

Bayesian Inference

Or: How to make sense of
what we see in the wild

Bjorn Larsen – bjorn.larsen@yale.edu



Thomas Bayes

Astronomy is like a safari: *Look but don't touch*

Working with limited observations (d)

1. What are the behaviors and properties of distant objects (pulsars, BHs, GWs)?

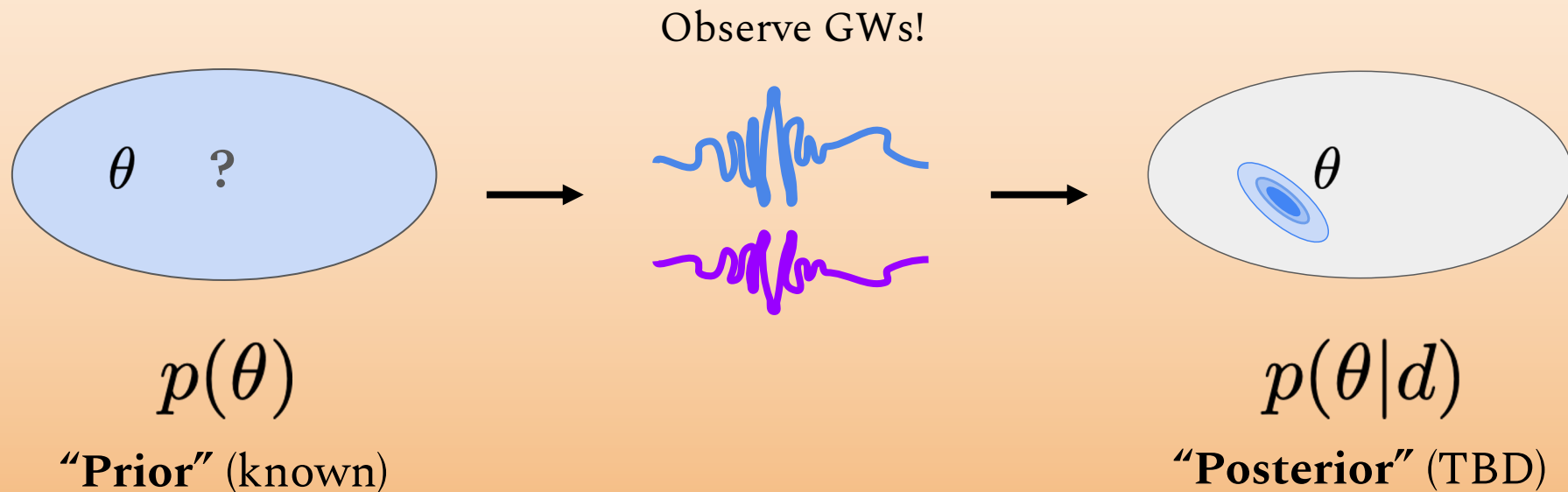
θ

2. Do I really see what I think I see?

$\mathcal{M}$

Bayesian parameter estimation: working with probability

Example: locate the origin of a gravitational wave signal



Basics with Bayes: Multivariate probabilities

Suppose two variables (x and y) are important

$p(y) \equiv$ probability over y independent of x

$p(y|x) \equiv$ probability over y conditioned on x

$p(x, y) \equiv$ joint probability over x and y

$$p(x, y) \neq p(x)p(y)$$

$$p(x, y) = p(y|x)p(x) \quad \textit{x and y may be correlated!}$$

Basics with Bayes: Marginalization

Recall probabilities are
always normalized

$$\int p(x)dx = 1$$

(Discrete version)

$$\sum_X P(X) = 1$$

$$p(y) = \int p(x, y)dx$$



$$p(y) = \int p(y|x)p(x)dx$$

Bayes' theorem

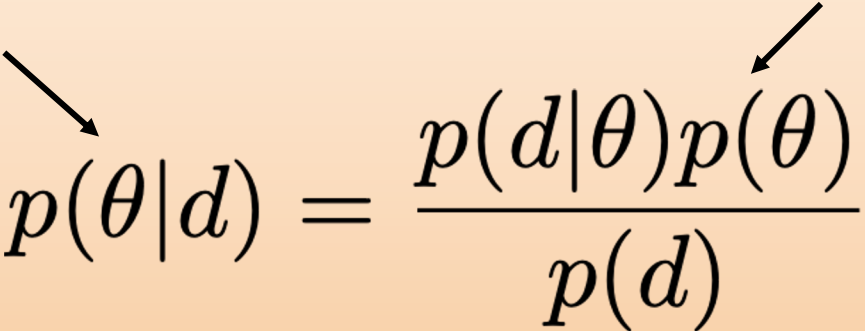
$$p(y|x) = \frac{p(x|y)p(x)}{p(y)}$$

(Derive in exercise #1 of the notebook
using the joint probability)

Bayes' theorem enables parameter estimation

Posterior

Prior

$$p(\theta|d) = \frac{p(d|\theta)p(\theta)}{p(d)}$$


Bayes' theorem enables parameter estimation

Posterior *New quantity:* **Data likelihood** **Prior**

$$p(\theta|d) = \frac{p(d|\theta)p(\theta)}{p(d)}$$

New quantity: **Evidence**

The diagram illustrates Bayes' theorem with the following components and annotations:

- Posterior**: Labeled with an arrow pointing to the term $p(\theta|d)$ on the left side of the equation.
- Data likelihood**: Labeled with an arrow pointing to the term $p(d|\theta)$ in the numerator.
- Prior**: Labeled with an arrow pointing to the term $p(\theta)$ in the numerator.
- Evidence**: Labeled with an arrow pointing to the term $p(d)$ in the denominator, with the text *New quantity:* preceding it.

