nested sampling does not get stuck at phase changes, unlike thermodynamic integration. In addition to efficient evaluation of the evidence, posterior samples are easily obtained as a by-product. One simply takes the full sequence of sampled points θ_i and weights the ith sample by $p_i = L_i w_i / E$. For example, if one were interested in deriving constraints on some quantity Q, then its mean and standard deviation are given by

$$\mu_Q = \sum_i p_i Q(\theta_i), \qquad \sigma_Q^2 = \sum_i (p_i Q(\theta_i) - \mu_Q)^2.$$

The main problem to address in the nested sampling algorithm is how to sample efficiently from the prior (which is often uniform) within some complicated, hard-edged likelihood constraint. This is not well-suited to standard MCMC techniques. The best published technique thus far [19] involves fitting an ellipse to the active points and selecting within it, but this is still problematic in a number of ways, particularly for multimodal posteriors. This difficulty is illustrated in Fig. 6 (left panel), from which it is clear that sampling from the single ellipse will have a very low acceptance rate. Moreover, this problem becomes rapidly worse as the number of dimensions increases.

In such a situation, one would instead wish to sample from the two separate ellipses illustrated in Fig. 6 (right panel), in which case the acceptance rate would remain very high. [20] propose a clustered nested sampling algorithm that can accommodate multimodal posteriors. In this approach k-means clustering (see [16]) is used to find exactly two clusters at each stage of the nest. The volumes of the enclosing ellipsoid(s) in the clustered and unclustered cases are calculated and the clusters accepted if the total enclosing volume is reduced by a specified fraction and the clustered ellipsoids do not overlap. The process is then repeated hierarchically. The advantage of this approach is that one need not know the number of clusters in advance and it provides an elegant method for parallelising the process. Nonetheless, the method would still be inefficient for elongated 'banana-shaped' degeneracies. Fig. 7 shows the samples obtained in applying the algorithm to a simple toy posterior consisting of three Gaussian peaks. The estimated log-evidence value is $\ln E_{\rm est} = -5.16 \pm 0.09$ as compared with the true value $\ln E_{\rm true} = -5.22$.

4. Cosmological applications of Bayesian model selection

Although the use of the evidence to perform Bayesian model selection is relatively new in

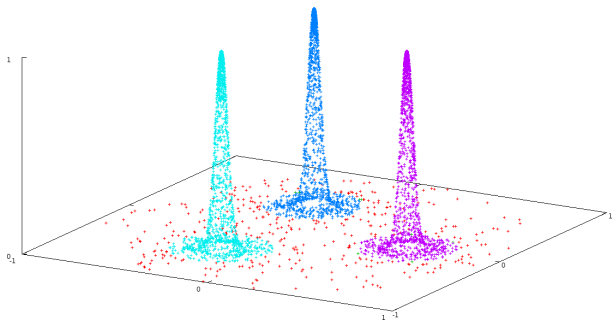


Figure 7: Illustration of cluster nested sampling applied to a posterior consisting of three Gaussians.

cosmology, there already exist a number of areas in which it has been applied. In this final section, we give a brief outline of some of these investigations, with the aim of illustrating the wide range of applications for the evidence.

4.1 Extending the cosmological parameter set

The most obvious use for the evidence in a cosmological context is in deciding whether the existing cosmological data imply the need for more free parameters than in the standard ΛCDM model. In this application, the hypotheses (models) are clearly nested. Care must be taken, however, since evidence values clearly depend on the choice of parameterisation and the associated priors. It must also be remembered that evidences for different models can only be compared when considering the same (combined) dataset.

