


STATISTICAL METHODS IN PRECISION COSMOLOGY

(NOTES FOR LECTURE 2)

WILLIAM GIARÈ 

School of Mathematics & Statistics, The University of sheffield
Hicks Building · Hounsfield Road · Sheffield S3 7RH

✉ w.giare@sheffield.ac.uk |  www.williamgiare.com

– Abstract –

In these brief notes, we review the main statistical methods employed to produce observational constraints on cosmological parameters.

1 INTRODUCTION

Cosmology is characterized by a strong synergy between the theoretical formulation and numerical analysis of models, and the planning of observational probes, as well as the collection and treatment of relevant data. With new theories being proposed every day, data analysis in cosmology has become essential to test and even falsify their statistical support, with many subtleties at play (theoretical biases, model independence of the analysis, overfitting due to an excess of parameters, degeneracies in the parameter space, etc.).

Ultimately, there are two radically different approaches to statistics:

- The **frequentist** or classical notion of probability: related to the frequency of particular outcomes for a large (ideally infinite) series of trials. Relies on repeatability of the experiment.
- The **Bayesian** approach: probability is based on the knowledge available prior to the experiment, encoding a notion of confidence or belief in particular conditions. Relies on confidence of information relevant for the experiment.

In cosmological data analysis, for obvious reasons, we are forced to follow the Bayesian approach.

To provide a quick and dirty example, in the context of cosmology, the Bayesian approach would be to ask: *given* the Planck CMB data, what is the probability that the density parameter of cold dark matter is found to be between 0.3 and 0.4? More quantitatively, the inferential use of the Bayes theorem for a model M with parameters θ fitting some data distribution d is expressed as:

$$p(\theta | d) = \frac{p(d | \theta) p(\theta)}{p(d)} = \frac{p(d | \theta) p(\theta)}{\int d\theta_i p(d | \theta_i) p(\theta_i)} \quad (1)$$

where

- $p(\theta | d, M)$ is the *posterior* probability density function (PDF) for θ under the model M which we wish to compute. From a Bayesian point of view it expresses the confidence degree in the values of θ according to the information given by the data d . In practice, it expresses the probability that the θ in the theory M can be explained by the data d .
- $p(d | \theta) = \mathcal{L}(\theta)$ is the *likelihood* function, corresponding instead to the probability/likeness of the observed data given a certain value of the parameter θ , given as a function of the parameter itself. Thus, it is the PDF of a given data set d to be explained by the parameters θ .
- $p(\theta)$ is the *prior* PDF, which encodes the degree of belief in the values of θ before we see the data or, in simple terms, the prior information/knowledge of the parameter before the experiment. This is a key element for Bayesian inference in physics, usually assessed from the theory or from past experiments. More commonly, either flat priors or Gaussian priors are chosen, but there may be situations that require more sophisticated choices.

- $p(d)$ is seen as the condition imposing the normalisation of the posterior probability dependent on the data distribution d known as the *evidence*:

$$p(\theta) = \int p(d | \theta) p(\theta) d\theta \quad (2)$$

It can be disregarded in the simplest cases for parameter inference, and only becomes important for the statistical *evidence* comparison between different models. This model selection criteria relies on the fact that the Bayes' theorem relates what we learn about θ *after* seeing the data (the posterior) to what we knew about θ *before* looking at the data (the likelihood and the prior). Therefore, in some way it measures how much we have learned or how much our knowledge has been updated from the prior to the posterior.

The basic principle of Bayesian inference is that our knowledge about the quantity we are analyzing is updated in cycles: the posterior from a previous iteration becomes the prior for the next one. This means that we need to have a starting point, which corresponds exactly to fixing an initial prior which must be given as an input and that should be a good representation of the current knowledge about the quantity we wish to analyze. In normal conditions the posterior will converge to a unique and "true" result independently of the prior, as long as the prior range includes regions of the parameter space with large likelihoods.

Once the posterior distribution has been computed, a lot of information about the parameters of the model $\{\theta_i\} = \{\theta_1, \theta_2, \dots, \theta_n\}$ given the data d can be extracted. For instance, if we wish to find the probability of a single chosen parameter θ_i we can integrate over the remaining parameters, a technique called *marginalization*:

$$p(\theta_i) = \int \dots \int p(\theta_i | d) d\theta_1 \dots d\theta_{i-1} d\theta_{i+1} \dots d\theta_n. \quad (3)$$

This is especially important in the case where there are multiple *nuisance parameters*, related with the experiment and the method itself, which are not informative for the cosmological analysis.

