PHYS 644 Lecture #24: Inference

What we have done so far is to talk about how measurements of the matter power spectrum and the CMB allow us to measure cosmological parameters:

$\Rightarrow$ Show cosmo param summary

But so far we only have the ability to "eyeball" these parameters. And how do we place rigorous error bars?

This is the subject of inference.

Inference Example

Suppose we have measured the heights of 100 people. We would like to use this dataset to infer to mean height $\hat{h}$ and variance σ_h^2.

Data: $\vec{d} = (d_1, d_2, \ldots, d_{100})$

Parameters: $\vec{\theta} = (\hat{h}, \sigma_h^2)$

Going from a statistical description of the data to parameters is what we call inference.

To connect the data to the parameters we need a theory or a model.

Here let's assume that the ~~measurement~~ data ~~error~~ is Gaussian, so that for one person's height we have

$$p(d_i | \hat{h}, \sigma_h^2) = \frac{1}{\sqrt{2\pi\sigma_h^2}} \exp\left[- \frac{(d_i - \hat{h})^2}{2\sigma_h^2} \right]$$

Hilbert

Probability density of measuring d_i given these theory parameters

If we assume that the measurements are independent (~~the sense of their errors~~) then the probability of measuring the collection of heights is just the product:

$$\mathcal{L}(\vec{d}; h, \sigma_h^2) \equiv p(\vec{d}\,|\,h,\sigma_h^2) = \prod_{i=1}^{N_h} \frac{\exp\left[-\frac{(d_i-h)^2}{2\sigma_h^2}\right]}{\sqrt{2\pi\sigma_h^2}}$$

"Likelihood fct"

$$= \frac{1}{(2\pi\sigma_h^2)^{N_h/2}} \exp\left[-\frac{\sum_{i=1}^{N_h}(d_i-h)^2}{2\sigma_h^2}\right].$$

The pdf $p(\vec{d}\,|\,\vec{\theta})$ is the key piece that connects data to theory and is so important that we call it by a special name: the likelihood.

The likelihood is a probability distribution for the possible data I can measure given the theory (and its parameters). Phrased in this way, we can think about it as capturing our forward model from theory to data. → "Theory fixed; data random"

This way of thinking is very natural in particle physics. We are essentially asking what is the distribution of data we could measure for an experiment we can repeat many times (Eg at a particle collider).

Even if we often pretend we can generate lots of Universes when we take ensemble averages ⟨...⟩

But in cosmology we only have one universe! What we would like to operate in is a methodology of

"Data fixed; theory random"

Fixed to what we measured

Usually by this we want a prob. dist. of parameters.

Hilroy

As a first guess, we could say the following: the likelihood $\mathcal{L}$ is a multivariate function of both the data $\vec{x}$ and the parameters $\vec{\theta}$. Before we held $\vec{\theta}$ fixed and plotted $\vec{x}$. Why don't we just hold $\vec{x}$ fixed and plot as a function of $\vec{\theta}$?

This is conceptually the right idea, and in some cases (see below) IS the right thing. To do this rigorously, though, we need Bayes' theorem:

$$p(A|B) = \frac{p(B|A)\,p(A)}{p(B)}.$$

In our case, we ~~have~~ want $p(\vec{\theta}|\vec{x})$, the probability distribution of parameters $\vec{\theta}$ given our measured data $\vec{x}$. Thus, $A \equiv \vec{\theta}$ and $B = \vec{x}$. And:

$$p(\vec{\theta}|\vec{x}) \propto p(\vec{x}|\vec{\theta})\,p(\vec{\theta})$$

"posterior"

$$= \mathcal{L}(\vec{x};\vec{\theta})\,p(\vec{\theta})$$

"likelihood" "prior"

To ~~be~~ get the posterior distribution we want, we do indeed keep data fixed and interpret the likelihood as a function of the parameters, but in principle we need to multiply by the prior.

The prior encodes our prior knowledge before taking the data. This knowledge could come from:
i) Physical arguments (Eg Ω_m cannot be negative)
ii) Previous experiments.
iii) Reasonable guesses.

Hilbury

This is not as innocent as it sounds! For example, uniform in σ_h is not the same as uniform in σ_h^2!

When the data is really good, the prior has little effect.

