Mon. Not. R. Astron. Soc. **378,** 72–82 (2007) doi:10.1111/j.1365-2966.2007.11738.x

Applications of Bayesian model selection to cosmological parameters

Roberto Trotta[1,2]★

[1]*Oxford University, Astrophysics, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH*
[2]*Département de Physique Théorique, Université de Genève, 24 quai Ernest Ansermet, 1211 Genève 4, Switzerland*

Accepted 2007 March 13. Received 2007 March 12; in original form 2007 February 18

ABSTRACT

Bayesian model selection is a tool for deciding whether the introduction of a new parameter is warranted by the data. I argue that the usual sampling statistic significance tests for a null hypothesis can be misleading, since they do not take into account the information gained through the data, when updating the prior distribution to the posterior. In contrast, Bayesian model selection offers a quantitative implementation of Occam's razor.

I introduce the Savage–Dickey density ratio, a computationally quick method to determine the Bayes factor of two nested models and hence perform model selection. As an illustration, I consider three key parameters for our understanding of the cosmological concordance model. By using *Wilkinson Microwave Anisotropy Probe* (*WMAP*) 3-year data complemented by other cosmological measurements, I show that a non-scale-invariant spectral index of perturbations is favoured for any sensible choice of prior. It is also found that a flat universe is favoured with odds of 29:1 over non-flat models, and that there is strong evidence against a cold dark matter isocurvature component to the initial conditions which is totally (anti)correlated with the adiabatic mode (odds of about 2000:1), but that this is strongly dependent on the prior adopted.

These results are contrasted with the analysis of *WMAP* 1-year data, which were not informative enough to allow a conclusion as to the status of the spectral index. In a companion paper, a new technique to forecast the Bayes factor of a future observation is presented.

Key words: methods: data analysis – methods: statistical – cosmic microwave background – cosmological parameters.

1 INTRODUCTION

In the epoch of precision cosmology, we often face the problem of deciding whether or not cosmological data support the introduction of a new quantity in our model. For instance, we might ask whether it is necessary to consider a running of the spectral index, an extra isocurvature mode, or a non-constant dark energy equation of state. The status of such additional parameters is uncertain, as often sampling (frequentist) statistics significance tests do not allow them to be ruled out with high confidence. There is a large body of work[1] that addresses the difficulties arising from the use of *p*-values (significance level) in assessing the need for a new parameter. Many weaknesses of significance tests are clarified, and some even overcome, by adopting a Bayesian approach to testing. In this work, we take the viewpoint of Bayesian model selection to determine whether a parameter is needed in the light of the data at hand.

The key quantity for Bayesian model comparison is the marginal likelihood, or evidence, whose calculation and interpretation is attracting increasing attention in cosmology and astrophysics (Drell, Loredo & Wasserman 2000; Saini, Weller & Bridle 2004; Lazarides, de Austri & Trotta 2004; Beltran et al. 2005; Kunz, Trotta & Parkinson 2006; Trotta 2007c), after it was introduced in the cosmological context by Jaffe (1996) and Slosar et al. (2003). The marginal likelihood has proved useful in other contexts, as well, for instance consistency checks between data sets (Hobson, Bridle & Lahav 2002; Marshall, Rajguru & Slosar 2006), the detection of galaxy clusters via the Sunayev–Zel'dovich effect (Hobson & McLachlan 2003) and neutrino emissions from type II supernovae (Loredo & Lamb 2002). In this paper we use the Savage–Dickey density ratio (SDDR) for an efficient computation of marginal likelihoods ratios (Bayes factor), while in a companion paper (Trotta 2007a) we present a new method to forecast the Bayes factor probability distribution of a future observation, called PPOD (for 'Predictive Posterior Odds Distribution').[2] We then illustrate applications to some important parameters of current cosmological model building.

This paper is organized as follows: we review the basics of Bayesian model comparison in Section 2 and we introduce the

★E-mail: rxt@astro.ox.ac.uk

[1] A good starting point is the collection of references available from the website of David R. Anderson, Department of Fishery and Wildlife Biology, Colorado State University.

[2] The method was called ExPO for 'Expected Posterior Odds' in a previous version of this work (Trotta, unpublished, astro-ph/0504022v1). I am grateful to Tom Loredo for suggesting the new, more appropriate name.

SDDR for the computation of the Bayes factor between two nested models. Section 3 is devoted to the application of model selection to three central parameters of the cosmological concordance model: the spectral tilt of scalar perturbations, the spatial curvature of the Universe and a totally (anti)correlated isocurvature cold dark matter (CDM) contribution to the initial conditions. We discuss our results and summarize our conclusions in Section 4.

Some complementary material is presented in the appendices. An explicit illustration of Lindley's paradox is given in Appendix A, the mathematical derivation of the SDDR is presented in Appendix B while a series of benchmark tests for the accuracy of the SDDR are carried out in Appendix C.

2 BAYESIAN MODEL COMPARISON

In this section, we first briefly review the basics of Bayesian inference and model comparison and introduce our notation. We then present the SDDR for a quick computation of the Bayes factor of two nested models.

2.1 Bayes factor

Bayesian inference (see e.g. Jaynes 2003; MacKay 2003) is based on Bayes' theorem, which is a consequence of the product rule of probability theory:

$$p(\theta \mid d, M) = \frac{p(d \mid \theta, M)\pi(\theta \mid M)}{p(d \mid M)}. \tag{1}$$

On the left-hand side, the posterior probability for the parameters θ given the data d under a model M is proportional to the likelihood $p(d \mid \theta, M)$, times the prior probability distribution function (PDF), $\pi(\theta \mid M)$, which encodes our state of knowledge before seeing the data. In the context of model comparison it is more useful to think of $\pi(\theta \mid M)$ as an integral part of the model specification, defining the prior available parameter space under the model M. The normalization constant in the denominator of (1) is the *marginal likelihood for the model M* (sometimes also called the 'evidence') given by

$$p(d \mid M) = \int_\Omega p(d \mid \theta, M)\pi(\theta \mid M)\,\mathrm{d}\theta, \tag{2}$$

where Ω designates the parameter space under model M. In general, θ denotes a multidimensional vector of parameters and d a collection of measurements (data covariance matrix, etc.), but to avoid cluttering the notation we will stick to the simple symbols introduced above.

Consider two competing models M_0 and M_1 and ask what is the posterior probability of each model given the data d. By Bayes' theorem we have

$$p(M_i \mid d) \propto p(d \mid M_i)\pi(M_i) \quad (i = 0, 1), \tag{3}$$

where $p(d \mid M_i)$ is the marginal likelihood for M_i and $\pi(M_i)$ is the prior probability of the ith model before we see the data. The ratio of the likelihoods for the two competing models is called the *Bayes factor*:

$$B_{01} \equiv \frac{p(d \mid M_0)}{p(d \mid M_1)}, \tag{4}$$

which is the same as the ratio of the posterior probabilities of the two models in the usual case when the prior is presumed to be non-committal about the alternatives and therefore $\pi(M_0) = \pi(M_1) = 1/2$. The Bayes factor can be interpreted as an automatic Occam's razor, which disfavours complex models involving many parameters (see e.g. MacKay 2003, for details). A Bayes factor $B_{01} > 1$ favours model M_0 and in terms of betting odds it would prefer M_0 over M_1 with odds of B_{01} against 1. The reverse is true for $B_{01} < 1$.

Table 1. Jeffreys' scale for the strength of evidence when comparing two models, M_0 versus M_1, with our convention for denoting the different levels of evidence. The probability column refers to the posterior probability of the favoured model, assuming non-committal priors on the two competing models, that is, $\pi(M_0) = \pi(M_1) = 1/2$ and that the two models exhaust the model space, $p(M_0|d) + p(M_1|\,d) = 1$.

$\lvert \ln B_{01} \rvert$	Odds	Probability	Notes
<1.0	$\lesssim 3:1$	< 0.750	Inconclusive
1.0	$\sim 3:1$	0.750	Positive evidence
2.5	$\sim 12:1$	0.923	Moderate evidence
5.0	$\sim 150:1$	0.993	Strong evidence

It is usual to consider the logarithm of the Bayes factor, for which the so-called 'Jeffreys' scale' gives empirically calibrated levels of significance for the strength of evidence (Jeffreys 1961; Kass & Raftery 1995), $|\ln B_{01}| > 1$; >2.5; >5.0. Different authors use different conventions to qualify the Jeffreys' levels of strength of evidence. In this work we will use the convention summarized in Table 1 – often in the literature one deems odds above $|\ln B_{01}| = 5$ to be 'decisive', but we prefer to avoid the use of the term because of the strong connotation of finality that it carries with it. If we assume that the two competing models are exhaustive, that is, that $p(M_0|\,d) + p(M_1|\,d) = 1$ and a non-committal prior $\pi(M_0) = \pi(M_1) = 1/2$, we can relate the strength of evidence to the posterior probability of the models,

$$p(M_0 \mid d) = \frac{B_{01}}{B_{01} + 1},$$
$$p(M_1 \mid d) = \frac{1}{B_{01} + 1}. \tag{5}$$

This probability is indicated in the third column of Table 1.

