Statistical Methods in Cosmology

L. Verde

Summary The advent of large data set in cosmology has meant that in the past 10 or 20 years our knowledge and understanding of the Universe has changed not only quantitatively but also, and most importantly, qualitatively. Cosmologists are interested in studying the origin and evolution of the physical Universe. They rely on data where a host of useful information is enclosed, but is encoded in a non-trivial way. The challenges in extracting this information must be overcome to make the most of the large experimental effort. Even after having analyzed a decade or more of data and having converged to a standard cosmological model (the so-called and highly successful ΛCDM model) we should keep in mind that this model is described by 10 or more physical parameters and if we want to study deviations from the standard model the number of parameters is even larger. Dealing with such a high-dimensional parameter space and finding parameters constraints is a challenge on itself. In addition, as gathering data is such an expensive and difficult process, cosmologists want to be able to compare and combine different data sets both for testing for possible disagreements (which could indicate new physics) and for improving parameter determinations. Finally, because experiments are always so expansive, cosmologists in many cases want to find out a priori, before actually doing the experiment, how much one would be able to learn from it. For all these reasons, more and more sophisticated statistical techniques are being employed in cosmology, and it has become crucial to know some statistical background to understand recent literature in the field. Here, I will introduce some statistical tools that any cosmologist should know about in order to be able to understand recently published results from the analysis of cosmological data sets. I will not present a complete and rigorous introduction to statistics as there are several good books which are reported in the references. The reader should refer to those. I will take a practical approach and I will touch upon useful tools such as statistical inference, Bayesians vs Frequentists approach, chisquare and goodness of fit, confidence regions, likelihood, Fisher matrix approach, Monte Carlo methods, and a brief introduction to model testing. Throughout, I will use practical examples often taken from recent

L. Verde (✉)
ICREA & ICE (IEEC-CSIC) and ICC UB, Bellaterra, Spain
e-mail: verde@ieec.uab.es

Verde, L.: *Statistical Methods in Cosmology*. Lect. Notes Phys. **800**, 147–177 (2010)
DOI 10.1007/978-3-642-10598-2_4

literature to illustrate the use of such tools. Of course this will not be an exhaustive guide: it should be interpreted as a "starting kit," and the reader is warmly encouraged to read the references to find out more.

1 Introduction

As cosmology has made the transition from a data-starved science to a data-driven science, the use of increasingly sophisticated statistical tools has increased. As explained in detail below, cosmology is intrinsically related to statistics, as theories of the origin and evolution of the Universe do not predict, for example, that a particular galaxy will form at a specific point in space and time or that a specific patch of the cosmic microwave background will have a given temperature; any theory will predict average statistical properties of our Universe, and we can only observe a particular realization of that.

It is often said that cosmology has entered the precision era: "precision" requires a good knowledge of the error bars and thus confidence intervals of a measurement. This is an inherently statistical statement. We should try, however, to go even further, and also achieve "accuracy" (although cosmology does not have a particularly stellar track record in this regard). This requires quantifying systematic errors (beyond the statistical ones) and it also requires statistical tools. For all these reasons, knowledge of basic statistical tools has become indispensable to understand the recent cosmological literature.

Examples of applications where probability and statistics are crucial in Cosmology are (i) Is the Universe homogenous and isotropic on large scales? (ii) Are the initial conditions consistent with being Gaussian? (iii) Is there a detection of non-zero tensor modes? (iv) What is the value of the density parameter of the Universe Ω_m given the WMAP data for a ΛCDM model? (v) What are the allowed values at a given confidence level for the primordial power spectrum spectral slope n? (vi) What is the best fit value of the dark energy equation of state parameter w? (vii) Is a model with equation of state parameter different from -1 a better fit to the data than a model with non-zero curvature? (viii) What will be the constraint on the parameter w for a survey with given characteristic?

The first three questions address the hypothesis-testing issue. You have an hypothesis and you want to check whether the data are consistent with it. Sometimes, especially for addressing issues of "detection" you can test the null hypothesis: assume the quantity is zero and test whether the data are consistent with it.

The next three questions are "parameter estimation" problems: we have a model, in this example, the ΛCDM model, which is characterized by some free parameters which we would like to measure.

The next question, (vii), belongs to "model testing"; we have two models and ask which one is a better fit to the data. Model testing comes in several different flavors: the two models to be considered may have different number of parameters or equal number of parameters, may have some parameters in common or not, etc.

Finally question (viii) is on "forecasting," which is particularly useful for or quickly forecasting the performance of future experiments and for experimental design.

Here we will mostly concentrate on the issue of parameter estimation but also touch upon the other applications.

2 Bayesians vs Frequentists

The world is divided into Frequentists and Bayesians. For Frequentists probabilities $\mathcal{P}$ are frequencies of occurrence:

$$\mathcal{P} = \frac{n}{N}, \tag{1}$$

where n denotes the number of successes and N the total number of trials. Frequentists define probability as the limit for the number of independent trials going to infinity. Bayesians interpret probabilities as *degree of belief in a hypothesis*.

Let us say that x is our random variable (event). Depending on the application, x can be the number of photons hitting a detector, the matter density in a volume, the Cosmic Microwave Background temperature in a direction in the sky, etc. The probability that x takes a specific value is $\mathcal{P}(x)$ where $\mathcal{P}$ is called probability distribution. Note that probabilities (the possible values of x) can be discrete or continuous. $\mathcal{P}(x)$ is a *probability density*: $\mathcal{P}(x)dx$ is the probability that the random variable x takes a value between x and $x+dx$. Frequentists only consider probability distributions of events while Bayesians consider hypothesis as events.

For both, the rules of probability apply.

1. $\mathcal{P}(x) \geq 0$
2. $\int_{-\infty}^{\infty} dx\mathcal{P}(x) = 1$. In the discrete case $\int \longrightarrow \sum$.
3. For mutually exclusive events $\mathcal{P}(x_1 U x_2) \equiv \mathcal{P}(x_1.OR.x_2) = \mathcal{P}(x_1) + \mathcal{P}(x_2)$
4. In general $\mathcal{P}(x_1,x_2) = \mathcal{P}(x_1)\mathcal{P}(x_1|x_2)$. In words, the probability of x_1 AND x_2 to happen is the probability of x_1 times the *conditional probability* of x_2 given that x_1 has already happened.

The last item deserves some discussion. For example, only for independent events where $\mathcal{P}(x_2|x_1) = \mathcal{P}(x_2)$ one can write $\mathcal{P}(x_1,x_2) = \mathcal{P}(x_1)\mathcal{P}(x_2)$. Of course in general one can always rewrite $\mathcal{P}(x_1,x_2) = \mathcal{P}(x_1)\mathcal{P}(x_1|x_2)$ by switching $x_1$and x_2. If then one makes the apparently tautological identification that $\mathcal{P}(x_1,x_2) = \mathcal{P}(x_2,x_1)$ and substitute $x_1 \longrightarrow D$ standing for *data* and $x_2 \longrightarrow H$ standing for *hypothesis*, one gets Bayes theorem :

$$\mathcal{P}(H|D) = \frac{\mathcal{P}(H)\mathcal{P}(D|H)}{\mathcal{P}(D)}, \tag{2}$$

where $\mathcal{P}(H|D)$ is called the *posterior*, $\mathcal{P}(D|H)$ is the *likelihood* (the probability of the data given the hypothesis) and $\mathcal{P}(H)$ is called the *prior*. Note that here explicitly we have probability and probability distribution of a hypothesis.

3 Bayesian Approach and Statistical Inference

Despite its simplicity, Bayes theorem is at the base of statistical inference. For the Bayesian point of view let us use D to indicate our data (or data set). The hypothesis H can be a model, say for example the ΛCDM model, which is characterized by a set of parameters $\boldsymbol{\theta}$. In the Bayesian framework what we want to know is "What is the probability distribution for the model parameters given the data?" i.e. $\mathcal{P}(\boldsymbol{\theta}|D)$. From this information we can extract the most likely value for the parameters and their confidence limits.[1] However, what we can compute accurately, in most instances, is the likelihood, which is related to the posterior by the prior. (At this point one assumes that one has collected the data and so $\mathcal{P}(D) = 1$). The prior, however, can be somewhat arbitrary. This is a crucial point to which we will return below. For now let us consider an example: the constraint from WMAP data on the integrated optical depth to the last scattering surface τ. One could do the analysis using the variable τ itself, however, one could also note that the temperature data (the angular power spectrum of the temperature fluctuations) on large scale depend approximately linearly on the variable $Z = \exp(-2\tau)$. A third person would note that the polarization data (in particular the EE angular power spectrum) depend roughly linearly on τ^2. So person one could use a uniform prior in τ, person two a uniform prior in $\exp(-2\tau)$, and person three in τ^2. What is the relation between $\mathcal{P}(\tau)$, $\mathcal{P}(Z)$, and $\mathcal{P}(\tau^2)$?

3.1 *Transformation of Variables*

We wish to transform the probability distribution of $\mathcal{P}(x)$ to the probability distribution of $\mathcal{G}(y)$ with y as a function of x. Recall that probability is a conserved quantity (we cannot create or destroy probabilities...) so

$$\mathcal{P}(x)dx = \mathcal{G}(y)dy, \tag{3}$$

thus

$$\mathcal{P}(x) = \mathcal{G}(y(x)) \left|\frac{dy}{dx}\right|. \tag{4}$$

Following the example above if x is τ and y is $\exp(\tau)$ then $\mathcal{P}$ is related to $\mathcal{G}$ by a factor 2τ and if y is τ^2 by a factor 2. In other words using different priors leads to different posteriors. This is the main limitation of the Bayesian approach.

---

[1] At this point many Frequentists stop reading this document...