Bayes' theorem enables parameter estimation

Posterior *New quantity:* **Data likelihood** **Prior**

$$\mathcal{P}(\theta|d) = \frac{\mathcal{L}(d|\theta)\pi(\theta)}{\mathcal{Z}(d)}$$

New quantity: **Evidence**

A diagram illustrating Bayes' theorem. The equation is $\mathcal{P}(\theta|d) = \frac{\mathcal{L}(d|\theta)\pi(\theta)}{\mathcal{Z}(d)}$. Annotations include: 'Posterior' with an arrow pointing to $\mathcal{P}(\theta|d)$; 'Data likelihood' with an arrow pointing to $\mathcal{L}(d|\theta)$; 'Prior' with an arrow pointing to $\pi(\theta)$; and 'Evidence' with an arrow pointing to $\mathcal{Z}(d)$. The words 'New quantity:' are placed above 'Data likelihood' and below 'Evidence'.

Bayes' theorem enables parameter estimation

Posterior *New quantity:* **Data likelihood** **Prior**

$$\mathcal{P}(\theta|d) = \frac{\mathcal{L}(d|\theta)\pi(\theta)}{\mathcal{Z}(d)}$$

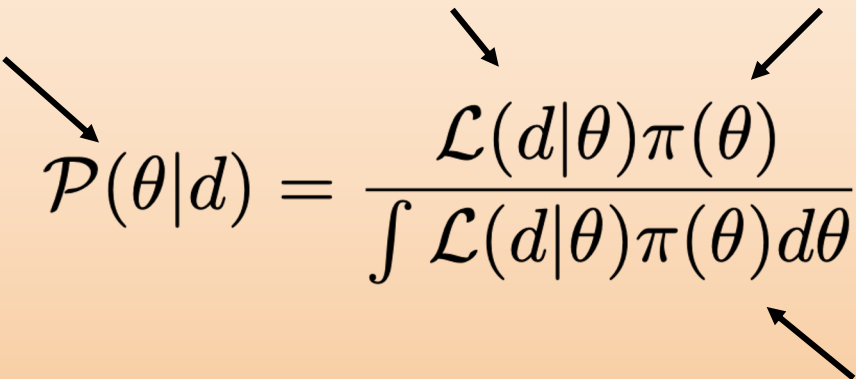
New quantity: **Evidence**

Apply marginalization:

$$p(y) = \int p(y|x)p(x)dx$$

Bayes' theorem enables parameter estimation

Posterior *New quantity:* **Data likelihood** **Prior**


$$\mathcal{P}(\theta|d) = \frac{\mathcal{L}(d|\theta)\pi(\theta)}{\int \mathcal{L}(d|\theta)\pi(\theta)d\theta}$$

New quantity: **Evidence**
(normalization constant)

Apply marginalization:

$$p(y) = \int p(y|x)p(x)dx \longrightarrow \mathcal{Z}(d) = \int \mathcal{L}(d|\theta)\pi(\theta)d\theta$$

How to compute the likelihood?

The Likelihood is computed for the different ways the data could be generated

Example:

- Flip a weighted coin 3x $\rightarrow$ observe $\mathbf{d} = H, H, T$
- $\boldsymbol{\theta}$ is the coin weight (probability to flip heads)

$$\mathcal{L}(d|\theta) = p(H) \times p(H) \times p(T)$$

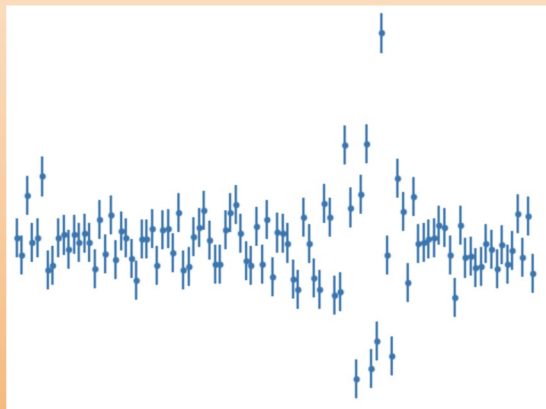
$$\mathcal{L}(d|\theta) = \theta^2(1 - \theta)$$

$\boldsymbol{\theta}$	$\mathcal{L}(d \theta)$
0.25	~ 0.05
0.5	0.125
0.75	~ 0.14
1	0