The subject of hypothesis testing has received an enormous amount of attention in the past, and the controversy on the subject is far from being resolved among statisticians. An illustration of the difference between Bayesian model selection and frequentist hypothesis testing is given in Appendix A, where Lindley's paradox is worked out with the help of a simple example. There it is shown that the Bayesian approach has the advantage of taking into account the information provided by the data, which is ignored by frequentist hypothesis testing.

Evaluating the marginal likelihood integral of equation (2) is in general a computationally demanding task for multidimensional parameter spaces. Thermodynamic integration is often the method of choice, whose computational burden can become fairly large, as it depends heavily on the dimensionality of the parameter space and on the characteristic of the likelihood function. In certain cosmological applications, thermodynamic integration can require up to 100 times more likelihood evaluation than parameter estimation (Beltran et al. 2005). An elegant algorithm called 'nested sampling' has been recently put forward by Skilling (2004), and implemented in the cosmological context by Bassett, Corasaniti & Kunz (2004) and Mukherjee, Parkinson & Liddle (2006). While nested sampling reduces the number of likelihood evaluations to the same order of magnitude as for parameter estimation, in the cosmological context this does not necessarily imply that the computing time can be reduced accordingly, see Mukherjee et al. (2006) for details.

2.2 The Savage–Dickey density ratio

Here we investigate the performance of the Savage–Dickey density ratio (SDDR), whose use is very promising in terms of reducing the computational effort needed to calculate the Bayes factor of two nested models, as we show below (for other possibilities, see e.g. DiCiccio et al. 1997).

Suppose we wish to compare a two-parameters model M_1 with a restricted submodel M_0 with only one free parameter, ψ, and with fixed $\omega = \omega_\star$ (for simplicity of notation we take a two-parameters case, but the calculations carry over trivially in the multidimensional case). Assume further that the prior is separable (which is usually the case in cosmology), that is, that

$$\pi(\omega, \psi \mid M_1) = \pi(\omega \mid M_1)\pi(\psi \mid M_0). \tag{6}$$

Then the Bayes factor B_{01} of equation (4) can be written as (see Appendix B)

$$B_{01} = \left.\frac{p(\omega \mid d, M_1)}{\pi(\omega \mid M_1)}\right|_{\omega=\omega_\star} \quad \text{(SDDR)}. \tag{7}$$

This expression goes back to Dickey (1971), who attributed it to L. J. Savage, and is therefore called Savage–Dickey density ratio (SDDR, see also Verdinelli & Wasserman 1995, and references therein). Thanks to the SDDR, the evaluation of the Bayes factor of two nested models only requires the properly normalized value of the marginal posterior at $\omega = \omega_\star$ under the extended model M_1, which is a by-product of parameter inference. We note that the derivation of (7) *does not involve any assumption about the posterior distribution*, and in particular about its normality.

For a Gaussian prior centred on $\omega_\star$ with s.d. $\Delta\omega$ and a Gaussian likelihood[3] with mean $\hat{\mu}$ and width $\hat{\sigma}$, equation (7) gives

$$\ln B_{01}(\beta, \lambda) = \frac{1}{2}\ln(1+\beta^{-2}) - \frac{\lambda^2}{2(1+\beta^2)}, \tag{8}$$

where we have introduced the number of sigma discrepancy $\lambda = |\hat{\mu} - \omega_\star|/\hat{\sigma}$ and the volume reduction factor $\beta = \hat{\sigma}/\Delta\omega$ (see Appendix A for details). For strongly informative data, $\beta^{-1} \gg 1$ and in terms of the information content $I = -\ln\beta \geqslant 0$, equation (8) is approximated by

$$\ln B_{01} \approx I - \lambda^2/2 \quad \text{(informative data)}. \tag{9}$$

The two terms on the right-hand side pull the Bayes factor in opposite directions: a large information content I signals a large volume of wasted parameter space under the prior, and acts as an Occam's razor term favouring the simpler model, while a large λ favours the more complex model because of the mismatch between the measured and the predicted value of the extra parameter. Evidence against the simpler model scales as λ^2, while evidence in its favour only accumulates as $I = -\ln\beta$. Furthermore, for strong odds against the simpler model ($\lambda \gg 1$) the prior choice becomes irrelevant unless $I \gg \lambda$, a situation which gives rise to Lindley's paradox (see Appendix A). For the case where $\lambda \ll 1$, that is, the prediction of the simpler model is confirmed by the observation, the odds in favour of the simpler model are determined by the information content I, and therefore by the prior choice.

The use of the SDDR for nested models has several advantages. A first important point is the analytical insight equation (7) gives into the working of model selection for two nested models, which we have briefly sketched above. Priors on the common parameters on both models are unimportant, as they factor out when computing the Bayes factor. The only relevant scales in the problem are the quantities λ and β, see equation (9), with the latter controlled by the prior width on the extra parameter. The volume effect arising from a change in the prior (e.g. when enlarging the prior range) can be easily estimated from the SDDR expression, without recomputing the posterior. Usually, the posterior PDF in equation (7) will be obtained by Monte Carlo Markov Chain (MCMC) techniques. In this case, even a change in the variables, or a more restrictive prior can usually be applied by simply posterior reweighting the MCMC samples without recomputing them. Secondly, the SDDR can be applied to existing MCMC chains, and therefore the model selection question can be dealt with easily after the parameter estimation step has already been performed. Finally, Appendix C demonstrates that in the benchmark Gaussian likelihood scenario the SDDR gives accurate results out to $\lambda \lesssim 3$. For larger value of λ the performance of the method is hindered by the fact that it becomes very difficult with conventional MCMC methods to obtain samples far out into the tails of the posterior. One could argue however that the most interesting regime for model comparison is precisely where the SDDR can yield accurate answers. This is also the region where most of the model selection questions in cosmology currently lie. Finally, often a high numerical accuracy in the Bayes factor does not seem to be central for most model comparison questions, especially in view of the fact that the uncertainty in the result can be strongly dominated by the prior range one assumes. This suggests that a quick and computationally inexpensive method such as the SDDR might be helpful in assessing the model comparison outcome for a broad range of priors. We therefore advocate the use of SDDR method for model selection questions involving nested models with moderate discrepancies between the prediction of the simple model and the posterior result, $\lambda \lesssim 3$. We now turn to the demonstration of the method on current cosmological observations.

3 APPLICATION TO COSMOLOGICAL PARAMETERS

In this section we apply the Bayesian model selection toolbox presented above to three cosmological parameters which are central for our understanding of the cosmological concordance model: the spectral index of scalar (adiabatic) perturbations, the spatial curvature of the Universe and an isocurvature CDM component to the initial conditions for cosmological perturbations.

3.1 Parameter space and cosmological data

We use the *Wilkinson Microwave Anisotropy Probe* (*WMAP*) 3-year temperature and polarization data (Hinshaw et al. 2006; Page et al. 2006) supplemented by small-scale cosmic microwave background (CMB) measurements (Kuo et al. 2004; Readhead et al. 2004). We add the *Hubble Space Telescope* (*HST*) measurement of the Hubble constant $H_0 = 72 \pm 8\,\mathrm{km\,s^{-1}\,Mpc^{-1}}$ (Freedman et al. 2001) and the Sloan Digital Sky Survey (SDSS) data on the matter power spectrum on linear ($k < 0.1\,h^{-1}$ Mpc) scales (Tegmark et al. 2004). Furthermore, we shall also consider the supernovae luminosity distance measurements (Riess et al. 2004). We denote all of the data sets but *WMAP* as 'external' for simplicity of notation. We are also interested in assessing the changes in the model comparison outcome in going from *WMAP* 1-year to *WMAP* 3-year data. We shall therefore compare our results using the 3-year *WMAP* data with the first-year *WMAP* data release (Bennett et al. 2003; Hinshaw et al.

[3] Notice that $\hat{\mu}$ and $\hat{\sigma}$ are referred to the likelihood, *not* the posterior PDF.

Table 2. Summary of model comparison results from *WMAP* data combined with small-scale CMB measurements, SDDS, *HST* and SNIa data. *WMAP*3+ext refers to *WMAP* 3-year data release, *WMAP*1 + ext to *WMAP* first-year data. The most spectacular improvement from *WMAP*1 to *WMAP*3 is the moderate evidence against a scale-invariant spectral index. Errors in the Bayes factor are obtained by computing the variance of the SDDR estimate from five subchains (see Appendix C for details). The 'estimate' column gives the value obtained by employing the Gaussian approximation to the likelihood, equation (A9) for a Gaussian prior or equation (A10) for a flat prior.

