AstroStatistics and Cosmology

Jacopo Tissino

2020-12-21

Contents

Introduction

In this course we will discuss

1. Bayesian statistics for parameter estimation and model comparison;

2. Application of these to practical data analysis problems in cosmology.

The goal of the first part will *not* be on mathematical proofs, but on *applications*. We are "customers" of statistics. We need a good understanding of the theory, but not necessarily a very *formal* one.

However, we will not take the "cookbook" approach: we need to understand the statistics in depth, blindly applying a technique is bad.

Lectures from the fifth of October will be in room P1C in the Paolotti building. We have the handwritten notes from the professor, and LaTeX notes from students of the earlier years.

Email: michele.liguori@unipd.it, or liguori.unipd@gmail.com (it's the same).

There will be **homework**. Some exercises to solve and other things. We should hand it in within 3 weeks of the assignment, this is not strict, but we should let him know if we cannot do it in time. The final homework will require some coding, making some Monte Carlo Markov chains, and it will not have the deadline since coding takes time. We should hand it in before the exam.

We can choose whatever programming language we want (Python is good, C++ is slightly worse since the professor is not so familiar with it, but it's fine). We should have a summary of the results with plots, and show the source code.

The exam is an oral, to do whenever we want. When we contact him we will be given a journal paper to read. At the exam, we will do a blackboard presentation of the paper and be asked questions about it, like in journal club.

A book which does things similarly to this course is "Data Analysis: a Bayesian tutorial" by Silvia and Skilling [SS06].

Usual COVID safety procedure if we come to class physically. There are 23 seats available. Of course, we can also follow the lectures online.

Chapter 1

Bayesian statistics

1.1 Inference

We need to apply inductive reasoning, since deductive reasoning cannot work in real life, since we cannot know anything with certainty.

We apply reasoning in the form: "if A, then B is more plausible" and "we see evidence for B": so, A is more plausible. If "we see evidence for $\neg B$", instead, then A is less plausible.

We then need to establish clear mathematical rules for this plausible reasoning.

1.1.1 Cox's theorem

This theorem gives constraint on our system of 'probability', by which we mean a function which associates a real number to each 'proposition/hypothesis', by which we mean a possible state of the world. We want the probability of a certain event to be 1, and the probability of an impossible event to be 0. The probability is denoted as $\mathbb{P}$, and we interpret it as a **degree of belief**.

We want the rules we use to be self-consistent:

1. if $\mathbb{P}(A) > \mathbb{P}(B)$ and $\mathbb{P}(B) > \mathbb{P}(C)$, then $\mathbb{P}(A) > \mathbb{P}(C)$;

2. for any events, $\mathbb{P}(A \text{ and } B) = \mathbb{P}(A|B)\mathbb{P}(B)$, where the notation $\mathbb{P}(A|B)$ denotes the probability of A, *given that B* has happened;

3. starting from the same information, we must arrive at the same conclusions.

Find examples of inconsistency if these are not verified.

Maybe write point 2 in a more verbose way...

If these hold, then we have the **Kolmogorov axioms**:

$$\mathbb{P}(X|I) + \mathbb{P}(\neg X|I) = 1 \qquad \text{and} \qquad \mathbb{P}(AB|I) = \mathbb{P}(A|B, I)\mathbb{P}(B|I). \qquad (1.1.1)$$

We write the probabilities as conditioned on *preexisting knowledge I*. This quickly becomes annoying in the notation, so I will stop writing it, but it is always implied: we never discuss probabilities in a vacuum.

A corollary of the second rue is **Bayes' theorem**:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}. \tag{1.1.2}$$

The procedure for hypothesis testing will look like

$$\mathbb{P}(\text{hypothesis}|\text{data}) = \frac{\mathbb{P}(\text{data}|\text{hypothesis})\mathbb{P}(\text{hypothesis})}{\mathbb{P}(\text{data})}. \tag{1.1.3}$$

The key difference from the frequentist approach is that, while there the parameters have certain fixed values, here we can describe our *belief* about their values through a probability distribution.

The things we will want to do can be classified into

1. **hypothesis testing**, "are CMB data consistent with gaussianity?";

2. **parameter estimation**, "what is the value of the mass of the Sun?";

3. **model selection**, "is GR the correct theory of gravity?".

1.1.2   Parameter estimation

We start with an example: the toss of a coin. The question is: we toss it N times and get R heads. Is it a fair coin?

If $H \in [0,1]$ is the probability of getting heads in a single coin flip, then

$$\mathbb{P}(R \text{ heads}|H, I) \propto H^R (1-H)^{N-R}. \tag{1.1.4}$$

Here, I is the other information we have about the coin: the fact that every throw is independent, the fact that there are no outcomes beyond heads or tails.

This is the **likelihood**, what we want to do is to invert the relation, finding a probability density function for H given the data. The probability $\mathbb{P}(\text{data})$, also called the **evidence**, is not something we need to calculate when doing parameter estimation: we can just write

$$\mathbb{P}(\text{parameters}|\text{data}) \propto \mathbb{P}(\text{data}|\text{parameters})\mathbb{P}(\text{parameters}), \tag{1.1.5}$$

since we are computing probability density functions, which need to be normalized in order to make sense. This is a useful parameter estimation toy problem, since we only have one parameter to estimate.

In order to use the formula we need a prior, $\mathbb{P}(\text{hypothesis})$. This is hard in general, and it depends on the problem. We might want a prior which is peaked around 0.5 for a regular coin, if we have no reason to think that it is unfair. Let us suppose we have doubts about the honesty of who gave us the coin: then, we might want a noninformative prior, like a flat one. Let us suppose we are in this case: $\mathbb{P}(\text{parameters}) = \text{const}$.

Since the prior is flat, the posterior is proportional to the likelihood:

$$\mathbb{P}(H|R \text{ heads}, I) \propto \mathbb{P}(R \text{ heads}|H, I) \propto H^R (1-H)^{N-R}. \tag{1.1.6}$$

4

We can simulate this experiment! We need a binomial random number generator.

> **Do the simulation!**

As N increases, the posterior "zeroes in" onto the correct value. If we split the N simulated throws in two, and use the posterior from the first batch as a prior for the second, we get the same result!

A completely agnostic prior in the coin-flip example would be flat.

As we toss many times, the prior becomes less and less relevant.

> **As long as we never set it to zero!**

Where is objectivity, if we include our beliefs in the analysis? There are fundamentalist bayesians and frequentists. In cosmology the Bayesian approach is best since we only have one realization of the universe.

Also, we want to include previous experience in our analysis. We can take an objective approach to priors, by selecting a *noninformative* one. These are *objective*, in the sense that they express the *a priori* ignorance about the process.

After our analysis in parameter estimation we find a PDF, which contains everything we need, but we want to give a number in our paper in order to summarize the result. This does not give us *more* information than the PDF.

A common approach, which works as long as the posterior is unimodal, is to give the maximum of the posterior: if the parameter is x, then the central value x_0 is calculated by setting

$$\left.\frac{\mathrm{d}P}{\mathrm{d}x}\right|_{x_0} = 0. \tag{1.1.7}$$

This is called *Maximum A-Posteriori parameter estimation*, MAP.

How do we provide error bars? The precision becomes higher as the peak becomes narrower. Let us consider the log of the posterior, $L = \log P$. This is typical, since the log is monotonic, and it is convenient since P is typically very "peaky".

Around the maximum, L is given by

$$L = \log \mathbb{P}(x|\text{data}, I) \tag{1.1.8}$$

$$= L(x_0) + \left.\frac{\mathrm{d}L}{\mathrm{d}x}\right|_{x_0}(x - x_0) + \frac{1}{2}\left.\frac{\mathrm{d}^2 L}{\mathrm{d}x^2}\right|_{x_0}(x - x_0)^2 + \mathcal{O}\left((x - x_0)^3\right). \tag{1.1.9}$$

If the peak is well-behaved, then we can approximate it at second order in a large enough region around the maximum: this is

$$P = \exp(L) \approx \underbrace{P_0}_{e^{L(x_0)}} \exp\left(\frac{1}{2}\left.\frac{\mathrm{d}^2 L}{\mathrm{d}x^2}\right|_{x_0}(x - x_0)^2\right), \tag{1.1.10}$$

which is a Gaussian whose variance is

$$\sigma^2 = -\left(\left.\frac{\mathrm{d}^2 L}{\mathrm{d}x^2}\right|_{x_0}\right)^{-1}, \tag{1.1.11}$$

which will be positive: if x_0 is a maximum the second derivative is negative.

> So the expansion in L is justified a posteriori through the Central Limit Theorem?

There will be a theorem telling us that the posterior converges to a Gaussian, and the estimate will converge to the maximum likelihood estimate.

Typically we have many parameters, not just one. Even if the theory only has a few, the experiment will also have several.

Integration is difficult in multidimensional contexts, we want to have something better than exponential time in the parameter number.

Suppose that our posterior is very asymmetric. Then, it might be better to give the mean of the posterior instead of the maximum:

$$\langle x \rangle = \int x \mathbb{P}(x|\text{data}, I) \, dx \, . \tag{1.1.12}$$

If there is symmetry, this is similar to the maximum. Quoting both if there is asymmetry might be good.

We then want to build a **credible interval**, which we define as the *shortest* interval $[x_1, x_2]$ containing the representative value we choose, say $\langle x \rangle$, and which integrates to a certain chosen value, often chosen to be 95 %:

$$\int_{x_1}^{x_2} \mathbb{P}(x|\text{data}, I) = 0.95 \, . \tag{1.1.13}$$

Let us apply this procedure to the coin toss problem. Recall that the posterior PDF, with a flat prior, was given by

$$P = \mathbb{P}(H|\text{data}, I) \propto H^R (1 - H)^{N-R} \, . \tag{1.1.14}$$

The derivative is given by

$$\frac{dL}{dH} = \frac{d \log P}{dH} = \frac{d}{dH} \left[R \log H + (N - R) \log(1 - H) \right] \tag{1.1.15}$$

$$= \frac{R}{H} - \frac{N - R}{1 - H} \, , \tag{1.1.16}$$

which we set to zero: this yields

$$\frac{R}{H_0} = \frac{N - R}{1 - H_0} \implies R - H_0 N = 0 \, , \tag{1.1.17}$$

so $H_0 = R/N$. This is what we get in the end, when we have many data.

The errorbar can be found by differentiating again:

$$\frac{d^2 L}{dH^2} = -\frac{R}{H^2} - \frac{N - R}{(1 - H)^2} \tag{1.1.18}$$

$$= \frac{R(2H - 1) - NH^2}{H^2(1 - H)^2} \, , \tag{1.1.19}$$

6

so we can find the errorbar by computing it in $H = R/N$: skipping a few steps, it is

$$\sigma^2 = \frac{(R/N)^2(1 - R/N)^2}{N(R/N)^2 - R(2R/N - 1)} = \frac{H_0(1 - H_0)}{N} \tag{1.1.20}$$

$$\sigma = \sqrt{\frac{H_0(1 - H_0)}{N}}. \tag{1.1.21}$$

As is expected, this scales like $1/\sqrt{N}$. The reason we're doing this with the full Bayesian machinery is that the procedure will not yield these simple results in general.

Let us solve a probability problem using Bayesian statistics: the Monty Hall problem.

There are three doors, one of which is desirable, the other two are not. We choose one door; the host knows which the desirable door is, he excludes a door as undesirable and asks us whether we want to change our choice.

Coming back to the Monty Hall problem. We picked A, the host picked C.

Let us denote the presence of the desirable object with 1, 0 its absence. So, we have three options: $(A, B, C) = (1, 0, 0)$, or $(0, 1, 0)$, or $(0, 0, 1)$. *A priori*, we assign a probability of $1/3$ to each: this is, then, our prior. Let us assume that WLOG A is the door we picked. This has probability 1, since we can make it true in any case by relabeling the doors.

We want to compute

$$\mathbb{P}(B = 1 | \text{host picked } C, \text{ we picked } A). \tag{1.1.22}$$

We will write the condition as $[C]$ for compactness. The complement to this probability will be

$$\mathbb{P}(A = 1 | [C]). \tag{1.1.23}$$

We can then apply Bayes' theorem:

$$\mathbb{P}(B = 1 | [C]) = \frac{\mathbb{P}([C] | B = 1) \mathbb{P}(B = 1)}{\mathbb{P}([C])}. \tag{1.1.24}$$

$\mathbb{P}(B = 1) = 1/3$ is our prior. Now, if $B = 1$ and we chose A, then the host is *forced* to pick C since otherwise he will uncover the prize: $\mathbb{P}([C] | B = 1) = 1$.

Then, what we are left with is the computation of $\mathbb{P}([C])$.

We can write this through *marginalization*, integrating over all the possible events which can happen:[1]

$$\mathbb{P}([C]) = \underbrace{\mathbb{P}([C] | A = 1)}_{= 1/2} \mathbb{P}(A = 1) + \underbrace{\mathbb{P}([C] | B = 1)}_{= 1} \mathbb{P}(B = 1) + \underbrace{\mathbb{P}([C] | C = 1)}_{= 0} \mathbb{P}(C = 1)$$

$$\tag{1.1.25}$$

---

[1] We assume that, if we selected the good door, the host chooses uniformly between the two: that is, $\mathbb{P}([C] | A = 1) = x = 1/2$. A host can actually be biased towards C or B, maybe he will choose the nearest door to him or something like that. In any case, if we leave x as a variable the final probability to find the prize by switching is found to be $1/(1 + x)$, which is always larger than $1/2$: we are always better off switching. An interesting fact, however, is that by changing x the probability can move from $1/2$ to 1.

$$= \frac{3}{2} \times \frac{1}{3} = \frac{1}{2}. \tag{1.1.26}$$

Another example. This is an investigation. The probability of anybody in a neighborhood to die of overdose is $\mathbb{P}(O) = 1/2$. Also, 30 % of murder victims are drug addicts:

$$\mathbb{P}(\text{addict}|\text{murder victim}) = 0.3. \tag{1.1.27}$$

We want to compute the probability that someone who was an addict was indeed murdered, and did not die of overdose: $\mathbb{P}(O|A)$: this will be given by

$$\mathbb{P}(O|A) = \frac{\mathbb{P}(A|O)\mathbb{P}(O)}{\mathbb{P}(A)}. \tag{1.1.28}$$

We do not have $\mathbb{P}(A|O)$, but we can estimate it, and then try to understand how much it affects the final result. Let us start out by estimating it as 0.9, since overdoses will likely most often happen to addicts.

We can calculate the probability of being an addict by marginalizing over the cause of a drug-induced death:

$$\mathbb{P}(A) = \mathbb{P}(A|M)\mathbb{P}(M) + \mathbb{P}(A|O)\mathbb{P}(O), \tag{1.1.29}$$

and since $\mathbb{P}(O) = 0.5$, we can also have $\mathbb{P}(M) = 0.5$. This then means that $\mathbb{P}(A) = 0.6$. If there were other possible events, we would need to sum over them.

With all of this, we have

$$\mathbb{P}(O|A) = 0.75. \tag{1.1.30}$$

If we were to change $\mathbb{P}(A|O)$ this would not change much, so it is ok to estimate this roughly.