Finally, we can also compute the *confidence level* for the inferred parameters as:

$$p(\theta_i) = \int_{R(r)} p(\theta_i | d) d^n \theta = r \quad (4)$$

where r is some fraction linked to the probability that the parameters will be contained in that parameter region. The most commonly studied cases are $r = 0.683, 0.954, 0.997$, corresponding to the 1σ , 2σ and 3σ confidence levels, respectively.

2 MARKOV CHAIN MONTE CARLO

In broad terms, Markov Chain Monte Carlo (MCMC) methods constitute a category of sampling algorithms to derive the posterior distribution taking random steps in parameter space according to some proposed distribution and an acceptance criteria. They are based on Monte Carlo methods that are a broad class of computational algorithms based on consecutive random sampling to achieve numerical solutions.

Overall, the aim of any MCMC algorithm is to generate a *sequence* of points in parameter space – usually referred to as *chain* – with a density proportional to the posterior PDF. More precisely, a Markov chain is defined as a sequence of $m > i$ random variables with probability density proportional to some known function. This is a stochastic method in the sense that the probability of the $(i + 1)$ th element in the chain depends only on the value of the i th element. The iteration method to generate repeated elements of the chain has a probabilistic nature, measured through a transition probability $T(\theta_i, \theta_{i+1})$ in parameter space. Having a stochastic nature, this means that in a Markov chain the transition probability satisfies the detailed *balance condition*

$$p(\theta_i | d) T(\theta_i, \theta_{i+1}) = p(\theta_{i+1} | d) T(\theta_{i+1}, \theta_i) \quad (5)$$

which simply means that the ratio of the probabilities for going from θ_i to θ_{i+1} is inversely proportional to the ratio of the corresponding posterior probabilities. As is the general *motto* in statistics, the more steps are included, the more closely will the sampled distribution resemble the actual target distribution.

In practice, a number of chains are generated, starting from arbitrarily chosen positions in the parameter space which should be sufficiently separated from each other. This will generate a *random walk* following an algorithm that assigns higher *jumping* probabilities to iterations in which the parameters give a considerable contribution to the distribution.

2.1 Metropolis-Hastings algorithm

The Metropolis-Hastings (MH) algorithm is arguably the simplest MCMC algorithm for conducting analyses, particularly suited for parameters with a non-symmetrical distribution. The algorithm works as follows:

1. Start from an *initial point in the parameter space* θ_0 randomly selected from the prior density, and *compute its posterior*, that is, $p_0 \equiv p(\theta_0|d)$.
2. *Propose a next candidate iteration* θ_c based on the proposal distribution $q(\theta_0, \theta_c)$, which does not necessarily need to satisfy the symmetry condition $q(x, y) = q(y, x)$. The most common and simple example of a symmetric distribution would be a Gaussian of fixed width σ centered around the current point.
3. In the same way, evaluate the value of the posterior at the candidate point, $p_c = p(\theta_c|d)$. The *probability of acceptance of the candidate point* is given as

$$\alpha(\theta_0, \theta_c) = \min \left(\frac{p_c q(\theta_c, \theta_0)}{p_0 q(\theta_0, \theta_c)}, 1 \right). \quad (6)$$

The *criteria for acceptance/rejection* is based on generating a uniform random number $u \in [0, 1]$ and accepting the candidate sample if $u < \alpha$, and rejecting if otherwise.

4. If the candidate point is accepted, it is *added to the chain* and becomes the new starting point for the next iteration. Otherwise, stay at the old point, which is registered again in the chain.
5. Repeat the procedure by starting again from 2.

Note how if the distribution is symmetric then the candidate sample is always accepted if it has a larger posterior than the current one, that is if $p_c > p_0$.

The normalization constant given by the evidence becomes irrelevant as it is eliminated in the ratio computed in the acceptance rate. This may be problematic for model comparison. However, it means that *marginalizing over the remaining parameters becomes trivial*, as we can simply discard the parameter to be marginalized in each sample.

However, this method has some well-known *limitations*. First, *the initial point is chosen arbitrarily*. This means that the starting point in the chain may be significantly far from the peak of the stationary distribution, resulting in a significantly slowing down of the process. This is usually accounted for by discarding some fraction of the initially accumulated chain samples (*burn-in fraction*).

