# Statistical methods for cosmological parameter selection and estimation

Andrew R. Liddle
Astronomy Centre, University of Sussex, Brighton BN1 9QH, UK

March 24, 2009

**Abstract**

The estimation of cosmological parameters from precision observables is an important industry with crucial ramifications for particle physics. This article discusses the statistical methods presently used in cosmological data analysis, highlighting the main assumptions and uncertainties. The topics covered are parameter estimation, model selection, multi-model inference, and experimental design, all primarily from a Bayesian perspective.

## 1 INTRODUCTION

During the last decade, cosmology has advanced from an era of largely qualitative questions — is the Universe flat, open or closed?, does dark energy exist in the Universe?, etc. — to one where precision determinations of many of the Universe's properties are possible. We have cosmological models capable of explaining the detailed observations available, and whose parameters are beginning to be pinned down at the ten percent, and in some cases one percent, level [1]. Nevertheless, quality cosmological data are an expensive resource and it is imperative to make the best possible use of them. This implies use of the best available statistical tools in order to obtain accurate and robust conclusions.

For around a decade now, the established leading cosmological model considers five material constituents: baryons (taken, imprecisely, to include electrons), photons, neutrinos, cold dark matter (CDM), and dark energy. The simplest model for dark energy, a cosmological constant $\Lambda$, is in excellent agreement with observations, and the model is then known as a $\Lambda$CDM model. The most important constraints come from the evolution of cosmic structures. These are seeded by small initial density perturbations, which in the standard cosmological model are taken as adiabatic, gaussian, and nearly scale-invariant, as predicted by the simplest models of cosmological inflation [2].

This model is supported and constrained by a series of cosmological observations. Most important are measurements of cosmic microwave background (CMB) anisotropies, particularly by the Wilkinson Microwave Anisotropy Probe (WMAP) as shown in Figure 1. Typical analyses also incorporate other data, such as galaxy clustering data, the luminosity distance–redshift relation of Type Ia supernovae, and direct measures of the Hubble constant. The region of parameter space where the $\Lambda$CDM model matches all those data is often referred to as the concordance model.

In its very simplest incarnation, the photon density is taken to be well measured by the CMB temperature, neutrinos to be nearly massless, and the Universe spatially flat. The model then features only four fundamental parameters: the Hubble parameter $h$, the densities of baryons $\Omega_{\rm b}$ and CDM $\Omega_{\rm CDM}$, and the amplitude of primordial density perturbations $A_{\rm S}$. In addition, a comparison with data will usually include extra phenomenological parameters, which in principle can be computed from the above but which in practice cannot be reliably. For cosmic microwave background studies, the optical depth $\tau$, measuring the fraction of CMB photons scattering from ionized gas at low redshift, is needed, while use of galaxy clustering data may require inclusion of the galaxy bias parameter $b$ which relates galaxy clustering to dark matter clustering.

Beyond those basic parameters, cosmologists hope that future investigations will uncover new physical processes, permitting extra parameters to be incorporated and measured. In some cases, it

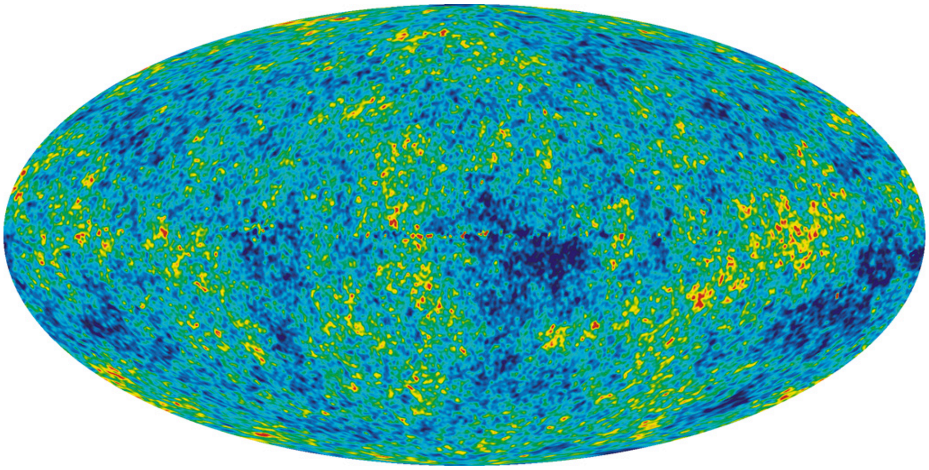

Figure 1: Cosmic microwave background anisotropies as imaged by WMAP from five years of accumulated data. [Figure courtesy NASA/WMAP Science Team.]

is more or less certain that the parameter is relevant and only a matter of time before observational sensitivity becomes sufficient. Examples here are neutrino masses and the cosmic Helium fraction (though the latter is again in principle computable from other parameters, independent verification of its value would be an important consistency check).

Much more numerous, though, are parameters describing effects which may or may not be relevant to our Universe. An extensive list is given, for instance, in Ref. [3]. Are the primordial perturbations precisely scale invariant, or do they have a scale dependence quantified by the spectral index $n$? Do primordial gravitational waves exist, as predicted by inflation? Does the dark energy density evolve with time? Are there cosmic strings in the Universe? Are the initial perturbations really adiabatic and gaussian? A fuller account of these parameters can be found for instance in Ref. [4].

In summary, creation of precision cosmological models is an ongoing process with two distinct goals. One is to determine the set of parameters, i.e. physical processes, necessary to describe the available observations. The second is to determine the preferred values of these parameters. We can then pursue the ultimate aim of relating cosmological observations to underlying fundamental physics.

# 2 INFERENCE

## 2.1 Orientation

Inference is the method by which we translate experimental/observational information into constraints on our mathematical models. The model is a representation of the physical processes that we believe are relevant to the quantities we plan to observe. To be useful, the model must be sufficiently sophisticated as to be able to explain the data, and simple enough that we can obtain predictions for observational data from it in a feasible time. At present these conditions are satisfied in cosmology, with the best models giving an excellent representation of that data, though the computation of theoretical predictions for a representative set of models is a supercomputer class problem. Particularly valued are models which are able to make distinctive predictions for observations yet to be made, though Nature is under no obligation to behave distinctively.

The data which we obtain may be subject only to experimental uncertainty, or they may also have a fundamental statistical uncertainty due to the random nature of underlying physical processes. Both types of data arise in cosmology. For instance, the present expansion rate of the Universe (the Hubble constant), could in principle be measured to near-arbitrary accuracy with sufficiently advanced instrumentation. By contrast, the detailed pattern of cosmic microwave anisotropies, as measured by WMAP, is not believed to be predictable even in principle, being attributed to a particular random realization of quantum processes occurring during inflation [2]. Observers at different locations in the Universe see different patterns in the CMB sky, the cosmological information being contained in statistical measures of the anisotropies such as the power spectrum. Observers at any particular location, such as ourselves, can witness only our own realization and there is an inherent statistical uncertainty, cosmic variance, that we cannot overcome, but which fortunately can be modelled and incorporated in addition to measurement uncertainty.

A model will typically not make unique predictions for observed quantities; those predictions will instead depend on some number of parameters of the model. Examples of cosmological parameters are the present expansion rate of the Universe, and the densities of the various constituents such as baryons, dark matter, etc. Such parameters are not (according to present understanding, anyway) predictable from some fundamental principle; rather, they are to be determined by requiring that the model does fit the data to hand. Indeed, determining the values of such parameters is often seen as the primary goal of cosmological observations, and Chapter 3 is devoted to this topic.