Several investigations have been performed using different evidence evaluation methods. The earliest use of evidence in this context was by [10] in relation to fixing or varying the Hubble parameter. In terms of full, multi-parameter cosmological model fitting to combined datasets, however, [23] presented the first account, in which thermodynanic integration was used to evaluate evidences for a set of models of increasing complexity, namely $\Lambda\mathrm{CDM} + \Omega_k + f_v + (R, n_t)$. Subsequently, [1] used the Gaussian approximation and thermodynamic integration to investigate the model set $\Lambda\mathrm{CDM} + 3$ isocurvature mode models; [27] used the Gaussian approximation and the Savage–Dickey density ratio to evaluate evidences for the model sets $\Lambda\mathrm{CDM\text{-}HZ} + n_s + \Omega_k$ and $\Lambda\mathrm{CDM}$ + simple isocurvature mode; and [19] used nested sampling to select from the model set $\Lambda\mathrm{CDM\text{-}HZ} + n_s + w$. All find the maximum evidence for the standard ΛCDM model with a variable power-law index n_s.

4.2 The form of the primordial power spectrum

More extensive investigations into the preferred model for the primordial power spectrum have been performed by [2] using Bayesian evidence. Fig. 8 shows the models of the primordial power spectrum considered. In particular, the 'Lasenby & Doran' spectrum results from a cosmological model with a novel boundary condition that restricts the total conformal time available to the universe [15]. Also investigated was a 'broken spectrum' model that consisted of two scale-invariant sections joined by a sloping line segment. Some example evidence results are given in Table 1, from which we see that the Harrison-Zel'dovich ($n_s = 1$) model is strongly disfavoured relative to

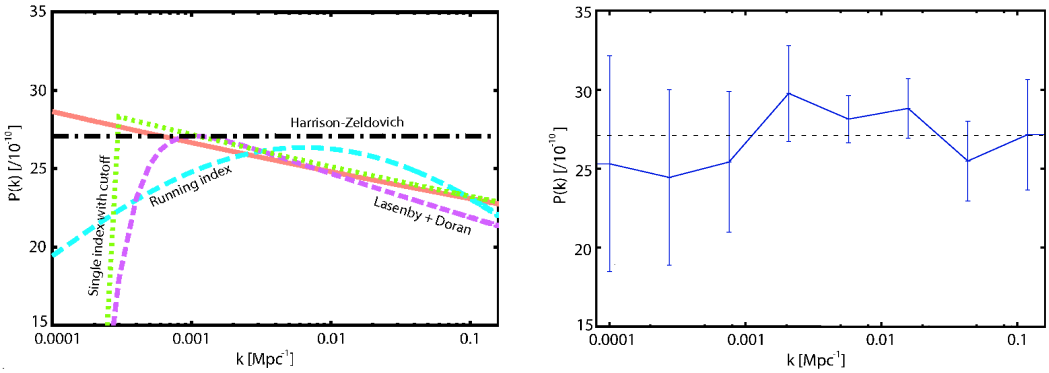


Figure 8: Models of the primordial power spectrum: parameterised models (left) and free-form fit in bins (right).

Model	$\ln E_i - \ln E_0$
Constant n	0.0 ± 0.5
H-Z	-4.4 ± 0.5
Running	-0.8 ± 0.6
Cutoff	0.4 ± 0.5
Broken	-2.7 ± 0.6
Binned	-6.1 ± 0.6
Lasenby & Doran	4.1 ± 0.5

Table 1: Differences of log evidences (for primordial parameters) for all models with respect to single index model within a current (near) concordance cosmology: $\Omega_0 = 1.024, \Omega_b h^2 = 0.0229, h = 0.61, \Omega_{cdm} h^2 = 0.118$, as compared to the Lasenby & Doran model (treated as a template).

a power law with variable n_s. The binned model is also disfavoured indicating it is unnecessaily complicated to explain the data. Interestingly, the Lasenby & Doran model is clearly the most favoured.

