LIGO–India Scientific Collaboration

Bayesian Inference in Gravitational Wave Astronomy

Mukesh Kumar Singh

ICTS Astrophysical Relativity Group

GW Data Analysis Workshop, DTU

14/02/2023

Outline

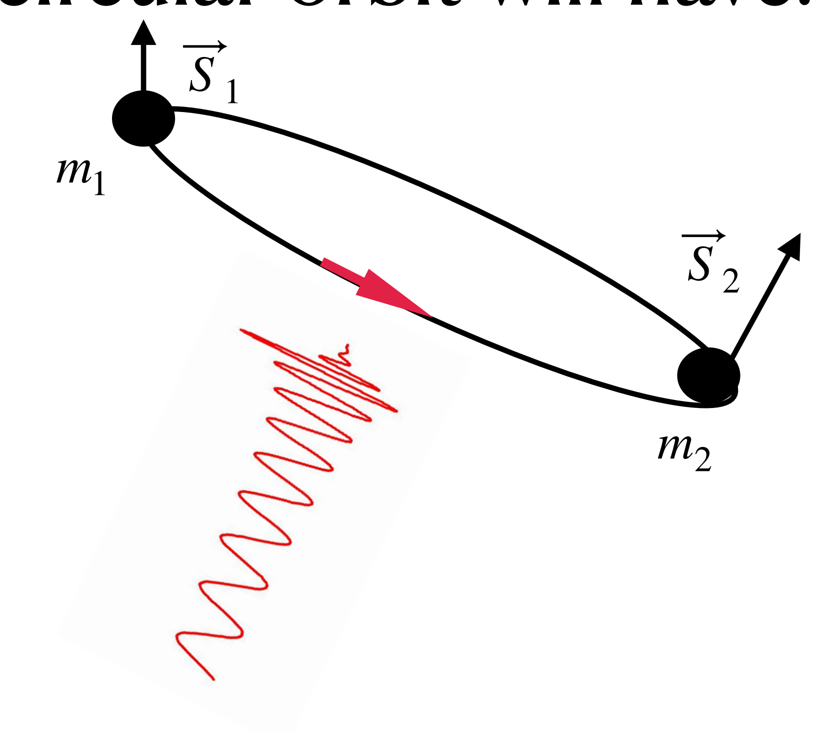
- Introduction
- Parameter estimation
- Conditional probability & Bayes' theorem
- Bayesian inference
- Stochastic sampling methods
- GW parameter estimation: `Bilby`
- Summary

Introduction

- Characteristic shape of the gravitational-wave signal encodes the information about the astrophysical properties of the source.
- In a compact binary merger with a quasicircular orbit will have:

Intrinsic parameters: $m_1, m_2, \overrightarrow{S_1}, \overrightarrow{S_2}$

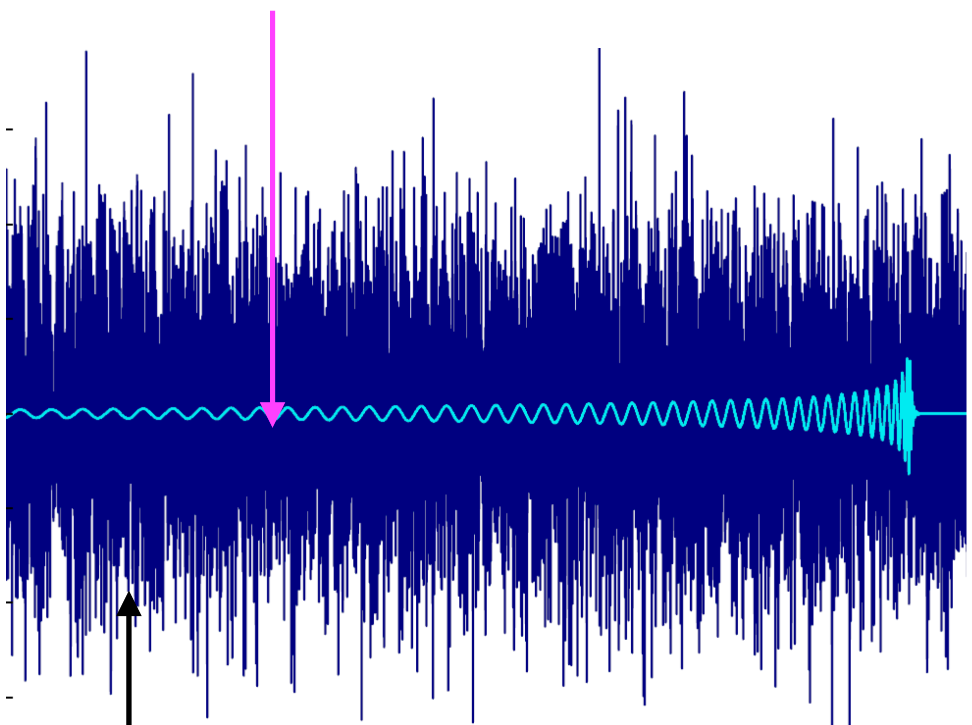
Extrinsic parameters: $\begin{cases} d_L, \alpha, \delta & \text{location} \\ \iota, \psi & \text{orientation} \\ \varphi_0, t_c & \text{arrival time \& phase} \end{cases}$