At a more fundamental level, several different models might be proposed as explanations of the observational data. These models would represent alternative physical processes, and as such would correspond to different sets of parameters that are to be varied in fitting to the data. It may be that the models are nested within one another, with the more complex models positing the need to include extra physical processes in order to explain the data, or the models may be completely distinct from one another. An example of nested models in cosmology is the possible inclusion of a gravitational wave contribution to the observed CMB anisotropies. An example of disjoint models would be the rival explanations of dark energy as caused by scalar field dynamics or by a modification to the gravitational field equations. Traditionally, the choice of model to fit to the data has been regarded as researcher-driven, hopefully coming from some theoretical insight, with the model to be validated by some kind of goodness-of-fit test. More recently, however, there has been growing interest in allowing the data to distinguish between competing models. This topic, model selection or model comparison, is examined in Chapter 4.

The comparison of model prediction to data is, then, a statistical inference problem where uncertainty necessarily plays a role. While a variety of techniques exist to tackle such problems, within cosmology one paradigm dominates — Bayesian inference. This article will therefore focus almost exclusively on Bayesian methods, with only a brief account of alternatives at the end of this section. The dominance of the Bayesian methodology in cosmology sets it apart from the traditional practice of particle physicists, though there is now increasing interest in applying Bayesian methods in that context (e.g. Ref. [5]).

## 2.2 Bayesian inference

The Bayesian methodology goes all the way back to Thomas Bayes and his theorem, posthumously published in 1763 [6], followed soon after by pioneering work on probability by Laplace. The technical development of the inference system was largely carried out in the first half of the 20th century, with Jeffreys' textbook [7] the classic source. For several decades afterwards progress was held up due to an inability to carry out the necessary calculations, and only in the 1990s did use of the methodology become widespread with the advent of powerful multiprocessor computers and advanced calculational algorithms. Initial applications were largely in the fields of social science and analysis of medical data, the volume edited by Gilks et al. [8] being a particularly important source. The publication of several key textbooks in the early 21st century, by Jaynes [9], MacKay [10] and Gregory [11], the last of these being particularly useful for physical scientists seeking to apply the methods, cemented the move of such techniques to the mainstream. An interesting history of the development of Bayesianism is given in Ref. [12].

The essence of the Bayesian methodology is to assign probabilities to all quantities of interest, and to then manipulate those probabilities according to a series of rules, amongst which Bayes theorem is the most important. The aim is to update our knowledge in response to emerging data. An important implication of this set-up is that it requires us to specify what we thought we knew *before* the data was obtained, known as the prior probability. While all subsequent steps are algorithmic, the specification of the prior probability is not, and different researchers may well have different views on what is appropriate. This is often portrayed as a major drawback to the Bayesian approach. I prefer, however, to argue the opposite — that the freedom to choose priors is the opportunity to express physical insight. In any event, one needs to check that one's result is robust under reasonable changes to prior assumptions.

An important result worth bearing in mind is a theorem of Cox [13], showing that Bayesian inference is the unique consistent generalization of Boolean logic in the presence of uncertainty. Jaynes in particular sees this as central to the motivation for the Bayesian approach [9].

In abstract form, Bayes theorem can be written as

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}\,, \tag{1}$$

where a vertical line indicates the conditional probability, usually read as 'the probability of B given A'. Here $A$ and $B$ could be anything at all, but let's take $A$ to be the set of data $D$ and $B$ to be the parameter values $\theta$ (where $\theta$ is the $N$-dimensional vector of parameters being varied in the model under consideration), hence writing

$$P(\theta|D) = \frac{P(D|\theta)P(\theta)}{P(D)}\,. \tag{2}$$

In this expression, $P(\theta)$ is the prior probability, indicating what we thought the probability of different values of $\theta$ was before we employed the data D. One of our objectives is to use this equation to obtain the posterior probability of the parameters given the data, $P(\theta|D)$. This is achieved by computing the likelihood $P(D|\theta)$, often denoted $\mathcal{L}(\theta)$ with the dependence on the dataset left implicit.

### 2.3 Alternatives to Bayesian inference

The principal alternative to the Bayesian method is usually called the frequentist approach, indeed commonly a dichotomy is set up under which any non-Bayesian method is regarded as frequentist. The underpinning concept is that of sampling theory, which refers to the frequencies of outcomes in random repeatable experiments (often caricatured as picking coloured balls from urns). According to MacKay [10], the principal difference between the systems is that frequentists apply probabilities only to random variables, whereas Bayesians additionally use probabilities to describe inference. Frequentist analyses commonly feature the concepts of estimators of statistical quantities, designed to have particular sampling properties, and null hypotheses which are set up in the hope that data may exclude them (though without necessarily considering what the preferred alternative might be).

An advantage of frequentist methods is that they avoid the need to specify the prior probability distribution, upon which different researchers might disagree. Notwithstanding the Bayesian point-of-view that one should allow different researchers to disagree on the basis of prior belief, this means that the frequentist terminology can be very useful for expressing results in a prior-independent way, and this is the normal practice in particle physics.

A drawback of frequentist methods is that they do not normally distinguish the concept of a model with a fixed value of a parameter, versus a more general model where the parameter happens to take on that value (this is discussed in greater detail below in Section 4), and they find particular difficulties in comparing models which are not nested.

## 3 COSMOLOGICAL PARAMETER ESTIMATION

### 3.1 Goals and methodology

In cosmological parameter estimation, we take for granted that we have a dataset D, plus a model with parameter vector $\theta$ from which we can extract predictions for those data, in the form of the likelihood $\mathcal{L}(\theta) \equiv P(D|\theta)$. Additionally, we will have a prior distribution for those parameters, representing our knowledge before the data was acquired. While this could be a distribution acquired from analyzing some previous data, more commonly cosmologists take the prior distribution to be flat, with a certain range for each parameter, and reanalyze from scratch using a compilation of all data deemed to be useful.

Our aim is then to figure out the parameter values which give the best fit to the data, or, more usefully, the region in parameter space within which the fit is acceptable, i.e. to find the maximum likelihood value and explore the shape of the likelihood in the region around it. In many cases one can hope that the likelihood takes the form of a multi-variate gaussian, at least as long as one doesn't stray too far from the maximum.

The task then is to find the high-likelihood regions of the function $\mathcal{L}(\theta)$, which sounds straightforward. However, there are various obstacles

- The likelihood function may be extremely sharply peaked, and it may have several maxima masquerading as the true maximum.
- The parameter space may have a high dimensionality, cosmological examples often having 6 to 10 parameters independently varying.
- There may be parameter degeneracies, where likelihood varies only weakly, or not at all, along some direction in parameter space.
- The evaluations of the likelihood may be computationally demanding, either in generating the theoretical predictions from the model, or in computing the likelihood of those predictions. A typical likelihood evaluation in a cosmological calculation involving CMB anisotropies is a few seconds of CPU time.

In combination, these seriously obstructed early data analysis efforts, even when the dataset was fairly limited, because available computer power restricted researchers to perhaps $10^5$ to $10^6$ likelihood evaluations. Once beyond five or six parameters, which is really the minimum for an interesting comparison, brute-force mapping of the likelihood on a grid of parameters becomes inefficient, as the resolution in each parameter direction becomes too coarse, and anyway too high a fraction of computer time ends up being used in regions where the likelihood turns out to be too low to be of interest.