4.3 A rotating universe and Bianchi models

Several authors have commented on a significant North/South asymmetry in the WMAP data, plus strange alignments between low multipoles. [11] fitted a Bianchi VII$_h$ template to the WMAP sky and found a best fit with $\Omega_0 = 0.5$. The coldest part of the template corresponds with a non-Gaussian spot found in [28] and investigated further in [4]. However, $\Omega_0 = 0.5$ is in conflict with most other astrophysical indicators. Can one achieve the same in models including Λ? This has recently been investigated by [12] and a full Bayesian exploration of the parameter space of this model has been performed by [3].

Generally, a non-zero Λ has the effect of the shortening conformal time available, and so one needs very small h values in order to get similar smaller scale effects. One discovers it is impossible to find a good model in which the Bianchi template cosmology values match those of a background cosmology that fits existing data (e.g. the acoustic peaks). Nevertheless, it is still interesting to

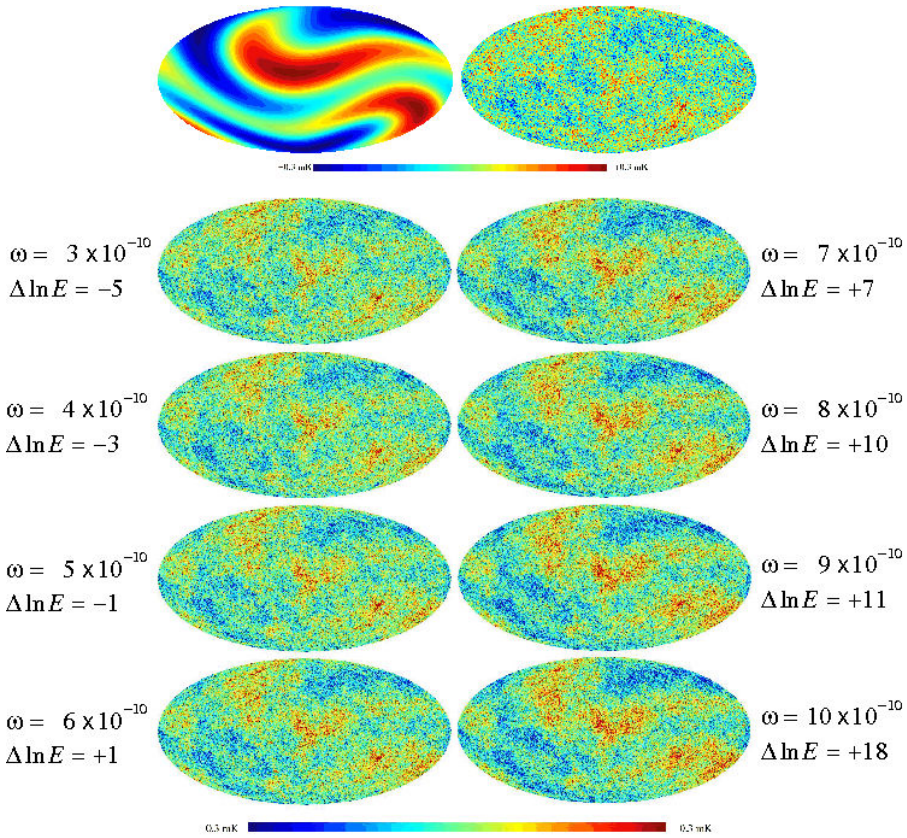


Figure 9: Simulated skies containing standard CMB anisotropies (top left) superposed on a Bianchi template (top left) with different amplitudes (bottom 8 panels). The Bianchi amplitude parameter ω, which measures the vorticity, and the log-evidence difference obtained are shown in each case.

evaluate the evidence for the Bianchi VII$_h$ model, treating it merely as a template. How much do we really need it in our data?

As an illustration, we can simulate maps containing Bianchi templates with different vorticities and see how well the evidence value can discriminate between models with and without a Bianchi component. From Fig. 9, we find that we start to be able to discriminate, at about the level of the original Bianchi template. Indeed, considering this for the real data (no Λ now), then in [3] evidence is found in favour of the introduction of a Bianchi template. Since the original version of this paper, however, new calculations have shown that the evidence difference is only weak (less then 1 unit in $\ln E$ for both the WMAP1 and WMAP3 data sets), so the jury is still out on whether the introduction of a template of this kind is really needed.