Data	$\ln B_{01}$ from SDDR (numerical)	(estimate)	Odds in favour of simpler model	Probability of simpler model	Comment
			Spectral index: $n_S = 1$ versus $0.8 \leqslant n_S \leqslant 1.2$ (Gaussian)		
*WMAP*3+ext	-2.86 ± 0.28	-3.00	1 to 17	0.05	Moderate evidence for non-scale invariance
*WMAP*1+ext	0.68 ± 0.04	0.71	2 to 1	0.66	Inconclusive result
			Spatial curvature: $\Omega_\kappa = 0$ versus $-1.0 \leqslant \Omega_\kappa \leqslant 1$ (flat)		
*WMAP*3+ext	3.37 ± 0.05	3.25	29 to 1	0.97	Moderate evidence for a flat universe
*WMAP*1+ext	2.70 ± 0.09	2.68	15 to 1	0.94	Moderate evidence for a flat universe
			Adiabaticity: $f_{\rm iso} = 0$ versus $-100 \leqslant f_{\rm iso} \leqslant 100$ (flat)		
*WMAP*3+ext	7.62 ± 0.02	7.63	2050 to 1	0.9995	Strong evidence for adiabatic conditions
*WMAP*1+ext	7.50 ± 0.03	7.53	1800 to 1	0.9994	Strong evidence for adiabatic conditions

2003; Verde et al. 2003), complemented by the 'external' data sets described above.[4]

We make use of the publicly available codes CAMB and COSMOMC (Lewis & Bridle 2002) to compute the CMB and matter power spectra and to construct MCMCs in parameter space. The Monte Carlo (MC) is performed using 'normal parameters' (Kosowsky, Milosavljevic & Jimenez 2002), in order to minimize non-Gaussianity in the posterior PDF. In particular, we sample uniformly over the physical baryon and CDM densities, $\omega_{\rm b} \equiv \Omega_{\rm b} h^2$ and $\omega_{\rm c} \equiv \Omega_{\rm c} h^2$, expressed in units of 1.88×10^{-29} g cm^{-3}; the ratio of the angular diameter distance to the sound horizon at decoupling, $\Theta_\star$, the optical depth to reionization τ_r (assuming sudden reionization) and the logarithm of the adiabatic amplitude for the primordial fluctuations, $\ln 10^{10} A_S$. When combining the matter power spectrum with CMB data, we marginalize analytically over a bias b considered as an additional nuisance parameter. Throughout we assume three massless neutrino families and no massive neutrinos (for constraints on these quantities, see instead e.g. Bowen et al. 2002; Lesgourgues & Pastor 2006; Spergel et al. 2006), we fix the primordial Helium mass fraction to the value predicted by big bang nucleosynthesis (see e.g. Trotta & Hansen 2004) and we neglect the contribution of gravitational waves to the CMB power spectrum.

3.2 Model selection from current data

The scalar spectral index

As a first application we consider the scalar spectral index for adiabatic perturbations, n_S. We compare the evidence in favour of a scale-invariant index ($M_0 : n_S = 1$), also called an Harrison–Zel'dovich (HZ) spectrum, with a more general model of single-field inflation, in which we do not require the spectral index to be scale invariant, $M_1 : n_S \neq 1$. The latter case is called for brevity 'generic inflation'.

Within the framework of slow-roll inflation, the prior allowed range for the spectral index can be estimated by considering that $n_S = 1 - 6\epsilon + 2\eta$, where η and ϵ are the slow-roll parameters. If we assume that ϵ is negligible, then $n_S = 1 + 2\eta$. If the slow-roll conditions are to be fulfilled, $\eta \ll 1$, then we must have $|\eta| \lesssim 0.1$, which gives $0.8 \lesssim n_S \lesssim 1.2$. Hence we take a Gaussian prior on n_S with mean $\mu = 1.0$ and width $\sigma = 0.2$.

The result of the model comparison is shown in Table 2. When employing *WMAP* 1-year data, the model comparison yields an inconclusive result ($\ln B_{01} = 0.68 \pm 0.04$), but the new, lower value for n_S from the *WMAP* 3-year data, enhanced by the small-scale CMB measurements and SDDS matter power spectrum data, does yield moderate evidence for a non-scale-invariant spectral index ($\ln B_{01} = -2.86 \pm 0.28$), with odds of about 17:1, or a posterior probability of a scale-invariant index of 5 per cent, when compared to the above alternative generic inflation model. This is a consequence of both the shift of the peak of the posterior to $n_S = 0.95$ and a reduction of its spread when using *WMAP* 3-year data, which places the scale-invariant value of $n_S = 1$ at about 3.3σ away from the posterior's peak (see however the discussion about possible systematic effects in Parkinson, Mukherjee & Liddle 2006). In Table 2 we also give the resulting value of the Bayes factor obtained by using the SDDR formula and a Gaussian approximation to the posterior, see equation (A9). Since the marginalized posterior for n_S is very well approximated by a Gaussian, we find a very good agreement between this crude estimate and the numerical result using the actual shape of the posterior, with a discrepancy of the order of 5 per cent. This supports the idea that for reasonably Gaussian PDFs using a Gaussian approximation to the SDDR might be a good way of obtaining a first estimate of the Bayes factor for nested models.

Our findings are in broad agreement with Parkinson et al. (2006), where it was found using nested sampling that a similar data compilation as the one employed here gives $\ln B_{01} = -1.99 \pm 0.26$ for the comparison between the HZ model and a generic inflationary model with a flat prior between $0.8 \leqslant n_S \leqslant 1.2$. For such a flat prior, we obtain, using the SDDR, $\ln B_{01} = -2.98 \pm 0.28$, where the difference with Parkinson et al. (2006) has to be ascribed to different constraining power of the different data compilations used, rather than to the methods for computing the Bayes factor. For a more detailed discussion of a series of possible systematic effects which might change the outcome of the model comparison, see section IIIC in Parkinson et al. (2006).

The spatial curvature

We now turn to the issue of the geometry of spatial sections. We evaluate the Bayes factor for $\Omega_\kappa = 0$ (flat universe) against a model

[4] A more detailed discussion on the *WMAP* first-year data model comparison result and the power of the external data sets can be found in the original version of the present work, Trotta (unpublished, astro-ph/0504022v1).

with $\Omega_\kappa \neq 0$. As discussed above, we only need to specify the prior distribution for the parameter of interest, namely Ω_κ. We choose a flat prior of width $\Delta\Omega_\kappa = 1.0$ on each side of $\Omega_\kappa = 0$, for we know that the universe is not empty (thus $\Omega_\kappa < 1.0$, setting aside the case of $\Lambda < 0$) nor largely overclosed (therefore $\Omega_\kappa \gtrsim -1$ is a reasonable range, see Section 3.3 for further comments).

Cosmic microwave background data alone cannot strongly constrain Ω_κ because of the fundamental geometrical degeneracy. Even CMB and SDSS data together allow for a wide range of values for the curvature parameter, which translates into approximately equal odds for the curved and flat models. Adding SNIa observations drastically reduces the range of the posterior, since their degeneracy direction is almost orthogonal to the geometrical degeneracy of the CMB. Further inclusion of the *HST* measurement for the Hubble parameter narrows down the posterior range considerably, since the handle on the value of the Hubble constant today breaks the geometrical degeneracy. When all of the data (*WMAP*3 + ext) are taken into account, we obtain for the Bayes factor $\ln B_{01} = 3.37 \pm 0.05$, favouring a flat universe model with moderate odds of about 29 : 1 (see Table 2). This corresponds to a posterior probability for a flat universe of 97 per cent, for our particular choice of prior. We notice the slight improvement in these odds from the result obtained using *WMAP*1 + ext data, where the odds were 15 : 1, which is to be ascribed mainly to the inclusion of polarization data that helps further tightening constraints around the geometrical degeneracy.

The CDM isocurvature mode

The third case we consider is the possibility of a CDM isocurvature contribution to the primordial perturbations. For a review of the possible isocurvature modes and their observational signatures, see, for example, Trotta (2004). Determining the type of initial conditions is a central question for our understanding of the generation of perturbations, and has far reaching consequences for the model building of the physical mechanisms which produced them. Constraints on the isocurvature fraction have been derived in several works, which considered different phenomenological mixtures of adiabatic and isocurvature initial conditions (Pierpaoli, Garcia-Bellido & Borgani 1999; Amendola et al. 2002; Trotta, Riazuelo & Durrer 2001, 2003; Crotty et al. 2003; Valiviita & Muhonen 2003; Beltran et al. 2004; Bucher et al. 2004; Kurki-Suonio, Muhonen & Valiviita 2005; Moodley et al. 2004; Trotta & Durrer 2006). Two recent studies making use of the latest CMB data (Bean, Dunkley & Pierpaoli 2006; Keskitalo et al. 2006) obtain different conclusions as to the level of isocurvature contribution. While both groups report a lower best-fitting chi square for a model with a large ($n \sim 3$) spectral index for the CDM isocurvature component, they give a different interpretation of the statistical significance of the improvement. It is precisely in such a context that a model selection approach as the one presented here might be helpful, in that it allows to account for the Occam's razor effect described above. The question of isocurvature modes has been addressed from a model comparison perspective by Beltran et al. (2005) and Trotta (2007b).

Since the goal of this work is not to present a detailed analysis of isocurvature contributions, but rather to give a few illustrative applications of Bayesian model selection, we restrict our attention to the comparison of a purely adiabatic model against a model containing a CDM isocurvature mode totally correlated or anticorrelated. For simplicity, we also take the isocurvature and adiabatic mode to share the same spectral index, n_S. This phenomenological set-up is close to what one expects in some realizations of the curvaton scenario, see, for example, Gordon & Lewis (2003), Lyth & Wands (2003) and Lazarides et al. (2004). For an extended treatment including all of the four different isocurvature modes, see Trotta (2007b).