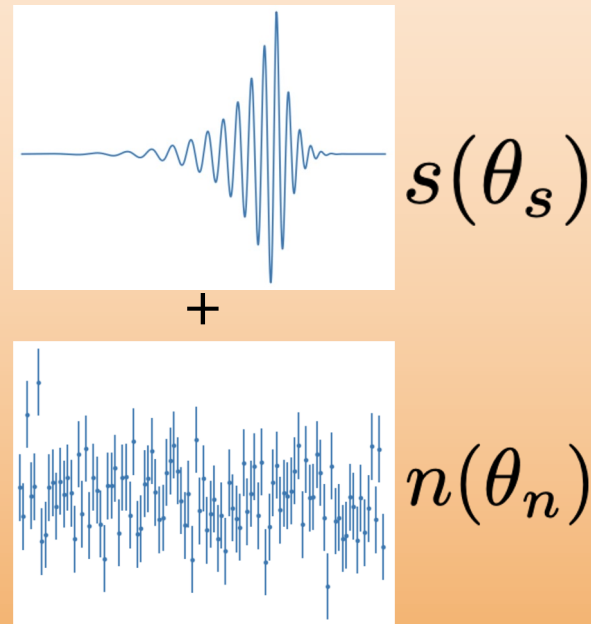
How to compute the likelihood?

The Likelihood is computed for the different ways the data could be generated

Data analysis: Typically model data as a signal with some noise



$$d = s + n$$

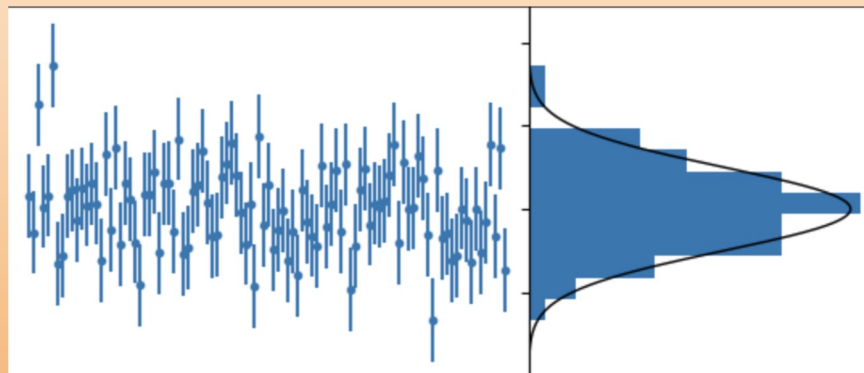


How to compute the likelihood?

Common choice: Noise is a draw from a Gaussian distribution

$$p(n_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{n_i^2}{2\sigma_i^2}\right)$$

1. Easy distribution
2. Often a good match to real data because of the central limit theorem



How to compute the likelihood?

Common choice: Noise is a draw from a Gaussian distribution

$$p(n_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{n_i^2}{2\sigma_i^2}\right)$$

Use $d = s + n$:

$$\mathcal{L}(d_i|\theta) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(d_i - s_i(\theta))^2}{2\sigma_i^2}\right)$$

Computational Bayes: How to compute the posterior

Posterior **Data likelihood** **Prior**
(we can compute) (we set this)

$$\mathcal{P}(\theta|d) = \frac{\mathcal{L}(d|\theta)\pi(\theta)}{\mathcal{Z}(d)}$$

“**Evidence**”
(still just a normalization constant)

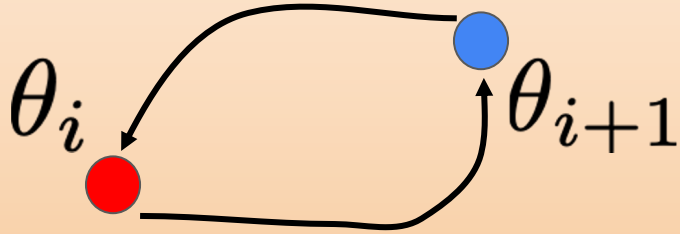
The diagram illustrates the components of Bayes' theorem. It features the equation $\mathcal{P}(\theta|d) = \frac{\mathcal{L}(d|\theta)\pi(\theta)}{\mathcal{Z}(d)}$. Above the equation, three labels are positioned: 'Posterior' on the left, 'Data likelihood' in the center, and 'Prior' on the right. Below these labels are their respective descriptions: '(we can compute)' under 'Data likelihood' and '(we set this)' under 'Prior'. Arrows point from 'Posterior' to $\mathcal{P}(\theta|d)$, from 'Data likelihood' to $\mathcal{L}(d|\theta)$, and from 'Prior' to $\pi(\theta)$. Below the denominator $\mathcal{Z}(d)$, the label '“Evidence”' is placed, with the description '(still just a normalization constant)' underneath it. An arrow points from '“Evidence”' to $\mathcal{Z}(d)$.