- The 15 dimensional parameter space.
- How to infer the these complex set of parameters?

Problem in hand

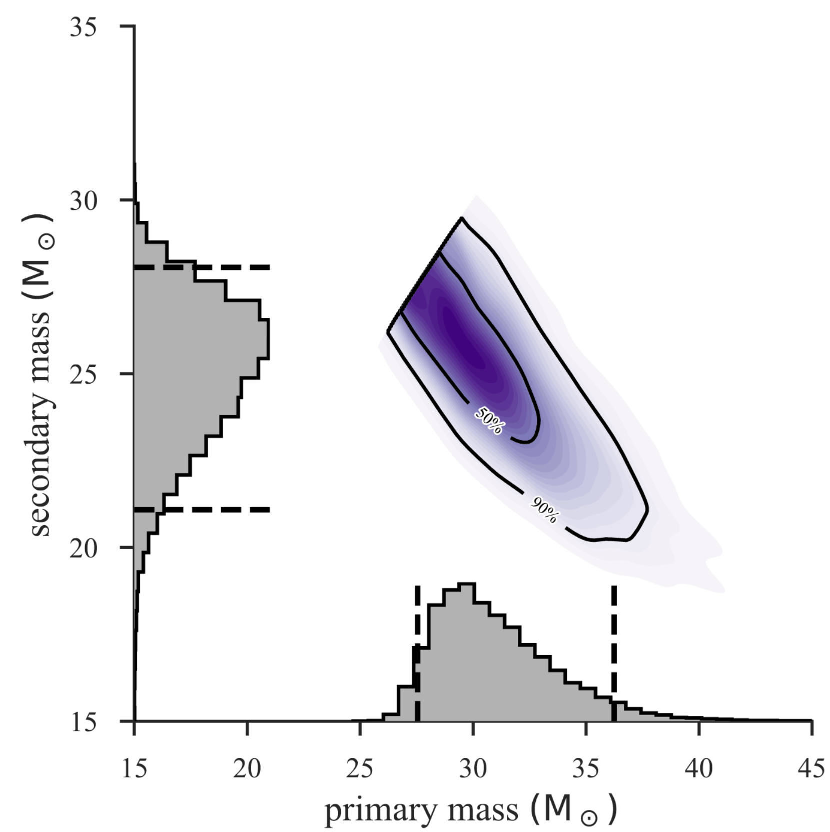
Gravitational-Wave Signal



Noise

Credit: H Gabbard et al

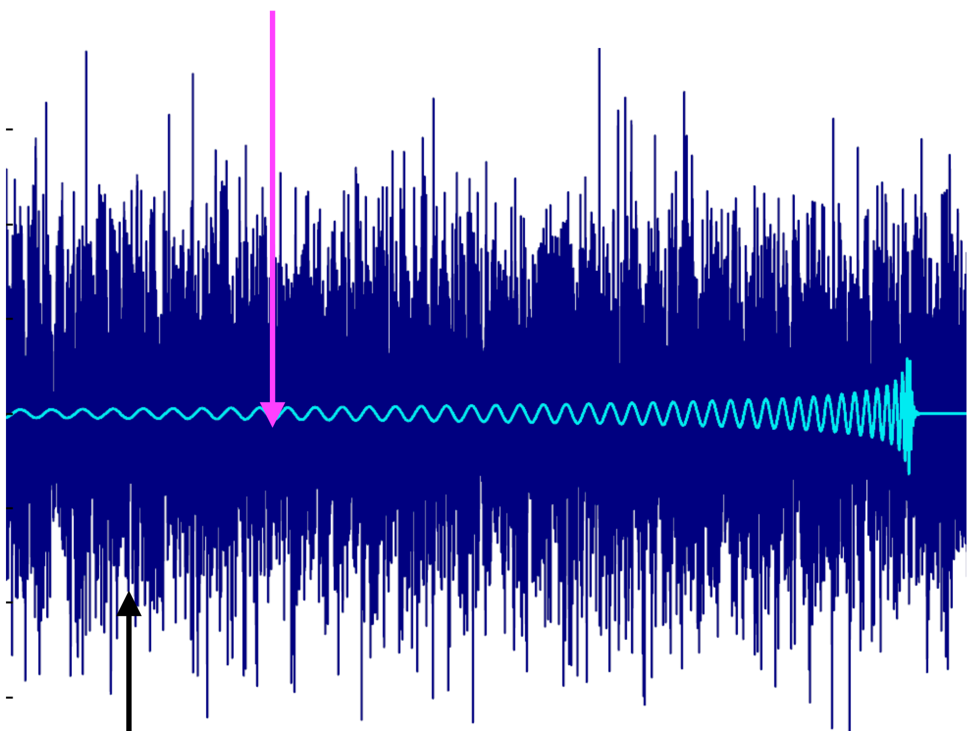
