What Can a Bayesian Say About y/x?

Lorne Whiteway
lorne.whiteway.13@ucl.ac.uk

Astrophysics Group
Department of Physics and Astronomy
University College London

FLATs
21 November 2025

Purpose of presentation

- I'll show how to do Bayesian inference in a context that is straightforward but non-trivial.
- Talk will be pedagogical and mostly mathematical.
- Prior knowledge of Bayesian ideas is useful but not necessary.
- Thanks to Niall Jeffrey for help.

The Original Cosmological Motivation (1)

- The *overdensity* of dark matter (DM), denoted δ_{DM}, is the percentage difference between the local density (in some volume) and the universal average density.
- Overdensities have existed since early times. Gravity draws DM into overdense regions, so the overdensities get bigger with time.
- There is a similar definition for baryons. Baryons fall into DM overdensities, so we get baryon overdensities in the same place.

The Original Cosmological Motivation (2)

- Physical effects mean that the baryon overdensity need not equal the DM overdensity; a simple model assumes a linear relationship:

$$b = \frac{\delta_{\mathrm{baryon}}}{\delta_{\mathrm{DM}}}$$

- In a recent paper (`https://arxiv.org/abs/2509.18967`), we (Ellen, Niall, LW, Ofer, Josh, et al.) measured the baryon overdensity (from galaxy counts) and the DM overdensity (from weak-lensing mass maps). But how then to infer something about b?

Restate as a Statistics Problem

- We make noisy measurements of x and y and we want to infer $b = y/x$.
- The answer should be probabilistic (due to the uncertainty arising from the measurement noise).

Noise model

- Let's assume a simple model for the measurement noise:

  The measurement noise in x and the measurement noise in y have uncorrelated Gaussian distributions with zero means and with standard deviations σ_x and σ_y (which we will assume to both be unity).

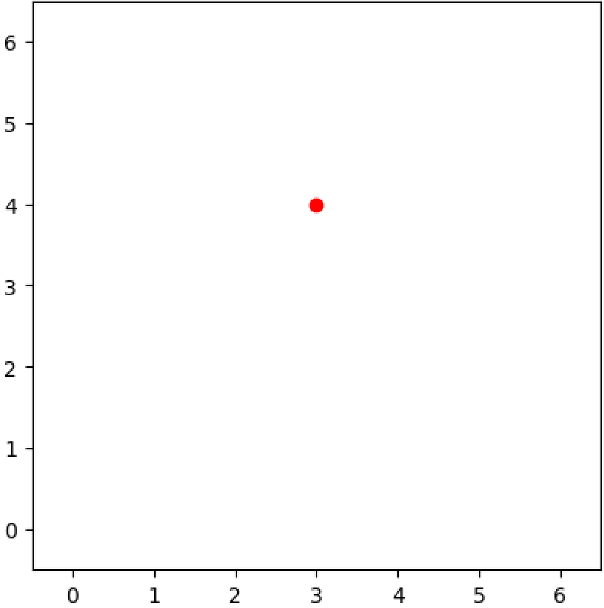
Data

- We make one observation of x (call the measured value $\tilde{x}$) and one of y (call the measured value $\tilde{y}$).
- Assume that we are in the low signal-to-noise regime, so $\tilde{x}$ and $\tilde{y}$ are 'a few' (e.g. not 'a few thousand'). For example, $\tilde{x} = 3$ and $\tilde{y} = 4$.