Second, most MCMC methods require a "proposal distribution," which determines the "length" of each step of the random walk through the parameter space. For example, a D -dimensional Gaussian (normal distribution) is often used to draw probabilities for moving from point to point. Even within this simple choice (which is one function among many possibilities), there are $\frac{D(D+1)}{2}$ parameters in the $D \times D$ symmetric covariance matrix to be set. The problem is that if the proposal distribution leads to too small steps, then almost all steps will be accepted at the cost of taking a long time to move anywhere. On the other hand, if the proposal distribution leads to too large steps, then the parameter space will be easily covered but almost no steps will be accepted, and we will tend to end up in much lower probability regions. Intuitively, the optimal step size should be the step size that makes for the shortest autocorrelation time, approximated by the *acceptance rate*. The golden ratio for the acceptance rate (fraction of accepted moves) is between a quarter and about a half, with the best performance argued for ~ 0.23 in high-dimensional problems. This tuning of the acceptance rate must take place during the *burn-in phase* (in which some part of the chain will be discarded later) of an MCMC run. This is because the access to the past history of the chains is a violation of the "Markovity" of the chains, for which each step only depends on the previous state. During this process, we aim to track the acceptance ratio and adjust the proposal distribution variance as you get acceptance ratios that are far from the ideal ratio, with this process being easily automated. However, some previous knowledge about the proposal distribution is needed, meaning that the results have to be taken with a grain of salt and carefully examined. This also implies that MH is very inefficient if the distribution is multi-modal, in which case other sampling methods may be necessary.

2.2 Convergence Criteria

We say that a Markov chain has converged when it reaches a stationary state (does not change for increasing i), where successive elements of the chain become samples from the target posterior distribution.

One simple and generally effective test of convergence is the *Gelman–Rubin diagnostic*, which compares the variance (in individual or in all parameters) within a chain to the variance across chains. This, of course, requires running multiple chains and looking at the variance of the parameter away from its mean within each individual chain, and comparing it to the variance in the mean of that parameter across chains. The Gelman–Rubin method involves complex calculations with these variances. However, the crucial aspect for convergence assessment is whether, as the chains lengthen, these variances ultimately stabilize and whether those stabilized values align. A usual criterion is to assume the MCMC has converged if testing whether the Gelman–Rubin factor $\max(R - 1) \approx 10^{-2}$ across all the sampled parameters.

2.3 Goodness of fit: χ^2 and *Likelihood*

If the PDFs are Gaussian then maximizing the likelihood is essentially equivalent to the least square method of minimizing the χ^2 (best fit of parameters), which corresponds to the exponent of the Gaussian distribution

$$\mathcal{L}(d|\theta) \propto \exp(-\chi^2/2). \quad (7)$$

The maximum likelihood estimation is more general than the minimum χ^2 and the best-fitting parameters will correspond to $\nabla_{\theta}\mathcal{L}(d|\theta) = 0$.

Simple models are built on the underlying assumption that the observed data should be normally distributed around the corresponding model values up to some standard deviation. This implies that the sum of squares of normalized residuals around the best-fit model should follow the χ^2 distribution:

$$\chi^2(d; \theta, M) = \sum_{n=1}^N \left[\frac{y_n - y_M(\theta)}{\sigma_{y,n}} \right]^2, \quad (8)$$

where we have assumed that there are N data points $d(n)$ with corresponding errors σ_n . In each case, the theory predictions can be estimated according to a given set of parameters Θ . Optimization of Θ implies minimization of the χ^2 . This statistic is proportional to the negative log-likelihood, up to a constant offset, and so for model fitting, minimizing the χ^2 is equivalent to maximizing the likelihood. The key information is that this statistic replaces the quantity $P(d|M)$ with an easily computable quantity $P(\chi^2|M_{\chi^2})$ (with known probability distribution). We are implicitly replacing the question of how likely it is to observe the data D assuming the model M by how likely we are to see the χ^2 statistic under the best-fit model M_{χ^2} , and the two likelihoods will be equivalent as long as our assumptions hold (namely that the data is normally distributed).

However, in most cases, the data are not independent and some given covariance matrix is also needed to compute χ^2 :

$$\chi^2 = \sum_{n=1}^N [y_n - y_M(\theta)]^T \text{COV}^{-1} [y_n - y_M(\theta)], \quad (9)$$

COV^{-1} is the inverse covariance matrix where the diagonal terms are $1/\sigma_{y,n}^2$.

2.4 Model Comparison Criteria

It is important to highlight the fundamental difference between model fitting and model selection. The process of model fitting relies on assuming that a particular model is the true model and extracting the constraints that provide the best possible fit to the available data. On the other hand, in model selection, we are interested in assessing the level of compatibility of each model with the data.