This changed with a paper by Christensen and Meyer [14], who pointed out that problems of this kind are best tackled by Monte Carlo methods, already extensively developed in the statistics literature, e.g. Ref. [8]. Subsequently, Lewis and Bridle wrote the CosmoMC package [15], implementing a class of Monte Carlo methods for cosmological parameter estimation. The code has been very widely adopted by researchers, and essentially all cosmological parameter estimation these days is done using one of a variety of Monte Carlo methods.

## 3.2 Monte Carlo methods

Monte Carlo methods are computational algorithms which rely on random sampling, with the algorithm being guided by some rules designed to give the desired outcome. An important subclass of Monte Carlo methods are Markov Chain Monte Carlo (MCMC) methods, defined as those in which the next 'step' in the sequence depends only upon the previous one. The sequence of steps is then known as a Markov chain. Each step corresponds to some particular value of the parameters, for which the likelihood is evaluated. The Markov chain can therefore be viewed as a series of steps (or jumps) around the parameter space, investigating the likelihood function shape as it goes.

You may find it convenient to visualize the likelihood surface as a mountainous landscape with one dominant peak. The simplest task that such a process could carry out would be to find the maximum of the likelihood: choose a random starting point, propose a random jump to a new point, accept the jump only if the new point has a higher likelihood, return to the proposal step and repeat until satisfied that the highest point has been found. Even this simple algorithm obviously needs some tuning: if the steps are too large, the algorithm may soon find it difficult to successfully find a higher likelihood point to jump to, whereas if they are small the chain may get stuck in a local maximum which is not the global maximum. That latter problem may perhaps be overcome by running a series of chains from different starting points.

Anyway, the maximum itself is not of great interest; what we want to know is the region around the maximum which is compatible with the data. To do this, we desire an algorithm in which the Markov chain elements correspond to random samples from the posterior parameter distribution of the parameters, i.e. that each chain element represents the probability that those particular parameter values are the true ones. The simplest algorithm which achieves this is the Metropolis–Hastings algorithm, which is a remarkably straightforward modification of the algorithm described in the previous paragraph.

### 3.2.1 Metropolis–Hastings algorithm

The Metropolis–Hastings algorithm is as follows:

1. Choose a starting point within the parameter space.

2. Propose a random jump. Any function can be used to determine the probability distribution for the length and direction of the jump, as long as it satisfies the 'detailed balance' condition that a jump back to the starting point is as probable as the jump away from it. This is most easily done using a symmetric proposal function, e.g. a multivariate gaussian about the current point. Evaluate the likelihood at the new point, and hence the probability by multiplying by the prior at that point. [If the prior is flat, the probability and likelihood become equivalent.]

3. If the probability at the new point is higher, accept the jump. If it is lower, we accept the jump with a probability given by the ratio of the probabilities at the new and old point. If the jump is not accepted, we stay at the same point, creating a duplicate in the chain.

4. Repeat from Step 2, until satisfied that the probability distribution is well mapped out. This may be done for instance by comparing several chains run from different starting points, and/or by using convergence statistics amongst which the Gelman–Rubin test [16, 8] is the most commonly used.

By introducing a chance of moving to a lower probability point, the algorithm can now explore the shape of the posterior in the vicinity of the maximum. The generic behaviour of the algorithm is to start in a low likelihood region, and migrate towards the high likelihood 'mountains'. Once near the top, most possible jumps are in the downwards direction, and the chain meanders around the maximum mapping out its shape. Accordingly, all the likelihood evaluations, which is where the CPU time is spent, are being carried out in the region where the likelihood is large enough to be interesting. The exception is the early stage, which is not representative of the posterior distribution as it maintains a memory of the starting position. This 'burn-in' phase is then deleted from the list of chain points.

Although any choice of proposal function satisfying detailed balance will ultimately yield a chain sampling from the posterior probability distribution, in practice, as with the simple example above, the algorithm needs to be tuned to work efficiently. This is referred to as the convergence of the chain (to the posterior probability). The proposal function should be tuned to the scale of variation of the likelihood near its maximum, and if the usual choice of a gaussian is made its axes should ideally be aligned to the principal directions of the posterior (so as to be able to navigate quickly along parameter degeneracies). Usually, a short initial run is carried out to roughly map out the posterior distribution which is then used to optimize the proposal function for the actual computation.[1] The resulting acceptance rate of new points tends to be around 25%.

The upshot of this procedure is a Markov chain, being a list of points in parameter space plus the likelihood/posterior probability at each point. A typical cosmological example may contain $10^4$ to $10^5$ chain elements, and some collaborations including WMAP make their chains public. There is usually some short-scale correlation of points along the chain due to the proposal function; some researchers 'thin' the chain by deleting elements to remove this correlation though this procedure appears unnecessary. By construction, the elements correspond to random samples from the posterior, and hence a plot of the point density maps it out.

The joy of having a chain is that marginalization (i.e. figuring out the allowed ranges of a subset of the full parameter set, perhaps just one parameter) becomes trivial; you just ignore the other

[1] One can also help things along by using variables which respect known parameter degeneracies in the data, e.g. that the CMB data is particularly good at constraining the angular-diameter distance to the last-scattering surface through the position of the peaks.

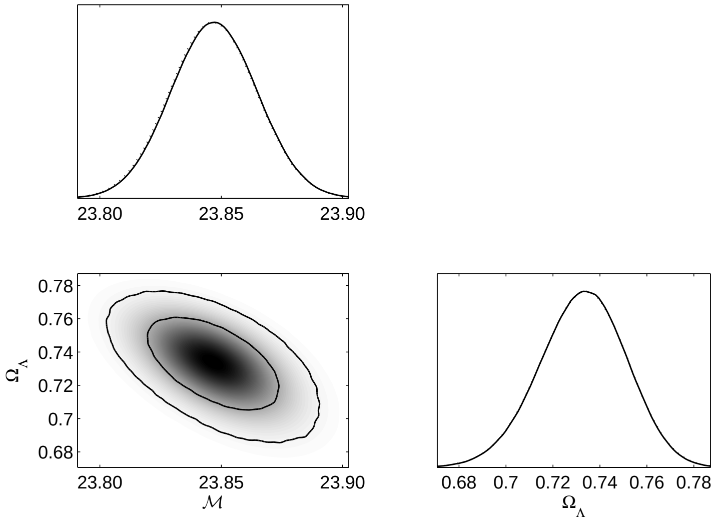

Figure 2: Output of a very simple two-parameter MCMC analysis. The data set is a combination of supernova, CMB, and galaxy correlation data, and the two parameters are the cosmological constant $\Omega_\Lambda$ and the standardized supernova brightness $\mathcal{M}$. The two-dimensional distribution is in the bottom left and the marginalized single-parameter constraints in the other two boxes. In this case, both parameters have accurately gaussian distributions and are correlated. [Figure reproduced from Ref. [17].]

parameters and plot the point density of the one you are interested in. By contrast, a grid evaluation of the marginalized posterior requires an integration over the uninteresting directions.

Figure 2 shows a typical outcome of a simple MCMC calculation.