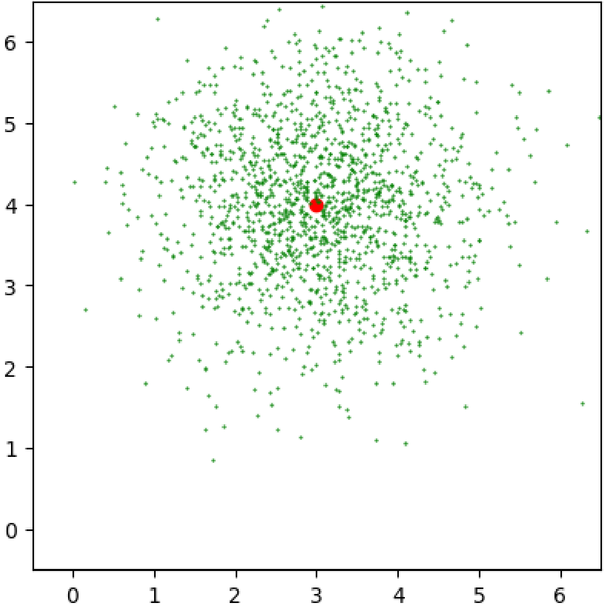
Naive calculation

- So here's a naive calculation that appears reasonable to do:
- By subtracting the (unknown) noise from $(\tilde{x}, \tilde{y})$ we get a Gaussian distribution of the true values (x, y) that is centred on $(\tilde{x}, \tilde{y})$. For each possible true (x, y) there is a true $b = y/x$, hence we have a distribution p of the true b.

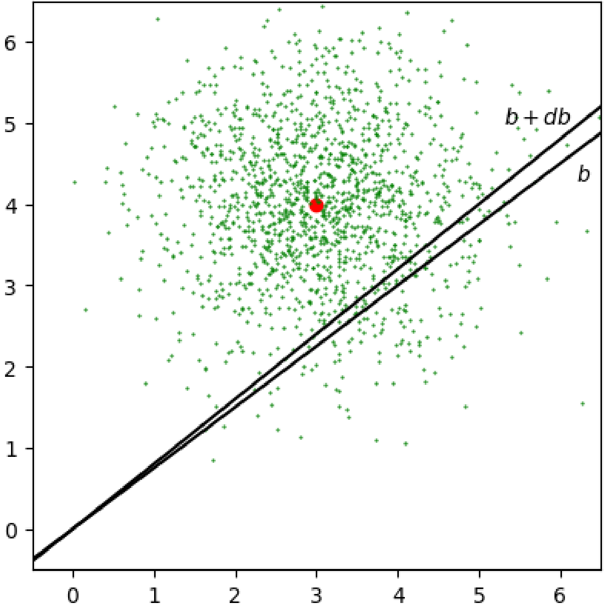
Naive calculation



Naive calculation



Naive calculation



Naive calculation

- So to calculate $p(b)\mathrm{d}b$ we need to calculate the probability mass within the thin pie-shaped region.
- Note that this region doesn't have constant angular width!
- Intuition: should we expect the distribution of b to be Gaussian? symmetric?
- Do you expect its mode to be at $\tilde{b} = \tilde{y}/\tilde{x}$?

Naive calculation

The calculation we need is

$$C \int_{\mathrm{pie}} \exp(-\frac{1}{2}(x - \tilde{x})^2 + (y - \tilde{y})^2)) \, \mathrm{d}x \, \mathrm{d}y \tag{1}$$

Naive calculation

Tricks for doing this calculation:

- Use polar coordinates $\mathrm{d}x \; \mathrm{d}y = r \; \mathrm{d}r \; \mathrm{d}\theta$.
- Don't forget the opposite quadrant! So $-\infty < r < \infty$.
- Complete the square.

What is the statistical theory behind this calculation?

Two main statistical paradigms, with different notions of what ‘probability’ means:

- Frequentist. Probability is *limiting ratio as the number of trials goes to infinity* e.g. probability of rolling 8 with two dice is $5/36$. The true value are fixed, and cannot be modelled probabilistically; randomness enters via measurement noise.
- Bayesian. Probability is *subjective degree of belief* (subjective as it depends on how much information we have). We don’t know the true value, hence we can speak probabilistically about it.

Naive calculation

- So the naive calculation can't be frequentist (as it uses the notion of a probability distribution of the true x and y).
- So it's Bayesian - but half-baked! Let's do it properly.

Bayesian theory requires *context*

- Context here means a model. The model will generally go all the way from (at the start) some model parameters, via various calculations (some mathematical, some physical, and certainly some modelling the effect of measurement noise) through to (at the end) observables i.e. measured data.
- Context also means that we already have existing results (e.g. from prior experiments) on the probability distributions of the model parameters.
- You have enough context when you can generate plausible new 'synthetic' measurement results.