4.4 Combining cosmological datasets

One often estimates cosmological parameters by a joint analysis of a number of datasets. The standard technique for independent datasets is simply to multiply likelihoods. Freedom exists, however, in the relative 'weight' given to each dataset; this weighting is usually ad-hoc – datasets are excluded (weight zero), or included (weight unity). Instead one can include weights as hy-

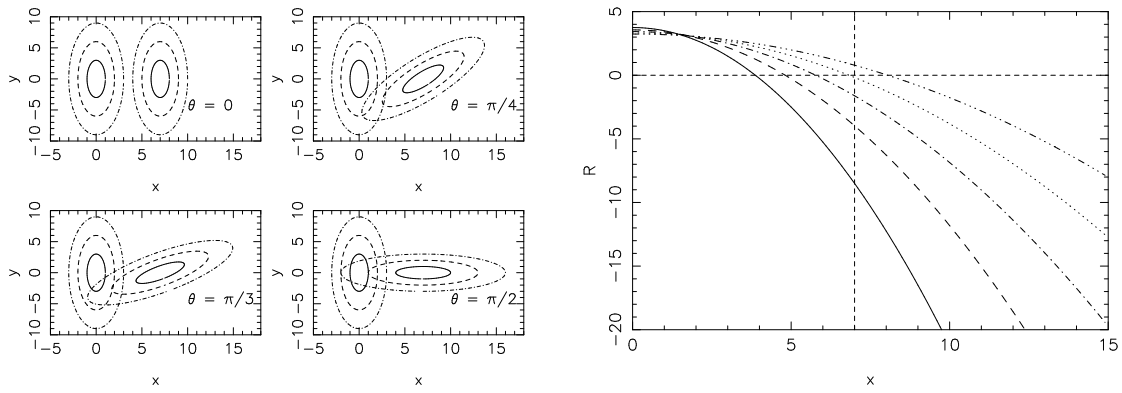


Figure 10: Left: posterior ellipses for joint data sets with differing geometric degeneracies with a centre separation of 7 units. Right: the corresponding ratio R for each case (and the case where the axis orientation $\theta = \pi/6$) as a function of the separate of the centres of the distributions (from [18]).

Basic parameter	Prior
$\omega_{\rm b}$	$(0.005, 0.05)$
$\omega_{\rm dm}$	$(0.01, 0.4)$
$\Omega_{\rm k}$	$(-0.3, 0.3)$
h	$(0.4, 0.9)$
n_s	$(0.8, 1.2)$
τ	$(0.01, 0.7)$
$\log 10^{10} A_s$	$(1, 5)$

Table 2: Prior ranges assumed in the test for mutual consistency of different cosmological datasets.

perparameters, then marginalise them out [14]. One can use the evidence to select between the models $H_0 =$ 'all weights unity' and $H_1 =$ 'each dataset has free weight (≥ 0)' to determine if the data support introduction of hyperparameters [7].

Another approach to testing the mutual consistency of different datasets [18] is to use the evidence to select between models $H_0 =$ 'all datasets generated by same cosmological parameters' and $H_1 =$ 'each dataset generated from different set of cosmological parameters', then calculate the ratio

$$R = \frac{E(\boldsymbol{D}|H_0)}{E(\boldsymbol{D}|H_1)} \frac{\Pr(H_0)}{\Pr(H_1)} = \frac{E(\boldsymbol{D}|H_0)}{\prod_k E(\boldsymbol{D}_k|H_1)} \frac{\Pr(H_0)}{\Pr(H_1)}.$$

A toy example is illustrated in Fig. 10. The method has been applied to real data (also by [18]), for the joint datasets CMB (WMAP1+VSA+CBI+ACBAR) + SDSS + SN1A, in the context of a ΛCDM model, using the priors listed in Table 2. For different dataset combinations one finds: $\ln R = 0.23$ for CMB + SDSS; $\ln R = 1.5$ for SDSS + SN1A; $\ln R = 1.6$ for SN1A + CMB; and $\ln R = 4.5$ for CMB + SDSS + SN1A. Thus, in general, the null hypothesis H_0 is favoured.