Suppose we picked a uniform prior ie $p(\vec{\theta}) = \text{constant}$ to be reasonably non-commital. Then our guessed prescription of just evaluating the likelihood as a function of $\vec{\theta}$ is basically correct!

(Note: I skiped the denominator ~~p(d)~~ $p(\vec{d})$ because it doesn't depend on $\vec{\theta}$. It normalizes the probability distribution but we can just do this ourselves afterwards).

Let's go back to our original example. Suppose we know σ_h (somehow) and we are just trying to infer h. Then:

$$p(h|\vec{d}) \propto \mathcal{L}(\vec{d};h) = \frac{1}{(2\pi\sigma_h^2)^{N_h/2}} \exp\left[-\frac{\sum_{i=1}^{N_h}(d_i-h)^2}{2\sigma_h^2}\right]$$

Let me manipulate the sum in the exponent:

$$\sum_{i=1}^{N_h}(d_i-h)^2 = \sum_i (h^2 - 2d_i h) + \text{const.}$$

(const. doesn't depend on h!)

$$= N_h h^2 - 2h\sum_i d_i + \text{const.}$$

Defined $\hat{h} \equiv \frac{\sum_i d_i}{N_h}$

$$= N_h(h^2 - 2h\hat{h}) + \text{const.}$$

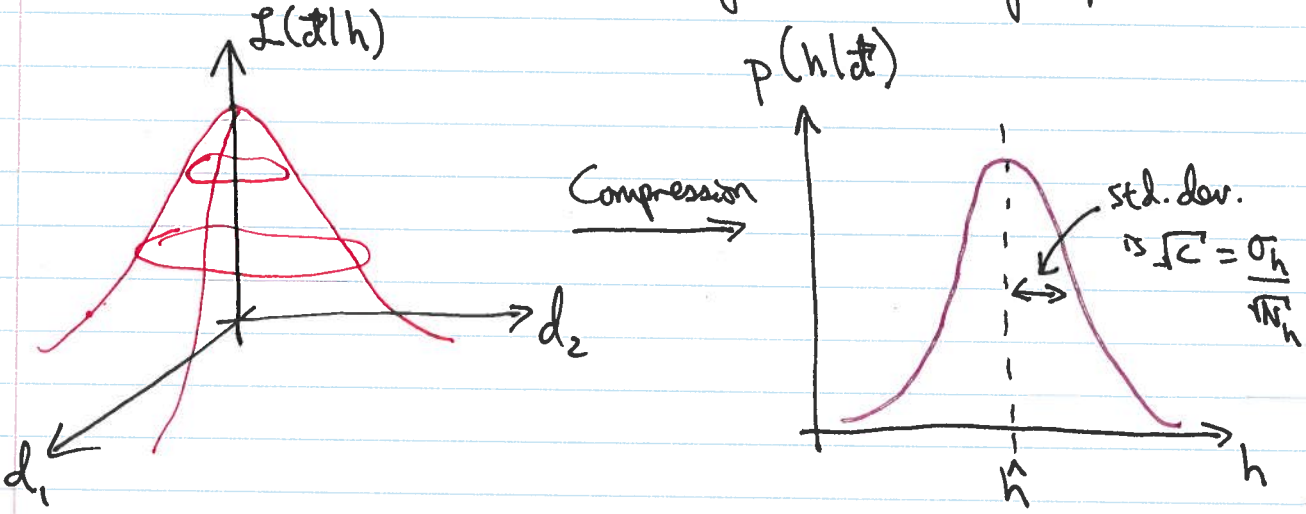
Complete the square

$$= N_h(h-\hat{h})^2 + (\text{different}) \text{ const.}$$

Reinserting this into our $p(h|\vec{d})$ expression and normalizing (that's why the constants didn't matter!)

$$p(h|\vec{d}) = \frac{1}{\sqrt{2\pi C}} \exp\left[-\frac{(h-\hat{h})^2}{2C}\right]$$

where $C \equiv \sigma_h^2 / N_h$. One way to think about this is that we have done a data compression operation, compressing the $N_h = 100$ datapoints into knowledge about a single parameter



Ideally we should just provide $p(h|\vec{d})$ when asked to summarize our results on the inference of h. But if asked to give one number and error bars, it would be sensible to say

$$\hat{h} = \frac{\sum_i d_i}{N_h} \quad \text{and} \quad \Delta\hat{h} = \frac{\sigma_h}{\sqrt{N_h}}$$

← Standard $1/\sqrt{N}$ result!