3.2 Marginalization

So far we have considered probability distributions of a random variable x, but one could analogously define *multi-variate distributions*, the joint probability distribution of two or more variables, e.g., $\mathcal{P}(x,y)$. A typical example is the description of the initial distribution of the density perturbations in the Universe. Motivated by inflation and by the central limit theorem, the initial distribution of density perturbation is usually described by a multi-variate Gaussian: at every point in space given by its spatial coordinates (x, y, z), $\mathcal{P}$ is taken to be a random Gaussian distribution. Another example is when one simultaneously constrains the parameters of a model, say, for example, $\boldsymbol{\theta} = \{\Omega_m, H_0\}$ (here H_0 denotes the Hubble constant). If you have $\mathcal{P}(\Omega_m,H_0)$ and want to know the probability distribution of Ω_m regardless of the values of H_0 then

$$\mathcal{P}(\Omega_m) = \int dH_0 \mathcal{P}(\Omega_m,H_0). \tag{5}$$

3.3 Back to Statistical Inference and Cosmology

Let us go back to the issue of statistical inference and follow the example from [1]. If you have an urn with N red balls and M blue balls and you draw one ball at the time then probability theory can tell you what are your chances of picking a red ball given that you have already drawn n red and m blue: $\mathcal{P}(D|H)$. However, this is not what you want to do, you want to make a few draws from the urn and use probability theory to tell you what is the red vs blue distribution inside the urn is, $\mathcal{P}(H|D)$. In the Frequentist approach all you can compute is $\mathcal{P}(D|H)$.

In the case of cosmology it gets even more complicated.

We consider that the Universe we live in is a random realization of all the possible Universes that could have been a realization of the true underlying model (which is known only to Mother Nature). All the possible realizations of this true underlying Universe make up the *ensemble*. In statistical inference one may sometime want to try to estimate how different our particular realization of the Universe could be from the true underlying one. Going back to the example of the urn with red and blue balls, it would as if we were to be drawing from one particular urn, but the urn is part of a large batch. On average, the batch distribution has 50% red and 50% blue, but each urn has only an odd number of balls and so any particular urn cannot reflect exactly the 50–50 spilt.

A crucial assumption of standard cosmology is that the part of the Universe that we can observe is a fair sample of the whole. But the peculiarity in cosmology is that we have just one Universe, which is just one realization from the ensemble (quite fictitious one: it is the ensemble of all possible Universes). The fair sample hypothesis states that samples from well-separated parts of the Universe are independent realizations of the same physical process, and that, in the observable part of the Universe, there are enough independent samples to be representative of the statistical ensemble.

In addition, experiments in cosmology are not like lab experiments: in many cases observations cannot be easily repeated (think about the observation of a particular supernova explosion or of a Gamma ray burst) and we cannot try to perturb the Universe to see how it reacts... After these considerations, it may be clearer why cosmologists tend to use the Bayesian approach.

4 Chisquare and Goodness of Fit

Say that you have a set of observations and have a model, described by a set of parameters $\boldsymbol{\theta}$, and want to fit the model to the data. The model may be physically motivated or a convenient function. One then should define a merit function, quantifying the agreement between the model and the data, by maximizing the agreement one obtains the best fit parameters. Any useful fitting procedure should provide: (1) best fit parameters (2) estimation of error on the parameters (3) possibly a measure of the goodness of fit. One should bear in mind that if the model is a poor fit to the data then the recovered best fit parameters are meaningless.

Following numerical recipes ([2], Chap. 15) we introduce the concept of model fitting (parameter fitting) using least squares. Let us assume we have a set of data points D_i, for example, these could be the band power galaxy power spectrum at a set of k values, and a model for these data $y(x, \boldsymbol{\theta})$ which depends on set of parameters $\boldsymbol{\theta}$ (e.g., the ΛCDM power spectrum, which depends on n_s-primordial power spectrum spectral slope, σ_8-present-day amplitude of rms mass fluctuations on scale of 8 Mpc/h–, $\Omega_m h$, etc.). Or it could be, for example, the supernovae type 1a distance modulus as a function of redshift; see, e.g., Fig. 1 [3, 4].

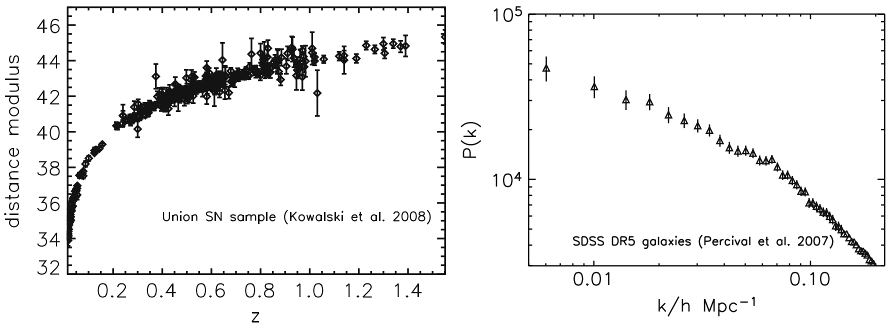

Fig. 1 *Left*: distance modulus vs redshift for supernovae type 1A from the UNion sample [3]. *Right*: bandpower $P(k)$ for DR5 SDSS galaxies, from [4]. In both cases one may fit a theory (and the theory parameters) to the data with the chisquare method. Note that in both cases errors are correlated. In the *right* panel the errors are also strictly speaking non-Gaussianly distributed

The least squares, in its simplest incarnation is

$$\chi^2 = \sum_i w_i [D_i - y(x_i|\boldsymbol{\theta})]^2, \tag{6}$$

where w_i are suitably defined weights. It is possible to show that the minimum variance weight is $w_i = 1/\sigma_i^2$ where σ_i denotes the error on data point i. In this case the least squares is called chisquare. If the data are correlated the chisquare becomes

$$\chi^2 = \sum_{ij} (D_i - y(x_i|\theta))\, Q_{ij}(D_j - y(x_j|\theta)), \tag{7}$$

where Q denotes the inverse of the so-called covariance matrix describing the covariance between the data. The best fit parameters are those that minimize the χ^2. See an example in Fig. 2.

For a wide range of cases the probability distribution for different values of χ^2 around the minimum of (7) is the χ^2 distribution for $\nu = n - m$ degrees of freedom where n is the number of independent data points and m the number of parameters. The probability that the observed χ^2 exceeds by chance a value $\widehat{\chi}$ for the correct model is $Q(\nu,\widehat{\chi}) = 1 - \Gamma(\nu/2,\widehat{\chi}/2)$ where Γ denotes the incomplete Gamma function. See the Numerical Recipes bible [2]. Conversely, the probability that the observed χ^2, even for the correct model, is less than $\widehat{\chi}$ is $1 - Q$. While this statement is strictly true if measurement errors are Gaussian and the model is a linear function of the parameters, in practice it applies to a much wider range of cases.

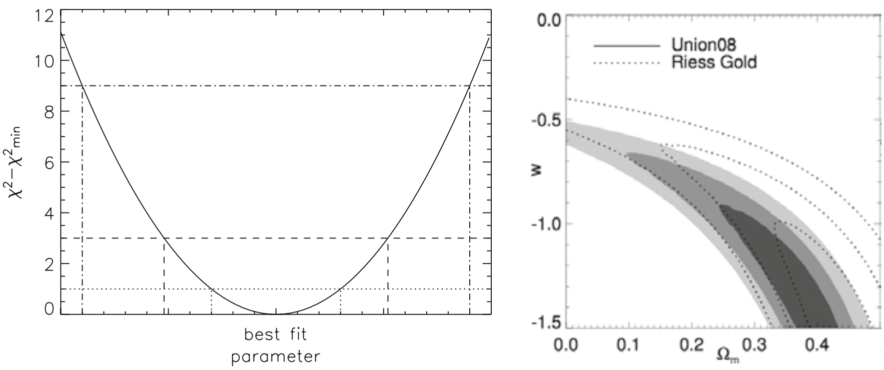

Fig. 2 *Left*: example of a one-dimensional chisquare for a Gaussian distribution as a function of a parameter and corresponding 68.3, 95.4, and 99.5% confidence levels. *Right*: a two-dimensional example for the union supernovae data. Figure from Kowalski et al. [3] reproduced with permission from the AAS. Note that in a practical application even if the data have Gaussian errors the errors on the parameter may not be well described by multi-variate Gaussians (thus the confidence regions are not ellipses)

The quantity Q evaluated that the minimum chisquare (i.e., at the best fit values for the parameters) gives a measure of the goodness of fit. If Q gives a very small probability then there are three possible explanations:

(1) the model is wrong and can be rejected. (Strictly speaking, the data are unlikely to have happened if the Universe was really described by the model considered)
(2) the errors are underestimated
(3) the measurement errors are non-Gaussianly distributed.

Note that in the example of the power spectrum we know a priori that the errors are non-Gaussianly distributed. In fact, even if the initial conditions were Gaussian and if the underlying matter perturbations were still evolving in the linear regime (i.e., $\delta\rho/\rho \ll 1$) and galaxies were nearly unbiased tracers of the dark matter, then the density fluctuation itself would obey Gaussian statistics and so would its Fourier transform, but *not* its power spectrum, which is a square quantity. In reality we know that by $z = 0$ perturbations grow non-linearly and that galaxies may not be nearly unbiased tracers of the underlying density field. Nevertheless, the central limit theorem comes to our rescue, if in each band power there is a sufficiently large number of modes.

If Q is too large (too good to be true) it is also cause for concern:

(1) errors have been overestimated
(2) data are correlated or non-independent
(3) the distribution is non-Gaussian

Beware: this last case is very rare.

A useful "chi-by-eye" rule is the minimum χ^2 should be roughly equal to ν (number of data–number of parameters). This is increasingly true for large ν. From this, it is easy to understand the use of the so-called reduced chisquare that is the $\chi^2_{\rm min}/m$: if $m \gg n$ (i.e., number of data much larger than the number of parameters to fit, which should be true in the majority of the cases) then $m \sim \nu$ and the rule of thumb is that reduced chisquare should be unity.

Note that the chisquare method, and the Q statistic, gives the probability for the data, given a model $\mathcal{P}(D|\theta)$ and not $\mathcal{P}(\theta|D)$. One can make this identification via the prior.

5 Confidence Regions

Once the best fit parameters are obtained, how can one represent the confidence limit or confidence region around the best fit parameters? A reasonable choice is to find a region in the m-dimensional parameter space (remember that m is the number of parameters) that contains a given percentage of the probability distribution. In most cases one wants a compact region around the best fit values. A natural choice is then given by regions of constant χ^2 boundaries. Note that there may be cases (when the χ^2 has more than one minimum) in which one may need to report a non-connected confidence region. For multi-variate Gaussian distributions, however, these are ellipsoidal regions. Note that the fact that the data have Gaussian errors does not imply that the parameters will have a Gaussian probability distribution . . .