Let us discuss marginalization in some more detail. We typically do it for all the "nuisance parameters", which we must account for but do not really care about in the end: typically, parameters connected to experimental noise.

Suppose we have a PDF like

$$\mathbb{P}(X, Y), \tag{1.1.31}$$

where Y can take values in the set of *exhaustive* and *mutually exclusive* events Y_k. Marginalization is the process of computing $\mathbb{P}(X)$ through

$$\mathbb{P}(X) = \sum_{k=1}^{N} \mathbb{P}(X, Y_k) = \sum_{k=1}^{N} \mathbb{P}(X|Y_k)\mathbb{P}(Y_k) \tag{1.1.32}$$

$$= \underbrace{\sum_{k=1}^{N} \mathbb{P}(Y_K|X)}_{=1} \mathbb{P}(X). \tag{1.1.33}$$

Thus, we have shown that $\mathbb{P}(X)$ is indeed given by the expression above. It is crucial to assume that the Y_k are exhaustive and mutually exclusive in order for the sum of $\mathbb{P}(Y_k|X)$ to equal 1.

8

Typically, the events are not discrete but continuous. Exhaustivity and exclusivity can still apply, however we need to turn the sum into an integral.

We are considering probability density functions of these continuous parameters: they are defined as

$$\frac{\mathrm{d}p}{\mathrm{d}y} = \lim_{\delta y \to 0} \frac{\mathbb{P}(X, y \leq Y \leq y + \delta y)}{\delta y}. \tag{1.1.34}$$

Then, we can compute the finite probability of X as

$$\mathbb{P}(X) = \int_{\mathbb{R}} \mathrm{d}Y \, \frac{\mathrm{d}p}{\mathrm{d}y}. \tag{1.1.35}$$

If we integrate keeping Y in a certain region we find the probability of finding a value for it in that range.

We now set up the problem for tomorrow: we do parameter estimation. We have a set of measurements of the same quantity: $\vec{d} = \{x_i\}$. Each of the measurements is affected by error, and by the "experimentalist's Central Limit Theorem" their sum will resemble a Gaussian.

Each measurement x_i will then be given by $x_i = \mu + n_i$, the mean value plus a noise term. In this example, we assume that n_i is Gaussian. We want to write a posterior distribution for μ: it will be given in terms of the likelihood of the data given the true value, assuming that σ^2 is a fixed and known quantity,

$$\mathscr{L}(x_i | \mu) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2}\right). \tag{1.1.36}$$

This is the likelihood from a single value x_i, while if we want to compute the joint likelihood of all the data $\{x_i\}$ it is harder. We can assume, in this specific case, that the errors are independent: therefore, the probability factors and we can write

$$\mathbb{P}(\{x_i\} | \mu) = \prod_i \mathbb{P}(x_i | \mu) = \frac{1}{\sigma^N \sqrt{2\pi}^N} \exp\left(-\frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2\right). \tag{1.1.37}$$

We come back to our noisy dataset $x_i = \mu + n_i$. If our noise comes from a Gaussian with a *known*, *constant* σ, then we have

$$\mathscr{L}(x_i | \mu, \sigma) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(x_i - \mu)^2}{\sigma^2}\right). \tag{1.1.38}$$

We also assume that the realizations of the noise are independent. So, the combined likelihood for all the data is given by

$$\mathscr{L}(\vec{x} | \mu, \sigma) = \prod_{i=1}^{N} \mathscr{L}(x_i, \mu, \sigma). \tag{1.1.39}$$

Now, choosing a prior as we mentioned is difficult in general. Until now we have assumed that our noninformative prior would be constant. For the type of parameter we

have now — μ is a *location* parameter — works; however we have an issue: a uniform distribution must have a certain range in which it is nonzero, elsewhere it is zero. This range must be finite.

Zero cannot be updated: this is a problem. We can, however take a prior that is so wide that it is nonzero in any place the likelihood is measurably nonzero. Practically speaking we only work up to finite precision, so this is not a problem.

We only case about proportionality: the posterior is proportional to the likelihood, so

$$P \propto \prod_{i=1}^{N} \exp\left(-\frac{1}{2}\frac{(x_i - \mu)^2}{\sigma^2}\right), \tag{1.1.40}$$

therefore the log-posterior is

$$L = \log P = -\frac{1}{2\sigma^2}\sum_{i=1}^{N}(x_i - \mu) + \text{const}, \tag{1.1.41}$$

so if we want to find the maximum (log)-posterior:

$$\frac{\mathrm{d}L}{\mathrm{d}\mu} = \sum_{i=1}^{N}\frac{x_i - \mu}{\sigma^2} = 0, \tag{1.1.42}$$

meaning that

$$N\mu = \sum_{i=1}^{N} x_i \implies \hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i. \tag{1.1.43}$$

We can also compute the estimate of the error on the estimate:

$$\frac{\mathrm{d}^2 L}{\mathrm{d}\mu^2} = -\sum_{i=1}^{N}\frac{1}{\sigma^2} = -\frac{N}{\sigma^2}, \tag{1.1.44}$$

therefore

$$\sigma_\mu = \left(-\frac{\mathrm{d}^2 L}{\mathrm{d}\mu^2}\right)^{-1/2} = \frac{\sigma}{\sqrt{N}}. \tag{1.1.45}$$

The great simplification we made here was that we are dealing with only one parameter. In any realistic astrophysics scenario we have at least a dozen.

1.2 Multiparameter estimation

1.2.1 A two-parameter example

We consider a photon counting experiment. We have a diffraction experiment: we can measure light in M frequency channels, labelled by k, and at different angles.

Our model (we need one for parameter estimation) is that for each frequency channel k we expect a Gaussian spatial profile centered around x_0 — for simplicity we assume x_0

is known, and that the standard deviation w is known as well. We are interested in the amplitude A of this Gaussian peak.

Also, we will have spatially constant noise with amplitude B. We need to estimate A and B jointly; B is a *nuisance parameter* we would ideally integrate over later.

We expect that the measured data at a frequency channel will look like

$$D_k = n_0 \left[A \exp\left(-\frac{(x_k - x_0)^2}{2w^2} + B \right) \right]. \tag{1.2.1}$$

The parameter n_0 accounts for the fact that if we measure for longer we see more photons.

Photons, both noise and signal, are expected to obey Poissonian statistics:

$$\mathbb{P}(N_k|D_k) = \frac{D_k^{N_k} e^{-D_k}}{N_k!}, \tag{1.2.2}$$

which is the likelihood: the probability of the data, given our model. Our actual likelihood will be given by $\mathscr{L} = \prod_k \mathbb{P}(N_k|D_k)$.

What is our prior $\mathbb{P}(A, B|I)$? We know that these parameters cannot be negative, so we just take a uniform bivariate prior on $A \geq 0$, $B \geq 0$. As before, this is just meant to say "uniform over all representable values".

We will discuss later that by the nature of A as a scale parameter it would be more reasonable to take $\log A$ to be uniform.

The posterior will then be

$$\mathbb{P}(A, B|N_k, I) \propto \prod_k \frac{D_k^{N_k} e^{-D_k}}{N_k!}. \tag{1.2.3}$$

This then depends on two parameters, and we can maximize it numerically. It is a difficult problem computationally, but not conceptually.

How can we model this in order to give an error? We can approximate it as a bivariate Gaussian.

In general, a multivariate Gaussian is given as

$$\mathbb{P}(\vec{x}) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\vec{x} - \vec{\mu})^\top \Sigma^{-1}(\vec{x} - \vec{\mu}) \right), \tag{1.2.4}$$

where Σ is the **covariance matrix** of the two variables. On the diagonal we have the variance of each variable: $\Sigma_{ii} = \sigma_i^2$, so $(\Sigma^{-1})_{ii} = 1/\sigma_i^2$ as long as the covariance (meaning, the nondiagonal elements) is zero.

Homework: we have on Moodle a sheet with the same homework as last year. This year's will likely be quite similar, although a couple exercises might change.

Coming back to two-parameter estimation: what is interesting now is to give error-bars on our parameters.

If our posterior for a single parameter is a Gaussian, then the errorbar for that parameter will be the σ for that Gaussian. We can compute the σ through the FWHM or any other height-interval.

For a multivariate Gaussian the reasoning is the same. We can intercept the distribution with a plane at a given height, and project the interception (which for a perfect Gaussian will be an ellipse) onto the plane of the two variables.

We can thus find the region corresponding to a 68 %, 95 % or whatever percent probability contained inside: a credible region.

We have a probability distribution $P = \mathbb{P}(x, y | \vec{D}, I)$, where $\vec{D}$ is the data vector while I represents previous information. We usually work with $L = \log P$.

The MAP estimate is the pair (x_0, y_0) such that

$$\vec{\nabla} L \Big|_{(x_0, y_0)} = \vec{0}. \tag{1.2.5}$$

We can then Taylor expand in two variables: the first derivative vanishes. Let us denote the parameter vector as $\vec{t} = (x, y)$

$$L(x, y) = L(x_0, y_0) + \frac{1}{2} \frac{\partial^2 L}{\partial x^2}\Big|_{\vec{t}_0} (x - x_0)^2 + \frac{1}{2} \frac{\partial^2 L}{\partial y^2}\Big|_{\vec{t}_0} (y - y_0)^2 + \frac{\partial^2 L}{\partial y \partial x}\Big|_{\vec{t}_0} (x - x_0)(y - y_0)$$

$$\tag{1.2.6}$$

$$= L(x_0, y_0) + \frac{1}{2}(\vec{t} - \vec{t}_0)^\top H\Big|_{\vec{t}_0} (\vec{t} - \vec{t}_0), \tag{1.2.7}$$

where H is the Hessian matrix, $H_{ij} = \partial_i \partial_j L$. Then, the probability reads

$$P = \exp(L) = \text{const} \times \exp\left[\frac{1}{2}(\vec{t} - \vec{t}_0)^\top H\Big|_{\vec{t}_0} (\vec{t} - \vec{t}_0)\right]. \tag{1.2.8}$$

This is precisely a multivariate Gaussian, with a covariance matrix Σ such that $H = -\Sigma^{-1}$.

Let us denote $Q = (\vec{t} - \vec{t}_0)^\top H\big|_{\vec{t}_0} (\vec{t} - \vec{t}_0)$.

This is the argument of the exponential, so having constant P means having constant L means having constant Q. Note that H is both *negative* definite and symmetric, since the covariance matrix must be positive definite and symmetric.

The matrix H will have two eigenvectors: these define the principal axes, knowing which we will be able to draw the contours. If we impose $Q = k$ for some real number k, the lengths of the eigenvectors $\vec{e}_i$ will satisfy $|\vec{e}_i| = k/\lambda_i$, where λ_i are the corresponding eigenvalues.

If $H_{12} = 0$, then the principal axes are aligned with the coordinate axes: the parameters are *uncorrelated*.

How do we compute an errorbar for a **marginalized** parameter? We start by marginalizing:

$$\mathbb{P}(x) = \int \exp(L) \, dy, \tag{1.2.9}$$

12

which will be a function of x only. The errorbar is then given by $\sigma_x^2 = \left\langle (x - x_0)^2 \right\rangle = \int \mathbb{P}(x)(x - x_0)^2 \, \mathrm{d}x$. So, the final expression is given by

$$\sigma_x^2 = \int (x - x_0)^2 \int \exp(L) \, \mathrm{d}y \, \mathrm{d}x \, . \tag{1.2.10}$$

For a bivariate Gaussian, this yields $\sigma_x^2 = (-H^{-1})_{11}$. Note that there is a big difference between $(H_{11})^{-1}$ and $(H^{-1})_{11}$.

We can also compute the covariance:

$$\mathrm{Cov}(x, y) = \left\langle (x - x_0)(y - y_0) \right\rangle = \int (x - x_0)(y - y_0) \exp(L) \, \mathrm{d}x \, \mathrm{d}y = -\left(H^{-1}\right)_{12}. \tag{1.2.11}$$

Now, let us discuss how $\sigma_x = \sqrt{(-H^{-1})_{11}}$ changes with the other parameters: in this bivariate case we can write it explicitly, it comes out to be

$$\sigma_x = \sqrt{\frac{-H_{yy}}{\det H}} \tag{1.2.12}$$

$$= \sqrt{\frac{-H_{yy}}{H_{xx}H_{yy} - H_{xy}^2}} \, . \tag{1.2.13}$$

If the two parameters are uncorrelated, this simplifies to $\sigma_x = \sqrt{-1/H_{xx}}$: the same as the one-parameter case, since the Hessian is minus the inverse of the covariance matrix. This is expected: if the parameters are uncorrelated the probability factors.

If, instead, we have correlation then the error on x increases.

If the determinant vanishes, the single-parameter errorbar diverges: the parameters are completely *degenerate*. This happens when the effect of the noise is undistinguishable from the effect of the signal. We can, though, often perform a change of variables in order to constrain some combination of the parameters.

The CMB has a temperature average of $\left\langle T(\hat{n}) \right\rangle \approx 2.725\,\mathrm{K}$, but there are anisotropies. These can be expanded in spherical harmonics, whose coefficients encode all the information: these represent the *angular power spectrum*.

The coefficients we compute are $\left\langle |a_{\ell m}|^2 \right\rangle = C_\ell$.

> Is there a sum there?

The problem is that the C_ℓ are sensitive to cosmological parameters. The features of the spectrum we change are different depending on which parameter we change. Changing Ω_{0b} raises the odd-numbered peaks and lowers the even-numbered ones, Ω_{0m} raises the third peak compared to the second...

We start off with a power spectrum $P(k) = Ak^n$, and then we have oscillations. Changing A has the effect of raising the whole spectrum.

After recombination there is *reionization*, when the first stars form. At $z \sim 6 \div 10$ the universe becomes ionized again, so there are free electrons messing up the CMB. This damps the primordial fluctuations.

The optical depth due to reionization is denoted as τ. Therefore, the parameters A and τ are very degenerate! The degeneracy is not complete, since τ is causal: it can not affect the first 10-so multipoles, while A can.

What must be done is finding some observable which breaks the degeneracy: one option is looking at CMB polarization. τ has a polarizing effect, A does not.

1.3 The multivariate gaussian

This will be a once-in-a-lifetime calculation, we need the result practically but it is good to have seen the derivation once.

The Multi-Variate Normal is in general

$$\mathcal{N}(\vec{x}, \vec{\mu}, C) = \frac{1}{(2\pi)^{n/2}\sqrt{\det C}} \exp\left(-\frac{1}{2}(\vec{x}-\vec{\mu})^\top C^{-1}(\vec{x}-\vec{\mu})\right). \tag{1.3.1}$$

Gaussians have many applications: with lots of data points, both the likelihood and the posterior converge to Gaussians with the same mean. We have many analytical results about Gaussians.