Here this "best fit" is a maximum likelihood solution. If we had multiplied by a non-flat prior, then the parameter values where $p(h|\vec{d})$ peaks would be called the maximum a posteriori solution.

I used a hat ($\hat{h}$) because we often use that to denote an estimator that is a recipe for some statistic. It's often a way of getting an estimate that's more lightweight than the full posterior.

This theme of data compression is quite common in cosmological analyses:

Time-ordered time-stream data from telescopes ($N \sim 10^8$ to 10^9?) ⟶ Maps (Eg of the CMB) ($N \sim 10^6$ to 10^7)

Maps ($N \sim 10^6$ to 10^7) ⟶ Power spectra (or other summary statistics) ($N \sim 10^3$)

Power spectra ($N \sim 10^3$) ⟶ Cosmological parameters ($N \sim 10$)

Let's look at an example of the last step. In this case, we might have our data be the CMB angular power spectrum, so $\vec{d} = \{C_\ell\}$, and we want cosmological parameters so $\vec{\theta} = \{\tau, n_s, \Omega_b h^2, \ldots\}$.

Ideally, we would plot $p(\vec{\theta}|\vec{d})$, but now we have a ~~visualization~~ visualization problem. Unlike our simple example, there are multiple parameters and their uncertainties may also be correlated, so in general we have a complicated higher-dimensional surface. How do we visualize this? One tool is the corner/triangle plot.

⇒ Show Planck corner plot

To understand this, we need to understand marginalization. Hilbert

Suppose I had a three-parameter probability distribution $p(\theta_1, \theta_2, \theta_3 | \vec{x})$. If I don't care about θ_3, I can integrate or marginalize over all its possible values:

$$p(\theta_1, \theta_2 | \vec{x}) = \int_{-\infty}^{+\infty} p(\theta_1, \theta_2, \theta_3 | \vec{x}) \, d\theta_3.$$

We could go even further and marginalize over everything except for one parameter:

$$p(\theta_1 | \vec{x}) = \int_{-\infty}^{\infty} p(\theta_1, \theta_2, \theta_3 | \vec{x}) \, d\theta_2 d\theta_3.$$

On a corner plot, we have on the diagonal the 1D marginalized posterior distributions for each parameter, and on the off-diagonals the pairwise distributions. Contours typically 68% and 95% credibility.

In fact, typically in these plots there has been additional marginalization that hasn't been shown, over nuisance parameters like instrument calibration parameters

The off-diagonal parts of the corner plot also shows the idea of a degeneracy — where we aren't able to determine a parameter very well because its value can be traded off another parameter and still fit the data well.

⇒ Show degeneracy plot and Q's

One thing that we haven't talked about (and won't) in this course is how I actually get these shapes. Brute force evaluation is not practical (the marginalization requires very high dimensional integration).

Typically, one resorts to Markov Chain Monte Carlo (MCMC) methods that allow one to draw samples from $p(\vec{\theta} | \vec{x})$. It is a way of walking through parameter space where the amount

of time spent in a location is proportional to $p(\vec{\theta}|\vec{x})$. Keeping track of the locations that our walker visited then gives us a chain of locations in parameter space. Histogramming these chains gives $p(\vec{\theta}|\vec{x})$! Marginalization is also trivial! Just ignore the variables you don't care about when histogramming!

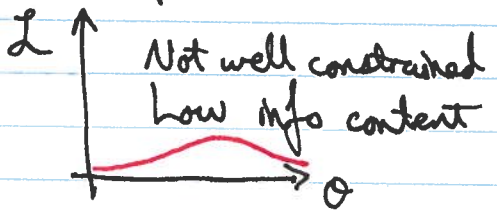
⇒ MCMC movie and marginalization slides

Fisher Matrices

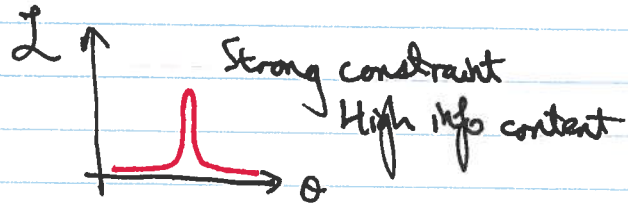
What I just outlined is what you might do if you had real data. But suppose I wanted just a quick and dirty way to forecast the performance of a survey. Is there an easy way to estimate how well an experiment can constrain parameters without simulating data?