We compare model M_0, with adiabatic fluctuations only, with M_1, which has a totally (anti)correlated isocurvature fraction

$$f_{\rm iso} \equiv \frac{\mathcal{S}}{\zeta}, \tag{10}$$

where ζ is the primordial curvature perturbation and $\mathcal{S}$ the entropy perturbation in the CDM component (see Lazarides et al. 2004; Trotta 2004, for precise definitions). The sign of the parameter $f_{\rm iso}$ defines the type of correlation. We adopt the convention that a positive (negative) correlation, $f_{\rm iso} > 0$ ($f_{\rm iso} < 0$), corresponds to a negative (positive) value of the adiabatic–isocurvature CMB correlator power spectrum on large scales. We choose $f_{\rm iso}$ as the relevant parameter for model comparison because of its immediate physical interpretation as an entropy-to-curvature ratio, but this is only one among several possibilities.

In the absence of a specific model for the generation of the isocurvature component, there is no cogent physical motivation for setting the prior on $f_{\rm iso}$. A generic argument is given by the requirement that linear perturbation theory be valid, that is, $\zeta, \mathcal{S} \ll 1$. This however does not translate into a prior on $f_{\rm iso}$, unless we specify a lower bound for the curvature perturbation. In general, $f_{\rm iso}$ is essentially a free parameter, unless the theory has some built-in mechanism to set a scale for the entropy amplitude. This however requires digging into the details of specific realizations for the generation of the isocurvature component. For instance, the curvaton scenario predicts a large $f_{\rm iso}$ if the CDM is produced by curvaton decay and the curvaton does not dominate the energy density, in which case $|f_{\rm iso}| \sim r^{-1} \gg 1$, since the curvaton energy density at decay compared with the total energy density is small, $r \equiv \rho_{\rm curv}/\rho_{\rm tot} \ll 1$ (Gordon & Lewis 2003; Lyth & Wands 2003). Once the details of the curvaton decay are formulated, it might be possible to argue for a theoretical lower bound on r, which gives the prior range for the predicted values of $f_{\rm iso}$.

In the absence of a compelling theoretical motivation for setting the prior, we can still appeal to another piece of information which is available to us before we actually see any data: the expected sensitivity of the instrument. By assessing the possible outcomes of a measurement given its forecasted noise levels we can limit the a priori accessible parameter space *for a specific observation* on the grounds that it is pointless to admit values which the experiment will not be able to measure. For the case of $f_{\rm iso}$, there is a lower limit to the a priori accessible range dictated by the fact that a small isocurvature contribution is masked by the dominant adiabatic part. Conversely, the upper range for $f_{\rm iso}$ is reached when the adiabatic part is hidden in the prevailing isocurvature mode. In order to quantify those two bounds, we carry out a Fisher Matrix (FM) forecast assuming noise levels appropriate for the measurement under consideration, thus determining which regions of parameter space is accessible to the observation. Such a prior is therefore motivated by the expected sensitivity of the instrument, rather then by theory. The prior range for a scale-free parameter thereby becomes a computable quantity which depends on our prior knowledge of the experimental apparatus and its noise levels.

We have performed an FM forecast in the $(\zeta, |\mathcal{S}|)$ plane, whose results are plotted in Fig. 1 for the *WMAP* expected sensitivity. We use a grid equally spaced in the logarithm of the adiabatic and isocurvature amplitudes, in the range $10^{-6} \leqslant \zeta \leqslant 5 \times 10^{-4}$ and $10^{-8} \leqslant |\mathcal{S}| \leqslant 10^{-2}$. For each pair $(\zeta, |\mathcal{S}|)$ the FM yields the

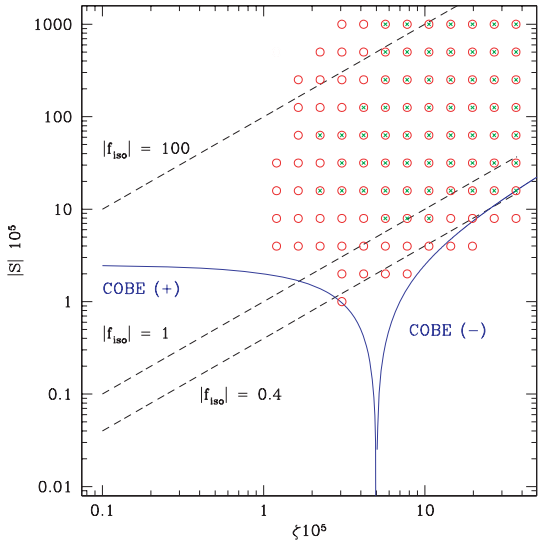

Figure 1. The parameter space accessible a priori to *WMAP* in the $(\zeta, |\mathcal{S}|)$ plane is obtained by requiring better than 10 per cent accuracy on $|f_{\rm iso}|$ in the Fisher Matrix error forecast (open circles for the best case, crosses for the worst case, depending on the fiducial values of $\tau_r, n_{\rm S}$ and on the sign of the correlation). This translates into a prior accessible range $0.4 \lesssim |f_{\rm iso}| \lesssim 100$ (diagonal, dashed lines), but only if $\zeta, |\mathcal{S}| \gtrsim 10^{-5}$. Models which roughly satisfy the *COBE* measurement of the large-scale CMB anisotropies ($\delta T/T \approx 10^{-5}$) lie on the blue/solid line and have positive (negative) correlation left-hand side (right-hand side) of the cusp.

expected error on the amplitudes as well as on $f_{\rm iso}$. The expected error however also depends on the fiducial values assumed for the remaining cosmological parameters. In order to take this into account, at each point in the $(\zeta, |\mathcal{S}|)$ grid we run 40 FM forecasts changing the type of correlation [sign$(\mathcal{S}) = \pm 1$], the spectral index ($n_{\rm S} = 0.8$–1.2 with a step of 0.1) and the optical depth to reionization ($\tau_r = 0.05$–0.35 with a step of 0.1). The other parameters $(\theta, \omega_{\rm c}, \omega_{\rm b})$ are fixed to the concordance model values, since ζ, $\mathcal{S}$ are mostly correlated with $\tau_r, n_{\rm S}$ and thus only the fiducial values assumed for the latter two parameters have a strong impact on the predicted errors of the amplitudes. We then select the best and worst outcome for the expected error on $f_{\rm iso}$, in order to bracket the expected result of the measurement independently on the fiducial value for $\tau_r, n_{\rm S}$. Notice that at no point we make use of real data. By requiring that the expected error on $f_{\rm iso}$ be of the order of 10 per cent or better, we obtain the a priori accessible area in amplitude space for *WMAP*, which is shown in Fig. 1.

It is apparent that $f_{\rm iso}$ cannot be measured by *WMAP* if either ζ or $|S|$ are below about 10^{-5}, in which case the signal is lost in the detector noise. For amplitudes larger than 10^{-5}, $|f_{\rm iso}| = 1$ is accessible to *WMAP* with high signal-to-noise ratio independently on the value of $\tau_r, n_{\rm S}$, while $|f_{\rm iso}| \approx 0.4$ can be measured only in a few cases for the most optimistic choice of parameters. As an aside, we notice that if we restrict our attention to models which roughly comply with the *COBE* measurement of the large-scale CMB power (blue/solid lines in Fig. 1), then *WMAP* can only explore the subspace of anticorrelated isocurvature contribution (right-hand side of the cusp) and only if $\zeta \gtrsim 7 \times 10^{-5}$. On the other end of the range, we can see that $|f_{\rm iso}| = 100$ is about the largest value accessible to *WMAP*, at least for $\zeta \geqslant 5 \times 10^{-4}, |\mathcal{S}| \geqslant 10^{-2}$. There is a simple physical reason for the asymmetry of the accessible range around $|f_{\rm iso}| = 1$: a small isocurvature contribution can be overshadowed by the adiabatic mode on large scales due to cosmic variance, but a subdominant adiabatic mode is still detectable even in the presence of a much larger isocurvature part, because the first adiabatic peak at $\ell \approx 200$ sticks out from the rapidly decreasing isocurvature power at that scale (at least if the spectral tilt is not very large, as in our case). In conclusion, the values of $|f_{\rm iso}|$ which *WMAP* can potentially measure with high signal-to-noise ratio are approximately bracketed by the range $0.4 \leqslant |f_{\rm iso}| \leqslant 100$, assuming that $\zeta \gtrsim 10^{-5}$. Given the fact that most of the prior volume lies above $|f_{\rm iso}| = 1$, we can take a flat prior on $f_{\rm iso}$ centred around $f_{\rm iso} = 0$, with a range $-100 \leqslant f_{\rm iso} \leqslant 100$, or $\Delta f_{\rm iso} = 100$. As we shall see below, it is this large range of a priori possible values compared with the small posterior volume which heavily penalizes an isocurvature contribution due to the Occam's razor behaviour of the Bayes factor.