Normalization We define $V = C^{-1}$, the precision matrix. This is diagonalized as $O^\top V O$, where Λ is diagonal with eigenvalues λ_k and O is orthogonal. Then, the integral of the exponential $\exp\left(-\vec{y}^\top V \vec{y}/2\right)$ can be expressed as

$$\int d^n y \exp\left(-\frac{1}{2}\vec{y}^\top O^\top V O \vec{y}\right) \det O = \int d^n y \exp\left(-\frac{1}{2}\lambda_k y_k^2\right) \tag{1.3.2}$$

$$= \prod_{k=1}^{n} \int dy_k \exp\left(-\frac{1}{2}\lambda_k y_k^2\right) \tag{1.3.3}$$

$$= \prod_{k=1}^{n} \sqrt{\frac{2\pi}{\lambda_k}} = \frac{(2\pi)^{n/2}}{\sqrt{\det \Lambda}} = (2\pi)^{n/2}\sqrt{\det C}. \tag{1.3.4}$$

Marginalization We have an n-variate MVN, which we want to marginalize over M parameters: the integral we want to perform is

$$\int dx_{n-M+1}\ldots dx_n \frac{1}{(2\pi)^{n/2}\sqrt{\det C}} \exp\left(-\frac{1}{2}\vec{y}^\top V \vec{y}\right). \tag{1.3.5}$$

We partition the precision matrix into M and $n-M$-dimensional blocks:

$$V = \begin{bmatrix} V_{aa} & V_{ab} \\ V_{ab} & V_{bb} \end{bmatrix}. \tag{1.3.6}$$

Note that there is no one-to-one correspondence with the inverse matrix: in general $V_{aa} \neq C_{aa}^{-1}$. The argument of the exponential can be written, with the notation $\vec{y} = \vec{x} - \vec{\mu}$:

$$-\frac{1}{2}\vec{y}^\top V \vec{y} = \vec{y}_a^\top V_{aa} \vec{y}_a + \vec{y}_b^\top V_{bb} \vec{y}_b + 2\vec{y}_a^\top V_{ab} \vec{y}_b. \tag{1.3.7}$$

14

We then use the square completion formula:

$$\frac{1}{2}\vec{z}^\top A\vec{z} + \vec{b}^\top\vec{z} + c = \frac{1}{2}\left(\vec{z} + A^{-1}\vec{b}\right)^\top A\left(\vec{z} + A^{-1}\vec{b}\right) - \frac{1}{2}\vec{b}^\top A^{-1}\vec{b} + c\,, \qquad (1.3.8)$$

which generalizes the procedure used to solve second-degree equations. We identify $A = V_{bb}$, $\vec{z} = \vec{y}_b$, $\vec{b} = V_{ab}^\top\vec{y}_a$, $c = \vec{y}_a V_{aa}\vec{y}_a/2$.

> Check calculation! something is wrong here

Finally, we get that the marginalized distribution is still Gaussian, and is written as

$$\mathcal{N}(\vec{x}_a|\vec{\mu}_a, (V_{aa} - V_{ab}V_{bb}^{-1}V_{ab})^{-1})\,. \qquad (1.3.9)$$

The new covariance can be expressed more simply after some matrix algebra (using the Woodbury formula for block matrix inversion): we can show that

$$C_{aa} = \left(V^{-1}\right)_{aa} = \left(V_{aa} - V_{ab}V_{bb}^{-1}V_{ab}\right)^{-1}\,, \qquad (1.3.10)$$

so the marginal PDF is just $\mathcal{N}(\vec{x}_a|\mu_a, C_{aa})$. Marginalizing for Gaussians can then be done fully analytically in $\mathcal{O}(1)$ time: we just discard the unnecessary parts of the covariance and mean.

The Hessian (calculated at the mean) is the inverse of the opposite of the covariance matrix: $H = -V = -C^{-1}$.

Then, $\sigma_m = \sqrt{(-H)_{mm}^{-1}}$.

Conditioning   Now we want to compute the conditional probability of a certain part of the parameter vector, $\vec{x}_b$, if we fix the rest of the vector to values $\vec{x}_a$. This can also be done analytically: now, the mean will be $\vec{\mu}_{b|a} = \vec{\mu}_b - V_{bb}^{-1}V_{ba}(\vec{x}_a - \vec{\mu}_a)$, while the covariance is $C_{b|a} = V_{bb}^{-1}$.

If we marginalize over all the other parameters, the error looks like $\sigma_m = \sqrt{(-H_{mm})^{-1}}$. This time we invert the single matrix element. This will usually be much smaller than what we would get when marginalizing.

Summing   If both $\vec{x}$ and $\vec{y}$ are MVN distributed and independent, then

$$\mathbb{P}(\vec{z}) = \mathcal{N}(\vec{z}|\vec{\mu}_x + \vec{\mu}_y, C_x + C_y)\,. \qquad (1.3.11)$$

The way to prove this is to start from the fact that

$$\mathbb{P}(z) = \int dx\,dy\,\mathbb{P}(x)\mathbb{P}(y)\delta(z - (x + y)) \qquad (1.3.12)$$

$$= \int dx\,\mathbb{P}(x)\mathbb{P}(z - x)\,, \qquad (1.3.13)$$

a convolution. The convolution of two Gaussians is a Gaussian, so we can just calculate the mean, covariance, and we are done:

$$\langle x + y \rangle = \langle x \rangle + \langle y \rangle\,, \qquad (1.3.14)$$

while the covariance is (implying a tensor product between the vectors):

$$\left\langle (\vec{x}+\vec{y})(\vec{x}+\vec{y}) \right\rangle - (\vec{x}+\vec{y})(\vec{x}+\vec{y}) = \langle \vec{x}\vec{x} \rangle - \langle \vec{x} \rangle^2 + \langle \vec{y}\vec{y} \rangle - \langle \vec{y} \rangle^2 , \tag{1.3.15}$$

using the property that $\langle \vec{x}\vec{y} \rangle - \langle \vec{x} \rangle \langle \vec{y} \rangle = 0$.

Note that independence $\implies$ uncorrelation, but not the other way around. For example, if x is normally distributed with mean zero, x^2 is *uncorrelated* to it (the correlation would be given by $\langle x^3 \rangle = 0$); however x and x^2 are related two-to-one, certainly not independent.

1.3.1 Correlation

The correlation between the component x_i and the component j is given by

$$\left\langle (x_i - \mu_i)(x_j - \mu_j) \right\rangle = \int \mathrm{d}^n x \, (x_i - \mu_i)(x_j - \mu_j) \mathcal{N}(\vec{x}|\vec{\mu}, C) = C_{ij}. \tag{1.3.16}$$

If $i = j$, this is the mean value of $(x_i - \mu_i)^2$: we can integrate out all the other parameters, so that we get the variance of that parameter.

We can show that the covariance is indeed given by the expression above: we define $\vec{y} = \vec{x} - \vec{\mu}$ as usual, so we have

$$\int \mathrm{d}^n y \, y_i y_j \mathcal{N}(\vec{y}|\vec{0}, C) = \frac{1}{N} \int \mathrm{d}^n y \, y_i y_j \exp\left(-\frac{1}{2} \vec{y}^\top V \vec{y} \right) = \left\langle y_i y_j \right\rangle , \tag{1.3.17}$$

for the usual normalization N and $V = C^{-1}$. We can compute it with a trick:

$$\left\langle y_i y_j \right\rangle = 2 \frac{\partial}{\partial V_{ij}} \left[\log \int \mathrm{d}^n y \exp\left(-\frac{1}{2} \vec{y}^\top V \vec{y} \right) \right] \tag{1.3.18}$$

$$= -2 \frac{\partial}{\partial V_{ij}} \left[\log \left(\frac{(2\pi)^{n/2}}{\sqrt{\det V}} \right) \right] \tag{1.3.19}$$

$$= 2 \frac{\partial}{\partial V_{ij}} \left[\frac{1}{2} \det V \right] \tag{1.3.20}$$

$$= \frac{1}{\det V} \frac{\partial}{\partial V_{ij}} \det V , \tag{1.3.21}$$

and we can express the determinant using the Laplace formula:

$$\det V = \sum_{p=1}^{n} V_{ip} K_{ip} , \tag{1.3.22}$$

where K_{ip} is a cofactor, which crucially does not depend on the coefficient V_{ip}. Now, since V is symmetric the cofactor also is; therefore we have

$$(V^{-1})_{ij} = \frac{1}{\det V}(K^\top)_{ij} = \frac{K_{ij}}{\det V} , \tag{1.3.23}$$

therefore

$$\langle y_i y_j \rangle = \frac{1}{\det V} \frac{\partial}{\partial V_{ij}} \underbrace{\left[\sum_{p=1}^{n} V_{ip} K_{ip} \right]}_{\det V} \tag{1.3.24}$$

$$= \frac{K_{ij}}{\det V} = \left(V^{-1} \right)_{ij} = C_{ij}. \tag{1.3.25}$$

1.4 Two-parameter estimation revisited

We have data $\vec{d}$, for which the likelihood with a Gaussian model is

$$\mathcal{L} = \mathbb{P}(\vec{d}|\mu,\sigma) = \prod_{i=1}^{N} \mathcal{L}_i = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{\sum_i (d_i - \mu)^2}{2\sigma^2} \right). \tag{1.4.1}$$

We want to estimate the average, μ, but we know neither μ nor σ: therefore, we will need to marginalize over σ.

We choose improper uniform priors, for $\sigma \in \mathbb{R}^+$ and $\mu \in \mathbb{R}$. Then, the posterior is proportional to the likelihood:

$$P = \mathbb{P}(\mu,\sigma|\vec{d}) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{1}{2}\frac{\sum_i (d_i - \mu)^2}{\sigma^2} \right) \tag{1.4.2}$$

for $\sigma > 0$, 0 for $\sigma < 0$. Integrating over σ to marginalize means we have to compute

$$\mathbb{P}(\mu|\vec{d}) = \int d\sigma \, \frac{1}{\sigma^n} \exp\left(-\frac{1}{2}\frac{\sum_i (d_i - \mu)^2}{\sigma^2} \right). \tag{1.4.3}$$

We make a change of variables $\sigma = 1/t$, so $d\sigma = -dt/t^2$ then

$$\mathbb{P}(\mu|\vec{d}) = -\int dt \, t^{n-2} \exp\left(-\frac{t^2}{2} \sum_i (d_i - \mu)^2 \right), \tag{1.4.4}$$

and now set $\tau^2 = t^2 \sum_i (d_i - \mu)^2$, so $d\tau = dt \sqrt{\sum_i (d_i - \mu)^2}$. The integral then reads

$$\mathbb{P}(\mu|\vec{d}) = \left[\sum_i (d_i - \mu)^2 \right]^{-\frac{n}{2}+1-\frac{1}{2}} \underbrace{\int d\tau \, \tau^{n-2} \exp\left(-\frac{\tau^2}{2} \right)}_{\text{constant}}. \tag{1.4.5}$$

Therefore, the log-posterior reads

$$L = -\frac{n-1}{2} \log\left(\sum_i (d_i - \mu)^2 \right). \tag{1.4.6}$$

17

The value μ_0 is defined so that $L(\mu)$ is stationary there, which means that

$$\frac{\mathrm{d}L}{\mathrm{d}\mu} = -\frac{n-1}{2}\frac{1}{\sum_i(d_i - \mu)^2}\frac{\mathrm{d}}{\mathrm{d}\mu}\left[\sum_i(d_i - \mu)^2\right] \overset{!}{=} 0 \tag{1.4.7}$$

$$= -\frac{n-1}{2}\frac{1}{\sum_i(d_i - \mu)^2}\sum_i 2(d_i - \mu)(-1) \tag{1.4.8}$$

$$= \frac{n-1}{2}\frac{\sum_i(d_i - \mu)}{\sum_i(d_i - \mu)^2}, \tag{1.4.9}$$

meaning that $\mu_0 = n^{-1}\sum d_i$, our estimator for the mean of the distribution is the arithmetic mean of the data.

> Unless $n = 1$! if we only made one measurement and we know *absolutely nothing* about σ, then the mean could be *whatever*.

The error is then defined by

$$\left.\frac{\mathrm{d}^2L}{\mathrm{d}\mu^2}\right|_{\mu_0} = \frac{\mathrm{d}}{\mathrm{d}\mu}\left[(n-1)\frac{\sum_i(d_i - \mu)}{\sum_i(d_i - \mu)^2}\right] \tag{1.4.10}$$

$$= (n-1)\frac{-n\sum_i(d_i - \mu)^2 + \sum_i(d_i - \mu)\times 2\overbrace{\left[\sum_i(d_i - \mu)\right]}^{=0}(n)}{\left[\sum_i(d_i - \mu)^2\right]^2} \tag{1.4.11}$$

$$= -\frac{(n-1)n}{\sum_i(d_i - \mu)^2}, \tag{1.4.12}$$

therefore

$$\hat{\sigma}^2 = -\frac{1}{\frac{\mathrm{d}^2L}{\mathrm{d}\mu^2}} = \frac{\sum_i(d_i - \mu)^2}{n(n-1)}. \tag{1.4.13}$$

This is the usual estimator for the standard deviation of a set of data with Gaussian errors.

1.5 Multiparameter estimation in the abstract

We always seek a parameter vector $\vec{x}_0$ in the form

$$\left.\frac{\partial \log P}{\partial x_i}\right|_{\vec{x}_0} = 0, \tag{1.5.1}$$

or more compactly $\nabla L(\vec{x}_0) = 0$. Typically, if this is linear we can do it, if it is nonlinear then it is a mess. Let us then start with the linear case, in which $\vec{\nabla}L = A\vec{x} + \vec{c}$.

The solution then reads $\vec{x}_0 = -A^{-1}\vec{c}$. Matrix inversion is an $\mathcal{O}\left(N^3\right)$ problem: often, even in this seemingly simple case, we cannot do the computation analytically.

We have seen if the gradient of the log-posterior is linear in the parameters: $\vec{\nabla}L(\vec{x}) = A\vec{x} + \vec{c}$ then the covariance matrix is $\Sigma_{ij} = \left(-A^{-1}\right)_{ij}$. Matrix inversion is slow, so we seek faster methods in general.

Are there situations in which the gradient of the log-posterior is indeed linear? We need the noise to be Gaussian — this is common, by the CLT this approximation is good if there are many concurring sources. For simplicity, we will assume that the measurements are statistically independent — this is not strictly necessary, what we will show holds even if this is not true. Formally, this means that for two measurements d_i and d_j we have $\mathbb{P}(d_i, d_j) = \mathbb{P}(d_i)\mathbb{P}(d_j)$.

Then, the likelihood reads

$$\mathscr{L}(d_i|\vec{x}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(d_i - \mu)^2}{\sigma^2}\right). \tag{1.5.2}$$

Let us assume that only the *mean* is a function of the parameters of our theory (and of the observation number!) $\mu = \mu_i(\vec{x})$. As usual, with the independence assumption we find

$$\mathscr{L}(d_i|\vec{x}) \propto \prod_i \exp\left(-\frac{1}{2}\frac{\left((d_i - \mu(\vec{x}))^2\right)}{\sigma^2}\right) \tag{1.5.3}$$

$$L \propto -\frac{1}{2}\sum_{i=1}^{n} \frac{(d_i - \mu_i(\vec{x}))^2}{\sigma_i^2}. \tag{1.5.4}$$

A sum of squares of Gaussian variables is distributed like a χ^2 variable (with n degrees of freedom, in our case): in this case, $L = -\chi^2/2$, so minimizing the χ^2 corresponds to minimizing the likelihood.