Bayesian inference

- With new data in hand, we use the rules of conditional probability to refine the probability distribution of the model parameters, moving from the *prior* probability that we had before our experiment to the new *posterior* probability after our experiment.
- The result from conditional probability is Bayes's Theorem:

$$p(\mathrm{params}|\mathrm{data}) \propto p(\mathrm{data}|\mathrm{params})\ p(\mathrm{params})$$

- This gives us the shape of the posterior distribution of the parameters, but not the overall normalisation; in many applications that doesn't matter.

Likelihood

- The first term on the right hand side, $p(\mathrm{data}|\mathrm{params})$, is the *likelihood*. It's a probability distribution for the data (e.g. if we integrate over all possible data, with the params held fixed, we will get unity).
- But we use it *the other way around*: we hold the data fixed, and see how the likelihood varies as the model parameters vary.

Model for our ratio problem

- We are interested in the ratio b, so let's make that a model parameter.
- We need to specify a prior probability distribution for b. Here we are departing from the naive calculation; it requires that we **be physicists** i.e. think about the actual meaning of the ratio. To keep the calculations easy, I make the following prior assumption: b must be positive, and cannot exceed 10, and the probability is flat between these values (so is a *top hat*).

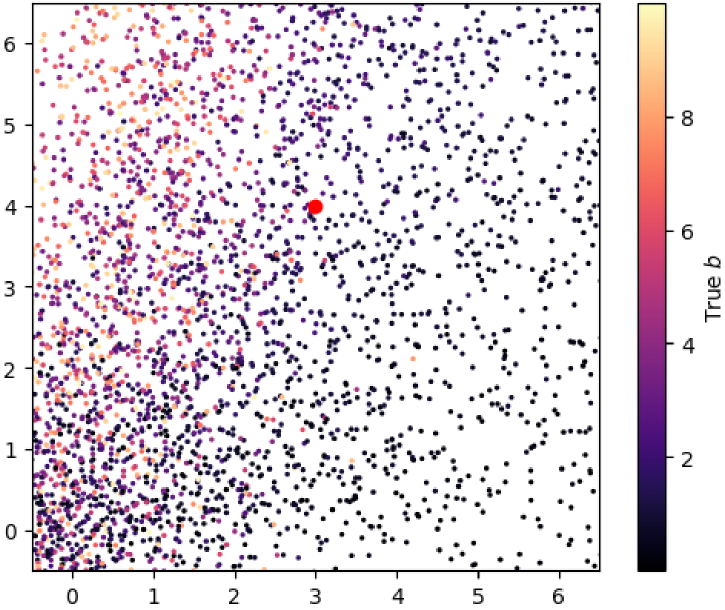
A nuisance parameter

- Is that enough? Well, can we generate new simulated data observations? No! All we can do is choose a plausable value of b from our prior distribution; that's not enough to generate new simulated data, as we have no idea what x and y need to be.
- So we need at least one new parameter. It will suffice e.g. to take x as a parameter. This is a *nuisance* parameter; it's needed but we don't actually care about its value. We also need a prior distribution on x; again here we need to know that actual physical meaning of x. For simplicity I will take a (broad) range: tophat between -10 and 30.

The model is now complete

- The model is now complete. If desired, we could generate new synthetic data: choose b and x from their priors, set $y = bx$, and add $N(0, 1)$ noise to x and y to get new data $\tilde{x}$ and $\tilde{y}$.
- So let's do it!

Synthetic data



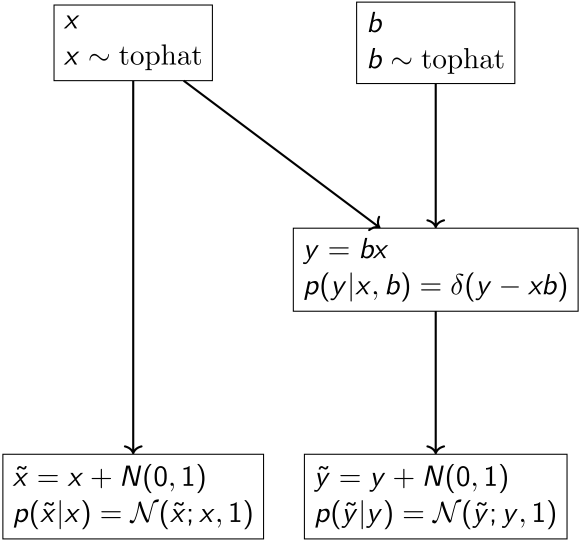
Simulation based inference

- If desired, we could stop there! We could look at the simulated data points that are ‘close to’ to the observed data point (big red dot), and make a histogram of the b values that were used to generate these synthetic points. This will give the posterior on b.
- This is *simulation based inference*.
- Note that the histogram will include some points with high b, corresponding to true (x, y) in the upper left corner, that were then scattered by measurement noise to be close to the red dot. So we should expect our posterior to have a tail to the right.