The marginalized posterior on $f_{\rm iso}$ from *WMAP*3 + ext data gives a 95 per cent interval $-0.06 \leqslant f_{\rm iso} \leqslant 0.10$, thus yielding only upper bounds on the CDM isocurvature fraction, in agreement with previous works using a similar parameterization (see Trotta 2007b for details). The spread of the posterior is of the order of 0.1, which lies an order of magnitude below the level ($|f_{\rm iso}| = 1$) at which an isocurvature signal would have stand out clearly from the *WMAP* noise. The Bayes factor corresponding to the above choice of prior ($-100 \leqslant f_{\rm iso} \leqslant 100$) is given in Table 2, and with $\ln B_{01} = 7.62$ it corresponds to a probability of 0.9995 (or odds of 2050 to 1) for purely adiabatic initial conditions. This is a consequence of the large volume of wasted parameter space under the large prior used here, and a fine example of automatic Occam's razor built into the Bayes factor. We notice that in order to obtain a model-neutral conclusion (odds of 1 : 1) one would have to choose a prior width below 0.1, that is, find a mechanism to strongly limit the available parameter space for the isocurvature amplitude (Trotta 2007b). In other words, the introduction of a new scale-free isocurvature amplitude is generically unwarranted by data, a feature already remarked by Lazarides et al. (2004).

This result differs from the findings of Beltran et al. (2005), who considered an isocurvature CDM admixture to the adiabatic mode with arbitrary correlation and spectral tilt and concluded that there is no strong evidence against mixed models (odds of about 3 : 1 in favour of the purely adiabatic model). While their set-up is not identical to the one presented here and thus a direct comparison is difficult, we believe that the key reason of the discrepancy can be traced back to the different basis for the initial conditions parameter space. Instead of the isocurvature fraction $f_{\rm iso}$, Beltran et al. (2005) employ the parameter α describing the fractional isocurvature power, which is related to $f_{\rm iso}$ by

$$\alpha = \frac{f_{\rm iso}^2}{1 + f_{\rm iso}^2}. \tag{11}$$

The infinite range $0 \leqslant |f_{\rm iso}| < \infty$ corresponds in this parametrization to a compact interval [0..1) for α [or $(-1..1)$ for $\sqrt{\alpha}$], over which they take a flat prior for the variable α (or $\sqrt{\alpha}$). Flat priors over α or $\sqrt{\alpha}$ correspond to the priors over $|f_{\rm iso}|$ depicted in Fig. 2, which cut away the region of parameter space where $|f_{\rm iso}| \gg 1$. As a consequence, the Occam's razor effect is suppressed and the resulting odds in favour of the purely adiabatic model are much smaller than in our case.

This example illustrates that model comparison results can depend crucially on the underlying parameter space. We now turn to discuss the dependence of our other results on the prior range one chooses to adopt.

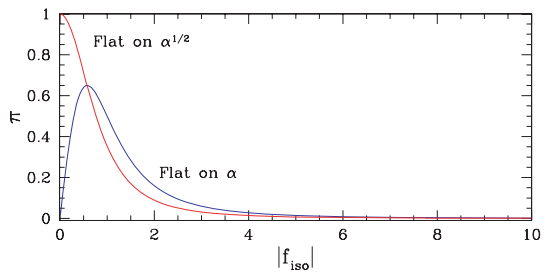

Figure 2. Equivalent priors on $|f_{\rm iso}|$ corresponding to the flat priors used in Beltran et al. (2005) for the parameters α and $\sqrt{\alpha}$. Both priors cut away the parameter space $|f_{\rm iso}| \gg 1$, thus reducing the Occam's razor effect caused by a scale-free parameter. The odds in favour of the purely adiabatic model thus become correspondingly smaller. Model comparison results can depend crucially on the variables adopted.

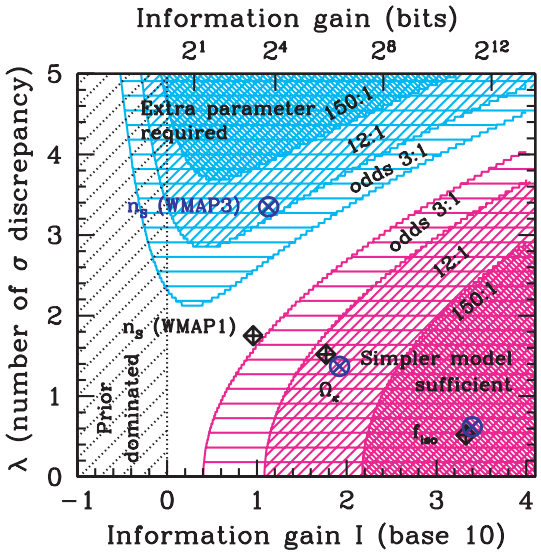

Figure 3. Regions in the (I, λ) plane (shaded) where one of the competing models is supported by positive (odds of 3 : 1), moderate (12 : 1) or strong (odds larger than 150 : 1) evidence. The white region corresponds to an inconclusive result (odds of about 1 : 1), while in the region $I < 0$ (dotted) the posterior is dominated by the prior and the measurement is non-informative. In the lower horizontal axis, I is given in base 10, that is, $I = -\log_{10}\beta$, while it is given in bits in the upper horizontal axis. The contours are computed from the SDDR formula assuming a Gaussian likelihood and a Gaussian prior. The location of the three parameters analysed in the text is shown by diamonds (circles) for *WMAP*1+ ext data (*WMAP*3+ ext data). Choosing a wider (narrower) prior range would shift the points horizontally to the right-hand side (left-hand side) of the plot.

3.3 Dependence on the choice of prior

As described in detail in Appendix A, the Bayes factor is really a function of two parameters, λ and the information content $I = -\ln\beta$, see equation (A9) for the case of a Gaussian prior and a Gaussian likelihood in the parameter of interest. Fig. 3 shows contours of $|\ln B_{01}| =$ constant for constant $= 1.0, 2.5, 5.0$ in the (I, λ) plane, as computed from equation (A9). The contours delimit significant levels for the strength of evidence, as summarized in Table 1. In the following, we will measure the information content I in base-10 logarithm. For moderately informative data ($I \approx 1$–2) the measured mean has to lie at least about 4σ away from $\omega_\star$ in order to robustly disfavour the simpler model (i.e. $\lambda \gtrsim 4$). Conversely, for $\lambda \lesssim 3$ highly informative data ($I \gtrsim 2$) do favour the conclusion that $\omega = \omega_\star$. In general, a large information content favours the simpler model, because Occam's razor penalizes the large volume of 'wasted' parameter space of the extended model. A large λ disfavours exponentially the simpler model, in agreement with the sampling theory result. The location on the plane of the three cases discussed in the text (the scalar spectral index, the spatial curvature and the CDM isocurvature component) is marked by diamonds (circles) for *WMAP*1 + ext (*WMAP*3 + ext). Even though the informative regions of Fig. 3 assume a Gaussian likelihood, they are illustrative of the results one might obtain in real cases, and can serve as a rough guide for the Bayes factor determination.

Another useful property of displaying the result of the model comparison in the (I, λ) plane as in Fig. 3 is that the impact of a change of prior can be easily estimated. A different choice of prior will amount to a *horizontal shift* of the points in Fig. 3, at least as long as $I > 0$ (i.e. posterior dominated by the likelihood). Thus we can see that given the results with the priors used in this paper, *no other choice of priors* for $f_{\rm iso}$ or Ω_κ within four orders of magnitude will achieve a reversal of the conclusion regarding the favoured model. At most, picking more restrictive priors (reflecting more predictive theoretical models) would make the points for $f_{\rm iso}$ or Ω_κ drift to the left-hand side of Fig. 3, eventually entering in the white, inconclusive region $I \lesssim 0.5$. For the spectral index from *WMAP* 3-year data, choosing a prior two orders of magnitude larger than the one employed here, that is, $-19 < n_{\rm S} < 20$ would reverse the conclusion of the model selection, favouring the model $n_{\rm S} = 1$ with odds of about 3 : 1. This choice of prior is however physically unmotivated. On the other hand, reducing the prior by one order of magnitude – that is, making it of the same order as the current posterior width ($I = 0$) – would still not alter the conclusion that $n_{\rm S} = 1$ is disfavoured with moderate odds.

The prior assignment is an irreducible feature of Bayesian model selection, as it is clear from its presence in the denominator of equation (7). There is a vast literature which addresses the problem of assigning prior probabilities (see footnote 1) in a way which reflects the state of knowledge before seeing the data. In applications to model selection, it might be more useful to regard the prior as expressing the available parameter space under the model, rather then a state of knowledge before seeing the data, as argued in Kunz et al. (2006). The underpinnings of the prior choice can be found in our understanding of model-specific issues. In this work we have offered two examples of priors stemming from theoretical motivations: the prior on the scalar spectral index is a consequence of assuming slow-roll inflation while the prior on the spatial curvature comes from our knowledge that the Universe is not empty (and therefore the curvature must be smaller than -1) nor overclosed (or it would have recollapsed). This simple observations set the correct scale for the prior on Ω_κ, which is of the order of unity. On the other hand, if one wanted to impose an inflation-motivated prior of width $\ll 1$, then the information content of the data would go to 0 and the outcome of the model selection would be non-informative. In general, it is enough to have an order of magnitude estimate of the a priori allowed range for the parameter of interest, since the logarithm of the model likelihood is proportional to the logarithm of the prior range. Furthermore, considerations of the type outlined above can help assessing the impact of a prior change on the model comparison outcome. Often one will find that most 'reasonable' prior choices will lead to qualitatively to the same conclusion, or else to a non-committal result of the model comparison.

For essentially scale-free parameters, such as the adiabatic and isocurvature amplitudes of our third application, model theoretical considerations of the type employed by Lazarides et al. (2004) can

lead to a limitation of the prior range. In the context of phenomenological model building, we have demonstrated that an analysis of the a priori parameter space accessible to the instrument can be used to define a prior encapsulating our expectations on the quality of the data we will be able to gather.