A crucial condition in order for the problem to be linear is that the mean $\mu_i(\vec{x})$ must be a linear function of the parameter vector: $\mu_i(\vec{x}) = A\vec{x} + \vec{c}$. Assuming a flat prior, our log-posterior reads

$$\log P = L \propto \chi^2 = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}\left[d_i - \sum_j A_{ij}x_j - c_i\right]^2. \tag{1.5.5}$$

What we want to show is that this implies that the gradient of the posterior, evaluated at the maximum, is linear in the parameter: it reads

$$\frac{\partial L}{\partial x_k} = \sum_i \frac{2}{\sigma_i^2}\left[d_i - \sum_j A_{ij}x_j - c_i\right]A_{ik}. \tag{1.5.6}$$

This is linear: one could see it by eye, but let us also compute its second derivative to make sure:

$$\frac{\partial^2 L}{\partial x_l \partial x_k} = \sum_i \frac{2}{\sigma_i^2}[-A_{il}A_{ik}] = -2\left(A^\top \Sigma^{-1}A\right)_{lk}. \tag{1.5.7}$$

This is a constant, so the initial expression was indeed linear. This is the usual least-squares fitting: people often simply do this without much care for the assumptions they are making.

1.5.1 Frequentist vs Bayesian

The conceptual meaning of a credible interval is different from that of a confidence interval. Sometimes, especially when we have few data points, they are also quantitatively different.

- For a frequentist a parameter T is a fixed unknown number, not a random variable;

- data x are random variables given by their frequency with which they occur in many repetitions;

- both the Bayesian and the frequentist build a *statistical estimator* f, which is a function giving an estimate for the parameter starting from the data: $f(x)$ yields t, an estimate for T;

- the distribution of the estimator — which is a random variable since its argument is — is called a *sampling distribution* (for example, a χ^2 distribution);

- the frequentist builds a $b\%$ confidence interval, which is an interval for t such that it will contain T $b\%$ of the time.

In the Gaussian limit the confidence interval and the credible interval coincide. In the frequentist case we write the Confidence Interval as

$$\mathbb{P}\left(-1.96 \leq \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) \approx 0.95\,, \tag{1.5.8}$$

while in the Bayesian case we write the Credible interval as

$$\mathbb{P}\left(\mu_0 - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \mu_0 + 1.96\frac{\sigma}{\sqrt{n}}\right) \approx 0.95\,. \tag{1.5.9}$$

These two are written in terms of different variables: μ_0, the estimate of the mean from the data, in the frequentist case, and the parameter μ in the Bayesian case.

If we also need to estimate the standard deviation we need to use a Student's t distribution: it converges to a Gaussian, but with few data it has fatter tails.

Now, we see an example of the failings of frequentist statistics [Van14]. The probability is given by $\mathbb{P}(x|T) = \exp(T - x)[x \geq T]$. We are given a dataset $[10, 12, 15]$. A frequentist will build an estimator: we have

$$\langle x \rangle = \int x\mathbb{P}(x)\,\mathrm{d}x = T + 1\,, \tag{1.5.10}$$

So the arithmetic mean of x converges to $T + 1$, therefore we average the results and subtract one. Our confidence interval will then look like $\hat{T} - 2\sqrt{N} \leq T \leq T + 2\sqrt{N}$. This is derived from an asymptotic estimate, but it is close to the true result. We find $[10.2, 12.5]$: but we know that $T \leq 10$!

In a Bayesian approach, we do not even need to take a uniform prior for $T \leq \min(\text{data})$. This is fixed by the data, since the likelihood includes a Heaviside theta. The likelihood then looks like

$$\mathbb{P}(T|x) = \exp(T - x)[T < x].\tag{1.5.11}$$

The α-interval we get is $T = x_{\min} + \log \alpha / N$.

> To include more of the discussion.

We were not very fair to the frequentists. We cherry-picked a very small dataset, which included an outlier. A frequentist might have chosen an estimator which was more *robust* to outliers, being less affected by them. Typically, the best estimator which is chosen is the Maximum Likelihood one.

We discuss the example in https://stats.stackexchange.com/a/2287/164421 [Win]. The frequentist uses the Neyman confidence belt. The construction of the CI is *not unique*.

1.5.2 Nonlinear parameter estimation

The problem, as usual, is to find the point $\vec{x}_0$ in parameter space where the log posterior $L = \log P$ is maximum:

$$\partial_i L(\vec{x}_0) = 0 \tag{1.5.12}$$
$$\partial_i \partial_j L(x_0) \text{ is negative definite.} . \tag{1.5.13}$$

Around such a point we can expand as

$$L(\vec{x}) \approx L(\vec{x}_0) + \frac{1}{2}(x - x_0)^i (x - x_0)^j \partial_i \partial_j L(x_0).\tag{1.5.14}$$

If we start from a point $\vec{x}$ which is reasonably close to $\vec{x}_0$ then we can expand up to second order there as well:

$$L(\vec{x}) \approx L(\vec{x}_1) + (x - x_1)^i \partial_i L(x_1).\tag{1.5.15}$$

We shall assume that here as well the function is reasonably close to the true posterior. Let us take the gradient of the previous equation: we have

$$\partial_k L(\vec{x}) \approx \delta_k^i \partial_i L(x_1) + (x - x_1)^i \partial_i \partial_k L(x_1) \tag{1.5.16}$$
$$\partial_k L(\vec{x}) \approx \partial_k L(x_1) + (x - x_1)^i \partial_i \partial_k L(x_1) \tag{1.5.17}$$
$$0 \approx \partial_k L(x_1) + (x_0 - x_1)^i \partial_i \partial_k L(x_1),\tag{1.5.18}$$

where in the last step we calculated the expression at $x = x_0$. We can solve this expression for x_0:

$$x_0^i \underbrace{\partial_i \partial_k L(x_1)}_{H_{ik}} \approx x_1^i \partial_i \partial_k L(x_1) - \partial_k L(x_1) \tag{1.5.19}$$

$$x_0 = \vec{x}_1 - (H^{-1})\Big|_{x_1} \nabla L(x_1),\tag{1.5.20}$$

which tells us that in order to move towards the solution we need to update our guess, x_1, by a factor $H^{-1}\nabla L$. This process is then the basis for an iterative procedure. It is called the **Newton-Rhapson** method.

The quality of our initial guess matters, but often in practice we have a good idea of where the parameter will approximately lie.

This has complications if the posterior is multimodal: we can find a local maximum instead of the global one. Also, if we start in the flat tail of the distribution we might not be able to tell that the flatness is not due to being in the maximum.

Even if everything works right, we still have $\mathcal{O}\left(N^3\right)$ matrix inversion to do.

1.6 Markov Chain Monte Carlo

We introduce the notation $\pi(\vec{\theta})$: this is the πrior, in terms of the parameter vector $\vec{\theta}$. We have the likelihood $\mathcal{L}(\vec{d}|\vec{\theta})$, in terms of the data vector $\vec{d}$.

The posterior reads

$$P(\vec{\theta}|\vec{d}) = \frac{\mathcal{L}(\vec{d}|\vec{\theta})\pi(\vec{\theta})}{\int \mathcal{L}(\vec{d}|\vec{\theta})\pi(\vec{\theta})\,\mathrm{d}\vec{\theta}}. \tag{1.6.1}$$

If we want to simulate the distribution we need the normalization, which is the reason why we wrote out the evidence.

The things we need to do are of the form of finding a marginal posterior for a single parameter θ_i,

$$P(\theta_i) = \int \mathrm{d}\theta_1 \ldots \widehat{\mathrm{d}\theta_i} \ldots \mathrm{d}\theta_n\, P(\vec{\theta}|\vec{d})\,, \tag{1.6.2}$$

where by $\widehat{\mathrm{d}\theta_i}$ we mean that we do *not* integrate over θ_i. This might very well be intractable.

The Monte Carlo approach is to try and solve the integral by sampling certain points, with a likelihood given by the posterior. We generate different realizations of the parameters, $\vec{\theta}_i$, as independent identically distributed random variables, with the same distribution as the posterior.

Then, we can estimate the expectation value of an arbitrary function $g(\vec{\theta})$ as

$$\left\langle g(\vec{\theta}) \right\rangle = \int g(\vec{\theta}) P(\vec{\theta}|\vec{d})\, \mathrm{d}\vec{\theta} \approx \frac{1}{N}\sum_{i=1}^{N} g(\vec{\theta}_i)\,. \tag{1.6.3}$$

Homework for the month of November: exercise 4, exercise 5, NOT exercise 6 nor 7, exercise 8, exercise 9.

We can compute the average value of some function g of our parameters $\vec{\theta}$, defined as

$$\mathbb{E}(g(\vec{\theta})) = \int g(\vec{\theta}) p(\vec{\theta})\, \mathrm{d}\vec{\theta}\,, \tag{1.6.4}$$

as

$$\hat{E}(g(\vec{\theta})) = \frac{1}{N}\sum_{i=1}^{N} g(\vec{\theta}_i)\,, \tag{1.6.5}$$

which converges as $N \rightarrow \infty$ to the true value, as long as the $\vec{\theta}_i$ are iid sampled according to their distribution $p(\vec{\theta})$. This converges to the true value with a variance $\mathrm{var}\left(\hat{E}\right) = \sigma^2/N$.

However, it is difficult to sample points from a generic multidimensional distribution. This is the problem we will treat now.

A possible solution is rejection sampling. See the numerical methods course for a detailed explanation of the method.

In short, in order to sample from a distribution $f(x)$ we choose a distribution $g(x)$ which *embeds* the generic one we have: this means that $g(x) \geq Mf(x)$ for some real number M. We generate numbers x according to $g(x)$, and then we reject the number we generated a fraction $Mf(x)/g(x)$ of the time (this can be done through a uniform distribution). This procedure yields samples distributed according to $f(x)$.

This can be wasteful, especially in high dimensions, due to the **curse of dimensionality**. Consider the volume of a D-dimensional sphere and a D-dimensional cube with a side equal to the diameter of the sphere. The ratio is $\pi/4$ in two dimensions, $\pi/6$ in three dimensions, and it approaches zero for $D \rightarrow \infty$.

So, rejection sampling fails in practice. The alternative approach is Markov Chain Monte Carlo. The idea here is to generate sequences of correlated variables: a random walk, with some specific rule giving us the next step. Starting from $\vec{\theta}_{(1)}$ we move to a point $\vec{\theta}_{(2)}$; each point is calculated with a rule which depends only on the previous one. The chain has very short-term memory.

It is possible to choose a rule so that after a long run of this chain the distribution of the points reached will be the same as that of the distribution.

How do we walk in parameter space? We start by discussing MCMC in general, as a mathematical tool.

A Markov Chain is a sequence of random variables; we start with a state space Ω, which we will assume to be discrete (since in any practical application we will not have infinite precision).

A sequence of random variables X parametrized by t, such that $X_t \in \Omega$ is a MC iff[2] the probability that the variable takes on a certain value at a certain moment is independent of all but the previous step:

$$\mathbb{P}(X_t = s_t | X_{t-1} = s_{t-1}) = \mathbb{P}(X_t = s_t | X_{t-1} = s_{t-1}, \{X_{t_i} = s_{t_i}\}_i), \qquad (1.6.6)$$

for any set of times satisfying $t_i \leq t - 2$.

To characterize a MC we need to provide all the probabilities which can be written like

$$\mathbb{P}(X_{t+1} = s_{t+1} | X_t = s_t). \qquad (1.6.7)$$

These will need to be $|\Omega|^2$ numbers (where $|\Omega|$ is the number of elements in Ω), and are called *transition probabilities* in a **transition matrix**.

The transition matrix can be written as

$$T_{ij} = \mathbb{P}(X_{t+1} = s_i | X_t = s_j). \qquad (1.6.8)$$

---

[2] If and only if.

In order for the probabilities to be normalized we need to impose $\sum_j T_{ij} = 1$ for any i. So, the degrees of freedom are actually $|\Omega|(|\Omega| - 1)$.

In a general MC these probabilities can change at each step; a *stationary* (or homogeneous) MC is one in which T_{ij} is constant.

An **ergodic** MC is one if any state can be reached from any other state (not necessarily in a single step).

A **stationary distribution** for a Markov Chain is a probability distribution $p(\vec{x})$ for $\vec{x}$ in Ω such that the following property holds:

$$p(\vec{y}) = \sum_{\vec{x} \in \Omega} \mathbb{P}(\vec{y}|\vec{x}) p(\vec{x}), \tag{1.6.9}$$

where $\mathbb{P}(\vec{y}|\vec{x})$ is the transition probability from $\vec{x}$ to $\vec{y}$: the probability that we are in $\vec{y}$, given that we were in $\vec{x}$ at the previous step. What this means is that after a step in the chain leaves $p(\vec{x})$ unchanged.

In terms of the transition matrix, the problem reads $p_i = T_{ij} p_j$, so p_i is an eigenvector of T_{ij} with eigenvalue 1.

In our case, we want to build a MC which converges to a stationary probability distribution which is our posterior. So, first we need to ask whether this can be done, and then how long it takes for us to converge to that distribution.

Theorem 1.6.1. *A finite, ergodic Markov chain has a unique stationary distribution.*

A state has **period** k if any return to it requires a multiple of k steps. Formally,

$$k(\vec{x}) = \gcd \left\{ n \in \mathbb{N} : \text{the probability of returning from } \vec{x} \text{ to } \vec{x} \text{ in } n \text{ steps is } > 0 \right\}. \tag{1.6.10}$$

If $k = 1$ the state is aperiodic. If a MC is ergodic and one state is aperiodic then all states are.

A state is **recurrent** if with probability 1 we will return to it at some point if we leave it. It is called **transient** otherwise.

A recurrent state is **positive recurrent** if the expected return time is finite; **null recurrent** if it diverges.

Theorem 1.6.2. *A finite, ergodic, positive recurrent, aperiodic Markov Chain converges to its stationary distribution as $n \to \infty$.*

So, eventually convergence is ensured, but we must also design a MC which converges *quickly*.

Let us now give a practical example, with $\Omega = \{A, B, C\}$. The transition matrix is

$$T_{ij} = \begin{bmatrix} 0.25 & 0.5 & 0.25 \\ 0 & 0.5 & 0.5 \\ 0.33 & 0.33 & 0.34 \end{bmatrix}, \tag{1.6.11}$$

as required all the rows are normalized to one. The initial condition is given by $q_0 = [0, 0.5, 0.5]$.

The following steps can be calculated as $q_n = q_0(T^n)$.

Let us define $m(s_i, s_j)$ as the mean time to go from s_i to s_j. How do we compute it? We need to make sure that it is finite. Let us fix a state, say C. What is $m(C, C)$?

Suppose that the first step is from C to B. Then, we will have $m(C, C) = 1 + m(B, C)$. Similarly, if we go from C to A we have $m(C, C) = 1 + m(A, C)$; while if we jump directly to C we have $m(C, C) = 1$.