4.5 Component separation

Observations of the CMB are contaminated by foregrounds, but multifrequency observations

allow a component separation to be performed. Several blind and non-blind methods have been proposed, but the current approach for Planck is to use spectral matching independent component analysis (SMICA) [5] to determine the component power spectra and mixing matrix, followed by the maximum-entropy method (MEM) or Wiener filter ([8],[24]) to obtain the component maps and refined power spectra. The evidence can be used to good effect in both stages.

In the SMICA algorithm the evidence provides a means of determining the number of components present in the data. The SMICA approach models the n_f frequency observations as a noisy mixture of n_c Gaussian random fields. The analysis is performed in harmonic space, using the model $\boldsymbol{d}_{\ell m} = \boldsymbol{B}_\ell \boldsymbol{A} \boldsymbol{s}_{\ell m} + \boldsymbol{n}_{\ell m}$, where $\boldsymbol{B}_\ell$ and $\boldsymbol{A}$ denote the beam and mixing matrices respectively. Thus the model data covariance matrices read $\boldsymbol{D}_\ell = \boldsymbol{B}_\ell \boldsymbol{A} \boldsymbol{S}_\ell \boldsymbol{A}^{\rm t} \boldsymbol{B}_\ell^{\rm t} + \boldsymbol{N}_\ell$, with $\boldsymbol{S}_\ell$ and $\boldsymbol{N}_\ell$ (block) diagonal. We may construct the corresponding data covariances $\widetilde{\boldsymbol{D}}_\ell$ and form the log-likelihood

$$\ln L = -\tfrac{1}{2} \sum_\ell \left[\mathrm{Tr}(\widetilde{\boldsymbol{D}}_\ell \boldsymbol{D}_\ell^{-1}) + \ln |\boldsymbol{D}_\ell| \right].$$

One maximises $\ln L$ using a combination of expectation maximisation and conjugate gradient algorithms to obtain estimates $\hat{\boldsymbol{S}}_\ell$ and $\hat{\boldsymbol{A}}$. Calculating the Hessian matrix at the peak allows one to calculate the Gaussian approximation to the evidence. One may then plot the evidence versus n_c to estimate the number of components [26].

In the MEM/Wiener filter algorithm the evidence can be used to determine the appropriate level of regularisation in performing the reconstruction. All regularised harmonic-space methods involve minimising some function of the form.

$$F(\boldsymbol{s}_{\ell m}) = \chi^2(\boldsymbol{s}_{\ell m}) - \alpha_\ell S(\boldsymbol{s}_{\ell m}).$$

At each ℓ one can determine α_ℓ by maximising the evidence (using a Gaussian approximation). This enables automatic, optimal, scale-dependent regularisation and stable iterative updating of power spectra. One can also use the Hessian $\boldsymbol{H}_{\ell m} = \nabla\nabla F(\boldsymbol{s}_{\ell m})$ at the peak to estimate the covariance matrix of the reconstruction errors. The general method is easily extended to accommodate anisotropic noise, cut-sky data and (weakly) spatially-varying spectral parameters [25]. The method has recently been applied to the Planck Working Group 2 component separation challenge with promising results (see Fig. 11); more sophisticated rounds are to come.

4.6 Object detection

An important issue in the analysis of CMB data is the detection and characterisation of discrete objects, such as SZ clusters and point sources. A number of Bayesian approaches to discrete object detection have been proposed by [9] and shown to outperform standard linear filtering techniques (see also [17]).

