AstroStatistics and Cosmology

Jacopo Tissino

2020-11-02

Contents

Introduction

Monday 2020-9-28, compiled 2020-11-02

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**:

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

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**:

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

Tuesday 2020-9-29, compiled 2020-11-02

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:

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

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!

Monday 2020-10-5, compiled 2020-11-02

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

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

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

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

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

which is a Gaussian whose variance is

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

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:

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

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

$$\frac{\mathrm{d}L}{\mathrm{d}H} = \frac{\mathrm{d}\log P}{\mathrm{d}H} = \frac{\mathrm{d}}{\mathrm{d}H}\big[R\log H + (N-R)\log(1-H)\big] \tag{1.1.15}$$

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

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:

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

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

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.

Tuesday 2020-10-6, compiled 2020-11-02

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.

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

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

Monday 2020-10-12, compiled 2020-11-02

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:

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

therefore

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

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.

Tuesday 2020-10-13, compiled 2020-11-02

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 $\vec{t} = (x, y)$

$$L(x,y) = L(x_0,y_0) + \frac{1}{2}\left.\frac{\partial^2 L}{\partial x^2}\right|_{\vec{t}_0}(x-x_0)^2 + \frac{1}{2}\left.\frac{\partial^2 L}{\partial y^2}\right|_{\vec{t}_0}(y-y_0)^2 + \left.\frac{\partial^2 L}{\partial y \partial x}\right|_{\vec{t}_0}(x-x_0)(y-y_0) \tag{1.2.6}$$

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

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

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

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

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

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:

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

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

Monday 2020-10-19, compiled 2020-11-02

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 $O^\top VO$, 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

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

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

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

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

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

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 $\vec{y} = \vec{x} - \vec{\mu}$:

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

We then use the square completion formula:

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

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

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

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

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

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

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 $\vec{\mu}_{b|a} = \vec{\mu}_b - V_{bb}^{-1}V_{ba}\left(\vec{x}_a - \vec{\mu}_a\right)$, 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

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

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

$$\mathbb{P}(z) = \int \mathrm{d}x\,\mathrm{d}y\,\mathbb{P}(x)\mathbb{P}(y)\delta(z - (x+y)) \tag{1.3.12}$$

$$= \int \mathrm{d}x\,\mathbb{P}(x)\mathbb{P}(z-x)\,, \tag{1.3.13}$$

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

$$\langle x + y\rangle = \langle x\rangle + \langle y\rangle\,, \tag{1.3.14}$$

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

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

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

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 $\left\langle x^3 \right\rangle = 0$); however x and x^2 are related two-to-one, certainly not independent.

1.3.1 Correlation

Tuesday 2020-10-20, compiled 2020-11-02

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:

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

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

The 2π factor disappears since the log of a ratio is the difference of the logs, so it becomes an additive constant.

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

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

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

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

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

1.4 Two-parameter estimation revisited

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

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

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

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

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

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

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

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

Therefore, the log-posterior reads

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

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

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

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

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

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

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

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

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

therefore

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

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.

Monday 2020-10-26, compiled 2020-11-02

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

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

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

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 distibution* (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.

Tuesday 2020-10-27, compiled 2020-11-02

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

$$\partial_k L(\vec{x}) \approx \delta^i_k\partial_i L(x_1) + (x-x_1)^i\partial_i\partial_k L(x_1) \tag{1.5.16}$$

$$\partial_k L(\vec{x}) \approx \partial_k L(x_1) + (x-x_1)^i\partial_i\partial_k L(x_1) \tag{1.5.17}$$

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

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

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

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

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

Monday 2020-11-2, compiled 2020-11-02

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 $N \to \infty$ 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 $D \to \infty$.

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:

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

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

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

These will need to be $|\Omega|^2$ numbers, and are called *transition probabilities* in a **transition matrix**.

The transition matrix can be written as

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

[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).

Bibliography

[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: lYMSDAAAQBAJ.

[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).

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