The overall recurrence time is obtained by integrating over all the possibilities:

$$m(C, C) = \mathbb{P}(C|A)[1 + m(A, C)] + \mathbb{P}(C|B)[1 + m(B, C)] + \mathbb{P}(C|C) \tag{1.6.12}$$

$$= \underbrace{\mathbb{P}(C|A) + \mathbb{P}(C|B) + \mathbb{P}(C|C)}_{1} + \mathbb{P}(C|A)m(A, C) + \mathbb{P}(C|B)m(B, C) + . \tag{1.6.13}$$

We can do a similar thing for the other recurrence times, this yields a linear system we can solve, in terms of the three unknowns $m(*, C)$. In this case we get $m(C, C) = 2.54$, $m(B, C) = 2.0$, $m(A, C) = 2.67$.

Therefore, this MC will admit a stationary distribution and converge to it.[3] We can calculate it through the left eigenvalue problem $\pi = \pi T$.

A Markov Chain is **reversible** if there exists a probability distribution π such that $\mathbb{P}(x|y)\pi(y) = \mathbb{P}(y|x)\pi(x)$. This is called the **detailed balance** equation, and if it holds it tells us that π is the stationary distribution of the Markov chain: if we sum over y we get

$$\sum_y \mathbb{P}(x|y)\pi(y) = \sum_y \mathbb{P}(y|x)\pi(x) = \pi(x) \sum_y \mathbb{P}(y|x) = \pi(x), \tag{1.6.14}$$

which is precisely the stationarity condition.

1.6.1 Metropolis-Hastings

Suppose we have a posterior $p(\vec{\theta})$, in terms of a parameter vector $\vec{\theta}$.

The current state is denoted as

$$X_\theta = (\theta_{1,t}, \ldots, \theta_{n,t}). \tag{1.6.15}$$

We choose a transition distribution Q; this can be some easy-to-sample distribution of our choosing.

Then, the new state is **proposed** by updating the old position

$$Y \sim Q(\cdot|X_t), \tag{1.6.16}$$

and choose to accept it or reject it depending on the following: we draw a uniform variable, and accept with probability

$$\alpha(X_{t+1} = y|X_t = x) = \min\left\{\frac{\mathscr{L}(\vec{y})}{\mathscr{L}(\vec{x})}\frac{Q(\vec{x}|\vec{y})}{Q(\vec{y}|\vec{x})}, 1\right\}. \tag{1.6.17}$$

---

[3] Check out the convergence by downloading the notebook at https://github.com/jacopok/notes/blob/master/ap_third_semester/astrostatistics_cosmology/figures/markov.ipynb!

Simply put, we favor steps which move us in a region of higher likelihood. The ratio of the Qs might well go away if we choose a symmetric transition matrix.

This satisfies the detailed balance equation: let us show it. One side of the equation reads

$$\underbrace{\alpha(\vec{y}|\vec{x})Q(\vec{y}|\vec{x})}_{\text{transition probability}} p(\vec{x}) = \min\left\{\frac{\mathscr{L}(\vec{y})}{\mathscr{L}(\vec{x})}\frac{Q(\vec{x}|\vec{y})}{Q(\vec{y}|\vec{x})}, 1\right\}Q(\vec{y}|\vec{x})p(\vec{x}) \tag{1.6.18}$$

$$= \min\left\{Q(\vec{x}|\vec{y})\mathscr{L}(\vec{y}), Q(\vec{y}|\vec{x})\mathscr{L}(\vec{x})\right\}\mathscr{L}(\vec{x}), \tag{1.6.19}$$

> check calculation

We apply Bayes' theorem... Finally, we show that

$$\alpha(\vec{y}|\vec{x})Q(\vec{y}|\vec{x})p(\vec{x}) = \alpha(\vec{x}|\vec{y})Q(\vec{x}|\vec{y})p(\vec{y}). \tag{1.6.20}$$

We have a parameter vector $\vec{x}$, and in a step in our MCMC we want to move to a new parameter $\vec{y}$. We start off with a transition proposal distributed according to the distribution Q, which can depend on $\vec{x}$. This Q may, for example, be a simple Gaussian.

Then, we have an acceptance/rejection step: we accept the step with a probability

$$\alpha(y|x) = \min\left(\frac{p(\vec{y})Q(\vec{x}|\vec{y})}{p(\vec{x})Q(\vec{y}|\vec{x})}, 1\right), \tag{1.6.21}$$

where p is the posterior. If we are using a Gaussian Q, it is symmetric, therefore we only have $\alpha = \min(p(\vec{y})/p(\vec{x}), 1)$. If the posterior is higher in $\vec{y}$ than in $\vec{x}$ we always accept the step, else we accept it sometimes.

We need to check whether the MC is ergodic: the transition probability reads

$$T(y|x) = \alpha Q = \min\left(\frac{p(\vec{y})Q(\vec{x}|\vec{y})}{p(\vec{x})Q(\vec{y}|\vec{x})}, 1\right)Q(\vec{y}|\vec{x}) \tag{1.6.22}$$

$$= \frac{1}{p(\vec{x})Q(\vec{y}|\vec{x})}\min\left[p(\vec{y})Q(\vec{x}|\vec{y}), p(\vec{x})Q(\vec{y}|\vec{x})\right]Q(\vec{y}|\vec{x}) \tag{1.6.23}$$

$$= \frac{1}{p(\vec{x})}\min\left[p(\vec{y})Q(\vec{x}|\vec{y}), p(\vec{x})Q(\vec{y}|\vec{x})\right], \tag{1.6.24}$$

therefore we have

$$T(\vec{y}|\vec{x})p(\vec{x}) = \min\left[p(\vec{y})Q(\vec{x}|\vec{y}), p(\vec{x})Q(\vec{y}|\vec{x})\right] = T(\vec{x}|\vec{y})p(\vec{y}), \tag{1.6.25}$$

which is precisely the **detailed balance** condition, with the stationary distribution being precisely p, the posterior.

Since we only need to compute ratios of the posterior, we never need to compute the evidence.

1.6.2 Gibbs sampling

We start from a parameter vector $\vec{x}$, and we only update the i-th parameter, keeping the others fixed, according to the conditional probability for that parameter given all the others: $\mathbb{P}(\theta_i|\theta_1, \hat{\theta}_i, \theta_n)$. Here we always accept the step.

We update each parameter separately (using the old values), and then move the full parameter vector together. It is a special case of Metropolis-Hastings with Q being the conditional probability and acceptance probability equal to 1. We might want to use it if we have easy access to the conditional distribution.

The issue is always convergence: we have convergence theorems for $N \to \infty$, but we need *fast* convergence practically.

We can compute averages and standard deviations as

$$\langle \theta_i \rangle \approx \frac{1}{N} \sum_{k=1}^{N} \theta_i^{(k)} \tag{1.6.26}$$

$$\sigma_{\theta_i}^2 \approx \frac{1}{N} \sum_{k=1}^{N} \left| \theta_i^{(k)} \right|^2 - \langle \theta_i \rangle^2 . \tag{1.6.27}$$

This is not done over *all* the points, but only over the ones calculated after convergence. Marginalization is done as such:

$$p(\theta_2) = \int d\theta_1 \, p(\theta_1, \theta_2) = \int d\theta_1 \, p(\theta_1|\theta_2)p(\theta_1) . \tag{1.6.28}$$

We can simply build a marginal histogram by neglecting a coordinate and then binning.

1.6.3 Convergence

How do we verify convergence? This is the complicated part. We can do a trace plot:

$$\text{Tr}\left[-\log \mathcal{P} \right] = -\sum_{k=1}^{M} \frac{1}{k} \log \mathcal{P}(\theta_i^{(k)}|\vec{d}) , \tag{1.6.29}$$

which should converge to a fixed maximum value after some time. However, we might not have covered a representative region of the full posterior.

A standard technique is to start M different chains, and for each compute the average separately.

We can compute the average of each chain, $\bar{\theta}^J$ (where $J = 1 \dots M$); and the total average $\bar{\theta}$ over all the chains. We can also estimate the variance of each chain, w^J:

$$w^J = \frac{1}{n-1} \sum_{i=1}^{n} \left(\theta_i^J - \bar{\theta}^J \right)^2 , \tag{1.6.30}$$

where the $n-1$ is the Bessel correction, needed because we use the estimated average instead of the true average.

Through manipulation, we can show that

$$w^J = \frac{n}{n-1}s_J - \frac{n}{n-1}\text{var}\left(\bar{\theta}^J\right),\tag{1.6.31}$$

where $\text{var}\left(\bar{\theta}^J\right) = (\bar{\theta}^J - \mu)^2$ is an estimator for the variance of the *mean* of the parameter vector, while s_J is the estimator of the true variance of the chain. Flipping the expression we can find

$$s_J = \frac{n-1}{n}w^J + \text{var}\left(\bar{\theta}^J\right).\tag{1.6.32}$$

We can estimate the variance of the chain mean as

$$\widehat{\text{var}\left(\bar{\theta}\right)} = \frac{1}{M-1}\sum_{J=1}^{M}(\bar{\theta}^J - \bar{\theta})^2 = \frac{B}{n},\tag{1.6.33}$$

and then we can average the single chain variance estimators:

$$W = \frac{1}{M}\sum_{J=1}^{M}w^J,\tag{1.6.34}$$

so then we get the estimator

$$\hat{V} = \frac{n-1}{n}W + \frac{B}{n},\tag{1.6.35}$$

and then we can define $\hat{R} = \hat{V}/W$: we expect B/n to become irrelevant asymptotically. If $\hat{R}$ is of order 1 then we can say that we have converged. Typically, we say that when $\hat{R} \lesssim 1.02$ we have converged.

Add more calculations

These tests are not fool-proof: we try to falsify the hypothesis of convergence, and at a certain point we stop trying to do so.

1.6.4 Autocorrelation

Successive points in a chain are correlated, and the Monte Carlo error scales like $1/\sqrt{N}$, where N is the number of *uncorrelated* points. How do we calculate the effective number of uncorrelated points?

The autocorrelation for points τ steps apart is

$$\rho_\theta(\tau) = \frac{1}{n-\tau}\frac{\sum_{i=1}^{n-\tau}(\theta_i - \bar{\theta})(\theta_{i+\tau} - \bar{\theta})}{\sum_{i=1}^{n-\tau}(\theta_i - \bar{\theta})^2}.\tag{1.6.36}$$

Then, the variance of the parameter q is

$$\text{var}(\theta) = \frac{\tau_f}{N}\sigma_\theta^2,\tag{1.6.37}$$

where τ_f is the integrated autocorrelation time:

$$\tau_f = \sum_{\tau=-\infty}^{+\infty} \rho_\theta(\tau) \approx 1 + 2 \sum_{\tau=1}^{N} \rho_\theta(\tau). \tag{1.6.38}$$

So, N/τ_f is the effective number of points. A note on numerical stability: in computing the previous expression one must be careful to not extend the sum to a region in which the estimation of $\rho_\theta(\tau)$ is imprecise: for very large τ we have very few samples of these large delays.

It is best if the proposal distribution Q is as close as possible to the true distribution. Preprocessing should include estimating the covariance matrix of the data, or even running a short chain.

Gibbs sampling does not work well if the parameters are heavily correlated, since it moves along the conditional distributions.

Examples of libraries.

Chapter 2

Real datasets

2.1 The CMB

At roughly 3×10^5 yr after the Big Bang the universe had a temperature of roughly 3×10^3 K, and electrons and protons combined into hydrogen atoms. Since the cross-section of Compton scattering is proportional to the inverse square mass of the particle, the optical depth of the gas decreased drastically: it became transparent.

We observe that light as a near-blackbody spectrum at $T_{CMB} \approx 2.7255$ K. We see slight anisotropies, of the order of $\Delta T / T \sim 10^{-5}$.

These fluctuations are very sensitive to inflationary cosmological parameters.

The quantum process which generates the anisotropies is inherently random. We must treat it as a random field. Given a random field $X(\vec{t})$ (where $\vec{t}$ is a position vector) with a pdf $P(X)$ we can compute its average value as

$$\mathbb{E}_X(\vec{t}) = \int X(\vec{t}) P(X) \, dX \,, \tag{2.1.1}$$

and its auto-covariance and auto-correlation functions:

$$C(\vec{t}, \vec{s}) = \text{cov}\left(X(\vec{t}), X(\vec{s}) \right) \tag{2.1.2}$$

$$\rho(\vec{t}, \vec{s}) = \frac{C(\vec{t}, \vec{s})}{\sqrt{C(\vec{t}, \vec{t}) C(\vec{s}, \vec{s})}} \,. \tag{2.1.3}$$

The hypothesis of stationarity can be stated by saying that $C(\vec{t}, \vec{s})$ is only a function of $\vec{t} - \vec{s}$. Isotropy means that it is only a function of $|\vec{t} - \vec{s}|$.

A Gaussian random field is such that the joint probability of the field attaining certain values at certain points is a multivariate Gaussian.

By homogeneity and isotropy many fields in cosmology can be treated as Gaussian, stationary and isotropic: they can then be described through their power spectrum alone.

We Fourier transform the field $F(\vec{x})$ into $F(\vec{k})$, and define the power spectrum P_F as

$$\left\langle F(\vec{k}_1) F(\vec{k}_2) \right\rangle = (2\pi)^3 P_F(k) \delta^{(D)}(\vec{k}_1 + \vec{k}_2) \,, \tag{2.1.4}$$

where $k = |k_1| = |k_2|$.

By the **Wiener-Kintchine** theorem the power spectrum is the 3D Fourier transform of the two-point correlation function.

We must work in spherical harmonics:

$$\frac{\Delta T}{T}(\theta, \varphi) = \sum_{\ell>2} \sum_{|m|\leq\ell} a_{\ell m} Y_{\ell m}(\theta, \varphi). \tag{2.1.5}$$

The angular frequency ℓ gives us the angular scale of the oscillations: roughly speaking, $\theta \sim \pi/\ell$. Then we define the power spectrum as

$$\left\langle a^*_{\ell_1 m_1} a_{\ell_2 m_2} \right\rangle = C_{\ell_1} \delta_{\ell_1 \ell_2} \delta_{m_1 m_2}. \tag{2.1.6}$$

We skip the monopole and the dipole: the monopole is just an offset, and the Doppler is given by the motion of the Solar System through the galaxy, and of the Galaxy through the CMB reference.

We start from the power spectrum estimation, neglecting the first parts of the pipeline.

The data is given by the signal plus the noise; we assume the latter to be given by a zero-mean multivariate normal with a covariance C_N. However, the signal is a zero-mean MVN as well, with covariance C_S. Then, the likelihood is given in terms of the array of temperature differences Δ as

$$\mathscr{L}(\Delta|C) = \frac{1}{\sqrt{(2\pi)^N \det C}} \exp\left[-\frac{1}{2}\Delta^\top C^{-1}\Delta\right], \tag{2.1.7}$$

where $C = C_S + C_N$, since we know that the sum of two Gaussian variables is another Gaussian distributed according to the sum of the covariances.