Computational Bayes: ~~How to compute the posterior~~
New goal: Draw samples from the posterior probability
distribution, normalize numerically

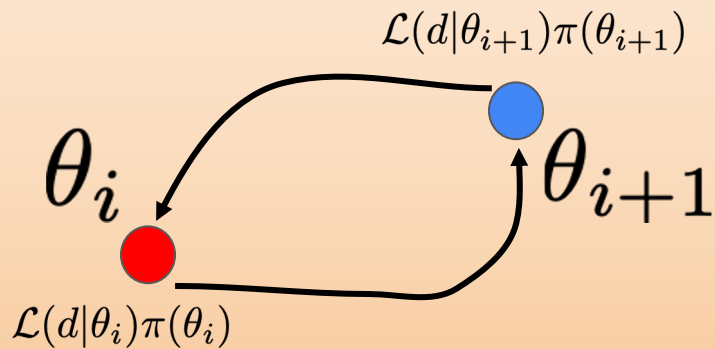
$$\theta|d \sim \mathcal{L}(d|\theta)\pi(\theta)$$

Markov Chain Monte Carlo in 1 slide: Sampling from an unknown distribution

1. Suppose I randomly sample a sequence of parameter values



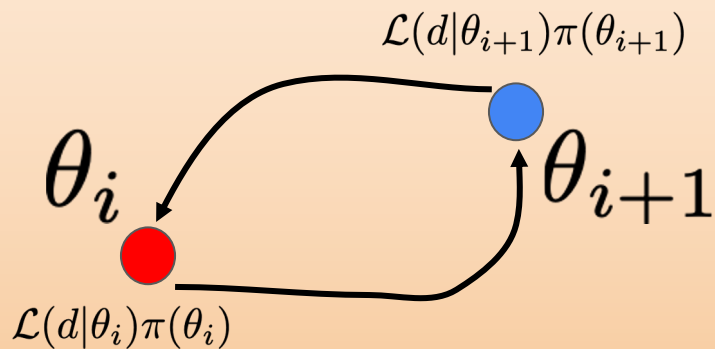
Markov Chain Monte Carlo in 1 slide: Sampling from an unknown distribution



1. Suppose I randomly sample a sequence of parameter values
2. Choose whether to accept samples with probability

$$P = \min \left[1, \frac{\mathcal{L}(d|\theta_{i+1})\pi(\theta_{i+1})}{\mathcal{L}(d|\theta_i)\pi(\theta_i)} \right]$$

Markov Chain Monte Carlo in 1 slide: Sampling from an unknown distribution



Process goes significantly
faster using
jump proposals



1. Suppose I randomly sample a sequence of parameter values
2. Choose whether to accept samples with probability

$$P = \min \left[1, \frac{\mathcal{L}(d|\theta_{i+1})\pi(\theta_{i+1})}{\mathcal{L}(d|\theta_i)\pi(\theta_i)} \right]$$

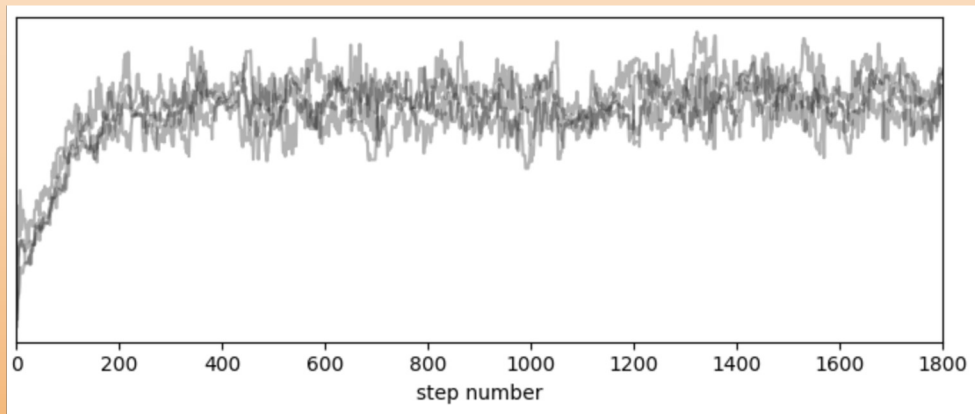
3. After enough time, my samples will eventually converge to the posterior

MCMC in action

<https://chi-feng.github.io/mcmc-demo/app.html>

MCMC in action

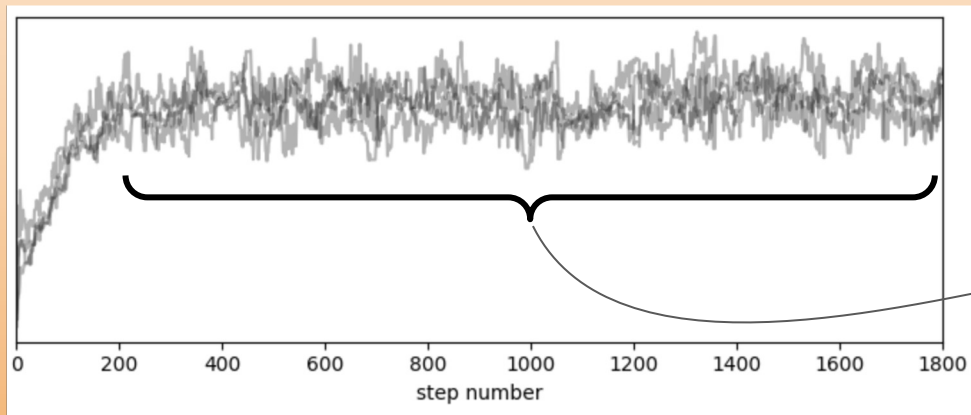
<https://chi-feng.github.io/mcmc-demo/app.html>