In mapping the posterior using the point density of the chains, one in fact ignores the actual values of the probability in the chains, since the chain point locations themselves encode the posterior. However, one can also plot the posterior by computing the average likelihood in a parameter-space bin (marginalized or not), which should of course agree. Whether it does or not is an additional guide to whether the chain is properly converged.

In addition to analyzing the posterior probability from the chains, both to plot the outcome and extract confidence levels to quote as headline results, there are a number of other ways of using them. Two examples are

- Importance sampling. If new data becomes available, rather than computing new chains from scratch one can importance sample existing chains, by reweighting the elements according to the new likelihood [15]. One can also use importance sampling to study the effect of varying the prior. These operations mean that the points now have non-integer weights, but this creates no new issue of principle. However importance sampling may result in an insufficient density of points in the new preferred region, making the sampling of the posterior noisier than would be possible with new chains.

- Bayesian complexity. This quantity measures the number of parameters actually constrained by the data in hand [18, 19], which may be less than the number of parameters of the model, either because overall the data are poorly constraining, or because of specific parameter degeneracies leaving some parameter combinations unconstrained. It can also be used to determine a semi-Bayesian model selection statistic known as the Deviance Information Criterion, discussed briefly in Section 4.5

### 3.2.2 Other sampling algorithms

Metropolis–Hastings achieves the desired goal, but may struggle to do so efficiently if it is difficult to find a good proposal function or if the assumption of a fixed proposal function proves disadvantageous. This has tended not to be a problem in cosmological applications to date, but it is nevertheless worthwhile to know that there are alternatives which may be more robust. Some examples, all discussed in MacKay's book [10], are

- Slice sampling: This method allows the proposal function to change during the calculation, tuning itself to an appropriate scale, though there is an additional computational cost associated with enforcing the detailed balance condition. The steps are made in a single parameter direction at a time, hence the name, and cycle through the parameter directions either sequentially or randomly. Slice sampling is implemented in CosmoMC as an alternative to Metropolis–Hastings, as are some other more specialized sampling algorithms.

- Gibbs sampling: This relies on obtaining a proposed step by sampling from conditional probability distributions, e.g. to step in the $\theta_1$ direction we sample from $P(\theta_1|\theta_2)$ and vice versa. It turns out that such proposals are always accepted, enhancing the efficiency. The method can however struggle to make progress along highly correlated parameter directions, traversing the diagonal through a series of short steps parallel to the axes.

- Hamiltonian sampling: This more sophisticated approach uses an analogy with Hamiltonian dynamics to define a momentum from derivatives of the likelihood. The momentum associated with a point enables large proposal steps to be taken along trajectories of constant 'energy' and is particularly well adapted to very high dimensionality problems. See Refs. [20, 21] for cosmological applications.

### 3.2.3 Machine learning

The slow likelihood evaluations, stemming mainly from the time needed to predict observational quantities such as the CMB power spectra from the models, remain a significant stumbling block in cosmological studies. One way around this may be to use machine learning to derive accurate theoretical predictions from a training set, rather than carry out rigorous calculations of the physical equations at each parameter point. Two recent attempts to do this are PICO [22] and CosmoNet [23], the former also allowing direct estimation of the WMAP likelihood from the training set. This is a promising method, though validation of the learning output when new physical processes are included may still mean that many physics-based calculations need to be done.

### 3.2.4 Overview

In conclusion, cosmological parameter estimation from data is increasingly regarded as a routine task, albeit one that requires access to significant amounts of computing resource. This is principally attributed to the public availability of the CosmoMC package, and the continued work by its authors and others to enhance its capabilities. On the theoretical side, an impressive range of physical assumptions are supported, and many researchers have the expertise to modify the code to further broaden its range. On the observational side, researchers have recognized the importance of making their data available in formats which can readily be ingested into MCMC analyses. The WMAP team were the first to take this aspect of delivery very seriously, by publically releasing the 'WMAP likelihood code' at the same time as their science papers, allowing cosmologists everywhere to immediately explore the consequences of the data for their favourite models. Indeed, I would argue that by now the single most important product of an observational programme would be provision of a piece of software calculating the likelihood as a function of input quantities (e.g. the CMB power spectra) computable by CosmoMC.

## 3.3 Uncertainties, biases and significance

The significance with which results are endowed appears to be strongly dependent on the science community obtaining them. At one extreme lies particle physics experimentalists, who commonly set a 'five-sigma' threshold to claim a detection (in principle, for gaussian uncertainties, corresponding to 99.99994% probability that the result is not a statistical fluke). By contrast, astrophysicists have been known to get excited by results at or around the 'two-sigma' level, corresponding to 95.4%

confidence for a gaussian. At the same time, there is clearly a lot of skepticism as to the accuracy of confidence limits; some possibly apocryphal quotes circulating in the data analysis community include “Once you get to three-sigma you have about a half chance of being right.” and “95% of 95% confidence results do not turn out to be right; if anything 95% of them turn out to be wrong”.

There are certainly good reasons to think that results are less secure than the stated confidence from a likelihood analysis would imply. Amongst these are

- In realistic cases, the probability may not fall as fast as a gaussian in the tails of the distribution even if accurately gaussian near the peak.
- There may be unmodelled systematic errors. The natural trend in a maturing observational field is to be initially dominated by statistical uncertainty, but as instrumental accuracy improves to then reach a systematic floor where it becomes difficult or impossible to model extraneous physical effects (e.g. population evolution in supernovae which one is planning to use as standard candles over cosmic epochs).
- The likelihood function may be uncertain in a way not included in the quoted uncertainty. For instance there are now several different treatments of the WMAP likelihood, differing in the way the beam profiles, source subtraction, or low multipole likelihoods are calculated.
- The researchers may, consciously or otherwise, have adopted a model motivated by having seen the data, and then attempted to verify it from the same data. It has recently been claimed that this problem is widespread in the neuroscience field [24], with the authors of that study suspecting the issue extends to other disciplines. An example of this would be to spot an unusual feature in, say, a cosmic microwave background map, and then attempt to assess the probability of such a feature using Monte Carlo simulations. This ignores the fact that there may have been any number of other no more unlikely features, that *weren't seen in the data*. Consider the well-publicized appearance of Stephen Hawking's initials in the WMAP maps, obviously massively improbable *a priori* but nonetheless visible in Fig. 1 (in blue, just above the middle axis, somewhat left of center).
- Publication bias: positive results are more likely to get published than negative ones, e.g. the 95% confidence result you just read about arose from one of 20 studies, the other 19 of which generated null results. This is widely recognized in the medical statistics community,[2] leading to introduction of costly treatments that may be ineffective or even harmful. In this context, the additional problem may exist that trials are funded by large companies whose profitability depends on the outcome, and who may be in a position to influence whether they are published. In cosmology, this may be a particular problem for cosmic non-gaussianity studies, where many different independent tests can be carried out.
- Model uncertainty: the possibility of different models, rather than just parameter values, describing the data has not be consistently allowed for.

I'm aware of situations where all of these have been important, and in my view 95% confidence is not sufficient to indicate a robust result. This appears to increasingly be the consensus in the cosmology community, perhaps because too much emphasis was put on two 95% confidence level results in the first-year WMAP data (a high optical depth and running of the spectral index) which lost support in subsequent data releases.

On the other hand, ‘five-sigma’ may be rather too conservative, intended to give a robust result in all circumstances. In reality, the confidence level at which a result becomes significant should depend on the nature of the model(s) under consideration, and the nature of the data obtained. Some guidance on where to draw the line may be obtained by exploring the issue of model uncertainty, the topic of the next section.