In exceptionally simple cases, we can select between models by comparing the value of the maximum likelihood, but this is not valid in general. If model A has more degrees of freedom (i.e., more parameters) than model B (and model B can be recovered from model A), then A will always give an equal or larger maximum likelihood compared to B, irrespective of the data.

For a simpler analysis, we can consider information criteria for approximate model comparison keeping in mind that these make use of considerably strong assumptions about the posterior distribution:

- Akaike Information Criterion (AIC): $\text{AIC} \equiv \chi_{\text{eff}}^2 + 2k$
- Bayesian Information Criterion (BIC): $\text{BIC} \equiv \chi_{\text{eff}}^2 + k \ln(N)$
- Deviance Information Criterion (DIC): $\text{DIC} \equiv \chi_{\text{eff}}^2 + 2 \left[\bar{\chi}_{\text{eff}}^2 - \chi_{\text{eff}}^2 \right]$

where k is the number of fitted parameters, N is the number of data points, and the terms $\chi_{\text{eff}}^2 \equiv -2 \ln(\mathcal{L}_{\text{max}})$ give the χ^2 of the best-fit, and the upper bar denotes quantities computed at the average of the posterior distribution.

In this case, the best model must minimize the AIC/BIC/DIC. However, each criterion penalizes models differently, with the BIC including a stronger penalty for models with a larger number of free parameters k when the number of data points N is sufficiently large and gives a better approximation to the full Bayesian evidence in the large N limit. On the other hand, by evaluating the difference according to the average of the posterior distributions, the DIC also considers whether information about the parameters is actually gained or not. Because of all these subtleties, it's always preferable to calculate the full Bayesian evidence, although this may be non-trivial for more complex cases, in which the information criteria already provide some useful insight.

As an intuitive example, the $\Delta\text{AIC} = \text{AIC}(M_2) - \text{AIC}(M_1)$ is an approximation to the Bayes factor. It penalizes models with more parameters (a correction for the sample size can be included as $\frac{2k(k+1)}{N-p-1}$, with N being the number of bins). The model with the smaller value of AIC is preferred, since the likelihood ratio is $e^{\Delta\text{AIC}/2}$.

2.4.1 Bayesian Criteria

Another way to assess preference for different models can be related to the Bayesian evidence or model likelihood:

$$p(d|M) = \int d\theta_i p(d|\theta_i, M) p(\theta_i|M) \quad (10)$$

which is simply expressed as the integral of the likelihood over the prior range and we have now made the model dependence explicit.

Basically, we determine the values of the χ^2 over a grid of n parameters and convert to probability $P(\theta_1, \dots, \theta_n) \propto e^{-\chi^2/2}$. Then we marginalize to obtain the posterior probability distributions for each parameter, $P(\theta_1), P(\theta_2), \dots$ and so on. The integration under these distributions leads to what is referred to as the confidence regions (such as 68% and 99%). Because this integral is computed over the entire parameter space of the model it can be extremely computationally expensive, especially as the dimension of the model grows beyond 2 or 3.

The model's posterior is expressed simply in terms of Bayes' theorem and it's this ratio that quantifies the model comparison:

$$\frac{p(M_1|d)}{p(M_2|d)} = \frac{p(d|M_1) p(M_1)}{p(d|M_2) p(M_2)} \quad (11)$$

which defining the Bayes factor as

$$B_{12} \equiv \frac{p(d|M_1)}{p(d|M_2)} \quad (12)$$

is equivalent to the statement *posterior odds* = *Bayes factor* \times *prior odds*.

The strength of the evidence is evaluated according to Jeffrey's scale for the favored model (different definitions in the literature):

$ \ln B $	fractional odds	model's probability	Evidence
< 1.0	$< 3 : 1$	< 0.750	inconclusive
< 2.5	$< 12 : 1$	0.923	weak
< 5.0	$< 150 : 1$	0.993	moderate
> 5.0	$> 150 : 1$	> 0.993	strong

Evaluated according to the logarithm as:

$$\ln B_{12} = \ln \frac{p(d|M_1)}{p(d|M_2)} = \ln p(d|M_1) - \ln p(d|M_2) \quad (13)$$

The Bayes factor gives a compromise between quality of fit and additional model complexity (such as additional free parameters). This means that it will "reward" highly predictive models, "penalizing" unnecessary extra parameter space. This is what is often referred to as *Occam's razor*.