Let us describe the signal covariance: it is given by

$$C_S^{ij} = \left\langle \frac{\Delta T}{T}(\hat{\gamma}_i) \frac{\Delta T}{T}(\hat{\gamma}_j) \right\rangle = \sum_{\ell} (2\ell + 1) C_\ell P_\ell(\cos\theta_{ij}). \tag{2.1.8}$$

This is not diagonal, since different pixels are correlated. However, in harmonic space the *full sky* covariance is diagonal:

$$\mathscr{L}(a_{\ell m}|C_\ell) = \prod_{\ell m} \frac{1}{\sqrt{2\pi C_\ell}} \exp\left[-\frac{1}{2}\frac{|a_{\ell m}|^2}{C_\ell}\right]. \tag{2.1.9}$$

We will make some simplifying but not too unrealistic assumptions: in order of unrealism we have

1. the noise is Gaussian (quite right);

2. the noise is uncorrelated in pixel space: $\left\langle n(\hat{\gamma}_i)n(\hat{\gamma}_j)\right\rangle = C_n \delta_{ij}$ (not completely true, the thermal fluctuations might change over time and there might be $1/f$ noise, however it is quite close to being true);

3. the noise variance is the same for each pixel (this is the most unrealistic: each pixel will be observed for different amounts of time for any realistic observing strategy).

What we are assuming, in short, can be stated as thinking of the noise as white and uniform.

The noise multipoles are given by

$$n_{\ell m} = \int \mathrm{d}\hat{\gamma}\, n(\hat{\gamma}) Y_{\ell m}^*(\hat{\gamma})\,, \tag{2.1.10}$$

and if we discretize with angular pixel area $\Delta\Omega$ we can compute the noise power spectrum as

$$\left\langle n_{\ell_i m_i} n_{\ell_j m_j} \right\rangle = (\Delta\Omega)^2 \sum_{ij} \left\langle n(\hat{\gamma}_i) n(\hat{\gamma}_j) \right\rangle Y_{\ell_i m_i}^*(\hat{\gamma}_i) Y_{\ell_j m_j}(\hat{\gamma}_j) \tag{2.1.11}$$

$$= (\Delta\Omega)^2 \sum_{ij} C_n \delta_{ij} Y_{\ell_i m_i}^*(\hat{\gamma}_i) Y_{\ell_j m_j}(\hat{\gamma}_j) \tag{2.1.12}$$

$$= C_n \Delta\Omega \sum_i \Delta\Omega Y_{\ell_i m_i}^*(\hat{\gamma}_i) Y_{\ell_j m_j}(\hat{\gamma}_j) \tag{2.1.13}$$

$$= C_n \Delta\Omega \delta_{\ell_i \ell_j} \delta_{m_i m_j}\,, \tag{2.1.14}$$

because of the orthonormality of the spherical harmonics.

In these experiments we have a finite resolution: small scales are spread out. This would be described by a Bessel function probably, but we can approximate it as convolution with a Gaussian:

$$b_\ell \propto \exp\left(-\ell(\ell+1)/\ell_{\text{beam}}^2\right)\,. \tag{2.1.15}$$

This means that at very small scales the signal is completely spread out. However, this only affects the signal, the noise fluctuations have no suppression at any scale: the total signal is

$$\Gamma_\ell = b_\ell^2 C_\ell + N_\ell\,. \tag{2.1.16}$$

Typically we plot $\ell(\ell+1)\Gamma_\ell$.

The problem becomes much harder without the assumptions we mentioned. Specifically, isotropy is broken: we need a sky mask because of the galaxy. To do it analytically we would need to invert an $N_{\text{pix}} \times N_{\text{pix}}$ matrix, with $N_{\text{pix}} \sim 10^7$.

A sky mask is nice in pixel space, but it is a convolution in Fourier space, and in order to deconvolve we need the *masking kernel*. This creates correlations in the sky.

Using the full $a_{\ell m}$ is not needed in the likelihood: all the information is in the C_ℓs.

So, we do an intermediate parameter estimation: calculating the C_ℓ from the $a_{\ell m}$.

We maximize the likelihood with respect to the C_ℓ. This is still a nonlinear problem, and the Newton-Rhapson method which was used in the nineties is no longer feasible. We can use the estimator

$$\hat{C}_\ell = \frac{1}{2\ell+1} \sum_{m=-\ell}^{\ell} |a_{\ell m}|^2\,. \tag{2.1.17}$$

This is an unbiased estimator, since by definition $\left\langle |a_{\ell m}|^2 \right\rangle = C_\ell$.

If we define

$$\hat{Y}_\ell = \sum_{|m|<\ell} \frac{|a_{\ell m}|^2}{C_\ell}, \tag{2.1.18}$$

this will have a χ^2 distribution if each $|a_{\ell m}|$ is Gaussian and independent. The estimator $\hat{C}_\ell$ can be written in terms of the $\hat{Y}_\ell$ as

$$\hat{C}_\ell = C_\ell \hat{Y}_\ell / (2\ell + 1). \tag{2.1.19}$$

This will not have a Gaussian distribution, but a Γ one: the likelihood will read

$$p(\hat{C}_\ell | C_\ell) \propto C_\ell^{-1} \left(\frac{\hat{C}_\ell}{C_\ell} \right)^{\nu/2 - 1} \exp\left(-\frac{\nu}{2} \frac{\hat{C}_\ell}{C_\ell} \right). \tag{2.1.20}$$

This is clearly maximized by $\hat{C}_\ell = C_\ell$. For large ($\gtrsim 30 \div 40$) we can approximate the Gamma distribution as a Gaussian; while for small C_ℓ we can smooth the map and have a reasonable amount of pixels.

In the cut sky case things get more complicated, since the $a_{\ell m}$ are correlated.

The multipoles $\widetilde{a}_{\ell m}$ on the cut sky are given as

$$\widetilde{a}_{\ell m} = \sum_{\ell' m'} a_{\ell' m'} K_{\ell m \ell' m'}(W), \tag{2.1.21}$$

where the kernel K is a complicated function of Wigner $3j$ symbols. We find that the pseudo-C_ℓ on the cut sky is biased, as

$$\left\langle \widetilde{C}_{\ell_1} \right\rangle = \sum_{\ell_2} M_{\ell_1 \ell_2} \left\langle C_{\ell_2} \right\rangle. \tag{2.1.22}$$

The function M is known, so we can recover the C_ℓ. This will converge to a Gaussian, which however will not be diagonal.

The log-likelihood will then be a χ^2 variable.

The papers Verde et al. [Ver+03] and Spergel et al. [Spe+03] are discussed.

2.2 Bayesian Hierarchical Modeling

An example from the lectures by A. Heavens, in the ICIC data analysis workshop.

Suppose we want to solve a linear regression problem for a line passing through the origin, and we have one data pair $(\hat{x}, \hat{y})$, with error on both.

We want the angular coefficient m:

$$\mathcal{P}(m|\hat{x}, \hat{y}) \propto \mathcal{L}(\hat{x}, \hat{y}|m) \mathbb{P}(m). \tag{2.2.1}$$

We know that the true x and y are related by $y = mx$. We can introduce these two as extra variables in the problem: applying $\mathbb{P}(a, b) = \mathbb{P}(a|b)\mathbb{P}(b)$ repeatedly we find

$$\mathbb{P}(\hat{x}, \hat{y}, x, y|m) \propto \mathbb{P}(\hat{x}, \hat{y}|x, y, m)\mathbb{P}(x, y|m) \propto \mathbb{P}(\hat{x}, \hat{y}|x, y)\underbrace{\mathbb{P}(y|x, m)}_{=\delta(y-mx)}\mathbb{P}(x|m). \qquad (2.2.2)$$

Then, we can get the final posterior by

$$\mathcal{P}(m|\hat{x}, \hat{y}) \propto \int dx\, dy\, \mathbb{P}(\hat{x}, \hat{y}, x, y|m)\mathbb{P}(m) \qquad (2.2.3)$$

$$\propto \int dx\, dy\, \underbrace{\mathbb{P}(\hat{x}, \hat{y}|x, y)}_{\text{error}}\underbrace{\mathbb{P}(y|x, m)}_{\text{theory}}\underbrace{\mathbb{P}(x|m)\mathbb{P}(m)}_{\text{prior}} \qquad (2.2.4)$$

$$\propto \int dx\, \mathbb{P}(\hat{x}, \hat{y}|x, mx)\mathbb{P}(x|m)\mathbb{P}(m), \qquad (2.2.5)$$

and assuming $\sigma = 1$ independent Gaussians for the error distribution, we get

$$\mathbb{P}(m|\hat{x}, \hat{y}) \propto \int dx \exp\left(-\frac{1}{2}(\hat{y} - mx)^2\right)\exp\left(-\frac{1}{2}(\hat{x} - x)^2\right) \qquad (2.2.6)$$

$$\propto \frac{1}{\sqrt{1 + m^2}}\exp\left(-\frac{1}{2(1 + m^2)}(-m\hat{x} + \hat{y})^2\right). \qquad (2.2.7)$$

Alternatively, we can marginalize with Monte Carlo integration.

This method works well for Gibbs sampling: it is easy to find the conditional distributions.

This can be applied to SN Ia data.

Next homework assignment: exercise 10 for December, exercise 7 just before the exam.

2.3   SN Ia observation

Standard candles have a uniform intrinsic luminosity L, so that measuring the flux f at Earth allows us to measure the luminosity distance as

$$d_L = \sqrt{\frac{L}{4\pi f}}, \qquad (2.3.1)$$

and we quantify fluxes in terms of magnitudes:

$$\mu = m - M = 5\log_{10}\left(\frac{d_L}{10\,\text{pc}}\right), \qquad (2.3.2)$$

where m and M are the apparent and intrinsic magnitudes.

The dispersion of the intrinsic magnitudes is not zero, but it is rather small, on the order of $0.5\,\text{mag}$.

There is a correlation between the peak luminosity of the SN light curve and the length of the decay; also there is a correlation between color and peak width. Correcting for both of these, there is a dispersion of the order of 0.2 mag. Therefore, the final error in luminosity distance is on the order of 10 %.

The magnitudes of the stretch and color corrections (α and β) are not precisely known, we need to include them as nuisance parameters in our model.

> How do we know that the spread is that small?

The distance moduli depend on the redshifts z_i and on the cosmological parameters $\vec{p} = h, \Omega_m, \Omega_\Lambda$; however h is fully degenerate with the intrinsic magnitude M. This degeneracy must be broken in some way, for example with Cepheids or red giants. With SN Ia only we can, however, measure Ω_m and Ω_Λ by marginalizing over α, β, H_0 and M.

Classically what was done was χ^2 minimization, however there are several problems with it. March et al. [Mar+11] improved the result through the use of Bayesian Hierarchical Models.

2.4 Fisher Matrix

Suppose we have a data vector $\vec{x}$, such that each element is distributed according to a Gaussian with mean μ and standard deviation σ. We could construct an estimator for the mean like $\hat{\theta} = f(\vec{x}) = x_3$. This is unbiased: $\langle x_3 \rangle = \mu = \langle x_i \rangle$ for any i.

However, this is not the best estimator we could choose, as its variance is σ^2; we could instead use the arithmetic mean, which is also unbiased but has a variance of σ^2/N.

An *unbiased* estimator θ for $\hat{\theta}$ is one such that $\langle \hat{\theta} \rangle = \theta$.

Given a likelihood $\mathscr{L}(\vec{x}|\vec{\theta})$, the question is: what is the **best unbiased estimator** for θ? In order to answer this, we need to define the **Fisher information matrix**:

$$F_{\alpha\beta} = \left\langle -\frac{\partial^2 \log \mathscr{L}}{\partial \theta_\alpha \partial \theta_\beta} \right\rangle . \tag{2.4.1}$$

The average is an ensemble one, over all the possible realizations of the data. This is precisely the Hessian of the log-posterior, computed at the maximum of the posterior, which estimates the precision matrix of the parameters. We have nice theorems for these.

Theorem 2.4.1. *For any unbounded estimator $\hat{\theta}_\alpha$ of θ_α, its error is bounded by*

$$\sqrt{\left\langle \hat{\theta}_\alpha^2 \right\rangle - \left\langle \hat{\theta}_\alpha \right\rangle^2} \geq \sqrt{F_{\alpha\alpha}^{-1}} . \tag{2.4.2}$$

This is known as the **Cramer-Rao** bound. So, if we can saturate the inequality, then we know we have the best possible estimator.

Theorem 2.4.2. *If there exists a Best Unbiased Estimator, then this is also the Maximum Likelihood estimator.*

Theorem 2.4.3. *The ML estimate is asymptotically the BUE.*

Consider a likelihood $P(\vec{d}|\theta)$; then the variance is given by

$$\left\langle (\hat{\theta} - \langle\hat{\theta}\rangle)^2 \right\rangle = \int d\vec{d}\, P(\vec{d}|\theta)(\theta - \langle\hat{\theta}\rangle)^2. \tag{2.4.3}$$

Let us define the product of two functions as

$$\left(u(\vec{d}), v(\vec{d}) \right) = \int d\vec{x}\, u(\vec{x})v(\vec{x})P(\vec{x}|\theta). \tag{2.4.4}$$

Then, the variance of the estimator $\hat{\theta}$ can be expressed by defining $v(\vec{d}) = \hat{\theta} - \langle\hat{\theta}\rangle$, and then $\mathrm{var}\left(\hat{\theta}\right) = \left(v(\vec{d}), v(\vec{d}) \right)$.