Hierarchical model

- Our model is *hierarchical*. For example, y sits in between b and x (upon which it depends) and $\tilde{y}$ (which depends upon it).
- We can make a graph of the model. It's a DAG (Directed (reflecting the dependence relationship) Acyclic (as cyclics would break causality) Graph) – just like with Git...

Bayes Model Diagram



Posterior

So now we can use Bayes's Theorem to write down the posterior. But if we have written out the graph correctly, then it's just a matter of multiplying all the probabilities in the graph (and you can skip to the third line):

$$\begin{aligned}
p(b, x|\tilde{x}, \tilde{y}) &\propto p(\tilde{x}, \tilde{y}|b, x)\ p(b)\ p(x) \\
&\propto p(\tilde{x}|x)\ p(\tilde{y}|y)\ p(y|x, b)\ p(b)\ p(x) \\
&\propto \mathcal{N}(\tilde{x}; x, 1)\ \mathcal{N}(\tilde{y}; y, 1)\ \delta(y - xb)\ U(b)\ U(x) \\
&\propto \mathcal{N}(\tilde{x}; x, 1)\ \mathcal{N}(\tilde{y}; xb, 1)\ U(b)\ U(x) \\
&\propto \exp(-\frac{1}{2}((\tilde{x} - x)^2 + (\tilde{y} - xb)^2))\ U(b)\ U(x)
\end{aligned}$$

Get rid of the nuisance parameter

- This is our answer - a posterior distribution for x and b.
- But we only care about b! So we need to *project* the probabilities from the two-dimensional (x, b) parameter space down to the one-dimensional b parameter space. Do this by integrating over x.

$$
\begin{aligned}
p(b|\mathrm{data}) &\propto \int p(b, x|\mathrm{data}))\mathrm{d}x \\
&\propto \int \exp(-\frac{1}{2}((\tilde{x} - x)^2 + (\tilde{y} - xb)^2))\ U(b)\ U(x)\ \mathrm{d}x
\end{aligned}
$$

A Useful Definite Integral

If $A > 0$ then

$$\int_{-\infty}^{\infty} \exp(-\frac{A}{2}x^2 + Bx + C)\,\mathrm{d}x = \sqrt{\frac{2\pi}{A}} \exp\left(\frac{B^2}{2A} + C\right)$$

This generalises the Gaussian integral (for which $A = 1$, $B = C = 0$).

One-dimensional posterior

Thus we get to our goal: the marginalised posterior for b (under this model):

$$p(b|\tilde{x}, \tilde{y}) \propto \frac{1}{\sqrt{1+b^2}} \exp\left(\frac{(\tilde{x}+b\tilde{y})^2}{2(1+b^2)}\right)$$

- Note a factor of $1/2$ in the exponential function, but no minus sign.
- Note the curious combination $\tilde{x}+b\tilde{y}$; you might have expected $\tilde{y}-b\tilde{x}$.

Widening the tophat priors

The left and right bounds on the priors for x and b were somewhat arbitrary. Can we get rid of these bounds, so that the prior density is the same non-zero value everywhere? Mathematically we do this by taking the limit as the bounds go to plus and minus infinity.

- For x there is no problem; the required limits exist. In fact we already assumed $-\infty < x < \infty$ when we did the marginalisation integral.
- But for b we see that the posterior looks like $(1 + b^2)^{-1/2}$ for large b; the integral of this over $0 < b < \infty$ is infinite. This is bad news for a probability distribution, as it means we can't normalise it so that the total probability (i.e. the area under the posterior function is unity. So it's crucial to have a prior with a right-hand cutoff!