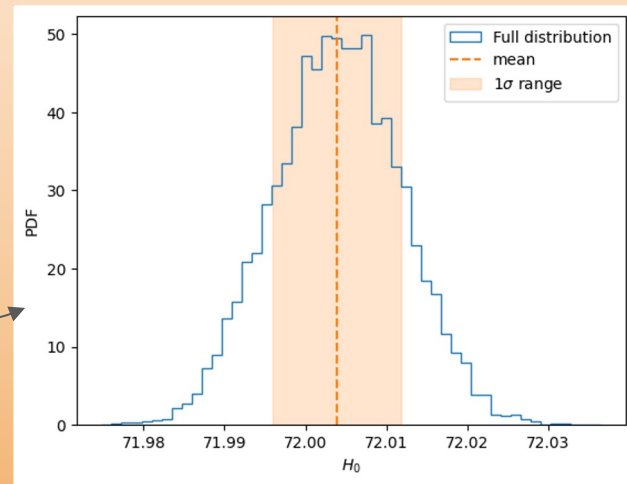
“Trace plot”

MCMC in action

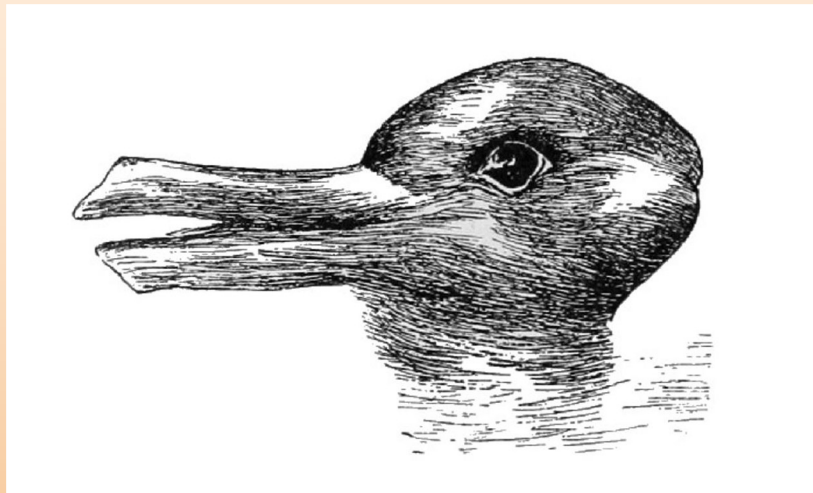
<https://chi-feng.github.io/mcmc-demo/app.html>



“Trace plot”

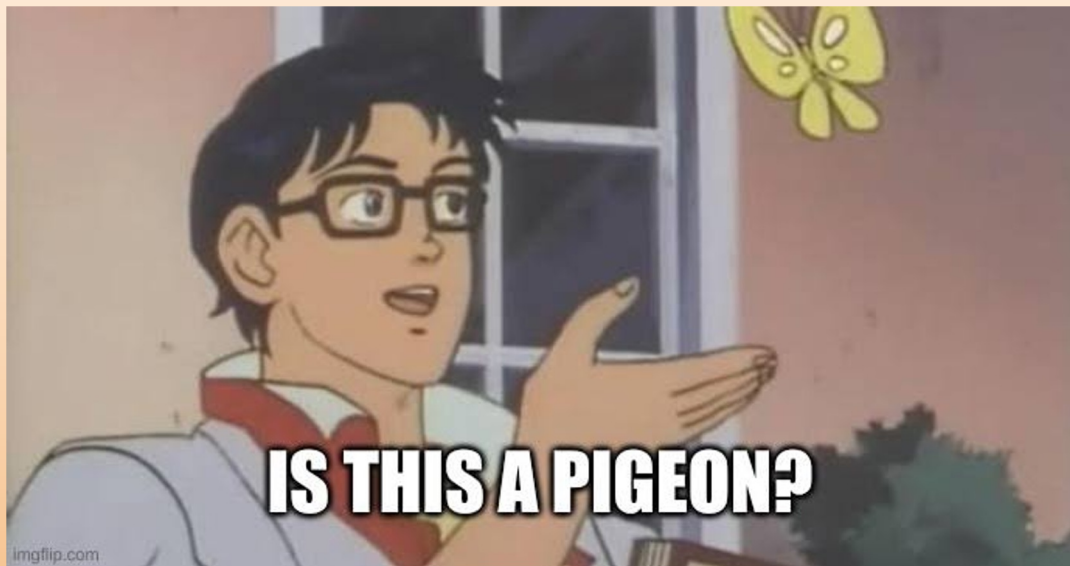


Bayesian model selection



$$\mathcal{P}(\theta|d) = \frac{\mathcal{L}(d|\theta)\pi(\theta)}{\mathcal{Z}(d)} \longrightarrow \mathcal{P}(\theta|d, \mathcal{M}) = \frac{\mathcal{L}(d|\theta, \mathcal{M})\pi(\theta|\mathcal{M})}{\mathcal{Z}(d|\mathcal{M})}$$