[2] Ioannidis [25] goes so far as to claim a proof that most published results are false, publication bias being partly responsible.

# 4 COSMOLOGICAL MODEL SELECTION

## 4.1 Model selection versus parameter estimation

Estimation of cosmological parameters, as described in the previous section, assumes that we have a particular model in mind to explain the data. More commonly, however, there tends to be competing models available to describe the data, invoking parametrizations of different physical effects. Each model corresponds to a different choice of variable parameters, accompanied by a prior distribution for those parameters. Indeed, the most interesting questions in cosmology tend to be those about models, because those are the qualitative questions. Is the Universe spatially flat or not? Does the dark energy density evolve? Do gravitational waves contribute to CMB anisotropies? We therefore need techniques not just for estimating parameters within a model, but also for using data to discriminate between models. Bayesian tools are particularly appropriate for such a task, though I also describe some non-Bayesian alternatives at the end of this section. A comprehensive review of Bayesian model selection as applied to cosmology was recently given by Trotta [26].

An important implication of model-level Bayesian analysis is that there is a clear distinction between a model where a quantity is fixed to a definite value, versus a more general model where that parameter is allowed to vary but happens to take on that special value. A cosmological example is the dark energy equation of state, $w$, which the cosmological constant model predicts (in good agreement with current observations) to be precisely $-1$, and which other models such as quintessence leave as a free parameter to be fit from data. Even if it is the cosmological constant which is the true underlying model, a model in which the equation of state can vary will be able to fit any conceivable data just as well as the cosmological constant model (assuming of course that its range of variation includes $w = -1$). What distinguishes the models is predictiveness.

As an example, consider a magician inviting you to “pick a card, any card”. Since you don’t know any better, your ‘model’ tells you that are equally likely to pick any card. The one you actually do pick is not particularly surprising, in the sense that whatever it was, it was compatible with your model. The magician, however, has a much more predictive model; by whatever means, they know that you will end up picking the queen of clubs. Both models are equally capable of explaining the observed card, but the magician’s much more predictive model lets them earn a living from it. Note in this example, any surprise that you might feel comes not from the card itself, but from the magician’s ability to predict it. Likewise, a scientist might find themselves surprised, or dismayed, as incoming data continue to lie precisely where some rival’s model said they would.

Model selection/comparison is achieved by choosing a ranking statistic which can be computed for each model, allowing them to be placed in rank order. Within the Bayesian context, where everything is governed by probabilities, the natural choice is the model probability, which has the advantage of having a straightforward interpretation.

## 4.2 The Bayesian evidence

The extension of the Bayesian methodology to the level of models is both unique and straightforward, and exploits the normalizing factor $P(D)$ in equation (2) which is irrelevant to and commonly ignored in parameter estimation.

We now assume that there are several models on the table, each with their own probability $P(M_i)$ and explicitly acknowledge that our probabilities are conditional not just on the data but on our assumed model $M$, writing

$$P(\theta|D, M) = \frac{P(D|\theta, M)P(\theta|M)}{P(D|M)} \,. \qquad (3)$$

This is just the previous equation with a condition on M written in each term. The denominator, the probability of the data given the model, is by definition the model likelihood, also known as the Bayesian evidence. Note that, unlike the other terms in this equation, it does not depend on specific values for the parameters $\theta$ of the model.

The evidence is key because it appears in yet another rewriting of Bayes theorem, this time as

$$P(M|D) = \frac{P(D|M)P(M)}{P(D)} \,. \qquad (4)$$

The left-hand side is the posterior model probability (i.e. the probability of the model given the data), which is just what we want for model selection. To determine it, we need to compute the Bayesian evidence $P(D|M)$, and we need to specify the prior model probability $P(M)$. It is a common convention to take the prior model probabilities to be equal (the model equivalent of a flat parameter prior), but this is by no means essential.

To obtain an expression for the evidence, consider Eq. (3) integrated over all $\theta$. Presuming we have been careful to keep our probabilities normalized, the left-hand side integrates to unity, while the evidence on the denominator is independent of $\theta$ and comes out of the integral. Hence

$$P(D|M) = \int P(D|\theta, M)P(\theta|M)d\theta \,, \tag{5}$$

or, more colloquially,

$$\text{Evidence} = \int (\text{Likelihood } \times \text{ Prior})\; d\theta \,. \tag{6}$$

In words, the evidence $E$ is the average likelihood of the parameters averaged over the parameter prior. For the distribution of parameter values you thought reasonable before the data came along, it is the average value of the likelihood.

The Bayesian evidence rewards model predictiveness. For a model to be predictive, observational quantities derived from it should not depend very strongly on the model parameters. That being the case, if it fits the actual data well for a particular choice of parameters, it can be expected to fit fairly well across a significant fraction of its prior parameter range, leading to a high average likelihood. An unpredictive model, by contrast, might fit the actual data well in some part of its parameter space, but because other regions of parameter space make very different predictions it will fit poorly there, pulling the average down. Finally, a model, predictive or otherwise, that cannot fit the data well anywhere in this parameter space will necessarily get a poor evidence.

Often predictiveness is closely related to model simplicity; typically the fewer parameters a model has, the less variety of predictions it can make. Consequently, model selection is often portrayed as tensioning goodness of fit against the number of model parameters, the latter being thought of as an implementation of Ockham's razor. However the connection between predictiveness and simplicity is not always a tight one. Consider for example a situation where the predictions turn out to have negligible dependence on one of the parameters (or a degenerate combination of parameters). This is telling us that our observations lack the sensitivity to tell us anything about that parameter (or parameter combination). The likelihood will be flat in that parameter direction and it will factorize out of the evidence integral, leaving it unchanged. Hence the evidence will not penalize the extra parameter in this case, because it does not change the model predictiveness.

The ratio of the evidences of two models $M_0$ and $M_1$ is known as the Bayes factor [27]:

$$B_{01} \equiv \frac{E_0}{E_1} \,, \tag{7}$$

which updates the prior model probability ratio to the posterior one. Some calculational methods determine the Bayes factor of two models directly. Usual convention is to specify the logarithms of the evidence and Bayes factor.

## 4.3 Calculational methods

Equation (5) tells us that to get the evidence for a model, we need to integrate the likelihood throughout the parameter space. In principle this is a very standard mathematical problem, but it is made difficult because the integrand is likely to be extremely highly peaked and we do not know in advance where in parameter space the peak might be. Further, the parameter space is multi-dimensional (between about 6 and 10 dimensions would be common in cosmological applications), and as remarked in Section 3 the individual likelihood evaluations of the integrand at a point in parameter space are computationally expensive (a few CPU seconds each), limiting practical calculations to $10^5$ to $10^6$ evaluations. Successful Bayesian model selection algorithms are therefore dependent on efficient algorithms for tackling this type of integral.

Model probabilities are meaningful in themselves and don't require further interpretation, but it is useful to have a scale by which to judge differences in evidence. The usual scale employed is the Jeffreys' scale [7] which, given a difference $\Delta \ln E$ between the evidences $E$ of two models, reads

| | |
|---|---|
| $\Delta \ln E < 1$ | Not worth more than a bare mention. |
| $1 < \Delta \ln E < 2.5$ | Significant. |
| $2.5 < \Delta \ln E < 5$ | Strong to very strong. |
| $5 < \Delta \ln E$ | Decisive. |

In practice the divisions at 2.5 (corresponding to posterior odds of about 13:1) and 5 (corresponding to posterior odds of about 150:1) are the most useful.