Thus, if the values of the parameters are perturbed from the best fit, the χ^2 will increase. One can use the properties of the χ^2 distribution to define confidence intervals in relation to χ^2variations or $\Delta\chi^2$. Table 1 reports the $\Delta\chi^2$ for 68.3, 95.4, and 99.5% confidence levels as function of number of parameters for the joint confidence level. In the case of Gaussian distributions these correspond to the conventional 1, 2, and 3σ. See an example of this in Fig. 2

Beyond these values here is the general prescription to compute constant χ^2 boundaries confidence levels. After having found the best fit parameters by minimizing the χ^2 and if Q for the best fit parameters is acceptable then

- Let m be the number of parameters, n the number of data, and p be the confidence limit
- Solve the following equation for $\Delta\chi^2$:

$$Q(n-m, \min(\chi^2) + \Delta\chi^2) = p \tag{8}$$

- Find the parameter region where $\chi^2 \le \min(\chi^2) + \Delta\chi^2$. This defines the confidence region.

If the actual error distribution is non-Gaussian but it is known then it is still possible to use the χ^2 approach, but instead of using the chisquare distribution and Table 1, the distribution needs to be calibrated on multiple simulated realization of the data as illustrated below in Sect. 13.

Table 1 $\Delta\chi^2$ for the conventionals 1, 2, and $3-\sigma$ as a function of the number of parameters for the joint confidence levels

p (%)	1	2	3
68.3	1.00	2.30	3.53
95.4	2.71	4.61	6.25
99.73	9.00	11.8	14.2

6 Likelihood

So far we have dealt with the frequentist quantity $\mathcal{P}(D|H)$. If we set $\mathcal{P}(D) = 1$ and ignore the prior then we can identify the likelihood with $\mathcal{P}(H|D)$ and thus by maximizing the likelihood we can find the most likely model (or model's parameters) given the data. However, having ignored $\mathcal{P}(D)$ and the prior this approach cannot give in general a goodness of fit and thus cannot give an absolute probability for a given model. However, it can give relative probabilities. If the data are Gaussianly distributed the likelihood is given by a multi-variate Gaussian:

$$\mathcal{L} = \frac{1}{(2\pi)^{n/2}|detC|^{1/2}} \exp\left[-\frac{1}{2}\sum_{ij}(D-y)_i C_{ij}^{-1}(D-y)_j\right], \tag{9}$$

where $C_{ij} = \langle (D_i - y_i)(D_j - y_j)\rangle$ is the covariance matrix.

It should be clear from this that the relation between χ^2 and likelihood is that, for Gaussian distributions, $\mathcal{L} \propto \exp[-1/2\chi^2]$ and minimizing the χ^2 is equivalent to minimizing the likelihood. In this case likelihood analysis and χ^2 coincide and by the end of this section, it will thus be no surprise that the Gamma function appearing in the χ^2 distribution is closely related to the Gaussian integrals.

The subtle step is that now, in Bayesian statistics, confidence regions are regions R in *model space* such that $\int_R \mathcal{P}(\theta|D)d\theta = p$ where p is the confidence level we request (e.g., 68.3, 95.4%). Note that by integrating the posterior over the model parameters, the confidence region depends on the prior information, as seen in Sect. 3.1 different priors give different posteriors and thus different regions R.

It is still possible to report results independently of the prior by using the *likelihood ratio*. The likelihood at a particular point in parameter space is compared with that at the best fit value, $\mathcal{L}_{\max}$, where likelihood is maximized. Thus a model is acceptable if the likelihood ratio,

$$\Lambda = -2\ln\left[\frac{\mathcal{L}(\theta)}{\mathcal{L}_{\max}}\right], \tag{10}$$

is above a given threshold. The connection to the χ^2 for Gaussian distribution should be clear. In general, the threshold can be calibrated by calculating the entire distribution of the likelihood ratio in the case that a particular model is the true model. Frequently this is chosen to be the best fit model.

There is a subtlety to point out here. In cosmology the data may be Gaussainly distributed and still the χ^2 and likelihood ratio analysis may give different results. This happens because in identifying likelihood and chisquare we have neglected the term $[(2\pi)^{n/2}|detC|^{1/2}]^{-1}$. If the covariance does not depend on the model or model parameters, this is just a normalization factor which drops out in the likelihood ratio. However, in cosmology often the covariance depends on the model; this happens, for example, when errors are dominated by cosmic variance, like in the case of the CMB temperature fluctuations on the largest scales, or on the galaxies power spectrum on the largest scales. In this case the cosmology dependence of the covariance cannot be neglected, but one can always define a pseudo-chisquare as $-2\ln\mathcal{L}$ and work with this quantity.

Let us stress again that the likelihood is linked to the posterior through the prior; the identification of the likelihood with the posterior is prior dependent (as we will see in an example below). In the absence of any data it is common to assume a flat (uniform) prior, i.e., all values of the parameter in question are equally likely, but other choices are possible and sometimes more motivated. For example, if a parameter is positive definite, it may be interesting to use a logarithmic prior (uniform in the log).

Priors may be assigned theoretically or from prior information gathered from previous experiments. If the priors are set by theoretical considerations, it is always a good practice to check how much the results depend on the choice of the prior. If the dependence is significant, it means that the data do not have much statistical power to constrain that (those) parameter(s). Information theory helps us quantify the amount of "information gain"; the information in the posterior relative to the prior is

$$\mathcal{I} = \int \mathcal{P}(\theta|D) \log\left[\frac{\mathcal{P}(\theta|D)}{\mathcal{P}(\theta)}\right] d\theta. \tag{11}$$

6.1 Marginalization: Examples

Some of the model parameters may be uninteresting. For example, in many analyses one wants to include nuisance parameters (calibration factors, biases, etc.) but then report the confidence level on the real cosmological parameters regardless of the value of the nuisance ones. In other cases the model may have say, 10 or more real cosmological parameters but we may be interested in the allowed range of only one or two of them, regardless of the values of all the others. Typical examples are, e.g., constraints on the curvature parameter Ω_k (which we may want to know regardless of the values of, e.g., Ω_m or Ω_Λ) or, say, the allowed range for the neutrino mass regardless of the power spectrum spectral index or the value of the Hubble constant. As explained in Sect. 3.2 one can marginalize over the uninteresting parameters.

It should be kept in mind that marginalization is a Bayesian concept: the results may depend on the prior chosen.

In some cases, the marginalization can be carried out analytically. An example is reported below, this applies to the case of, e.g., calibration uncertainty, point sources amplitude, overall scale independent galaxy bias, magnitude intrinsic brightness, or beam errors for CMB studies. In this case it is useful to know the following results for Gaussian likelihoods:

$$\begin{aligned}\mathcal{P}(\theta_1..\theta_{m-1}|D) = \int & \frac{dA}{(2\pi)^{\frac{m}{2}}||C||^{\frac{1}{2}}} e^{\left[-\frac{1}{2}(C_i-(\hat{C}_i+AP_i))\Sigma_{ij}^{-1}(C_j-(\hat{C}_j+AP_j))\right]} \\ & \times \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\frac{(A-\hat{A})^2}{\sigma^2}\right],\end{aligned} \tag{12}$$

where repeated indices are summed over and $||C||$ denotes the determinant. Here, A is the amplitude of, say, a point source contribution P to the C_ℓ angular power spectrum, A is the mth parameter which we want to marginalize over with a Gaussian prior with variance σ^2 around $\hat{A}$. The trick is to recognize that this integral can be written as

$$\mathcal{P}(\theta_1..\theta_{m-1}|D) = C_0 \exp\left[-\frac{1}{2}C_1 - 2C_2A + C_3A^2\right] dA, \tag{13}$$

(where $C_{0...3}$ denote constants and it is left as an exercise to write them down explicitly) and that this kind of integral is evaluated by using the substitution $A \longrightarrow A - C_2/C_3$ giving something $\propto \exp[-1/2(C_1 - C_2^2/C_3)]$.

In cases where the likelihood surface (describing the value of the likelihood as a function of the parameters) is not a multi-variate Gaussian, the location of the maximum likelihood before marginalization may not coincide with the location after marginalization. An example is shown in Fig. 3. The figure shows the probability distribution for Ω_k from WMAP5 data for a model where curvature is free and the equation of state parameter for dark energy w is constant in time but not fixed at -1. The red line shows the N-dimensional maximum posterior value and the black line is the marginalized posterior over all other cosmological parameters.

It should also be added that, even in the case where we have a single-peaked posterior probability distribution there are two common estimators of the "best" parameters: the peak value (i.e., the most probable value) or the mean, $\hat{\theta} = \int d\theta\theta\mathcal{P}(\theta|D)$. If the posterior is non-Gaussian these two estimates need not coincide. In the same spirit, slightly different definitions of confidence intervals need not coincide for non-Gaussian likelihoods, as illustrated in the right panel of Fig. 3: for example, one can define the confidence interval $[\theta_{low}, \theta_{high}]$, such that equal fractions of the

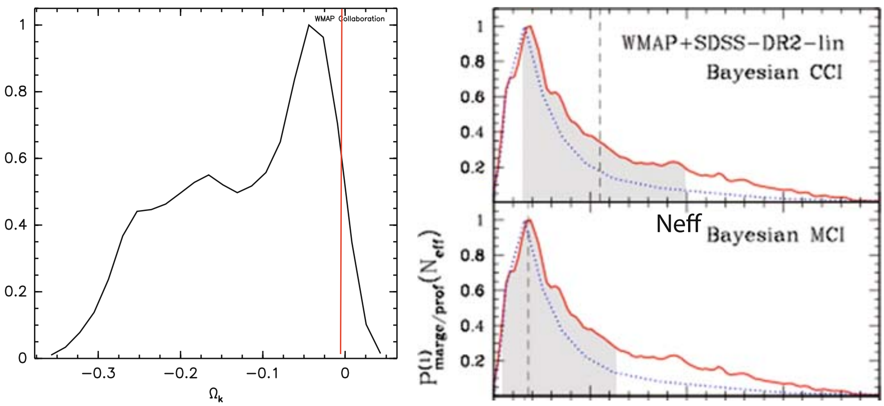

Fig. 3 Marginalization effects. *Left panel*: We consider the posterior distribution for the cosmological parameters of a dark energy + cold dark matter model where curvature is a free parameter and so is a (constant) equation of state parameter for dark energy. The data are the WMAP 5-year data. The *solid line* shows the N-dimensional maximum posterior value and the *black line* is the marginalized posterior over all other cosmological parameters. Figure courtesy of LAMBDA [5]. *Right panel*: figure from [6]. Illustration of central credible interval (CCI) and minimum credible interval (MCI), for the case of a ΛCDM model with free number of effective neutrino species (ignore *dotted line* for this example, *red line* is the marginalized posterior)

posterior volume lie in $(-\infty, \theta_{low})$ and (θ_{high}, ∞). This is called central credible interval and is connected to the median. Another possibility (minimum credible interval) is to consider the region so that the posterior at any point inside it is larger than at any point outside and so that the integral of the posterior in this region is the required fraction of the total. Thus remember, it is always a good practice to declare what confidence interval one is using. This subject is explored in more details in, e.g., [6].