Why Model Selection is Important



Need to compute $\mathcal{P}(\text{Pigeon}|d)$
and compare with $\mathcal{P}(\text{Butterfly}|d)$

Bayesian Model Selection: Compare two models head to head

$\mathcal{P}(\mathcal{M}_1|d) \equiv$ Probability of Model 1 given the observations

$\mathcal{P}(\mathcal{M}_2|d) \equiv$ Probability of Model 2 given the observations

$\mathcal{O}_2^1 \equiv$ Odds ratio for $\mathcal{M}_1$ vs $\mathcal{M}_2$

$$\mathcal{O}_2^1 = \frac{\mathcal{P}(\mathcal{M}_1|d)}{\mathcal{P}(\mathcal{M}_2|d)}$$

Bayesian Model Selection: Compare two models head to head

$\mathcal{P}(\mathcal{M}_1|d) \equiv$ Probability of Model 1 given the observations

$\mathcal{P}(\mathcal{M}_2|d) \equiv$ Probability of Model 2 given the observations

$\mathcal{O}_2^1 \equiv$ Odds ratio for $\mathcal{M}_1$ vs $\mathcal{M}_2$

$$\mathcal{O}_2^1 = \frac{\mathcal{P}(\mathcal{M}_1|d)}{\mathcal{P}(\mathcal{M}_2|d)} = \frac{\frac{\mathcal{L}(d|\mathcal{M}_1)\pi(\mathcal{M}_1)}{\mathcal{Z}(d)}}{\frac{\mathcal{L}(d|\mathcal{M}_2)\pi(\mathcal{M}_2)}{\mathcal{Z}(d)}}$$

(apply Bayes' theorem)

Bayesian Model Selection: Compare two models head to head

$\mathcal{P}(\mathcal{M}_1|d) \equiv$ Probability of Model 1 given the observations

$\mathcal{P}(\mathcal{M}_2|d) \equiv$ Probability of Model 2 given the observations

$\mathcal{O}_2^1 \equiv$ Odds ratio for $\mathcal{M}_1$ vs $\mathcal{M}_2$

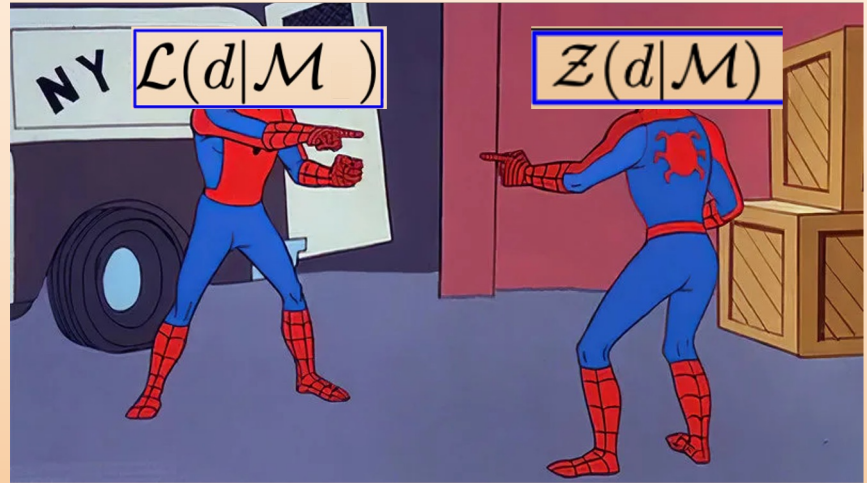
$$\mathcal{O}_2^1 = \frac{\mathcal{P}(\mathcal{M}_1|d)}{\mathcal{P}(\mathcal{M}_2|d)} = \frac{\frac{\mathcal{L}(d|\mathcal{M}_1)\pi(\mathcal{M}_1)}{\mathcal{Z}(d)}}{\frac{\mathcal{L}(d|\mathcal{M}_2)\pi(\mathcal{M}_2)}{\mathcal{Z}(d)}} = \underbrace{\frac{\pi(\mathcal{M}_1)}{\pi(\mathcal{M}_2)}}_{\text{Prior odds}} \underbrace{\frac{\mathcal{L}(d|\mathcal{M}_1)}{\mathcal{L}(d|\mathcal{M}_2)}}_{\text{Bayes Factor}} \mathcal{B}_2^1$$

(apply Bayes' theorem)

Bayes Factor is ratio of model evidences
(which are marginalized likelihoods)

$$\mathcal{B}_2^1 = \frac{\mathcal{L}(d|\mathcal{M}_1)}{\mathcal{L}(d|\mathcal{M}_2)}$$

$$\mathcal{P}(\theta|d, \mathcal{M}) = \frac{\mathcal{L}(d|\theta, \mathcal{M})\pi(\theta|\mathcal{M})}{\mathcal{Z}(d|\mathcal{M})}$$



$$p(y) = \int p(y|x)p(x)dx \longrightarrow \mathcal{Z}(d|\mathcal{M}) = \int \mathcal{L}(d|\theta, \mathcal{M})\pi(\theta|\mathcal{M})d\theta$$