As the main steps in the Jeffreys' scale are 2.5, this sets a target accuracy in an evidence calculation; it should be accurate enough that calculational uncertainties do not move us amongst these different categories. The accuracy should therefore be better than about $\pm 1$. However there is no point in aiming for an accuracy very much better than that, which would not tell us anything further about the relative merits of the models. Hence an accuracy of a few tenths in $\ln E$ is usually a good target.

### 4.3.1 Exact methods

Computing the evidence is more demanding than mapping the dominant part of the posterior, as we now have to be able to accurately integrate the likelihood over the entire parameter space. There are several methods which become exact in the limit of infinite computer time, and are capable of offering the desired accuracy with finite resources. Those used so far in cosmological settings are

- Thermodynamic integration. This variant on Metropolis–Hastings MCMC varies the effective temperature of the chain, allowing it to explore lower likelihood regions and hence fully probe the prior space. While well regarded in the statistics community, cosmological applications to date have proven extremely CPU demanding, e.g. Ref. [28].

- Nested sampling. Introduced by Skilling [29], this algorithm explores parameter space with a large collection of points, typically hundreds. The lowest likelihood point is deleted and replaced by a randomly-drawn point of higher likelihood, the cluster of points in this way migrating to the high likelihood regions as shown in Figure 3. As a byproduct it generates a set of posterior samples that are suitable for parameter estimation. It was first implemented for cosmology in Refs. [30, 31], though effectively limited to cases with a single strong likelihood peak. A more powerful implementation, MultiNest [32] is capable of handling multi-peaked likelihoods. Nested sampling is currently the method of choice in cosmology and has also been applied to supersymmetry [33].

- VEGAS. This is a multi-dimensional integrator much used by particle physicists [34]. It has only been used once in cosmological model selection thus far [35] but shows promise as an alternative to nested sampling.

### 4.3.2 Approximate or restricted methods

In addition to brute force numerical methods to compute the evidence, various approximate or restricted methods exist which may be useful in some circumstances. The most important three, discussed in more detail in Refs. [26, 3], are

1. The Laplace approximation: one assumes that the likelihood is adequately described by a multi-variate gaussian, expands around the maximum and carries out the evidence integral analytically. This may work well at least in the absence of strong degeneracies, but it can be hard to assess its accuracy.

2. The Savage–Dickey ratio: This computes the Bayes factor for the case where one model is nested within another. It amounts to a careful evaluation of the marginalized posterior of the more complex model, evaluated at the parameter value of the embedded model. It is in

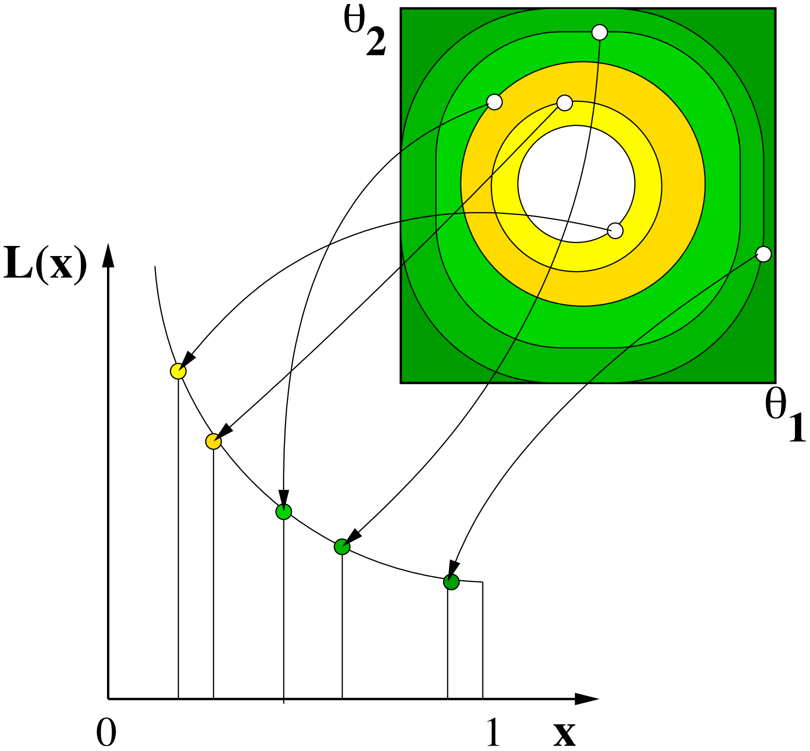

Figure 3: A schematic of the nested sampling algorithm. The two-dimensional parameter space is shown at the top right. The points within it are considered to represent contours of constant likelihood, which sit within each other like layers of an onion (there is however no need for them to be simply connected). The volume corresponding to each thin shell of likelihood is computed by the algorithm, allowing the integral for the evidence to be accumulated as shown in the graph. [Figure courtesy David Parkinson.]

principle exact and can be estimated from Markov chains. However, if the embedded model is not near the actual maximum likelihood, so that one is trying to judge whether the simple model is excluded (likely to be the case most of interest), there may be few Markov chain samples near the embedded model and the estimate becomes noisy. For this reason a general implementation of the Savage–Dickey method has not yet been made available, though it has been successfully applied in specific examples, e.g. Ref. [36].

3. The Bayesian Information Criterion (BIC): Introduced by Schwarz [37], this approximates the Bayes factor without requiring the models to be nested, using a Laplace approximation and further assumptions on the nature of the data. The BIC is given by

$$\mathrm{BIC} = -2\ln\mathcal{L}_{\mathrm{max}} + k\ln N\,, \tag{8}$$

where $\mathcal{L}_{\mathrm{max}}$ is the maximum likelihood achievable by the model, $k$ the number of parameters, and $N$ the number of datapoints. Models are ranked according to their BIC values. Despite the name, it has nothing to do with information theory, but rather was named by analogy to information theory measures discussed in Section 4.5. It has been used quite widely in cosmology, but the validity of the approximation needs careful consideration and full evidence evaluation is to be preferred whenever possible.

## 4.4 Multi-model inference

Ultimately, a successful implementation of model selection should result in all models being eliminated except for a single survivor. This model will encode all the physics relevant to our observations, and we can then proceed to estimate the parameters of the surviving model to yield definitive values of quantities of interest, e.g. the baryon density, neutrino masses, etc. Unfortunately, in practice it is quite likely that more than one model might survive in the face of the data, as has always been the case so far in any analysis I've done. Despite that, we might well want to know what the best current knowledge of some of the parameters is.

Fortunately, the Bayesian framework again supplies a unique procedure for extracting parameter constraints in the presence of model uncertainty. One simply combines the probability distribution of the parameters within each individual model, weighted by the model probability. This is very much analogous to quantum mechanics, the superposition of model states being equivalent to a superposition of eigenstates of an observable such as position. The combination is particularly straightforward if one already has a set of samples from the likelihood, such as a Markov chain, for each model; we then just concatenate the chains giving each point a fractional weighting according to the model probability of the chain it came from.