One approach is to detect objects simultaneously. One assumes an unknown number N of objects in the model of the data: $\boldsymbol{D} = \boldsymbol{n} + \sum_{k=1}^{N} \boldsymbol{s}(\boldsymbol{a}_k)$, where $\boldsymbol{a}_k$ are the parameters characterising the kth object. Assuming the background emission and noise to be Gaussian, the likelihood function is simply

$$L(\boldsymbol{D}|\theta) = \frac{\exp\left\{-\tfrac{1}{2}[\boldsymbol{D} - \boldsymbol{s}(\theta)]^{\rm t} \boldsymbol{N}^{-1} [\boldsymbol{D} - \boldsymbol{s}(\theta)]\right\}}{(2\pi)^{N_{\rm pix}/2} |\boldsymbol{N}|^{1/2}},$$

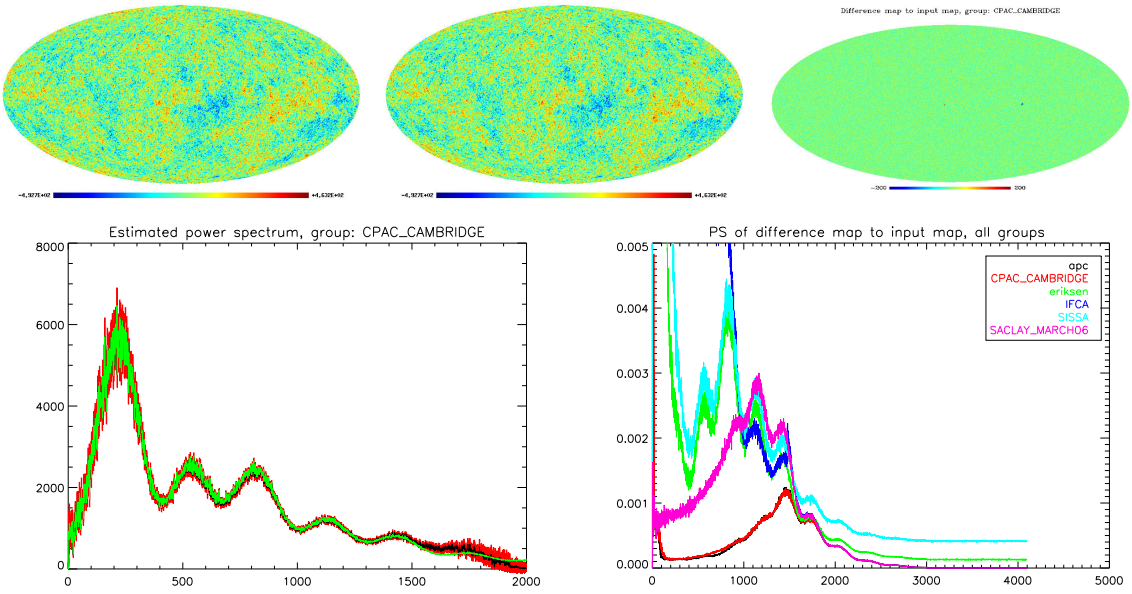


Figure 11: Application of MEM component separation to the WG2 challenge, consisting of 9 Planck channels, all with 10 arcmin beams and containing CMB, dust, free-free and synchrotron emission. Top row: CMB map input (left), output (middle), residuals (right). Bottom left: CMB power spectrum input (green), output (black) and errors bars (red). Bottom right: CMB power spectrum residuals for various methods.

where $\theta = (\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_N, N)$ is the total parameter vector. One can set priors on θ-space $\pi(\theta) = \pi(\boldsymbol{a}_1)\cdots\pi(\boldsymbol{a}_N)$ and N, e.g. $\Pr(N) = \mu^N e^{-\mu}/N!$. One then explores the posterior distribution using MCMC sampling to obtain optimal values of parameters, and associated errors, in a single step. The number of objects present is determined by maximising evidence with respect to N. A toy example is shown in Fig. 12. The main problem with this approach is that it is computationally very demanding, taking nearly 1 hour on a 1 GHz intel processor to obtain the results shown.