7 Why Gaussian Likelihoods?

Throughout these lectures we always refer to Gaussian likelihoods. It is worth mentioning that if the data errors are Gaussianly distributed then the likelihood function for the data will be a multi-variate Gaussian. If the data are not Gaussianly distributed (but still are drawn from a distribution with finite variance!) we can resort to the central limit theorem: we can bin the data so that in each bin there is a superposition of many independent measurements. The central limit theorem will tell us that the resulting distribution (i.e., the error distribution for each bin) will be better approximated by a multi-variate Gaussian. However, as mentioned before, even if the data are Gaussianly distributed this does not ensure that the likelihood surface for the parameters will be a multi-variate Gaussian; for this to be always true the model needs to depend linearly on the parameters. Even without resorting to the central limit theorem, the Gaussian approximation is in many cases recovered even when starting from highly non-Gaussian distribution. A neat example is provided by Cash [7] which we follow here.

Let us say you want to constrain cosmology by studying cluster number counts as a function of redshift. The observation of a discrete number N of clusters is a Poisson process, the probability of which is given by the product

$$\mathcal{P} = \Pi_{i=1}^{N} [e_i^{n_i} \exp(-e_i)/n_i!], \tag{14}$$

where n_i is the number of clusters observed in the *i-th* experimental bin and e_i is the expected number in that bin in a given model, $e_i = I(x)\delta x_i$ with i being the proportional to the probability distribution. Here δx_i can represent an interval in clusters mass and/or redshift. Note: this is a product of Poisson distributions, thus one is assuming that these are independent processes. Clusters may be clustered, so when can this be used?

For unbinned data (or for small bins so that bins have only 0 and 1 counts) we define the quantity:

$$C \equiv -2 \ln \mathcal{P} = 2(E - \sum_{i=1}^{N} \ln I_i), \tag{15}$$

where E is the total expected number of clusters in a given model. The quantity ΔC between two models with different parameters has a χ^2 distribution! (so all that was said in Sect. 4 applies, even though we started from a highly non-Gaussian distribution.)

8 The Effect of Priors: Examples

Let us consider the two figures in Fig. 4. On the left, WMAP first-year data constraints in the Ω_m, Ω_Λ plane. On the right, models consistent with the WMAP 3-year data. In both cases the model is a non-flat ΛCDM model. So why the addition of more data (the two extra years of WMAP observations) gives worst constraints? The key is that what is reported in the plots is a representation of the posterior probability distribution. In the left panel a flat prior on Θ_A (angular size distance to the last scattering surface, giving by the position of the first peak) was assumed. In the figure on the right a flat prior on the Hubble constant H_0 was assumed. Remember: always declare the priors assumed!

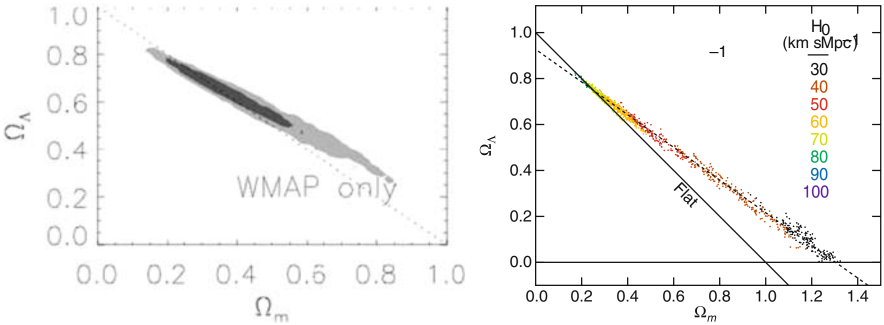

Fig. 4 *Left*: WMAP first-year data constraints in the Ω_m, Ω_Λ plane, from Spergel et al. [9]. *Right*: models consistent with the WMAP 3-year data, from Spergel et al. [8]. In both cases the model is a non-flat ΛCDM model. Figures reproduced with permission from the AAS

9 Combining Different Data Sets: Examples

It has become common to "combine data sets" and explore the constraints from the "data set combination." This means in practice that the likelihoods can be multiplied if the data sets are independent (if not the one should account for the appropriate covariance). It is important to note that *If the data-sets are inconsistent, the resulting constraints from the combined data set are nonsense*. An example is shown in Fig. 5.

On the left panel we show a figure from [8] constraints in the Ω_m, σ_8 plane for a flat ΛCDM model for WMAP 3-year data (blue), weak lensing constraints (orange), and combined constraints. On the right panel the figure shows the constraints in the

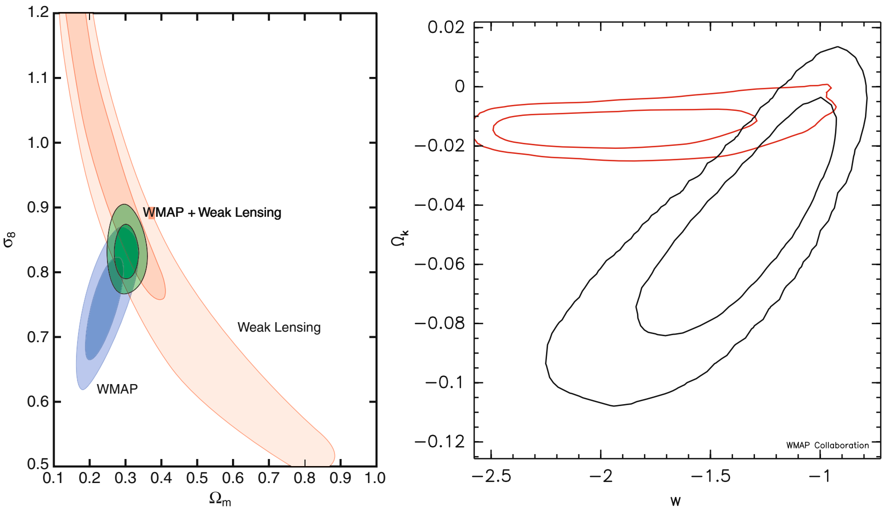

Fig. 5 *Left*: constraints in the Ω_m, σ_8 plane for a flat ΛCDM model for WMAP 3-year data, weak lensing constraints, and combined constraints. Figure from Spergel et al. [8], reproduced with permission from the AAS. *Right*: Constraints in the Ω_k,w plane for non-flat dark energy models with constant w for WMAP5+supernovae data (*lower curves*) and WMAP5+BAO (*upper curves*). Figure courtesy of LAMBDA [5]

Ω_k,w plane for non-flat dark energy models with constant w for WMAP5+ supernovae data (in black) and WMAP5+BAO (in red). Even though the WMAP data are in common there is some tension in the resulting constraints. The two data sets (supernovae and BAO, WMAP and weak lensing) are not fully consistent, as the authors themselves, note, they should not be combined.

10 Forecasts: Fisher Matrix

Before diving into the details let us re-examine the error estimates for parameters from the likelihood. Let us assume a flat prior in the parameter so we can identify the posterior with the likelihood. Close to the peaks we can expand the log likelihood in Taylor series

$$\ln \mathcal{L} = \ln \mathcal{L}(\theta_0) + \frac{1}{2}\sum_{ij}(\theta_i - \theta_{i,0})\left.\frac{\partial^2 \ln \mathcal{L}}{\partial\theta_i \partial\theta_j}\right|_{\theta_0}(\theta_j - \theta_{j0}) + \dots. \tag{16}$$

By truncating this expansion to the quadratic term (remember that by expanding around the maximum we have the first derivative equal to zero) we say the likelihood surface is locally a multi-variate Gaussian. The Hessian matrix is defined as

$$\mathcal{H}_{ij} = -\frac{\partial^2 \ln \mathcal{L}}{\partial\theta_i \partial\theta_j}. \tag{17}$$

It encloses the information of the parameters' errors and their covariance. If this matrix is not diagonal it means that the parameters' estimates are correlated. Loosely speaking we said "the parameters are correlated": it means that they have a similar effect on the data and thus the data have hard time in telling them apart. The parameters may or may not be physically related with each other.

More specifically if all parameters are kept fixed except one (parameter i, say), the error on that parameter would be given by $1/\sqrt{\mathcal{H}_{ii}}$. This is called conditional error but is almost never used or interesting.

Having understood this, we can move on to the Fisher information matrix [10]. The Fisher matrix plays a fundamental role in forecasting errors from a given experimental set up and thus is the work-horse of experimental design. It is defined as

$$F_{ij} = -\left\langle \frac{\partial^2 \ln \mathcal{L}}{\partial\theta_i \partial\theta_j} \right\rangle. \tag{18}$$

It should be clear that $F = \langle \mathcal{H} \rangle$.

Here the average is the ensemble average over observational data (those that would be gathered if the real Universe was given by the model – and model parameters – around which the derivative is taken). As we have seen the likelihood for independent data sets is the product of the likelihoods, it follows that the Fisher matrix for independent data sets is the sum of the individual Fisher matrices. This will become useful later on.