Note that in this superposition, some parameters may have fixed values in some of the models. For example, if we are interested in the dark energy equation of state $w$, we will no doubt be including a cosmological constant model in which $w$ is fixed to $-1$. Within this model, the probability distribution is an appropriately normalized delta function, and the combined probability distribution is therefore neither continuous nor differentiable. This can have the interesting consequence that some confidence limits may have zero parameter width. See Ref. [38] for some actual calculations.

## 4.5 Other approaches to model selection

Within the Bayesian framework, the evidence is the natural model selection statistic. However, other paradigms suggest alternatives. The most prevalent is information theory approaches, where the statistics are known as *information criteria.* The first and most widespread is the Akaike Information Criterion (AIC), defined by [39]

$$\mathrm{AIC} = -2\ln\mathcal{L}_{\mathrm{max}} + 2k\,, \tag{9}$$

using the same terminology as equation (8). This sets up a tension between goodness of fit and complexity of the model, with the best model minimizing the AIC. There is also a slightly modified version applicable to small sample sizes, which ought to be used in any case [40]. It is not always clear how big a difference in AIC is required for the worse model to be significantly disfavoured, but a typical guide (actually obtained by attempting a Bayesian-style probabilistic interpretation) is a difference of 6 or more should be taken seriously.

Beyond the AIC, there are a large variety of different information criteria in use. Burham and Anderson [40] give an excellent textbook account of information theory approaches to model selection, and cosmological applications of some of them are discussed in Refs. [3, 41].

Perhaps most worthy of special mention is the Deviance Information Criterion of Spiegelhalter et al. [18], which combines aspects of the Bayesian and information theory approaches. It replaces the $k$ in equation (9) with the Bayesian complexity mentioned towards the end of Section 3.2 (for technical reasons it also replaces the maximum likelihood with the likelihood at the mean parameter values, but this is not so significant). The complexity measures the number of model parameters actually constrained by the data. In doing so it overcomes the main difference between the Bayesian and Information Criterion approaches, which is their handling of parameter degeneracies. The Information Criterion approach penalizes such parameters, but the Bayesian method does not as the integral of the likelihood over the degenerate parameter factors out of the evidence integral. The Bayesian view is that in such cases the more complex model should not be penalized, because the data are simply not good enough to say whether the degenerate parameter is needed or not. As this seems more reasonable, the DIC may be preferable to the AIC.

An alternative model selection approach comes from algorithmic information theory, and has variants known as minimum message length and minimum description length [42]. These interpret the best model as being the one which offers the maximal compression of the data, i.e. that the combination of model plus data can be described in the smallest number of bits. As a concept, this idea remarkably originates with Leibniz in the 17th century.

# 5 FORECASTING AND EXPERIMENTAL DESIGN

The above discussion concerned analysis of data which was in hand. Another important application of statistical techniques is to forecast the outcomes of future experiments, which is increasingly expected by funding agencies wishing to compare the capabilities of competing proposals. A Figure of Merit (FoM) is defined for each experiment, and used to rank experiments. More ambitiously, the same techniques can be used to optimize the design of a survey in order to maximize the likely science return (e.g. Ref. [43]).

## 5.1 Fisher matrix approaches

The leading approach presently is the Fisher information matrix approach, introduced to cosmology in Ref. [44] and popularized especially when adopted by the initial report of the Dark Energy Task Force [45]. The Fisher matrix measures the second derivatives of the likelihood around its maximum, and in a gaussian approximation is used to estimate the expected uncertainty in parameters around some selected fiducial model. For instance, the DETF FoM considers the uncertainty on the two parameters of the dark energy equation of state defined by the form $w = w_0 + (1-a)w_a$, taking the fiducial model to be ΛCDM ($w_0 = -1$, $w_a = 0$), defining the FoM to be the inverse area of the 95% confidence region. The likelihood is determined via a model of how well the instrument will perform, possibly through analysis of a simulated data stream.

An important caveat is the following: an experiment capable of reducing the volume of permitted parameter space by say a factor of 10 should in no way be considered as having a 90% chance of measuring those parameters as different from some special value. An example is whether upcoming data can exclude the cosmological constant model in favour of dark energy. What the Fisher FoM shows is the reduction in allowed parameter volume *provided that the dark energy model is correct.* However that assumption is actually what we wish to test; it is perfectly possible that it is the cosmological constant model which is correct and then no amount of improvement to the uncertainty will rule it out. Once again, to develop a full picture we need to consider multiple models, and hence model selection statistics.

## 5.2 Model selection approaches

Model selection forecasting, pioneered for cosmology in Refs. [36, 46], instead forecasts the ability of upcoming experiments to carry out model comparisons. Let's restrict to the simplest case of two models, one nested within another, though the generalization to other circumstances is straightforward.

The very simplest model selection question is to assume that the nested model (say ΛCDM) is true, and ask whether a given experiment would be able to rule out the more complex alternative. Data is simulated only for the ΛCDM model, as with the Fisher analysis, and the Bayes factor between ΛCDM and the parametrized dark energy model calculated. This shows how strongly the experiment will rule out the dark energy model if it is wrong.

More generally, one may wish to consider the dark energy model as the correct one, and ask whether the simpler model can be excluded. This is more complex, as the outcome depends on the actual parameters of the dark energy model (known as the fiducial parameters), and hence has to be considered as a function of them. This leads to the concept of the Bayes factor plot, showing the expected Bayes factor as a function of fiducial parameters [46]. A suitable FoM may be to minimize the area in fiducial parameter space in which the wrong model cannot be ruled out by the proposed experiment. Alternatively, one can study the distribution of the Bayes factor weighted by present knowledge of the parameters, to predict the probability distribution of expected outcomes of the experiment [36].

Examples of cosmological model selection forecasts can be found in Refs. [36, 46, 38].

# 6 THE END

I was going to write a brief summary. Decided not to.

# Acknowledgments

This work was supported by STFC (UK). I thank my collaborators on these topics — Maria Beltran, Pier Stefano Corasaniti, Marina Cortês, Juan Garcia-Bellido, Martin Kunz, Ofer Lahav, Samuel Leach, Julien Lesgourgues, Pia Mukherjee, Cédric Pahud, David Parkinson, Martin Sahlén, Anze Slosar, Roberto Trotta, and Yun Wang — for numerous discussions, with special appreciation for Pia and David's work with me on the nested sampling method. I also thank Sarah Bridle, George Efstathiou, Alan Heavens, Mike Hobson, Andrew Jaffe, Eric Linder, and John Peacock for many debates concerning Bayesian data analysis.