An alternative approach is to detect objects iteratively or sequentially. At each iteration or pixel, the model contains only a single object. One then maximises the posterior in the object parameters $\boldsymbol{a}$ using some optimiser or MCMC sampling. For each 'identified' object, one then selects between the models H_0 = 'there is no object centred in this pixel' and H_1 = 'there is an object centred in this pixel'. This is performed by calculating the ratio

$$R = \frac{E(\boldsymbol{D}|H_1)}{E(\boldsymbol{D}|H_0)}\frac{\Pr(H_1)}{\Pr(H_0)} = \frac{E(\boldsymbol{D}|H_1)}{E(\boldsymbol{D}|H_0)}\frac{\langle N_{\rm obj}\rangle}{N_{\rm pix}},$$

and one only accept objects with R above some threshold (usually zero). The iterative/sequential approach is very fast, detecting many 100s of objects in just less than a minute.

5. Conclusions

The Bayesian framework provides a unified approach to data analysis, providing two levels of inference: parameter estimation and confidence limits by maximising or exploring the posterior;

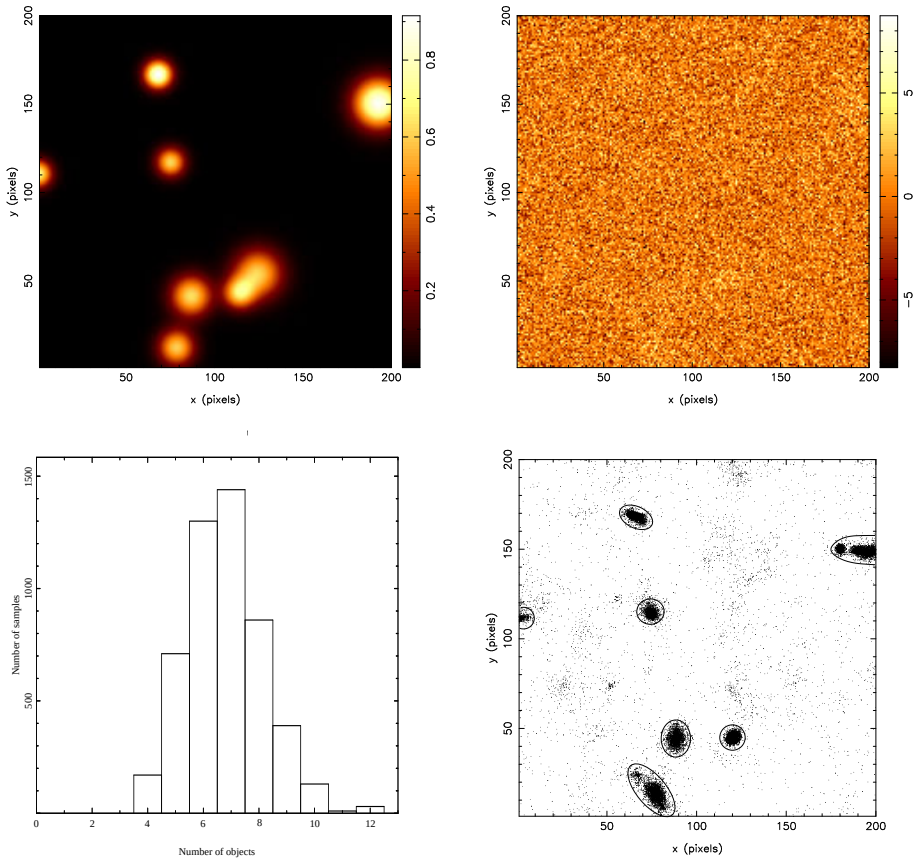


Figure 12: Simultaneous Bayesian object detection: true map (top left); data map (top right); evidence versus N (bottom left); posterior samples for $N = 7$ model.

and model selection by integrating the posterior to obtain the evidence. There exist several methods for evaluating the evidence: Gaussian approximation and the Savage–Dickey density ratio (which are approximate or restricted in their applicability); and thermodynamic integration and nested sampling (which are generally applicable).