In the one-parameter case, say only i component of θ, thinking back at the Taylor expansion around the maximum of the likelihood we have that

$$\Delta \ln \mathcal{L} = \frac{1}{2} F_{ii} (\theta_i - \hat{\theta}_i)^2 \tag{19}$$

when $2\Delta \ln \mathcal{L} = 1$ and by identifying it with the $\Delta\chi^2$ corresponding to 68% confidence level, we see that $1/\sqrt{F_{ii}}$ yields the $1 - \sigma$ displacement for θ_i. This is the analogous to the conditional error from above. In the general case

$$\sigma_{ij}^2 \geq (F^{-1})_{ij}. \tag{20}$$

Thus when all parameters are estimated simultaneously from the data the marginalized error is

$$\sigma_{\theta_i} \geq (F^{-1})_{ii}^{1/2}. \tag{21}$$

Let's spell it out for clarity: this is the square root of the element ii of the inverse of the Fisher information matrix.[2] This assumes that the likelihood is a Gaussian

---

[2] i.e., you have to perform a matrix inversion first.

around its maximum (the fact that the data are Gaussianly distributed is no guarantee that the likelihood will be Gaussian, see, e.g., Fig. 2). The terrific utility of the Fisher Information matrix is that, if you can compute it, it enables you to estimate the parameters errors *before you do the experiment*. If it can be computed quickly, it also enables one to explore different experimental setups and optimize the experiment. This is why the Fisher matrix approach is so useful in survey design. Also complementarity of different, independent, and uncorrelated experiments (i.e., how in combination they can lift degeneracies) can be quickly explored: the combined Fisher matrix is the sum of the individual matrices. This is of course extremely useful; however, read below for some caveats.

The $\geq$ is the Kramer–Rao inequality: the Fisher matrix approach always gives you an optimistic estimate of the errors (reality is only going to be worst). And this is not only because systematic and real-world effects are often ignored in the Fisher information matrix calculation, but for a fundamental limitation: only if the likelihood is Gaussian that $\geq$ becomes $=$. In some cases, when the Gaussian approximation for the likelihood does not hold, it is possible to make non-linear transformation of the parameter that makes the likelihood Gaussian. Basically, if the data are Gaussianly distributed and the model depends linearly on the parameters then the likelihood would be Gaussian. So the key is to have a good enough understanding of the theoretical model to be able to find such a transformation. See [11] for a clear example.

10.1 Computing Fisher Matrices

The simplest, brute force approach to compute a Fisher matrix is as follows: write down the likelihood for the data given the model. Instead of the data values (which are not known) use the theory prediction for a fiducial model. This will add a constant term to the log likelihood which does not depend on cosmology. In the covariance matrix include expected experimental errors. Then take derivatives with respect to the parameters as indicated in (18).

In the case where the data are Gaussianly distributed it is possible to compute explicitly and analytically the Fisher matrix, in a much more elegant way than above:

$$F_{ij} = \frac{1}{2} Tr\left[C^{-1}C_{,i}C^{-1}C_{,j} + C^{-1}M_{ij}\right], \tag{22}$$

where $M_{ij} = y_{,i}y_{,j}^T + y_{,j}y_{,i}^T$ and $,i$ denotes derivative with respect to the parameter θ_i. This is extremely useful, you need to know the covariance matrix (which may depend on the model and need not be diagonal) and you need to have a fiducial model y which you know how it depends on the parameter θ. Then the Fisher matrix gives you the expected (forecasted) errors. Priors or forecasts results from other experiments can be easily included by simply adding their Fisher before

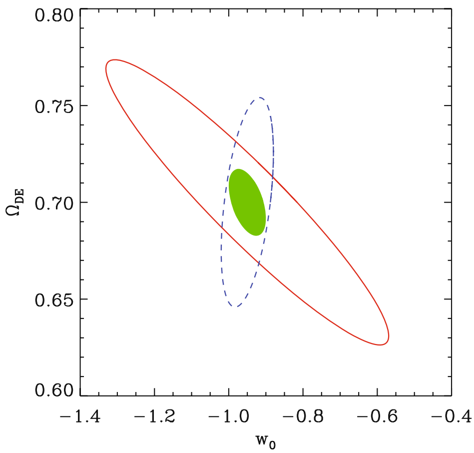

Fig. 6 Marginalized 68% CL constraints on the dark energy parameters expected for the DUNE weak lensing (*dashed*), a full sky BAO survey (*solid*), and their combination (*solid filled*). This figure was derived using the Fisher matrix routines of iCosmo. Figure from Refregier et al. [12]

performing the matrix inversion to obtain the marginal errors. This is illustrated in Fig. 6, from [12] and produced using the icosmo (http://www.icosmo.org/) software.

Before we finish this section let us spell out the following prescription.

Imagine you have computed a large Fisher matrix, varying all parameters Ω_k, w_0, neutrino mass m_ν, number of neutrino species N_ν, running of the spectral index α, etc. Now you want to compute constraints for a standard flat ΛCDM model. Simply ignore row and columns corresponding to the parameters that you want to keep fixed at the fiducial value before inverting the matrix.

Imagine now that you have a six parameters' Fisher matrix (say H_0, Ω_m,τ, Ω_Λ, n, Ω_b, σ_8), and want to produce 2D plots for the confidence regions for parameters 2 and 4, say, marginalized over all other (1,3,5,6) parameters. Invert F_{ij}. Take the sub-matrix made by rows and columns corresponding to the parameters of interest (2 and 4 in this case) and invert back this sub-matrix.

The resulting matrix, let us call it $\mathcal{Q}$, describes a Gaussian 2D likelihood surface in the parameters 2 and 4 or, in other words, the chisquare surface for parameters 2,4 – marginalized over all other parameters – can be described by the equation

$$\widetilde{\chi}^2 = \sum_{ij} (\theta_i - \theta_i^{fid.})\mathcal{Q}_{ij}(\theta_j - \theta_j^{fid.})\,. \tag{23}$$

From this equation, getting the errors corresponds to finding the quadratic equation solution $\widetilde{\chi}^2 = \Delta\chi^2$. For correspondence between $\Delta\chi^2$ and confidence region see the earlier discussion. If you want to make plots, the equation for the elliptical boundary for the joint confidence region in the sub-space of parameters of interest is $\Delta = \delta\theta\,\mathcal{Q}^{-1}\delta\theta$.

11 Example of Fisher Approach Applications

Here we are going to consider two cases of application of Fisher forecasts that are extensively used in the literature. This section assumes that the reader is familiar with basic CMB and large-scale structure concepts, such as power spectra, error on power spectra, cosmic variance, window and selection function, instrumental noise and shot noise, redshift space. Some readers may find this section more technical than the rest of this document; it is possible to skip it and continue reading from Sect. 12.

11.1 CMB

The CMB has become the single data set that gives most constraints on cosmology. As the recently launched Planck satellite will yield the ultimate survey for primary CMB temperature anisotropies, doing Fisher matrix forecasts of CMB temperature data may very soon be obsolete. There remains the scope for forecasting constraints from polarization experiments, however, systematic effects (e.g., foreground subtraction) will likely dominate the statistical errors (see, e.g., [13] for details). It is still, however, a good exercise to see how one can set up a Fisher matrix analysis for CMB data.

If we have a noiseless full sky survey and the initial conditions are Gaussian we can write that the signal in the sky (i.e., the spherical harmonic transform of the anisotropies) is Gaussianly distributed. We can write the signal as

$$\mathbf{s}_\ell = (a_\ell^T, a_\ell^E, a_\ell^B), \tag{24}$$

where a_{ell} denotes the spherical harmonic coefficients for temperature and E and B model polarization. The covariance matrix $\mathbf{C}_\ell$ is then given by

$$\mathbf{C}_\ell = \begin{pmatrix} C_\ell^{TT} & C_\ell^{TE} & 0 \\ C_\ell^{TE} & C_\ell^{EE} & 0 \\ 0 & 0 & C_\ell^{BB} \end{pmatrix}, \tag{25}$$

where C_ℓ denotes the angular CMB power spectrum. Using (22) and considering that, for rotational invariance, for every ℓ there are $(2\ell+1)$ modes, it is possible to show that the Fisher matrix for CMB experiments can be rewritten as

$$F_{ij}^{CMB} = \sum_{XY} \sum_\ell \frac{\partial C_\ell^X}{\partial \theta_i} \left(\mathcal{C}_\ell^{XY}\right)^{-1} \frac{\partial C_\ell^Y}{\partial \theta_j}, \tag{26}$$

where for the matrix $\mathcal{C}_\ell$ the elements are $\mathcal{C}_\ell^{XY}$, where X,Y=TT, TE, EE, BB, etc., is given by[3]

$$\mathcal{C}_\ell = \frac{2}{2\ell+1}\begin{pmatrix} (C_\ell^{TT})^2 & (C_\ell^{TE})^2 & C_\ell^{TT}C_\ell^{TE} & 0 \\ (C_\ell^{TE})^2 & (C_\ell^{EE})^2 & C_\ell^{EE}C_\ell^{TE} & 0 \\ C_\ell^{TT}C_\ell^{TE} & C_\ell^{EE}C_\ell^{TE} & 1/2[(C_\ell^{TE})^2 + C_\ell^{TT}C_\ell^{EE}] & 0 \\ 0 & 0 & 0 & (C_\ell^{BB})^2 \end{pmatrix}. \tag{27}$$

Note that this matrix is more complicated that what one would have obtained by assuming a Gaussian distribution for the C_ℓ and no correlation between TT, TE, and EE. Nevertheless, (26) is simple enough and allows one to quickly compute forecasts from ideal CMB experiments.