# References

[1] Dunkley J et al. 2008. Five-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: likelihoods and parameters from the WMAP data. arXiv:0803.0586;
Komatsu E et al. 2008. Five-year Wilkinson Microwave Anisotropy Probe (WMAP) observations: cosmological interpretation. arXiv:0803.0547

[2] Liddle AR, Lyth DH. 2000. *Cosmological inflation and large-scale structure*. Cambridge University Press

[3] Liddle AR. 2004. How many cosmological parameters? *Mon. Not. Roy. Astr. Soc.* 351: L49–L53

[4] Lahav O, Liddle AR. 2008. The Cosmological Parameters. In Amsler C et al. 2008. Reviews of Particle Physics, http://pdg.lbl.gov/

[5] See e.g. the Mathematical Methods articles by Cowan G in Amsler C et al. 2008. Reviews of Particle Physics, http://pdg.lbl.gov/

[6] Bayes T. 1763. An essay towards solving a problem in the doctrine of chances. *Phil. Trans.* 53:370–418 (translated to modern notation in *Biometrika* 45: 296–315)

[7] Jeffreys H. 1961. *Theory of probability*, 3rd ed. Oxford University Press

[8] Gilks WR, Richardson S, Spiegelhalter DJ (eds). 1996. *Markov Chain Monte Carlo in practice.* Chapman and Hall/CRC (London)

[9] Jaynes ET. 2003. *Probability theory: the logic of science.* Cambridge University Press

[10] MacKay DJC. 2003. *Information theory, inference, and learning algorithms.* Cambridge University Press

[11] Gregory P. 2005. *Bayesian logical data analysis for the physical sciences.* Cambridge University Press

[12] Feinberg SE. 1996. When did Bayesian inference become "Bayesian"? *Bayesian Analysis* 1:1–40

[13] Cox RT. 1946. Probability, frequency, and reasonable expectation. *Am. Jour. Phys.* 14:1–13

[14] Christensen N, Meyer R. 2000. Bayesian methods for cosmological parameter estimation from cosmic microwave measurements. arXiv:astro-ph/0006401. Extended and published under the same title, with additional authors Knox L and Luey B, 2001. *Class. Quant. Grav.* 18:2677–2688

[15] Lewis A, Bridle S. 2002. Cosmological parameters from CMB and other data: a Monte Carlo approach. *Phys. Rev. D* 66:103511, code at http://cosmologist.info

[16] Gelman A, Rubin D. 1992. Inference from iterative simulation using multiple sequences. *Statistical Science* 7:457–511

[17] Sahlén M, Liddle AR, Parkinson D. 2007. Quintessence reconstructed: new constraints and tracker viability. *Phys. Rev. D* 75:023502

[18] Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. 2002. Bayesian measures of model complexity and fit. *J. Roy. Statist. Soc. B* 64:583–640

[19] Kunz M, Trotta R, Parkinson D. 2006. Measuring the effective complexity of cosmological models. 2006. *Phys. Rev. D* 74:023503

[20] Haijin A. 2007. Efficient cosmological parameter estimation with Hamiltonian Monte Carlo. *Phys. Rev. D* 75:083525

[21] Taylor JF, Ashdown MAJ, Hobson MP. 2007. Fast optimal CMB power spectrum estimation with Hamiltonian sampling. arXiv:0708.2989

[22] Fendt WA, Wandelt BD. 2007. Computing high accuracy power spectra with Pico. arXiv:0712.0194, code at http://cosmos.astro.uiuc.edu/pico

[23] Auld T, Bridges M, Hobson MP. 2007. CosmoNet: fast cosmological parameter estimation in non-flat models using neural nets. arXiv:astro-ph/0703445, code at http://www.mrao.cam.ac.uk/software/cosmonet

[24] Vul E, Harris C, Winkielman P, Pashler H. 2009. Voodoo correlations in social neuroscience. In press, *Perspectives in psychological science*

[25] Ioannidis JPD. 2005. Why most published research fundings are false. *PLoS Medicine* 2(8):e124

[26] Trotta R. 2008. Bayes in the sky: Bayesian inference and model selection in cosmology. *Contemp. Phys.* 49(2): 71–104

[27] Kass RE, Raftery AE. 1995. Bayes factors. *J. Am. Stat. Assoc.* 90:773–795

[28] Beltran M, Garcia-Bellido J, Lesgourgues J, Liddle AR, Slosar A. 2005. Bayesian model selection and isocurvature perturbations. *Phys. Rev. D* 71:063532

[29] Skilling J. 2006. Nested sampling for general Bayesian computation. *Bayesian Anal.* 1:833–860

[30] Mukherjee P, Parkinson D, Liddle AR. 2006. A nested sampling algorithm for cosmological model selection. *Astrophys. J. Lett.* 638:L51–L55

[31] Parkinson D, Mukherjee P, Liddle AR. 2006. A Bayesian model selection analysis of WMAP3. *Phys. Rev. D* 73: 123523, code at http://cosmonest.org

[32] Feroz F, Hobson MP, Bridges M. 2008. MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics. arXiv:0809.3437, code at http://www.mrao.cam.ac.uk/software/multinest

[33] Trotta R, Feroz F, Hobson MP, Roszkowski L, de Austri RR. 2008. The impact of priors and observables on parameter inferences in the constrained MSSM. *JHEP* 0812:024

[34] Lepage GP. 1978. A new algorithm for adaptive multidimensional integration. *J. Comput. Phys.* 27:192–203

[35] Serra P, Heavens A, Melchiorri A. 2007. Bayesian evidence for a cosmological constant using new high-redshift supernovae data. *Mon. Not. Roy. Astr. Soc.* 379:169–175

[36] Trotta R. 2007. Applications of Bayesian model selection to cosmological parameters. *Mon. Not. Roy. Astr. Soc.* 378:72–82

[37] Schwarz G. 1978. Estimating the dimension of a model. Ann. Statist. 5:461–464

[38] Liddle AR, Mukherjee P, Parkinson D, Wang Y. 2006. Present and future evidence for evolving dark energy. *Phys. Rev. D* 74:123506

[39] Akaike H. 1974. A new look at the statistical model identification. *IEEE T. Automat. Contr.* 19, 716–723

[40] Burnham KP, Anderson DR. 2002. *Model selection and multimodel inference*, 2nd ed., Springer–Verlag (New York)

[41] Liddle AR. 2007. Information criteria for astrophysical model selection. *Mon. Not. Roy. Astr. Soc.* 377: L74–L78

[42] Wallace CS. 2005. *Statistical and inductive inference by minimum message length*, Springer

[43] Bassett BA. 2005. Eyes wide open — optimizing cosmological surveys in a crowded market. *Phys. Rev. D* 71:083517
Parkinson D, Blake C, Kunz M, Bassett BA, Nichol RC, Glazebrook K. 2007. Optimizing baryon acoustic oscillation surveys I: testing the concordance LCDM cosmology. *Mon. Not. Roy. Astr. Soc.* 377:185–197

[44] Knox L. 1995. Determination of inflationary observables by cosmic microwave background anisotropy experiments. *Phys. Rev. D* 52:4307–4318
Jungman G, Kamionkowski M, Kosowsky A, Spergel DN. 1996. Cosmological parameter determination with microwave background maps. *Phys. Rev. D* 54: 1332–1344
Zaldarriaga M, Spergel D, Seljak U. 1997. Microwave background constraints on cosmological parameters. *Astrophys. J.* 488:1–13

[45] Albrecht A et al. 2006. Report of the dark energy task force. arXiv:astro-ph/0609591

[46] Mukherjee P, Parkinson D, Corasaniti PS, Liddle AR, Kunz M. 2006. Model selection as a science driver for dark energy surveys. *Mon. Not. Roy. Astr. Soc.* 369:1725–1734

# LIST OF ACRONYMS

| | |
|---|---|
| AIC | Akaike information criterion |
| BIC | Bayesian information criterion |
| CDM | Cold dark matter |
| CMB | Cosmic microwave background |
| CosmoMC | [A software package for MCMC analysis] |
| DETF | Dark energy task force |
| FoM | Figure of merit |
| ΛCDM | Lambda cold dark matter (model) |
| MCMC | Markov Chain Monte Carlo |
| WMAP | Wilkinson Microwave Anisotropy Probe |