How to compute the evidence?

$$\mathcal{Z}(d|\mathcal{M}) = \int \mathcal{L}(d|\theta, \mathcal{M})\pi(\theta|\mathcal{M})d\theta$$

⇒ Big multidimensional integral $\theta \in \mathbb{R}^{N_p}$:(

⇒ Have to recompute likelihood at each point :(

⇒ Regions of the parameter space matter even where likelihood is small

⇒ Good method to compute is **nested sampling**

Nested Sampling (overview)

$$\mathcal{Z}(d|\mathcal{M}) = \int \mathcal{L}(d|\theta, \mathcal{M}) \pi(\theta|\mathcal{M}) d\theta$$

$$\downarrow$$
$$\mathcal{Z} = \sum \mathcal{L}(X) \Delta X$$

$\Delta X \equiv$ Volume elements

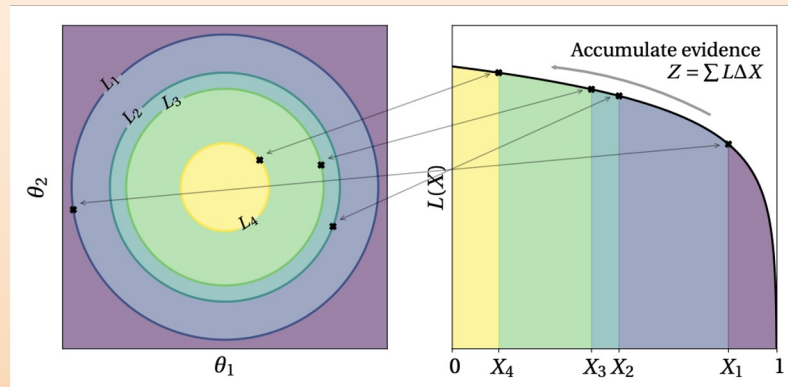


Fig. from Ashton et al. 2022

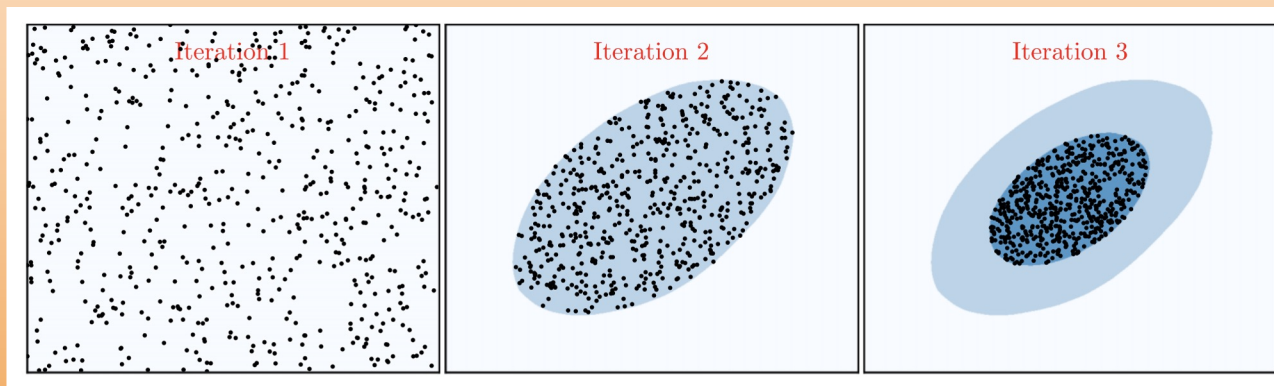
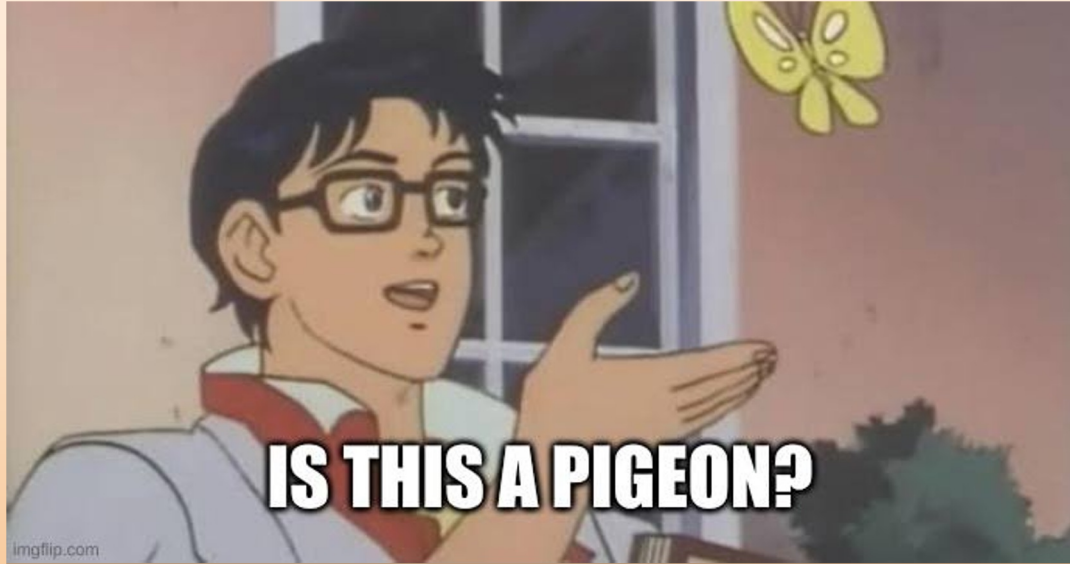
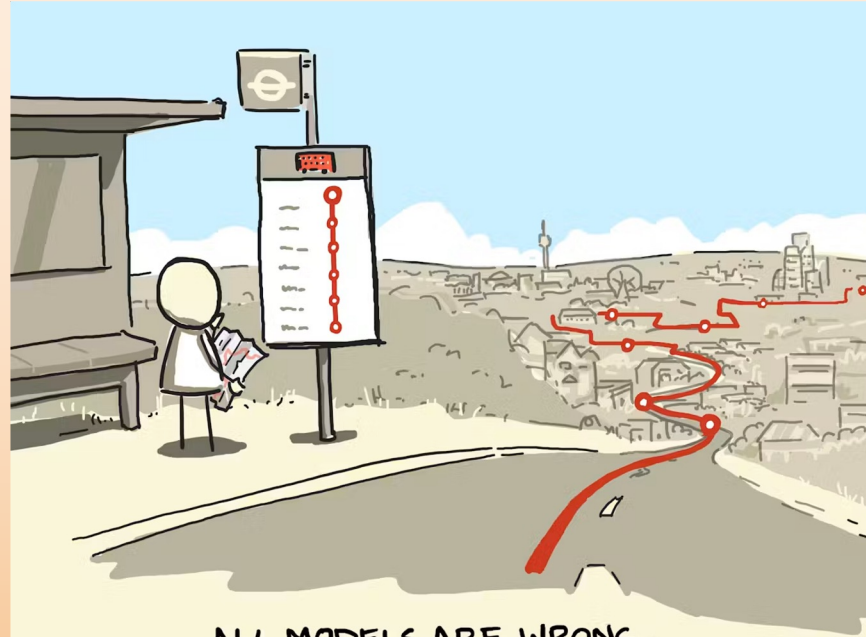