Model selection using the evidence has many (cosmological) applications, including: the inclusion of additional free cosmological parameters; the form of primordial power spectrum; Bianchi models and other exotic models; combining cosmological data-sets and consistency checks; number of components and regularisation in component separation; and object detection. Further (cosmological) applications remain – you might like to try it for yourself!

Acknowledgments

The authors would like to thank the organisers very much for an excellent and stimulating meeting, and for the opportunity to contribute to these proceedings.

References

[1] Beltran M., Garcia-Bellido J., Lesgourgues J., Liddle A.R., Slosar A., Phys. Rev. D, 71, 063532

[2] Bridges M.L., Lasenby A.N., Hobson M.P., MNRAS, 369, 1123

[3] Bridges M.L., McEwen J.M., Lasenby A.N., Hobson M.P., MNRAS, submitted (astro-ph/0605325)

[4] Cruz M., Martinez-Gonzalez E., Vielva P., Cayon L., MNRAS, 356, 29

[5] Delabrouille J., Cardoso J.-F., Patanchon, G., 2003, MNRAS, 346, 1089

[6] Dickey J.M., 1971, Ann. Math. Stat., 42, 204

[7] Hobson M.P., Bridle S.L., Lahav O., 2002, MNRAS, 335, 377

[8] Hobson M.P., Jones A.W., Lasenby A.N., Bouchet F.R., MNRAS, 300, 1

[9] Hobson M.P., McLachlan C.I., 2003, MNRAS, 338, 765

[10] Jaffe A., 1996, ApJ, 471, 24

[11] Jaffe T.R., Banday A.J., Eriksen H.K., Go'rski K.M., Hansen F.K., 2005, ApJ, 629, L1

[12] Jaffe T.R., Hervik S., Banday A.J., Go'rski K.M., 2006, ApJ, 644, 701

[13] Kosowsky A., Milosavljevis M., Jimenez R., 2002, Phys. Rev. D, 66, 063007

[14] Lahav O., Bridle S.L., Hobson M.P., Lasenby A.N., Sodré L., Jr., 2000, MNRAS, 315, L45

[15] Lasenby A.N., Doran C., Phys. Rev. D, 71, 063502

[16] MacKay D.J.C., 2003, Information theory, inference and learning algorithms, Cambridge University Press

[17] Marshall P.J., Hobson M.P., Slosar A., 2003, MNRAS, 346, 489

[18] Marshall P.J., Rajguru N., Slosar A., 2006, Phys. Rev. D, 73, 067302

[19] Mukherjee P., Parkinson D., Liddle A.R., ApJ, 638, L51

[20] Shaw R, Hobson M.P., Bridges M.L, in preparation.

[21] Skilling J, 2004a, Bayesys and MassInf, unpublished, available from `http://www.inference.phy.cam.ac.uk/bayesys`

[22] Skilling J, 2004b, Nested sampling for general Bayesian computation, unpublished, available from `http://www.inference.phy.cam.ac.uk/bayesys`

[23] Slosar A. et al., 2003, MNRAS, 341, L29

[24] Stolyarov V., Hobson M.P., Ashdown M.A.J., Lasenby A.N., 2002, MNRAS, 336, 97

[25] Stolyarov V., Hobson M.P., Lasenby A.N., Barreiro R.B., 2005, MNRAS, 357, 145

[26] Taylor J., Ashdown M.A.J., Hobson M.P., in preparation

[27] Trotta R., 2005, astro-ph/0504022

[28] Vielva P., Martinez-Gonzalez E., Barreiro R.B., Sanz J.L., Cayon L., 2004, ApJ, 609, 22