An important *caveat* is the dependence of the Bayes factor on the basis one adopts in parameter space, which sets the natural measure on the parameters. A flat prior on θ does not correspond to a flat prior on some other set $\alpha(\theta)$ obtained via a non-linear transformation, since the two prior distributions are related via

$$\pi(\theta \mid M) = \pi(\alpha \mid M) \left| \frac{\mathrm{d}\alpha(\theta)}{\mathrm{d}\theta} \right|. \quad (12)$$

As illustrated by the case of the isocurvature amplitude, this is especially relevant for parameters which can vary over many orders of magnitudes. We put forward that the choice of the parameter basis can be guided by our physical insight of the model under scrutiny and our understanding of the observations. This principle would suggest that one should adopt flat priors along 'normal variables' or principal components, because those are directly probed by the data and usually can be interpreted in terms of physically relevant and meaningful quantities. A general principle of consistency can be invoked to select the most appropriate variable for cases where many apparently equivalent choices are present (e.g. $f_{\rm iso}$, α or $\sqrt{\alpha}$). We leave further exploration of this very relevant issue to a future publication.

4 CONCLUSIONS

We have argued that frequentist significance tests should be interpreted carefully and in particular that Bayesian model selection reasoning should be used to decide whether the introduction of a parameter is warranted by data. The main strengths of the Bayesian approach are that it does consider the information content of the data and that it allows one to confirm predictions of a model, instead of just disproving them as in the sampling theory approach.

We have investigated the use of the SDDR as a tool to compute the Bayes factor of two nested models, at no extra computational cost than the MC sampling of the parameter space. The technique is likely to be accurate for cases where the estimated value of the extra parameter under the larger model lies less than about 3σ away from the predicted value under the simpler model, as shown in Appendix C. In a companion paper (Trotta 2007a) a complementary technique is introduced, called PPOD, which produces forecasts for the probability distribution of the Bayes factor from future experiments.

We have applied this Bayesian model selection point of view to three central ingredients of present-day cosmological model building. Regarding the spectral index of scalar perturbations, we found that *WMAP* 3-year data disfavour a scale-invariant spectral index with moderate evidence, and that this result holds true for all reasonable choices of priors. This is a significant change with respect to the inconclusive result one obtained using the *WMAP* first-year data release instead. We found that the odds in favour of a flat universe have doubled (from 15 : 1 to 29 : 1) in going from *WMAP*1 + ext to *WMAP*3 + ext, and we have stressed that this conclusion can only be obtained if the Hubble parameter is measured independently or if supernovae luminosity distance measurements (or other low-redshift rulers, such as baryonic acoustic oscillations, see Eisenstein et al. (2005)) are employed. Finally, purely adiabatic initial conditions are strongly preferred to a mixed model containing a totally (anti)correlated CDM isocurvature contribution (odds larger than 1000 : 1), on the grounds of an Occam's razor argument, that the prior available parameter space is much larger than the small surviving posterior volume. This is however crucially dependent on the variable one chooses to impose flat priors on.

In the light of these findings, it seems to us that model comparison tools offer complementary insight in what the data can tell us about the plausibility of theoretical speculations regarding cosmological parameters, and can provide useful guidance in the quest of a cosmological concordance model.

ACKNOWLEDGMENTS

I am grateful to Chiara Caprini, Ruth Durrer, Samuel Leach, Julien Lesgourgues and Christophe Ringeval for useful discussions. I thank Martin Kunz, Andrew Liddle, Tom Loredo, Pia Mukherjee and David Parkison for many enlightening discussions and valuable comments on earlier drafts. I thank an anonymous referee for many helpful suggestions. This research is supported by the Tomalla Foundation, by the Royal Astronomical Society through the Sir Norman Lockyer Fellowship and by St Anne's College, Oxford. The use of the Myrinet cluster (University of Geneva) and of the Glamdring cluster (Oxford University) is acknowledged. I acknowledge the use of the package COSMOMC, available from cosmologist.info, and the use of the Legacy Archive for Microwave Background Data Analysis (LAMBDA). Support for LAMBDA is provided by the NASA Office of Space Science.

REFERENCES

Amendola L., Gordon C., Wands D., Sasaki M., 2002, Phys. Rev. Lett., 88, 211302
Bassett B. A., Corasaniti P. S., Kunz M., 2004, ApJ, 617, L1
Bean R., Dunkley J., Pierpaoli E., 2006, Phys. Rev. D, 74, 063503
Beltran M., Garcia-Bellido J., Lesgourgues J., Riazuelo A., 2004, Phys. Rev. D, 70, 103530
Beltran M., Garcia-Bellido J., Lesgourgues J., Liddle A. R., Slosar A., 2005, Phys. Rev. D, 71, 063532
Bennett C. L. et al., 2003, ApJS, 148, 1
Bowen R., Hansen S. H., Melchiorri A., Silk J., Trotta R., 2002, MNRAS, 334, 760
Bucher M., Dunkley J., Ferreira P. G., Moodley K., Skordis C., 2004, Phys. Rev. Lett., 93, 081301
Crotty P., Garcia-Bellido J., Lesgourgues J., Riazuelo A., 2003, Phys. Rev. Lett., 91, 171301
DiCiccio T., Kass R., Raftery A., Wasserman L., 1997, J. Am. Stat. Assoc., 92, 903
Dickey J. M., 1971, Ann. Math. Stat., 42, 204
Drell P. S., Loredo T. J., Wasserman I., 2000, ApJ, 530, 593
Eisenstein D. J. et al., 2005, ApJ, 633, 560
Freedman W. L. et al., 2001, ApJ, 553, 47
Gordon C., Lewis A., 2003, Phys. Rev. D, 67, 123513
Hinshaw G. et al., 2003, ApJS, 148, 135
Hinshaw G. et al., 2006, preprint (astro-ph/0603451)
Hobson M. P., McLachlan C., 2003, MNRAS, 338, 765
Hobson M. P., Bridle S. L., Lahav O., 2002, MNRAS, 335, 377
Jaffe A. H., 1996, ApJ, 471, 24
Jaynes E., 2003, Probability Theory. The Logic of Science. Cambridge Univ. Press, Cambridge
Jeffreys H., 1961, Theory of Probability, 3rd edn. Oxford Univ. Press, Oxford
Kass R., Raftery A., 1995, J. Am. Stat. Assoc., 90, 773
Keskitalo R., Kurki-Suonio H., Muhonen V., Valiviita J., 2006, preprint (astro-ph/0611917)
Kosowsky A., Milosavljevic M., Jimenez R., 2002, Phys. Rev. D, 66, 063007
Kunz M., Trotta R., Parkinson D., 2006, Phys. Rev. D, 74, 023503
Kuo C.-l. et al., 2004, ApJ, 600, 32

Kurki-Suonio H., Muhonen V., Valiviita J., 2005, Phys. Rev. D, 71, 063005
Lazarides G., de Austri R. R., Trotta R., 2004, Phys. Rev. D, 70, 123527
Lesgourgues J., Pastor S., 2006, Phys. Rep., 429, 307
Lewis A., Bridle S., 2002, Phys. Rev. D, 66, 103511
Lindley D., 1957, Biometrika, 44, 187
Loredo T. J., Lamb D. Q., 2002, Phys. Rev. D, 65, 063002
Lyth D. H., Wands D., 2003, Phys. Rev. D, 68, 103516
MacKay D., 2003, Information Theory, Inference, and Learning Algorithms. Cambridge Univ. Press, Cambridge
Marshall P., Rajguru N., Slosar A., 2006, Phys. Rev. D, 73, 067302
Moodley K., Bucher M., Dunkley J., Ferreira P. G., Skordis C., 2004, Phys. Rev. D, 70, 103520
Mukherjee P., Parkinson D., Liddle A. R., 2006, ApJ, 638, L51
Page L. et al., 2006, preprint (astro-ph/0603450)
Parkinson D., Mukherjee P., Liddle A. R., 2006, Phys. Rev. D, 73, 123523
Pierpaoli E., Garcia-Bellido J., Borgani S., 1999, J. High Energy Phys., 10, 015
Readhead A. C. S. et al., 2004, ApJ, 609, 498
Riess A. G. et al., 2004, ApJ, 607, 665
Saini T. D., Weller J., Bridle S. L., 2004, MNRAS, 348, 603
Skilling J., 2004, Nested sampling for general Bayesian computation, available from: http://www.inference.phy.cam.ac.uk/bayesys
Slosar A. et al., 2003, MNRAS, 341, L29
Spergel D. N. et al., 2006, preprint (astro-ph/0603449)
Tegmark M. et al., 2004, ApJ, 606, 702
Trotta R., 2004, PhD thesis, Univ. Geneva, Switzerland
Trotta R., 2005, archived as preprint (astro-ph/0504022v1)
Trotta R., 2007a, MNRAS, in press (astro-ph/0703063, doi:10.1111/j.1365-2966.2007.11861.x)
Trotta R., 2007b, MNRAS, 375, L26
Trotta R., 2007c, New Astron. Rev., 51, 316
Trotta R., Durrer R., 2006, in Novello M., Perez Bergliaffa S., eds, Proc. MG10 Meeting. World Scientific, Singapore
Trotta R., Hansen S. H., 2004, Phys. Rev. D, 69, 023509
Trotta R., Riazuelo A., Durrer R., 2001, Phys. Rev. Lett., 87, 231301
Trotta R., Riazuelo A., Durrer R., 2003, Phys. Rev. D, 67, 063520
Valiviita J., Muhonen V., 2003, Phys. Rev. Lett., 91, 131302
Verde L. et al., 2003, ApJS, 148, 195
Verdinelli I., Wasserman L., 1995, J. Am. Stat. Assoc., 90, 614

APPENDIX A: AN ILLUSTRATION OF LINDLEY'S PARADOX

Lindley's paradox describes a situation where frequentist significance tests and Bayesian model selection procedures give contradictory results. As we demonstrate below, it arises because the information content of the data is neglected in the frequentist approach.