Now we will use the Schwarz inequality, which says that in an inner product space:

$$(v, v) \geq \frac{(u, v)^2}{(u, u)}. \tag{2.4.5}$$

We will apply this, with the aforementioned v, and with the so-called *score function*:

$$u(\vec{d}) = \frac{\partial}{\partial\theta} \log P(\vec{d}|\theta). \tag{2.4.6}$$

Then,

$$(u, w) = \int d\vec{x}\, u(\vec{x})w(\vec{x})P(\vec{x}|\theta) \tag{2.4.7}$$

$$= \int d\vec{x}\, \frac{\partial}{\partial\theta} \log P(\vec{x}|\theta)w(\vec{x})P(\vec{x}|\theta) \tag{2.4.8}$$

$$= \int d\vec{x}\, \frac{1}{P(\vec{x}|\theta)}\frac{\partial P}{\partial\theta}w(\vec{x})P(\vec{x}|\theta) \tag{2.4.9}$$

$$= \frac{\partial}{\partial\theta} \langle w \rangle. \tag{2.4.10}$$

Then,

$$(u, v) = \int (\hat{\theta} - \langle\hat{\theta}\rangle)\frac{\partial \log P}{\partial\theta}P(\vec{x}|\theta)\, d\vec{x} \tag{2.4.11}$$

$$= \int \hat{\theta}\frac{\partial \log P}{\partial\theta}P(\vec{x}|\theta)\, d\vec{x} - \langle\hat{\theta}\rangle \underbrace{\int \frac{\partial \log P}{\partial\theta}P(\vec{x}|\theta)\, d\vec{x}}_{=\frac{\partial}{\partial\theta}\int P d\vec{x}=0} \tag{2.4.12}$$

$$= \int \hat{\theta}\frac{\partial \log P}{\partial\theta}P(\vec{x}|\theta)\, d\vec{x} \tag{2.4.13}$$

$$= \frac{\partial}{\partial\theta} \langle\hat{\theta}\rangle = 1, \tag{2.4.14}$$

therefore, we can apply Schwarz's inequality so that

$$\mathrm{var}\left(\hat{\theta}\right) = (v, v) \geq \frac{(u, v)^2}{(u, u)} = \frac{1}{(u, u)} = \left(\int dx \left(\frac{\partial}{\partial\theta} \log P \right)^2 P(\vec{x}|\theta) \right)^{-1} = \left\langle \left(\frac{\partial \log P}{\partial\theta} \right)^2 \right\rangle^{-1},$$

$$\tag{2.4.15}$$

so what we are left to prove is that

$$\left\langle \left(\frac{\partial \log P}{\partial \theta} \right)^2 \right\rangle = \left\langle \frac{\partial^2 \log P}{\partial \theta^2} \right\rangle . \tag{2.4.16}$$

The Fisher Information Matrix is defined as

$$F_{\alpha\beta} = \left\langle -\frac{\partial^2 \log \mathscr{L}}{\partial \theta_\alpha \partial \theta_\beta} \right\rangle_{\vec{\theta} = \vec{\theta}_{\max}} . \tag{2.4.17}$$

We almost proved the Cramer-Rao bound, however we still need to show that

$$\left\langle \left[\frac{\partial}{\partial \theta} \log \mathscr{L}(\vec{x}|\theta) \right]^2 \right\rangle = \left\langle -\frac{\partial^2}{\partial \theta^2} \log \mathscr{L}(\vec{x}|\theta) \right\rangle . \tag{2.4.18}$$

In order to do so, we can use the fact that the likelihood is normalized, therefore we can do what amounts to integrating by parts:

$$0 = \frac{\partial^2}{\partial \theta^2} \int \mathrm{d}^n x \, \mathscr{L} = \frac{\partial}{\partial \theta} \int \mathrm{d}^n x \, \frac{\partial}{\partial \theta} \mathscr{L} \tag{2.4.19}$$

$$= \frac{\partial}{\partial \theta} \int \mathrm{d}^n x \, \mathscr{L} \frac{\partial}{\partial \theta} \log \mathscr{L} = \int \mathrm{d}^n x \left[\mathscr{L} \frac{\partial^2}{\partial \theta^2} \log \mathscr{L} + \frac{\partial}{\partial \theta} \log \mathscr{L} \frac{\partial}{\partial \theta} \mathscr{L} \right] \tag{2.4.20}$$

$$= \left\langle \frac{\partial^2}{\partial \theta^2} \log \mathscr{L} \right\rangle + \left\langle \left[\frac{\partial}{\partial \theta} \log \mathscr{L} \right]^2 \right\rangle . \tag{2.4.21}$$

In the multivariate case, this becomes $F_{\alpha\beta} = \left\langle (\partial_\alpha \log \mathscr{L})(\partial_\beta \log \mathscr{L}) \right\rangle$; and the bound is $\mathrm{cov}\, \hat{\theta}_\alpha, \hat{\theta}_\beta \geq (F^{-1})_{\alpha\beta}$.

We would like to saturate this bound for our experiment, so we want to impose the condition

$$\frac{\partial}{\partial \theta} \log \mathscr{L}(\vec{x}|\theta) = q(\theta) \left[\hat{\theta} - \left\langle \hat{\theta} \right\rangle \right] . \tag{2.4.22}$$

This tells us that the likelihood must look like

$$\mathscr{L}(\vec{x}|\theta) = \frac{m(\vec{x})}{Z(\theta)} \exp \left[-\ell(\theta) \hat{\theta}(\vec{x}) \right] . \tag{2.4.23}$$

If the noise is Gaussian, then the likelihood will be a MVN:

$$\mathscr{L}(\vec{x}|\theta) = \frac{1}{(2\pi)^{N/2} \sqrt{\det C}} \exp \left(-\frac{1}{2} (\vec{x} - \vec{\mu})^\top C^{-1} (\vec{x} - \vec{\mu}) \right) . \tag{2.4.24}$$

In general, both μ and C may depend on the parameter(s) θ. We may compute the Fisher Matrix in this case: the log-likelihood looks like

$$-2 \log \mathscr{L} = N \log(2\pi) + \log \det C + (\vec{x} - \vec{\mu})^\top C^{-1} (\vec{x} - \vec{\mu}), \tag{2.4.25}$$

37

and now we will use a few linear algebra tricks: we start with $\log \det C = \text{Tr} \log C$, then we define the data array

$$D_{ij} = (x_i - \mu_i)(x_j - \mu_j),\tag{2.4.26}$$

then we rewrite the term as

$$(\vec{x} - \vec{\mu})^\top C^{-1}(\vec{x} - \vec{\mu}) = \text{Tr}\left[C^{-1}D\right],\tag{2.4.27}$$

so that $-2 \log \mathcal{L} = \text{const} + \text{Tr}\left[\log C + C^{-1}D\right]$.

From here we can start taking derivatives, and we will find an expression in terms of D and C, of which we must take the average; the averages we will need are

$$\left\langle D_{ij} \right\rangle = C_{ij}\tag{2.4.28}$$

$$\left\langle D_{ij,\alpha} \right\rangle = 0\tag{2.4.29}$$

$$\left\langle D_{ij,\alpha\beta} \right\rangle = \mu_{i,\alpha}\mu_{j,\beta} + \mu_{i,\beta}\mu_{j,\alpha}.\tag{2.4.30}$$

Finally, we get

$$F_{\alpha\beta} = \frac{1}{2} \text{Tr}\left[C^{-1}C_{,\alpha}C^{-1}C_{,\beta} + 2C^{-1}\vec{\mu}_{,(\alpha}^\top \vec{\mu}_{,\beta)}\right].\tag{2.4.31}$$

We find that the log-posterior can be expressed as

$$\log \mathcal{P}(\vec{\theta}|\vec{d}) = -\sum_{i=1}^{n} \frac{1}{2\sigma_i^2}\left(n^i - \frac{1}{\mathcal{R}}s_\alpha^i \Delta\bar{\theta}_\alpha - \frac{1}{2\mathcal{R}^2}s_{\alpha\beta}^i \Delta\bar{\theta}_\alpha \Delta\bar{\theta}_\beta + \dots\right)^2,\tag{2.4.32}$$

where $\mathcal{R}$ is the signal-to-noise ratio: if it is large, then we can linearize, and we are fine.

> Missing the last bit of the discussion

2.4.1 Galaxy surveys

Nowadays we can have 3D surveys, which make use of photometric and spectroscopic data to give distance information. The galaxy density field we observe is non-Gaussian, and we would expect it to be so even starting from Gaussian initial conditions since gravitational evolution is nonlinear.

Therefore, not all the information lies in the power spectrum. This means that it is hard to do the analysis on arbitrary scales; on the other hand if we average over $\gtrsim 10\,\text{Mpc}$ the field is indeed close to Gaussian, so we can analyze it in a simpler way. This is due to the hierarchical nature of the gravitational collapse, which is "bottom up".

So, on these large scales we can linearize. We work in Fourier space, defining:

$$\delta_g(k \ll 1) = \int \text{d}^3x\, \delta_g(\vec{x}) W_\lambda e^{-i\vec{k}\cdot\vec{x}},\tag{2.4.33}$$

where W_λ is our filter function, which has a wavelength of $\gtrsim 10\,\text{Mpc}$.

The power spectrum is defined as

$$\left\langle \delta_g(\vec{k}_1)\delta_g^*(\vec{k}_2) \right\rangle = (2\pi)^3 \delta^{(3)}(\vec{k}_1 - \vec{k}_2) P_g(k)\,, \tag{2.4.34}$$

where $k = \left|\vec{k}_1\right| = \left|\vec{k}_2\right|$.

The density will really be defined as

$$\Delta_i = \int \mathrm{d}^3x\, \psi_i(\vec{x}) \left(\frac{n(\vec{x}) - \bar{n}}{\bar{n}} \right)\,. \tag{2.4.35}$$

The function ψ can be chosen in real space to be the count-in-cells one:

$$\psi_i^{\text{CIC}}(\vec{x}) = \bar{n}[\vec{x} \in i\text{-th cell}]\,, \tag{2.4.36}$$

or the Fourier one:

$$\psi_{\vec{k}}(\vec{x}) = \frac{1}{V}e^{i\vec{k}\cdot\vec{x}}[\vec{x} \in V]\,. \tag{2.4.37}$$

In the latter case, we get

$$\Delta(\vec{k}) = \frac{1}{V\bar{n}}\int \mathrm{d}^3x\, e^{i\vec{k}\cdot\vec{x}}n(\vec{x}) - \frac{1}{V}\underbrace{\int \mathrm{d}^3x\, e^{i\vec{k}\cdot\vec{x}}}_{=\delta_{\vec{k},0}}\,, \tag{2.4.38}$$

where $\delta_{\vec{k},0}$ is a Kronecker delta between $\vec{k}$ and 0. From Δ we want to compute the power spectrum.

The galaxy density field can be written as

$$n(\vec{x}) = \sum_i \delta^{(3)}(\vec{x} - \vec{x}_i)\,, \tag{2.4.39}$$

therefore

$$\Delta(\vec{k}) = \frac{1}{\bar{n}V}\sum_i e^{i\vec{k}\cdot\vec{x}_i} - \delta_{\vec{k},0}\,. \tag{2.4.40}$$

We make a grid that is so fine that each cell contains at most 1 galaxy: then, we can write

$$\Delta(\vec{k}) = \frac{1}{\bar{n}V}\sum_{\alpha}^{N_{\text{cells}}} N_\alpha e^{i\vec{k}\cdot\vec{x}_\alpha} - \delta_{\vec{k},0}\,, \tag{2.4.41}$$

where $N_\alpha = 0, 1$ for each α.

Then, we can compute the power spectrum as

$$\begin{aligned}
\left\langle \Delta(\vec{k}_1)\Delta^*(\vec{k}_2) \right\rangle = {}& \frac{1}{(\bar{n}V)}\sum_{\alpha,\beta}\left\langle N_\alpha N_\beta \right\rangle e^{i\vec{k}_1\cdot\vec{x}_\alpha - i\vec{k}_2\cdot\vec{x}_\beta} + \delta_{\vec{k}_1,0}\delta_{\vec{k}_2,0} \\
& - \frac{1}{\bar{n}V}\delta_{\vec{k}_2,0}\sum_\alpha \left\langle N_\alpha \right\rangle e^{i\vec{k}_1\cdot\vec{x}_\alpha} - \frac{1}{\bar{n}V}\delta_{\vec{k}_1,0}\sum_\beta \left\langle N_\beta \right\rangle e^{i\vec{k}_2\cdot\vec{x}_\beta}\,.
\end{aligned} \tag{2.4.42}$$

Now, the last two terms can be computed making use of the fact that in a small cell of volume δV the expected number of galaxies is $\bar{n}\delta V$; therefore we can write

$$\frac{1}{\bar{n}V}\delta_{\vec{k}_2,0}\sum_\alpha \langle N_\alpha \rangle\, e^{i\vec{k}_1\cdot\vec{x}_\alpha} = \frac{1}{\bar{n}V}\delta_{\vec{k}_2,0}\sum_\alpha \bar{n}\delta V e^{i\vec{k}_2\cdot\vec{x}_\alpha} \tag{2.4.43}$$

$$\approx \delta_{\vec{k}_2,0}\left(\frac{1}{V}\int \mathrm{d}^3x\, e^{i\vec{k}_1\cdot\vec{x}_\alpha}\right) = \delta_{\vec{k}_2,0}\delta_{\vec{k}_1,0}, \tag{2.4.44}$$

which means that

$$\left\langle \Delta(\vec{k}_1)\Delta^*(\vec{k}_2)\right\rangle = \frac{1}{(\bar{n}V)^2}\sum_{\alpha,\beta}\left\langle N_\alpha N_\beta \right\rangle e^{i\vec{k}_1\cdot\vec{x}_\alpha - i\vec{k}_2\cdot\vec{x}_\beta} - \delta_{\vec{k}_1,0}\delta_{\vec{k}_2,0}. \tag{2.4.45}$$

If the positions of galaxies were purely random and unclustered we would have

$$\left\langle N_\alpha N_\beta \right\rangle = \langle N_\alpha\rangle\left\langle N_\beta\right\rangle = (\bar{n}V)^2, \tag{2.4.46}$$

so we define

$$\left\langle N_\alpha N_\beta \right\rangle = (\bar{n}\delta V)^2\left(1 + \left\langle \delta_g(\vec{x}_1)\delta_g(\vec{x}_2)\right\rangle\right) = (\bar{n}\delta V)^2\left(1 + \xi_{12}(|\vec{x}_1 - \vec{x}_2|)\right). \tag{2.4.47}$$

Also, we know that the autocorrelation can be written through the fact $N_\alpha = N_\alpha^2$, since $N_\alpha = 0,1$. Therefore,

$$\left\langle \Delta(\vec{k}_1)\Delta^*(\vec{k}_2)\right\rangle = \frac{1}{(\bar{n}V)}\left(\sum_\alpha \langle N_\alpha\rangle\, e^{i(\vec{k}_1-\vec{k}_2)\cdot\vec{x}_\alpha} + \sum_{\alpha\neq\beta}\left\langle N_\alpha N_\beta \right\rangle e^{i\vec{k}_1\cdot\vec{x}_\alpha - i\vec{k}_2\cdot\vec{x}_\beta}\right), \tag{2.4.48}$$

where the cross-term is given by

$$\frac{1}{(\bar{n}V)^2}\sum_{\alpha\neq\beta}\left\langle N_\alpha N_\beta \right\rangle e^{i\vec{k}_1\cdot\vec{x}_\alpha - i\vec{k}_2\cdot\vec{x}_\beta} = \left\langle \delta_g(\vec{k}_1)\delta_g^*(\vec{k}_2)\right\rangle = \frac{1}{V}P_g(k_1)\delta_{\vec{k}_1,\vec{k}_2}, \tag{2.4.49}$$

The final term represents shot noise, so this finally yields

$$\left\langle \Delta(\vec{k}_1)\Delta^*(\vec{k}_2)\right\rangle = \frac{1}{V}\left(P_g(k) + \frac{1}{\bar{n}}\right)\delta_{\vec{k}_1,\vec{k}_2}. \tag{2.4.50}$$

The shot noise term would be there even for a pure random process, while the $P_g(k)$ is our signal.

[Argument through characteristic functions as to why the total number of galaxies in each volume is Poisson distributed.]

For a small volume

$$Z_s(k) = 1 + \rho\delta V(e^{ik} - 1), \tag{2.4.51}$$

so for a large volume

$$Z_l(k) = [Z_s(k)]^{V/\delta V} \approx \exp\left[\rho V\left(e^{ik} - 1\right)\right], \tag{2.4.52}$$

40

therefore

$$Z_l(k) = \sum_{n=0}^{\infty} \frac{\lambda^n e^{-\lambda}}{n!} e^{ikn}, \tag{2.4.53}$$

where $\lambda = \rho V$; therefore the probability can be recovered as an inverse Fourier transform:

$$P_n = \frac{e^{-\lambda} \lambda^n}{n!}. \tag{2.4.54}$$

Then, we can compute the Fisher matrix from our general expression: the process is zero mean, so the mean-dependent terms cancel, so the expression is

$$F_{\alpha\beta} = \frac{1}{2} C_{ij,\alpha} C_{jm}^{-1} C_{mn,\beta} C_{ni}^{-1}, \tag{2.4.55}$$

where the covariance matrix and its inverse are diagonal:

$$C_{ij} = C(\vec{k}_i, \vec{k}_j) = \delta_{ij} \left[P_g(k_i) + \frac{1}{\bar{n}} \right] \tag{2.4.56}$$

$$C_{ij}^{-1} = \frac{\delta_{ij}}{P_g(k) + 1/\bar{n}}. \tag{2.4.57}$$

The derivative of the covariance matrix is

$$C_{ij,\alpha} = \frac{\partial C_{ij}}{\partial P_\alpha} = \delta_{ij} \frac{\partial P_g(k_i)}{\partial P_\alpha} = \delta_{ij} d_{i\alpha}, \tag{2.4.58}$$

where $d_{i\alpha} = [|\vec{k}_i| \in [k_\alpha, k_\alpha + \Delta k]]$ (using the Iverson bracket).