In this formalism effects of partial sky coverage and of instrumental noise can be included (approximatively) by the following substitutions:

$$C_\ell \longrightarrow C_\ell + N_\ell, \tag{28}$$

where N_ℓ denotes the effective noise power spectrum. Note that N_ℓ depends on ℓ even for a perfectly white noise because of beam effects. In addition the partial sky coverage can be accounted for by considering that the number of independent modes decreases with the sky coverage: if f_{sky} denotes the fraction of sky covered by the experiment, then

$$\mathcal{C}_\ell \longrightarrow \mathcal{C}_\ell / f_{sky}. \tag{29}$$

11.2 Baryon Acoustic Oscillations

Cosmological perturbations in the early Universe excite sound waves in the photon–baryon fluid. After recombination, these baryon acoustic oscillations (BAO) became frozen into the distribution of matter in the Universe imprinting a preferred scale, the sound horizon. This defines a standard ruler whose length is the distance sound can travel between the Big Bang and recombination. The BAO are directly observed in the CMB angular power spectrum and have been observed in the spatial distribution of galaxies by the 2dF GRS survey and the SDSS survey [14–16]. The BAO, observed at different cosmic epochs, act as a powerful measurement tool to probe the expansion of the Universe, which in turn is a crucial handle to constrain the nature of dark energy. The underlying physics which sets the sound horizon scale ($\sim$150 Mpc comoving) is well understood and involves only linear perturbations in the early Universe. The BAO scale is measured in surveys of galaxies from the statistics of the three-dimensional galaxy positions. Only recently have galaxy

[3] I owe this proof to P. Adshead.

surveys such as SDSS grown large enough to allow for this detection. The existence of this natural standard measuring rod allows us to probe the expansion of the Universe. The angular size of the oscillations in the CMB revealed that the Universe is close to flat. Measurement of the change of apparent acoustic scale in a statistical distribution of galaxies over a large range of redshift can provide stringent new constraints on the nature of dark energy. The acoustic scale depends on the sound speed and the propagation time. These depend on the matter to radiation ratio and the baryon-to-photon ratio. CMB anisotropy measures these and hence fixes the oscillation scale. A BAO survey measures the acoustic scale along and across the line of sight. At each redshift, the measured angular (transverse) size of oscillations, $\Delta\theta$, corresponds with the physical size of the sound horizon, where the angular diameter distance D_A is an integral over the inverse of the evolving Hubble parameter, $H(z)$. $r_\perp = (1+z)D_A(z)\delta\theta$. In the radial direction, the BAO directly measure the instantaneous expansion rate $H(z)$, through $r_\parallel = (c/H(z))\Delta z$, where the redshift interval (Δz) between the peaks is the oscillation scale in the radial direction. As the true scales $r_\perp$ and $r_\parallel$ are known (given by r_s, the sound horizon at radiation drag, well measured by the CMB) this is not an Alcock–Paczynski test but a "standard ruler" test. Note that in this standard ruler test the cosmological feature used as the ruler is not an actual object but a statistical property: a feature in the galaxy correlation function (or power spectrum). An unprecedented experimental effort is undergoing to obtain galaxy surveys that are deep, larger, and accurate enough to trace the BAO feature as a function of redshft. Before these surveys can even be designed it is crucial to know how well a survey with given characteristic will do. This was illustrated very clearly in [17], which we follow closely here. We will adopt the Fisher matrix approach. To start we need to compute the statistical error associated to a determination of the galaxy power spectrum $P(k)$. In what follows we will ignore the effects of non-linearities and complicated biasing between galaxies and dark matter: we will assume that galaxies, at least on large scales, trace the linear matter power spectrum in such a way that their power spectrum is directly proportional to the dark matter one: $P(k) = b^2 P_{DM}(k)$ where b stands for galaxy bias. At a given wavevector k, the statistical error of the power spectrum is a sum of a cosmic variance term and a shot noise term:

$$\frac{\sigma_P(k)}{P(k)} = \frac{P(k) + 1/n}{P(k)}. \tag{30}$$

Here n denotes the average density of galaxies and $1/N$ is the white noise contribution from the fact that galaxies are assumed to be a Poisson sampling of the underlying distribution. When written in this way this expression assumes that n is constant with position. While in reality this is not true for forecasts, one assumes that the survey can be divided in shells in redshifts and that the selection function is such that n is constant within a given shell. Since $P(k)$ is also expected to change in redshift then one should really implicitly assume that there is a z dependence in (30). In general $P(k,z) = b(z)^2 G^2(z) P_{DM}(k)$ where $G(z)$ denotes the linear growth factor, i.e., the bias is expected to evolve with redshift as well as clustering does, not only because galaxy bias changes with redshift but also because at different redshifts one may be seeing different type of galaxies which may have different bias parameter. We do not know a priori the form of $b(z)$ but given a fiducial cosmological model we know $G(z)$. Preliminary observations seem to indicate that the z evolution of b tends to cancel that of $G(z)$, so it is customary to assume that $b(z)G(z) \sim$ constant, but we should bear in mind that this is an assumption.

An extra complication arises because galaxy redshift surveys use the redshift as distance indicator, and deviations from the Hubble flow therefore distort the clustering. If the Universe was perfectly uniform and galaxies were test particles these deviations from the Hubble flow would not exist and the survey would not be distorted. But clustering does perturb the Hubble for and thus introduces the so-called redshift-space distortions in the clustering measured by galaxy redshift surveys. Note that redshift-space distortions only affect the line-of-sight clustering (it is a perturbation to the distances) not the angular clustering. Since these distortions are created by clustering they carry, in principle, important cosmological information. To write this dependence explicitly

$$P(k,\mu,z) = b(z)^2 G(z)^2 P_{DM}(k)(1+\beta\mu)^2, \tag{31}$$

where μ denotes the cosine of the angle between the line-of-sight and the wavevector $\beta = f/b = d\ln G(z)/d\ln a/b \simeq \Omega_m(z)^{0.6}/b$. In the linear regime, the cosmological information carried by the redshift-space distortions is enclosed in the $f(z) = \beta(z)b(z)$ combination.

For finite surveys, $P(k)$ at nearby wavenumbers are highly correlated, the correlation length is related to the size of the survey volume; for large volumes the cell size over which modes are correlated is $(2\pi)^3/V$ where V denotes the comoving survey volume. Only over distances in k-space larger than that modes can be considered independent. If one therefore wants to count over all the modes anyway (for example, by transforming discrete sums into integrals in the limit of large volumes) then each k needs to be downweighted, to account the fact that all k are not independent. In addition one should keep in mind that Fourier modes $\mathbf{k}$ and $-\mathbf{k}$ are not independent (the density field is real valued!), giving an extra factor of 2 in the weighings. We can thus write the error on a band power centered around k,

$$\frac{\sigma_P}{P} = 2\pi\sqrt{\frac{2}{Vk^2\delta k\Delta\mu}}\left(\frac{1+nP}{nP}\right). \tag{32}$$

In the spirit of the Fisher approach we now assume that the likelihood function for the band powers $P(k)$ is Gaussian, thus we can approximate the Fisher matrix by

$$F_{ij} = \int_{k_{\min}}^{k_{\max}} \frac{\partial \ln P(\mathbf{k})}{\partial\theta_i}\frac{\partial \ln P(\mathbf{k})}{\partial\theta_j} V_{\rm eff}(\mathbf{k})\frac{d\mathbf{k}}{2(2\pi)^3}. \tag{33}$$

The derivatives should be evaluated at the fiducial model and $V_{\rm eff}$ denotes the effective survey volume given by

$$V_{\rm eff}(\mathbf{k}) = V_{\rm eff}(k,\mu) = \int \left[\frac{n(z)P(k,\mu)}{n(z)P(k,\mu)+1}\right]^2 dz = \left[\frac{nP(k,\mu)}{nP(k,\mu)+1}\right]^2 V, \tag{34}$$

where $n = \langle n(z)\rangle$. Equation (33) can be written explicitely as a function of k and μ as

$$F_{ij} = \int_{-1}^{1}\int_{k_{\rm min}}^{k_{\rm max}} = \frac{\partial \ln P(k,\mu)}{\partial \theta_i}\frac{\partial \ln P(k,\mu)}{\partial \theta_j} V_{\rm eff}(k,\mu)\frac{k^2 dk d\mu}{2(2\pi)^2}\,. \tag{35}$$

In writing this equation we have assumed that over the entire survey extension the line-of-sight direction does not change: in other words, we made the flat sky approximation. For forecasts this encloses all the statistical information anyway, but for actual data analysis application the flat sky approximation may not hold. In this equation $k_{\rm min}$ is set by the survey volume; for future surveys where the survey volume is large enough to sample the first BAO wiggling the exact value of $k_{\rm min}$ does not matter, however, recall that for surveys of typical size L (where $L \sim V^{1/3}$), the largest scale probed by the survey will be corresponding to $k = 2\pi/L$. Keeping in mind that the first BAO wiggle happens at $\sim 150\,$Mpc the survey size needs to be $L \gg 150\,$Mpc for $k_{\rm min}$ to be unimportant and for the "large volume approximation" made here to hold. As anticipated above, one may want to sub-divide the survey into independent redshift shells, compute the Fisher matrix for each shell, and then combine the constraints. In this case L will be set by the smallest dimension of the volume (typically the width of the shell), so one needs to make sure that the width of the shell still guarantees a large volume and large L. $k_{\rm max}$ denotes the maximum wavevector to use. One could, for example, impose a sharp cut to delimit the range of validity of linear theory. In [18] this is improved as we will see below.

Before we do that, let us note that there are two ways to interpret the parameters θ_{ij} in (35). One could simply assume a cosmological model, say, for example, a flat quintessence model where the equation of state parameter $w(z)$ is parameterized by $w(z) = w(0) + w_a(1-a)$ and take derivatives of $P(k,\mu)$ with respect to these parameters. Alternatively, one could simply use as parameters the quantities $H(z_i)$ and $D_A(z_i)$, where z_i denotes the survey redshift bins. These are the quantities that govern the BAO location and are more general; they allow one not to choose a particular dark energy model until the very end. Then one must also consider the cosmological parameters that govern the $P(k)$ shape $\Omega_m h^2$, $\Omega_b h^2$, and n_s. Of course one can also consider $G(z_i)$ as free parameters and constrain these either through the overall $P(k)$ amplitude (although one would have to assume that $b(z)$ is known, which is dicey) or through the determination of $G(z)$ and $\beta(z)$. The safest and most conservative approach, however, is to ignore any possible information coming from $G(z)$, $\beta(z)$, or n_s and to only try to constrain expansion history parameters.

The piece of information still needed is how the expansion history information is extracted from $P(k,\mu)$. When one converts **ra, dec**, and redshifts into distances and positions of galaxies of a redshift survey, one assumes a particular reference cosmology. If the reference cosmology differs from the true underlying cosmology, the inferred distances will be wrong and so the observed power spectrum will be distorted:

$$P(k_\perp, k_\parallel) = \frac{D_a(z)^2_{ref} H(z)_{\rm true}}{D_A(z)^2_{\rm true} H(z)_{\rm ref}} P_{\rm true}(k_\perp, k_\parallel). \tag{36}$$

Note that since distances are affected by the choice of cosmology and k vectors are $k_{{\rm ref},\parallel} = H(z)_{\rm ref}/H(z)_{\rm true} k_{{\rm true},\parallel}$ and $k_{{\rm ref},\perp} = D_A(z)_{\rm true}/D_A(z)_{\rm ref} k_{{\rm true},\perp}$. Note that therefore in (36) we can write

$$P_{\rm true}(k_\perp, k_\parallel, z) = b(z)^2 \left(1 + \beta(z) \frac{k^2_{{\rm true},\parallel}}{k^2_{{\rm true},\perp} + k^2_{{\rm true},\parallel}}\right)^2 \left[\frac{G)(z)}{G(z_o)}\right]^2 P_{DM}(k, z_o), \tag{37}$$

where z_o is some reference redshift where to normalize $P(k)$ typical choices can be the CMB redshift or redshift $z = 0$. Not that from these equations it should be clear that what the BAO actually measure directly is $H(z) r_s$ and D_A/r_s where r_s is the BAO scale, the advantage is that r_s is determined exquisitely from the CMB.