Let us consider the toy example of a measurement of a Gaussian distributed quantity, ω, by drawing n independent and identically distributed samples with known s.d. σ. Then the likelihood function is the normal distribution

$$p(\hat{\mu}, \hat{\sigma} \mid \omega) = N_{\hat{\mu}\hat{\sigma}}(\omega), \tag{A1}$$

where $\hat{\mu}$ is the estimated mean and $\hat{\sigma} = \sigma/\sqrt{n}$ its uncertainty. From the point of view of frequentist statistics, a significance test is performed on the null hypothesis $\mathcal{H}_0 : \omega = \omega_\star$. We define λ as dimensionless number which indicates 'how many sigma away' is our estimate of the mean, $\hat{\mu}$, from its value under $\mathcal{H}_0$ in units of the estimated s.d.:

$$\lambda \equiv \frac{|\hat{\mu} - \omega_\star|}{\hat{\sigma}}. \tag{A2}$$

This 'number of sigma' difference is interpreted as a measure of the confidence with which one can reject $\mathcal{H}_0$. The 'p-value',

$$\int_\lambda^\infty p(\hat{\mu}, \hat{\sigma} \mid \omega)\, \mathrm{d}\omega, \tag{A3}$$

is compared to a number α, called the 'significance level' of the test and the hypothesis $\mathcal{H}_0$ is rejected at the $1 - \alpha$ confidence level if p-value $< \alpha$. If we pick a (fixed) confidence level, say $\alpha = 0.05$, then the frequentist significance test rejects the null hypothesis if

$$Z(\lambda) \equiv \frac{1}{\sqrt{2\pi}} \int_\lambda^\infty \exp\left(-\frac{t^2}{2}\right) \mathrm{d}t \leqslant \alpha/2. \tag{A4}$$

(for a two-tailed test). For $\alpha = 0.05$ the equality in equation (A4) holds for $\lambda = 1.96$. In other words, sampling statistics reject the null hypothesis at the 95 per cent confidence level if the measured mean is more than $\lambda = 1.96\sigma$ away from the predicted $\omega_\star$ under $\mathcal{H}_0$.

This conclusion can be in strong disagreement with the Bayesian evaluation of the Bayes factor, that is, a value $\omega_\star$ rejected under a frequentist test can on the contrary be favoured by Bayesian model comparison (Lindley 1957). In the Bayesian model comparison approach, the two competing models are M_0, with no free parameters, in which the value of ω is fixed to $\omega = \omega_\star$, and model M_1, with one free parameter $\omega \neq \omega_\star$. Under M_1, our prior belief before seeing the data on the probability distribution of ω is explicitly represented by the prior PDF $\pi(\omega)$. This prior PDF is then updated to the posterior via Bayes theorem,[5] equation (1).

A formal measure of the information gain obtained through the data is the cross-entropy between prior and posterior, the Kullback–Leibler divergence

$$D_{\mathrm{KL}}(p, \pi) = \int p(\theta \mid d) \ln \frac{p(\theta \mid d)}{\pi(\theta)}\, \mathrm{d}\theta. \tag{A5}$$

For a Gaussian prior of s.d. $\Delta\omega$ centred on $\omega_\star$ and a Gaussian likelihood with mean $\hat{\mu}$ and s.d. $\hat{\sigma}$, the information gain is given by

$$D_{\mathrm{KL}} + \frac{1}{2} = -\ln\beta + \frac{1}{2}\beta^2(\lambda^2 - 1), \tag{A6}$$

where we have defined

$$\beta \equiv \hat{\sigma}/\Delta\omega, \tag{A7}$$

the factor by which the accessible parameter space under M_1 is reduced after the arrival of the data (remember that $\hat{\sigma}$ is the s.d. of the likelihood). For totally uninformative data, $\beta = 1$ and $\lambda = 0$, and thus $D_{\mathrm{KL}} = 0$. Unless $\lambda \gg 1$ (in which case the null hypothesis is rejected with many sigma and there is hardly any need for model comparison) we can usually neglect the second term on the right-hand side of equation (A6). We are therefore led to define a simpler measure of the *information content* of the data, I, as

$$I \equiv -\ln\beta. \tag{A8}$$

The choice of the logarithm base is only a matter of convenience, and this sets the units in which the entropy is measured. Had we chosen base-2 logarithm instead, the information would have been measured in bits. In Fig. 3, the choice of using the base-10 logarithm for the bottom horizontal axis means that I describes the order of magnitude by which our prior knowledge has improved after the arrival of the data.

[5] Notice that, after applying Bayes theorem, the posterior probability is attached to the parameter ω itself, not to the estimator $\hat{\mu}$ as in sampling theory. In the Bayesian framework we only deal with observed data, never with properties of estimators based on a (fictional) infinite replication of the data. In cosmology one only has one realization of the Universe and there is not even the conceptual possibility of reproducing the data *ad infinitum* and therefore the Bayesian standpoint seems better suited to such a situation.

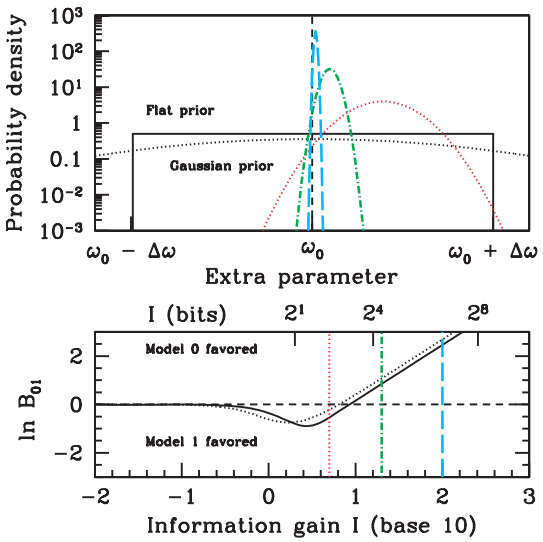

Figure A1. Illustration of Lindley's paradox. Sampling statistics hypothesis testing rejects the hypothesis that $\omega = \omega_\star$ with 95 per cent confidence in all three cases (coloured curves) illustrated in the top panel ($\lambda = 1.96$ in all cases). Bayesian model selection does take into account the information content of the data I, and correctly favours the simpler model (predicting that $\omega = \omega_\star$) for informative data (right-hand vertical line in the bottom panel, $I = 2$ expressed in base-10 logarithm), with odds of 14 : 1 (for a Gaussain prior, dotted black line). Using a flat prior of the same width (solid black line) instead reduces $\ln B_{01}$ by a geometric factor $\ln(2/\pi)/2 = 0.22$ in the informative ($I \gg 1$) regime. Notice that for non-informative data ($I \ll 0$) the Bayes factor reverts to equal odds for the two models.

We now compute the Bayes factor B_{01} in favour of model M_0 from equation (7), using again the above Gaussian prior, obtaining

$$B_{01}(\beta, \lambda) = \sqrt{1+\beta^{-2}}\exp\left[-\frac{\lambda^2}{2(1+\beta^2)}\right]. \tag{A9}$$

The model comparison result thus depends not only on λ, but also on the quantity β, which is proportional to the volume occupied by the posterior in parameter space and describes the information gain in going from prior to posterior. If instead of a Gaussian prior one takes a flat prior around $\omega_\star$ of width $2\Delta\omega$ (the factor of 2 being chosen to facilitate the comparison with the case of a Gaussian prior of s.d. $\Delta\omega$) one obtains instead

$$B_{01}(\beta, \lambda) = \sqrt{\frac{2}{\pi}}\beta^{-1}\exp\left(-\frac{\lambda^2}{2}\right) \times [Z(\lambda-\beta^{-1}) - Z(\lambda+\beta^{-1})]^{-1}, \tag{A10}$$

where the function $Z(y)$ is defined in (A4), a consequence of the top-hat prior. For $\beta^{-1} \gg \lambda$, the posterior is well localized within the boundaries of the prior and the term in square brackets in (A10) tends to 1.