Fig. from Lange et al. 2023



β^{Pigeon}
 Polar Bear will not get us the answer!



ALL MODELS ARE WRONG,
BUT SOME ARE USEFUL

— GEORGE BOX

sketchplanations

Other topics I didn't have time to talk about

- Advanced MCMC methods
 - Hamiltonian Monte Carlo
 - Parallel Tempering
 - Transdimensional sampling
- Other Bayes Factor/Odds Ratio computation methods
 - Savage-Dickey Density Ratio
 - Product-Space Sampling
- Hierarchical Bayesian Modeling
- Probabilistic Programming Libraries

Tutorial time!

DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)

THIS NEUTRINO DETECTOR MEASURES
WHETHER THE SUN HAS GONE NOVA.

THEN, IT ROLLS TWO DICE. IF THEY
BOTH COME UP SIX, IT LIES TO US.
OTHERWISE, IT TELLS THE TRUTH.

LET'S TRY:
DETECTOR! HAS THE
SUN GONE NOVA?

ROLL
YES.

FREQUENTIST STATISTICIAN:

THE PROBABILITY OF THIS RESULT
HAPPENING BY CHANCE IS $\frac{1}{36} \approx 0.027$.
SINCE $p < 0.05$, I CONCLUDE
THAT THE SUN HAS EXPLODED.

BAYESIAN STATISTICIAN:

BET YOU \$50
IT HASN'T.

MODIFIED BAYES' THEOREM:

$$P(H|X) = P(H) \times \left(1 + P(C) \times \left(\frac{P(X|H)}{P(X)} - 1 \right) \right)$$

H: HYPOTHESIS

X: OBSERVATION

$P(H)$: PRIOR PROBABILITY THAT H IS TRUE

$P(X)$: PRIOR PROBABILITY OF OBSERVING X

$P(C)$: PROBABILITY THAT YOU'RE USING
BAYESIAN STATISTICS CORRECTLY

xkcd.com

$$P\left(\text{I'M NEAR THE OCEAN} \mid \text{I PICKED UP A SEASHELL}\right) =$$

$$\frac{P\left(\text{I PICKED UP A SEASHELL} \mid \text{I'M NEAR THE OCEAN}\right) P\left(\text{I'M NEAR THE OCEAN}\right)}{P\left(\text{I PICKED UP A SEASHELL}\right)}$$



STATISTICALLY SPEAKING, IF YOU PICK UP A SEASHELL AND DON'T HOLD IT TO YOUR EAR, YOU CAN PROBABLY HEAR THE OCEAN.