How would then one convert these constraints on those on a model parameter? Clearly, one then projects the resulting Fisher matrix on the dark energy parameters space. In general if you have a set of parameters θ_i with respect to which the Fisher matrix has been computed, but you would like to have the Fisher matrix for a different set of parameters ϕ_i, where the θ_i are functions of the ϕ_i, the operation to implement is

$$F_{\phi_i,\phi_j} = \sum_{mn} \frac{\partial \theta_n}{\partial \phi_i} F_{\theta_n,\theta_m} \frac{\partial \theta_m}{\partial \phi_j}. \tag{38}$$

The full procedure for the BAO survey case is illustrated in Fig. 7. The slight complication is that one starts off with a Fisher matrix (for the original parameter set θ_i) where some parameters are nuisance and need to be marginalized over, so some matrix inversions are needed.

So far non-linearities have been just ignored. It is, however, possible to include then at some level in this description. Reference [18] proceed by introducing a distribution of Gaussianly distributed random displacements parallel or perpendicular to the line-of-sight coming from non-linear growth (in all directions) and from non-linear redshift-space distortions (only in the radial direction). The publicly available code that implements all this (and more) is at http://cmb.as.arizona.edu/ eisenste/acousticpeak/bao_forecast.html. In order to use the code keep in mind that in Ref. [18] the authors model the effect of non-linearities by convolving the galaxy distribution with a redshift dependent and μ-dependent smoothing kernel. The

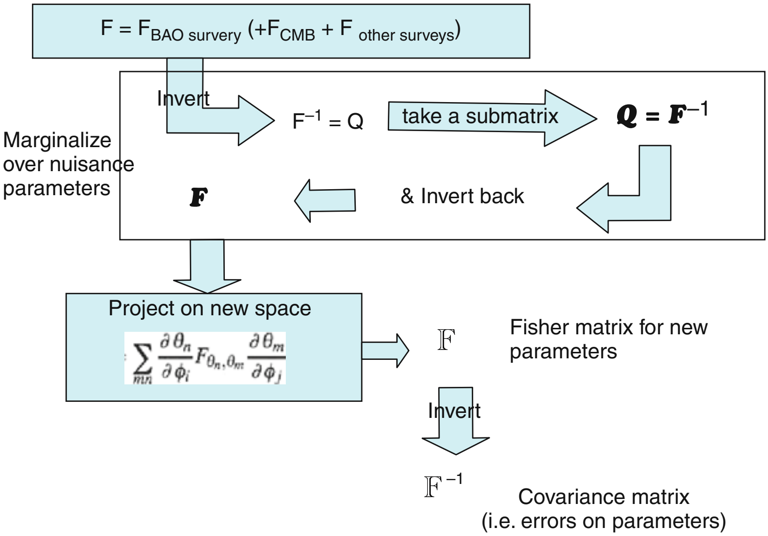

Fig. 7 Steps to implement once the Fisher matrix of (35) has been computed to obtain error on dark energy parameters

effect on the power spectrum is to multiply $P(k)$ by $\exp[-k^2\Sigma(k,\mu)/2]$, where $\Sigma(k,\mu) = \Sigma_\perp^2 - \mu^2(\Sigma_\parallel^2 - \Sigma_\perp^2)$. As a consequence the integrand of the Fisher matrix expression of (35) is multiplied by

$$\exp[-k^2\Sigma_\perp^2 - k^2\mu^2(\Sigma_\parallel^2 - \Sigma_\perp^2)], \tag{39}$$

where, to be conservative, the exponential factor has been taken outside the derivatives, which is equivalent to marginalize over the parameters $\Sigma_\parallel$ and $\Sigma_\perp$ with large uncertainties.

Note that $\Sigma_\parallel$ and $\Sigma_\perp$ depend on redshift and on the chosen normalization for $P_{DM}(k)$. In particular,

$$\Sigma_\perp(z) = \Sigma_0 G(z)/G(z_0), \tag{40}$$

$$\Sigma_\parallel(z) = \Sigma_0 G(z)/G(z_0)(1+f(z)), \tag{41}$$

$$\Sigma_0 \propto \sigma_8\,. \tag{42}$$

If in your convention $z_0 = 0$ then $\Sigma_0(z=0) = 8.6h^{-1}\sigma_{8,DM}(z=0)/0.8$.

As an example of an application of this approach for survey design, it may be interesting to ask the question of what is the optimal galaxy number density for a given survey. Taking redshifts is expensive and for a given telescope time allocated, only a certain number of redshifts can be observed. Thus is it better to survey more volume but have a low number density or survey a smaller volume with higher number density? You can try to address this issue using the available code. For a cross check, Fig. 8 shows what you should obtain. Here we have assumed $\sigma_8 = 0.8$

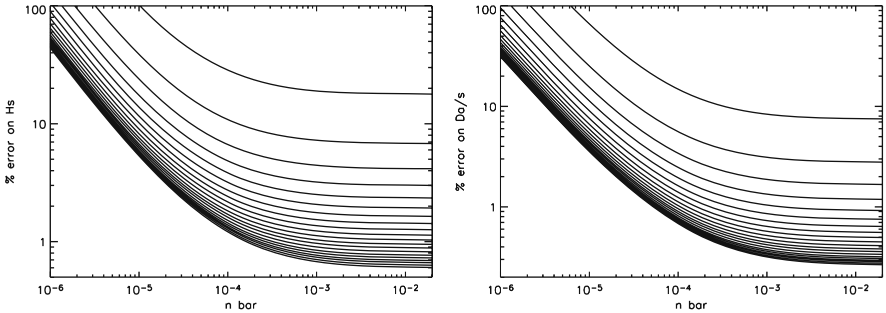

Fig. 8 Percent error on $H(z)r_s$ and D_a/r_s as a function of the galaxy number density of a BAO survey. This figure assumes full sky coverage $f_{sky} = 1$ (errors will scale like $1/\sqrt{f_{\rm sky}}$) and redshift range from $z = 0$ to $z = 2$ in bins of $\Delta z = 0.1$

at $z = 0$, $b(z = 0) = 1.5$ and we have assumed that $G(z)b(z) =$ constant. To interpret this figure note that with the chosen normalizations, $P(k)$ in real space at the BAO scale $k \sim 0.15$ h/Mpc is 6241(Mpc/h)^3, boosted up by large-scale redshift-space distortions to roughly 10^4(Mpc/h)^3 so $n = 10^{-4}$ corresponds to $nP(k = 0.15) = 1$. Note that the "knee" in this figure is therefore around $nP = 1$. This is where this "magic number" of reaching $nP \sim> 1$ in a survey comes from. Of course, there are other considerations that would tend to yield an optimal nP bigger than unity and of the order of few.

12 Model Testing

So far we have assumed a cosmological model characterized by a given set of cosmological parameters and used statistical tools to determine the best fit for these parameters and confidence intervals. However, the best fit parameters and confidence intervals depend on the underlying model, i.e., what set of parameters are allowed to vary. For example, the estimated value for the density parameter of baryonic matter Ω_b changes depending whether in a ΛCDM model the Universe is assumed flat or not (Fig. 9 right panel) or the recovered value for the spectral slope of the primordial power spectrum changed depending if the primordial power spectrum is assumed to be a power law or is allowed to have some "curvature" or "running" (Fig. 9 left panel). It would be useful to be able to allow the data to determine which combination of parameters gives the preferred fit to the data; this is the problem of *model selection*. Here we start by following [19] which is a clear introduction to the application of this subject in cosmology. Model selection relies on the so-called information criteria and the goal is to make an objective comparison of different *models* which may have a different number of parameters. The models considered in the example of Fig. 9 are "nested" as one model (the ΛCDM one) is completely specified by a sub-set of the parameters of the other

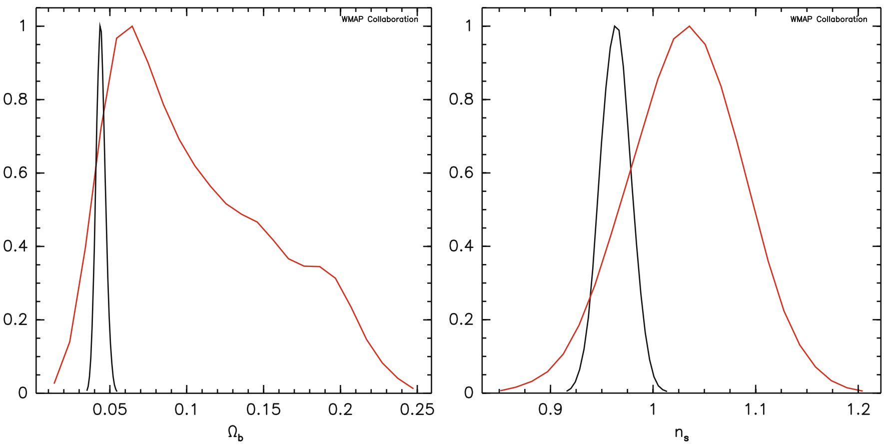

Fig. 9 Effect of the choice of the cosmological model in the recovered values for the parameters. Here we used WMAP5 data only: in both panels the *narrow curve* is for a standard flat ΛCDM model. In the *left panel* we show the posterior for Ω_b, the *broad curve* is for a non-flat ΛCDM model. In the *right panel* we show the posterior for n_s: the *broad curve* is for a ΛCDM model where the primordial power spectrum is not a perfect power law but is allowed to have some "curvature" also called "running" of the spectral index. Figure courtesy of LAMBDA [5]

(more general) model. In cosmology one is almost always concerned with nested models.