In order to clarify the role of the information content and the difference with frequentist hypothesis testing, consider the following example (see Fig. A1). For a fixed choice of prior width $\Delta\omega$, imagine performing three different measurements, each with a different value of β (i.e. with different information content I) but with outcomes such that λ is the same in all three cases. This is depicted in the top panel of Fig. A1, where the likelihood mean is $\lambda = 1.96\sigma$ away from $\omega_\star$ for all three cases. Under sampling statistics, all three measurements equally reject the null hypothesis, that $\omega = \omega_\star$, at the 95 per cent confidence level. And yet common sense clearly tells us that this cannot be the right conclusion in all three cases. Indeed, the Bayes factor, equation (A9) or (A10), correctly recovers the intuitive result (bottom panel of Fig. A1): the measurement with the larger error ($\beta = 1/5$, or $I = 0.7$, expressed in base-10 logarithm) corresponds to the least informative data, and the Bayes factor slightly disfavours the simpler model ($\ln B_{01} = -0.2$, or odds of about 5 : 4 against M_0 and $p(M_0|\,d) = 0.44$). For $\beta = 1/20$ or $I = 1.3$ (moderately informative data), evidence starts to accumulate *in favour* of M_0 ($\ln B_{01} = 1.08$, odds of 3 : 1 in favour and $p(M_0|\,d) = 0.75$). For very informative data, $\beta = 1/100, I = 2$, Bayesian reasoning correctly deduces that the simpler M_0 should be favoured ($\ln B_{01} = 2.68$, odds of 14 : 1 in favour of M_0 and a posterior probability $p(M_0|\,d) = 0.94$). The above numbers are for a Gaussian prior, but those conclusion are largely independent of the choice of a Gaussian or of a flat prior, provided the bulk of the prior volume is the same (compare the dotted and solid line in the bottom panel of Fig. A1 for a Gaussian and a flat prior, respectively).

This illustration shows that the Bayes factor can correctly favour models which would be rejected with high confidence by hypothesis testing in a sampling theory approach. While in sampling theory one is only able to disprove models by rejecting hypothesis, it is important to highlight that the Bayesian evidence can and does accumulate *in favour* of simpler models, scaling as $1/\beta$. While it is easier to disprove $\omega = \omega_\star$, since model rejection is exponential with λ, the Bayesian approach allows to evaluate what the data have to say *in favour* of a hypothesis, as well.

In summary, quoting the number of sigma away from $\omega_\star$ (the λ parameter) is not always an informative statement to decide whether or not a parameter ω differs from $\omega_\star$. Answering this question is a model comparison issue, which requires the evaluation of the Bayes factor.

APPENDIX B: DERIVATION OF THE SDDR

The Bayes factor B_{01} of equation (2) can be evaluated by computing the integrals

$$p(M_0\,|\,d) = \int \mathrm{d}\psi\,\pi_0(\psi)p(d|\psi,\omega_\star), \tag{B1}$$

$$p(M_1\,|\,d) = \int \mathrm{d}\psi\,\mathrm{d}\omega\,\pi_1(\psi,\omega)p(d|\psi,\omega) \equiv q. \tag{B2}$$

Here $\pi_0(\psi)$ denotes the prior over ψ in model M_0, and $\pi_1(\psi, \omega)$ the prior over (ψ, ω) under model M_1. Note that, since the models are nested, the likelihood function for M_0 is just a slice at constant $\omega = \omega_\star$ of the likelihood function in model M_1, $p(d\,|\,\psi, \omega)$.

Now multiply and divide B_{01} by the number $p(\omega_\star\,|\,d) \equiv p(\omega = \omega_\star\,|\,d, M_1)$, which is the marginalized posterior for ω under M_1 evaluated at $\omega_\star$, and using that $p(\omega_\star\,|\,d) = p(\omega_\star, \psi\,|\,d)/p(\psi\,|\,\omega_\star, d)$ at all points ψ, we obtain

$$B_{01} = p(\omega_\star\,|\,d)\int \mathrm{d}\psi\,\frac{\pi_0(\psi)p(d\,|\,\psi,\omega_\star)p(\psi\,|\,\omega_\star,d)}{qp(\omega_\star,\psi\,|\,d)} \tag{B3}$$

$$= p(\omega_\star\,|\,d)\int \mathrm{d}\psi\,\frac{\pi_0(\psi)p(\psi\,|\,\omega_\star,d)}{\pi_1(\omega_\star,\psi)}, \tag{B4}$$

where in the second equality we have used the definition of posterior, namely that $p(\omega_\star, \psi\,|\,d) = p(d\,|\,\omega_\star, \psi)\pi_1(\omega_\star, \psi)/q$. Up to this point we have not made any assumption nor approximation. We now assume that the prior satisfies

$$\pi_1(\psi\,|\,\omega_\star) = \pi_0(\psi), \tag{B5}$$

which always holds in the (usual in cosmology) case of separable priors, that is,

$$\pi_1(\omega, \psi) = \pi_1(\omega)\pi_0(\psi). \quad \text{(B6)}$$

Under this assumption, and since $p(\psi \mid \omega_\star, d)$ in (B4) is the normalized marginal posterior, equation (B4) simplifies to the SDDR given in equation (7).

APPENDIX C: BENCHMARK TESTS FOR THE SDDR

In order to explore the accuracy of the SDDR, we have tested its performance for the benchmark case of a Gaussian likelihood. A D-dimensional likelihood is generated by choosing a random D-dimensional, diagonal covariance matrix. The correlations can be set to 0 without loss of generality since in the Gaussian case it is always possible to rotate to the principal axis of the covariance ellipse. The mean of the likelihood is set to 0 for the last $D - 1$ dimensions, while for the first parameter (the one we are interested in testing) the mean is chosen to lie $\lambda\sigma_1$ away from 0, where λ is selected below and σ_1^2 is the covariance along direction 1. We then compare the two following nested models: M_0 predicts that the first parameter $\theta_1 = 0$, while M_1 has a Gaussian prior centred around 0 and of width $\Delta w = \sigma_1/\beta$, where β is fixed.

The posterior is then reconstructed using a MCMC algorithm and the Bayes factor computed using the SDDR. The results are shown in Fig. C1 as a function of the number of samples for parameter spaces of dimension $D = 5, 10, 20$ and for $\lambda = 1, 2, 3$. We have fixed $\beta = 0.2$ throughout (changing the value of β only rescales the Bayes factor without affecting the accuracy, as long as $\beta < 1$, i.e. for informative data). The errors on the Bayes factor are computed as in the text using a bootstrap technique: the full sample set is divided in $R = 5$ subsets, then the mean and s.d. of the SDDR are computed from those subsets. The error thus only reflect the statistical noise within the chain and it does not take into account a possible systematic underexploration of the tails of the likelihood.

It is clear that the SDDR performs extremely well for $\lambda \leqslant 2$ while it becomes less accurate for $\lambda = 3$. This is because it is rather difficult to explore regions further out in the tails of the distribution using conventional MCMC methods. For $\lambda > 3$ it becomes very unpractical to obtain sufficient samples in the tail. For models that lie less than about 3σ away from each other, the SDDR gives a satisfactory accuracy in the model comparison result at no extra cost than the parameter estimation step, requiring less than 10^5 samples. Furthermore, the scaling with the dimensionality of the parameter space appears to be rather favourable, and the error increases only mildly from $D = 5$ to 20 at a given number of samples.

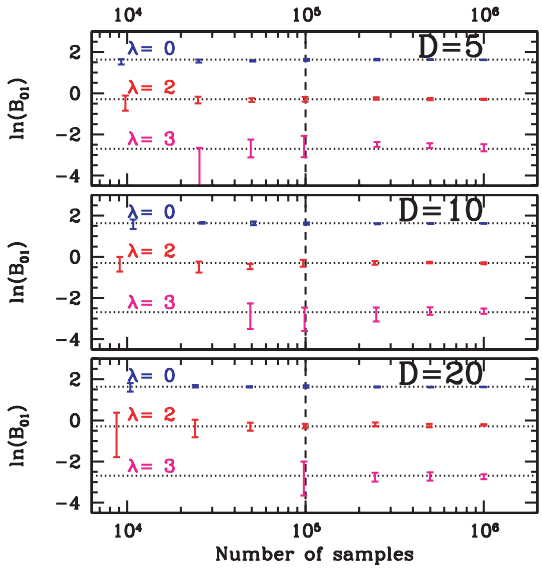

Figure C1. Benchmark test for the SDDR formula for a Gaussian likelihood and prior, for parameter spaces of dimensionality D. The horizontal, dotted lines give the exact value. The SDDR performs extremely well for comparing models lying $\lambda < 3\sigma$ away from each other. In this case, less than 10^5 samples are required to achieve a satisfactory agreement with the exact result. For $\lambda \gtrsim 4$ the tails of the likelihood are not sufficiently explored to apply the SDDR. The missing points for $\lambda = 3$ indicate that the given number of samples are insufficient to achieve coverage of the simpler model prediction.

Clearly, for likelihoods that are close to Gaussian, the approximations (A9) and (A10) can still give a useful order of magnitude estimate of the result. Finally, we stress that in the regime where the SDDR works well ($\lambda \lesssim 3$) its accuracy is not limited by the assumption of normality of the likelihood, but only by the efficiency and accuracy of the MCMC reconstruction of the posterior. Particular care must be exercised in exploring accurately distributions presenting heavier tails than Gaussians, and further work is required to extend the MCMC sampling to the regime $\lambda \gtrsim 4$. In this case, sampling at a higher temperature could help in obtaining sufficient samples in the tail, an issue whose exploration we leave for future work.

This paper has been typeset from a TEX/LATEX file prepared by the author.