Yes — the Fisher matrix.

The "peakier" the likelihood, the better we are able to constrain a parameter



Low curvature



High curvature.

The Fisher matrix quantifies information via the curvature. Taylor expand $\ln \mathcal{L}$ around the peak:

$$\ln \mathcal{L} = \ln \mathcal{L}\Big|_{\substack{\text{max} \\ \text{likelihood}}} + \sum_i (\theta_i - \theta_i^{max}) \frac{\partial \ln \mathcal{L}}{\partial \theta_i}\Big|_{max}$$

$$+ \frac{1}{2} \sum_{i,j} (\theta_i - \theta_i^{max})(\theta_j - \theta_j^{max}) \frac{\partial^2 \ln \mathcal{L}}{\partial \theta_i \partial \theta_j}\Big| + \cdots$$

Hilbroy

The 1st term is constant. The 2nd term is zero because the first derivative is zero at a maximum.

What controls the width of the distribution is the second derivative term, i.e. the Hessian matrix. This is basically the Fisher matrix:

$$F_{\alpha\beta} = \left\langle -\frac{\partial^2 \ln \mathcal{L}}{\partial\theta_\alpha \partial\theta_\beta} \right\rangle.$$

Especially when errors are small

Note that a Gaussian is a quadratic in $\ln \mathcal{L}$, so by discarding higher order terms, we are basically saying that we are approximating the likelihood surface as Gaussian. This is usually not awful, but be careful

⇒ Show banana slide

That said, Fisher matrices can be very helpful for a number of reasons:

① The covariance matrix of parameters $\Sigma \equiv \mathrm{Cov}(\theta_i, \theta_j) = \langle (\vec{\theta} - \vec{\theta}_{true})(\vec{\theta} - \vec{\theta}_{true})^t \rangle$ is well approximated by F^{-1}.

② We can study degeneracies:

Error bar on θ_i

$$\Delta\theta_i = \begin{cases} \sqrt{(F^{-1})_{ii}} & \text{(marginalizing all others)} \\ 1/\sqrt{F_{ii}} & \text{(fixing all others)} \end{cases}$$

③ Cramer-Rao bound says that the Fisher estimate of errors is the best possible if you do a perfect, optimal analysis.

Hilbert

④ Can be computed without data. For example, if the data is also Gaussian distributed with mean $\vec{\mu}$ and covariance C, then one can show that

$$F_{\alpha\beta} = \frac{\partial \vec{\mu}^t}{\partial \theta_\alpha} C^{-1} \frac{\partial \vec{\mu}}{\partial \theta_\beta} + \frac{1}{2} \mathrm{tr}\left[C^{-1} \frac{\partial C}{\partial \theta_\alpha} C^{-1} \frac{\partial C}{\partial \theta_\beta} \right]$$

⑤ For independent surveys with independent errors trying to measure the same parameters, Fisher information adds.

$$F^{\text{total}} = F^{\text{CMB}} + F^{\text{galaxies}} + F^{\text{lensing}} + \ldots$$

⇒ Show cosmo slide

As an example, for the CMB the Fisher matrix is

$$F_{\alpha\beta} = \sum_{l=2}^{l_{\max}} f_{\text{sky}} \left(\frac{2l+1}{2} \right) \left[C_l + \frac{4\pi\sigma^2}{N} e^{\theta_b^2 l(l+1)} \right]^{-2} \left(\frac{\partial C_l}{\partial \theta_\alpha} \right) \left(\frac{\partial C_l}{\partial \theta_\beta} \right)$$

beam smearing width

fraction of sky observed

CMB power spectrum

Notice how this depends on derivatives of C_l on parameters — makes sense, because the more the observable quantity (C_l) changes when we wiggle parameters, the more sensitive our probe

⇒ Show CMB derivatives

Thinking of each curve as a vector, the higher their "dot product" the more degenerate they are.