This then yields

$$F_{\alpha\beta} = \frac{1}{2} \delta_{ij} d_{i\alpha} \frac{\delta_{im}}{P_g(k_j) + 1/\bar{n}} (\delta_{mn} d_{n\beta}) \frac{\delta_{ni}}{P_g(k_i) + 1/\bar{n}} \qquad = \frac{1}{2} \frac{\sum_i d_{i\alpha}}{(P_\alpha + 1/\bar{n})^2}. \tag{2.4.59}$$

Now, we are working with a finite survey of volume V, therefore

$$\sum_i d_{i\alpha} = \frac{4\pi k_\alpha^2 \Delta k}{1/V} = 4\pi k_\alpha^2 \Delta k V. \tag{2.4.60}$$

This yields

$$F_{\alpha\beta} = \frac{1}{2} \delta_{\alpha\beta} \frac{4\pi k_\alpha^2 \Delta k V}{(P_g(k_\alpha) + 1/\bar{n})^2}, \tag{2.4.61}$$

so the Cramer-Rao bound is

$$\sigma_\alpha = \sqrt{F_{\alpha\alpha}^{-1}} = \left(P_g(k_\alpha) + \frac{1}{\bar{n}} \right) \frac{1}{\sqrt{2\pi k_\alpha^2 \Delta k V}}. \tag{2.4.62}$$

So, in order to decrease the error bar we want to increase $\bar{n}$ as much as possible (i.e. measure very dim galaxies), and increase the number of k modes: this is achieved by taking a survey with a volume which is as large as possible.

41

Chapter 3

Model selection

The Bayes factor between two models M' and M, depending on parameters θ' and θ respectively, having observed the data x, is

$$B = \frac{\int d\theta' \, \mathbb{P}(x|\theta' M')\mathbb{P}(\theta'|M')}{\int d\theta \, \mathbb{P}(x|\theta M)\mathbb{P}(\theta|M)} , \tag{3.0.1}$$

which, if the parameters are uniformly distributed in some range $\Delta\theta_i$, we will have

$$B = \frac{\int d\theta' \, \mathbb{P}(x|\theta'|M')}{\int d\theta \, \mathbb{P}(x|\theta, M)} \frac{\Delta\theta_1 \dots \Delta\theta_n}{\Delta\theta'_1 \dots \Delta\theta'_{n'}} . \tag{3.0.2}$$

This can account for having nested models, in which $n < n'$ and the prime parameters are a superset of the non-prime ones. The ratio of the likelihoods will favor the more general model (with θ'), while the ratio of the $\Delta\theta$ will reduce to $\Delta\theta'_{n+1} \dots \Delta\theta'_{n'}$, which will favor the simpler model.

In general computing the integrals must be done numerically, typically through MCMC. We will instead take the analytic approach. We consider the likelihoods as being MVNs around their maxima:

$$\mathcal{L}^{(i)} = \mathcal{L}_0^{(i)} \exp\left[\left(\vec{\theta}_i - \vec{\theta}_i^{ML}\right)^\top \left(\nabla\nabla\chi_i^2\right)\left(\vec{\theta}_i - \vec{\theta}_i^{ML}\right) \right]. \tag{3.0.3}$$

Now, we can calculate the integrals analytically:

$$\int d\theta \, \mathcal{L} = (2\pi)^{N/2} \left(\det\left(\nabla\nabla\chi^2\right) \right)^{-1/2} \mathcal{L}_0 , \tag{3.0.4}$$

so the likelihood ratio is just

$$\frac{\int d\theta_1 \, \mathcal{L}_1}{\int d\theta_2 \, \mathcal{L}_2} = (2\pi)^{(N_2 - N_1)/2} \frac{\mathcal{L}_I^{(\infty)}}{\mathcal{L}_I^{(2)}} \left[\frac{\det\left(\nabla\nabla\chi_{(2)}^2\right)}{\det\left(\nabla\nabla\chi_{(2)}^2\right)} \right]^{1/2} . \tag{3.0.5}$$

Therefore, for nested models

$$B_{12} = (2\pi)^{(N_2 - N_1)/2} \frac{\mathcal{L}_I^{(\infty)}}{\mathcal{L}_I^{(2)}} \left[\frac{\det\left(\nabla\nabla\chi_{(2)}^2\right)}{\det\left(\nabla\nabla\chi_{(2)}^2\right)} \right]^{1/2} \Delta\theta_2^{N_1+1} \dots \Delta\theta_2^{N_2} . \tag{3.0.6}$$

3.0.1 Savage-Dickey density ratio

This is an analytical tool which can be applied as long as we have nested models, with separable priors, and the more complex model only has one extra parameter.

Let us say that the likelihood for the complex model is $p_2(\vec{x}|\phi, \psi)$, while the likelihood of the simpler model is $p_1(\vec{x}|\phi) = p_2(\vec{x}|\phi, \psi_*)$. The Bayes factor is then

$$B_{12} = \frac{\int \mathrm{d}\phi \, p(\vec{x}|\phi, \psi_*)\pi_1(\phi)}{\int \mathrm{d}\phi \, \mathrm{d}\psi \, p(\vec{x}|\phi, \psi)\pi_2(\phi, \psi)} \, . \tag{3.0.7}$$

We call the bottom integral q. Let us consider $p(\psi_*|\vec{x})$: by Bayes' theorem, it is

$$p(\psi_*|\vec{x}) = \frac{\int \mathrm{d}\phi \, p(\vec{x}|\phi, \psi_*)\pi_2(\phi, \psi_*)}{q} \, , \tag{3.0.8}$$

therefore

$$B_{12} = p(\psi_*, \vec{x})\frac{\int \mathrm{d}\phi \, p(\vec{x}|\phi, \psi_*)\pi_1(\phi)}{p(\psi_*, \vec{x})} \tag{3.0.9}$$

$$= \frac{\int \mathrm{d}\phi \, p(\phi, \psi_*|\vec{x})}{\pi_2(\psi_*)} \, . \tag{3.0.10}$$

> Missing some calculation steps

Chapter 4

Choice of priors

If we have a location parameter θ, such that

$$\mathcal{L}_\theta(x) = \mathcal{L}(x - \theta),$$ (4.0.1)

then it makes sense to use a uniform prior $\pi(\theta) = \text{const}$, which is invariant under shifts in the form $\theta \to \theta + X$. On the other hand, if we have a scaling parameter

$$\mathcal{L}_\theta = \frac{1}{\theta}\mathcal{L}_\theta(x/\theta),$$ (4.0.2)

the requirements are different: we want θ to be invariant under rescalings in the form $\theta \to \theta/C = \phi$; so we ask

$$\pi(\theta)\,\mathrm{d}\theta = \pi(\phi)\,\mathrm{d}\phi$$ (4.0.3)

$$\pi(\theta) = \frac{1}{C}\pi(\theta/C),$$ (4.0.4)

which is satisfied by $\pi(\theta) \propto 1/\theta$. This prior is called a *log-uniform*: it is flat when plotted on a log scale, since if we set $\rho = \log\theta$ we have

$$\pi(\rho) = \pi(\theta)\left|\frac{\mathrm{d}\theta}{\mathrm{d}\rho}\right| = \frac{e^\rho}{\theta} = 1.$$ (4.0.5)

4.0.1 Jeffreys prior

In general, if we want to perform a change of variables from θ to ϕ then the Fisher matrix becomes

$$F(\phi) = -\left\langle \frac{\mathrm{d}^2 \log\mathcal{L}(\vec{x}|\phi)}{\mathrm{d}\phi^2} \right\rangle$$ (4.0.6)

$$= -\left\langle \frac{\mathrm{d}^2 \log\mathcal{L}(\vec{x}|\theta)}{\mathrm{d}\theta^2}\left(\frac{\mathrm{d}\theta}{\mathrm{d}\phi}\right)^2 + \frac{\mathrm{d}\log\mathcal{L}(\vec{x}|\theta)}{\mathrm{d}\theta}\frac{\mathrm{d}^2\theta}{\mathrm{d}\phi^2} \right\rangle$$ (4.0.7)

$$= -\left\langle \frac{\mathrm{d}^2 \log \mathcal{L}(\vec{x}|\theta)}{\mathrm{d}\theta^2} \right\rangle \left(\frac{\mathrm{d}\theta}{\mathrm{d}\phi} \right)^2 - \left\langle \frac{\mathrm{d} \log \mathcal{L}(\vec{x}|\theta)}{\mathrm{d}\theta} \right\rangle \frac{\mathrm{d}^2\theta}{\mathrm{d}\phi^2} \qquad (4.0.8)$$

$$= \underbrace{\left\langle \frac{\mathrm{d}^2 \log \mathcal{L}(\vec{x}|\theta)}{\mathrm{d}\theta^2} \right\rangle}_{F(\theta)} \left(\frac{\mathrm{d}\theta}{\mathrm{d}\phi} \right)^2. \qquad (4.0.9)$$

The prior is defined as the square root of the Fisher information matrix:

$$\sqrt{F(\phi)} = \sqrt{I(\theta)} \left| \frac{\mathrm{d}\theta}{\mathrm{d}\phi} \right| \qquad (4.0.10)$$

$$\pi(\phi) = \pi(\theta) \left| \frac{\mathrm{d}\theta}{\mathrm{d}\phi} \right|. \qquad (4.0.11)$$

If we use this definition, we get the correct law for the change of variable.

We can apply this to a univariate Gaussian with fixed σ: we get

$$\frac{\mathrm{d}^2 \log \mathcal{L}}{\mathrm{d}\mu^2} = -\frac{1}{\sigma^2}, \qquad (4.0.12)$$

therefore $F(\mu)$ is a constant, meaning that the Jeffreys prior $\pi_J(\mu)$ is also a constant.

Suppose instead that we also let σ vary: then

$$\frac{\mathrm{d} \log \mathcal{L}}{\mathrm{d}\sigma} = \frac{\mathrm{d}}{\mathrm{d}\sigma} \left(\log \sigma - \frac{1}{2} \frac{(x-\mu)^2}{\sigma^2} \right) \quad \frac{\mathrm{d}^2 \log \mathcal{L}}{\mathrm{d}\sigma^2} = \frac{1}{\sigma^2} - \frac{3(x-\mu)^2}{\sigma^4} \quad (4.0.13)$$

$$F(\sigma) = -\left\langle \frac{\mathrm{d}^2 \log \mathcal{L}}{\mathrm{d}\sigma^2} \right\rangle = -\frac{1}{\sigma^2} + \frac{3\sigma^2}{\sigma^4} = \frac{2}{\sigma^2}. \qquad (4.0.14)$$

Therefore, $F(\sigma) \propto 1/\sigma^2$, meaning that $\pi_J(\sigma) \propto 1/\sigma$.

Uniform priors may pose an issue in high-dimensional space: especially if the mean $\vec{\mu}$ is high-dimensional, a uniform prior on its entries gives a lot of weight to high-length vector. Often people resort to information-theory based criteria.

We discuss Lahav et al. [Lah+00].

There will be a couple of optional additional lectures uploaded to the Moodle.

Today we will discuss [HBL02]. The idea is to combine the χ^2 for the various experiments with various weights, which are our hyperparameters.

Do we actually need these hyperparameters, or can we avoid using them by simply adding the χ^2? This is a model selection problem.

In the hyperparameter model the likelihood is not a Gaussian anymore. The weights must be larger than 0, we can take them to be larger than 1 if the corresponding experiment was overestimating its errorbars.

Bibliography

[HBL02]   M. P. Hobson, S. L. Bridle, and O. Lahav. "Combining Cosmological Datasets: Hyperparameters and Bayesian Evidence". In: *Monthly Notices of the Royal Astronomical Society* 335.2 (Sept. 2002), pp. 377–388. ISSN: 0035-8711, 1365-2966. DOI: 10.1046/j.1365-8711.2002.05614.x. arXiv: astro-ph/0203259. URL: http://arxiv.org/abs/astro-ph/0203259 (visited on 2020-12-21).

[Lah+00] O. Lahav et al. "Bayesian 'hyper-Parameters' Approach to Joint Estimation: The Hubble Constant from CMB Measurements". In: *Monthly Notices of the Royal Astronomical Society* 315.4 (July 11, 2000), pp. L45–L49. ISSN: 0035-8711, 1365-2966. DOI: 10.1046/j.1365-8711.2000.03633.x. URL: https://academic.oup.com/mnras/article-lookup/doi/10.1046/j.1365-8711.2000.03633.x (visited on 2020-12-15).

[Mar+11] M. C. March et al. *Improved Constraints on Cosmological Parameters from SNIa Data*. Aug. 8, 2011. DOI: 10.1111/j.1365-2966.2011.19584.x. arXiv: 1102.3237 [astro-ph]. URL: http://arxiv.org/abs/1102.3237 (visited on 2020-11-23).

[SS06]   Devinderjit Sivia and John Skilling. *Data Analysis: A Bayesian Tutorial*. Oxford University Press, June 2006. 259 pp. ISBN: 978-0-19-856831-5. Google Books: 1YMSDAAAQBAJ.

[Spe+03] D. N. Spergel et al. "First Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Determination of Cosmological Parameters". In: *The Astrophysical Journal Supplement Series* 148.1 (Sept. 2003), pp. 175–194. ISSN: 0067-0049, 1538-4365. DOI: 10.1086/377226. arXiv: astro-ph/0302209. URL: http://arxiv.org/abs/astro-ph/0302209 (visited on 2020-11-17).

[Van14]   Jake Vanderplas. *Frequentism and Bayesianism III: Confidence, Credibility, and Why Frequentism and Science Do Not Mix | Pythonic Perambulations*. June 12, 2014. URL: http://jakevdp.github.io/blog/2014/06/12/frequentism-and-bayesianism-3-confidence-credibility/ (visited on 2020-10-26).

[Ver+03] L. Verde et al. "First Year Wilkinson Microwave Anisotropy Probe (WMAP) Observations: Parameter Estimation Methodology". In: *The Astrophysical Journal Supplement Series* 148.1 (Sept. 2003), pp. 195–211. ISSN: 0067-0049, 1538-4365. DOI: 10.1086/377335. arXiv: astro-ph/0302218. URL: http://arxiv.org/abs/astro-ph/0302218 (visited on 2020-11-17).

[Win]   Keith Winstein. *What's the Difference between a Confidence Interval and a Credible Interval?* URL: https://stats.stackexchange.com/q/2287.