Typically the introduction of extra parameters will yield an improved fit to the data set, so a simple comparison of the maximum likelihood value will always favor the model with more parameters, regardless of whether the extra parameters are relevant. There are several different approaches often used in the literature. The simplest is the likelihood ratio test [20] see Sect. 6. Consider the quantity $2\ln[\mathcal{L}_{\rm simple}/\mathcal{L}_{\rm complex}]$ where $\mathcal{L}_{\rm simple}$ denotes the maximum likelihood for the model with less parameters and $\mathcal{L}_{\rm complex}$ the maximum likelihood for the other model. This quantity is approximately chisquare distributed and thus the considerations of Sect. 4 can be applied.

The Akaike information criterion (AIC) [21] is defined as $\mathrm{AIC} = -2\ln\mathcal{L} + 2k$ where $\mathcal{L}$ denotes the maximum likelihood for the model and k the number of parameters of the model. The best model is the one that minimizes AIC.

The Bayesian information criterion (BIC)[22] is defined as $\mathrm{BIC} = -2\ln\mathcal{L} + k\ln N$ where N is the number of data points used in the fit.

It should be clear that all these approaches tend to downweigh the improvement in the likelihood value for the more complex model with a penalty that depends on how complex is the model. Each of these approaches has its pros and cons and there is no silver bullet.

However, it is possible to place model selection on firm statistical grounds within the Bayesian approach by using the *Bayesian factor* which is the Bayesian evidence ratio (i.e., the ratio of probabilities of the data given the two models).

Recalling the Bayes theorem (2) we can write $\mathcal{P}(D) = \sum_i \mathcal{P}(D|M_i)\mathcal{P}(M_i)$ where i runs over the models M we are considering. Then the Bayesian evidence is

$$\mathcal{P}(D|M_i) = \int d\theta \mathcal{P}(D|\theta,M_i)\mathcal{P}(\theta|M_i), \tag{43}$$

where $\mathcal{P}(D|\theta,M_i)$ is the likelihood. Given two models (i and j), the Bayes factor is

$$B_{ij} = \frac{\mathcal{P}(D|M_i)}{\mathcal{P}(D|M_j)}. \tag{44}$$

A large B_{ij} denotes preference for model i. In general this requires complex numerical calculations, but for the simple case of Gaussian likelihoods it can be expressed analytically. The details can be found, e.g., in [23] and references therein. For a didactical introduction see also [24].

13 Monte Carlo Methods

With the recent increase in computing power, in cosmology we resort to the application of Monte Carlo methods ever more often. There are two main applications of Monte Carlo methods: Monte Carlo error estimations and Markov Chains Monte Carlo. Here I will concentrate on the first as there are several basics and detail explanations of the second (see e.g., [25] and references therein).

Let us go back to the issue of parameter estimation and error calculation. Here is the conceptual interpretation of what it means that an experiment measures some parameters (say cosmological parameters). There is some underlying true set of parameters $\theta_{\rm true}$ that are only known to Mother Nature but not to the experimenter. There true parameters are statistically realized in the observable Universe and random measurement errors are then included when the observable Universe gets measured. This realization gives the measured data D_0. Only D_0 is accessible to the observer (you). Then you go and do what you have to do to estimate the parameters and their errors (chisquare, likelihood, etc.) and get θ_0. Note that D_0 is not a unique realization of the true model given by $\theta_{\rm true}$: there could be infinitely many other realizations as hypothetical data sets, which could have been the measured one: $D_2, D_2, D_3 \ldots$ each of them with a slightly different fitted parameters θ_1, $\theta_2 \ldots \theta_0$ is one parameter set drawn from this distribution. The hypothetical ensemble of Universes described by θ_i is called ensemble, and one expects that the expectation value $\langle \theta_i \rangle = \theta_{\rm true}$. If we knew the distribution of $\theta_i - \theta_{\rm true}$ we would know everything we need about the uncertainties in our measurement θ_0. The goal is to infer the distribution of $\theta_i - \theta_{\rm true}$ without knowing $\theta_{\rm true}$. Here is what we do we say that hopefully θ_0 is not too wrong and we consider a fictitious world where θ_0 was the true one. So it would not be such a big mistake to take the probability distribution of $\theta_i - \theta_0$ to be that of $\theta_i - \theta_{\rm true}$. In many cases we know how to simulate $\theta_i - \theta_0$ and so we can simulate many synthetic realization of "worlds" where θ_0 is the true underlying model. Then mimic the observation process of these fictitious Universes replicating all the observational errors and effects and from each of these fictitious Universe estimate the parameters. Simulate enough of them and from $\theta_i^S - \theta_0$ (where S stands for "synthetic" or "simulated") you will be able to map the desired multi-dimensional probability distribution. With the advent of fast computers this technique has become increasingly widespread. As long as you believe you know the underlying distribution and that you believe you can mimic the observation replicating all the observational effects this technique is extremely powerful and, I would say, indispensable. This is especially crucial when complicated effects such as instrumental and or systematic effects can be simulated but not described analytically by a model.

14 Conclusions

I have given a brief overview of statistical techniques that are frequently used in the cosmological literature. I have presented several examples often from the literature to put these techniques into context. This is not an exhaustive list nor a rigorous treatment, but a starter kit to "get you started." As more and more sophisticated statistical techniques are used to make the most of the data, one should always remember that they need to be implemented and used correctly:

- data gathering is an expensive and hard task; statistical techniques make possible to make the most of the data
- always beware of systematic effects
- an incorrect treatment of the data will give non-sensical results
- there will always be things that are beyond the statistical power of a given data set

Remember: "Treat your data with respect!"

15 Some Useful References

There are many good and rigorous statistics books out there. In particular Kendall's advanced theory of statistics made of three volumes are

- *Distribution Theory* (Stuart and Ort 1994 [26])
- *Classical Inference* (Stuart and Ort 1991 [27]) and
- *Bayesian Inference* (O'Hagan 1994 [20]).

For astronomical and cosmological applications in many cases one may need a practical manual rather than a rigorous textbook. Although it is important to note that a practical manual is no substitute for a rigorous introduction to the subject.

- *Practical Statistics for Astronomers*, by Wall and Jenkins, (2003) is a must have [1].
- *Numerical Recipes* is also an indispensable "bible": Press et al. [2]

It also provides a guide to the numerical implementation of the "recipes" discussed. Complementary information to what presented here can be found in

- Verde, in XIX Canary Island Winter School "*The Cosmic Microwave Background: From Quantum Fluctuations to the Present Universe*" [25]. In the form of lecture notes, and
- Martinez, Saar, "*Statistics of the Galaxy Distribution*" [28], with a slant on large scale structure and data analyis in cosmology, Martinez, Saar, Martinez-Gonzalez, Pons-Porteria, Lecture Notes in Physics 665, Springer, 2009

Acknowledgments LV is supported by FP7- PEOPLE-2002IRG4-4-IRG#202182 and MICINN grant AYA2008-03531. I acknowledge 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

1. J. P. Wall, and C. R. Jenkins, *Practical Statistics for Astronomers*, (Cambridge University Press, Cambridge, 2003). 151, 175
2. W. H. Press et al., *Numerical Recipes: The Art of Scientific Computing*, Cambridge University Press, Cambridge, 1992). 152, 153, 175
3. M. Kowalski, et al., Astronphys. J. **686**, 749 (2008). 152, 153
4. W. J. Percival, et al., Astronphys. J. **657**, 645 (2007). 152
5. Legacy Archive for Microwave Background Data Analysis http://lambda.gsfc.nasa.gov/ cosmological parameters plotter. 158, 161, 173
6. J. Hamann, S. Hannestad, G. Raffelt and Y. Y. Wong, JCAP **0708**, 021 (2007). 158, 159
7. W. Cash, Astronphys. J. **228**, 939 (1979). 159
8. D. N. Spergel, et al., Astron Phys. J. Suppl. **170**, 377 (2007). 160, 161
9. D. N. Spergel, et al., Astron Phys. J. Suppl. **148**, 175 (2003). 160
10. R.A. Fisher, J. Roy. Stat. Soc. **98**, 39 (1935). 162
11. A. Kosowsky, M. Milosaljevic and R. Jimenez, Phys. Rev. **D66**, 063007 (2002). 163
12. A. Refregier, A. Amara, T. Kitching, and A. Rassat, arXiv:0810.1285 (2008). 164
13. L. Verde, H. Peiris, and R. Jimenez, JCAP **0601**, 019 (2006). 165
14. D. J. Eisenstein, et al. Astronphys. J. **633**, 574 (2005). 166
15. S. Cole et al., MNRAS, 362, 505 (2005). 166
16. W. Percival, et al. Astronphys. J. **657**, 645 (2007). 166
17. H. Seo, and D. J. Eisenstein, Astronphys. J. **598**, 720 (2003). 167
18. H. Seo and D. J. Eisenstein, Astronphys. J. **664**, 679 (2007). 169, 170
19. A. R. Liddle, Mon. Not. Roy. Astron. Soc. **351**, L49–L53 (2004). 172
20. A. O'Hagan, *Kendall's Advanced theory of statistics, Volume 2b, Bayesian Inference*, (Arnold, Huntinston Beach, 1992). 173, 175
21. H. Akaike, IEEE Trans. Auto. Contol. **19**, 716 (1974). 173
22. G. Schwarz, Ann. Stat. **5**, 461 (1978). 173
23. A. F. Heavens, T. D. Kitching, and L. Verde, Mon. Not. Roy. Astron. Soc. **380**: 1029–1035 (2007). 174
24. A. F. Heavens, Lectures given at the "Francesco Lucchin" School of Astrophysics, Bertinoro, Italy, 25–29 (May 2009), arXiv:0906.0664 174
25. L. Verde, XIX Canary Island Winter School "*The Cosmic Microwave Background: From Quantum Fluctuations to the Present Universe*," (Cambridge University press, Cambridge 2009) 174, 176

26. A. Stuart and J. K. Ord, *Kendall's Advanced Theory of Statistics, Volume 1, Distribution Theory*, (Arnold, Huntinston Beach, 1924). 175
27. A. Stuart and J. K. Ord, *Kendall's Advanced theory of statistics, Volume 2a, Classical Inference and Relationship*, (Arnold, Huntinston Beach, 1992). 175
28. V. J. Martinez, E. Saar, O.Lahav, and *Statistics of the Galaxy Distribution*, (Chapman & Hall, New York, 2002). 176