Credit: LSC



Probability distribution for component masses of the binary

Problem in hand

Gravitational-Wave Signal



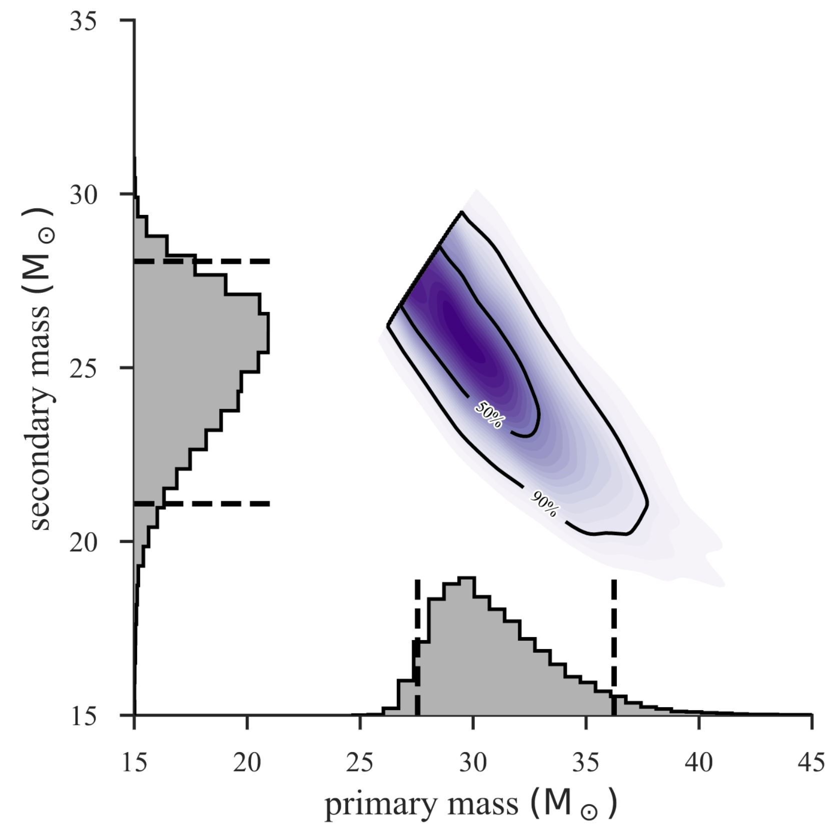
Noise

Credit: H Gabbard et al

Parameter

Estimation

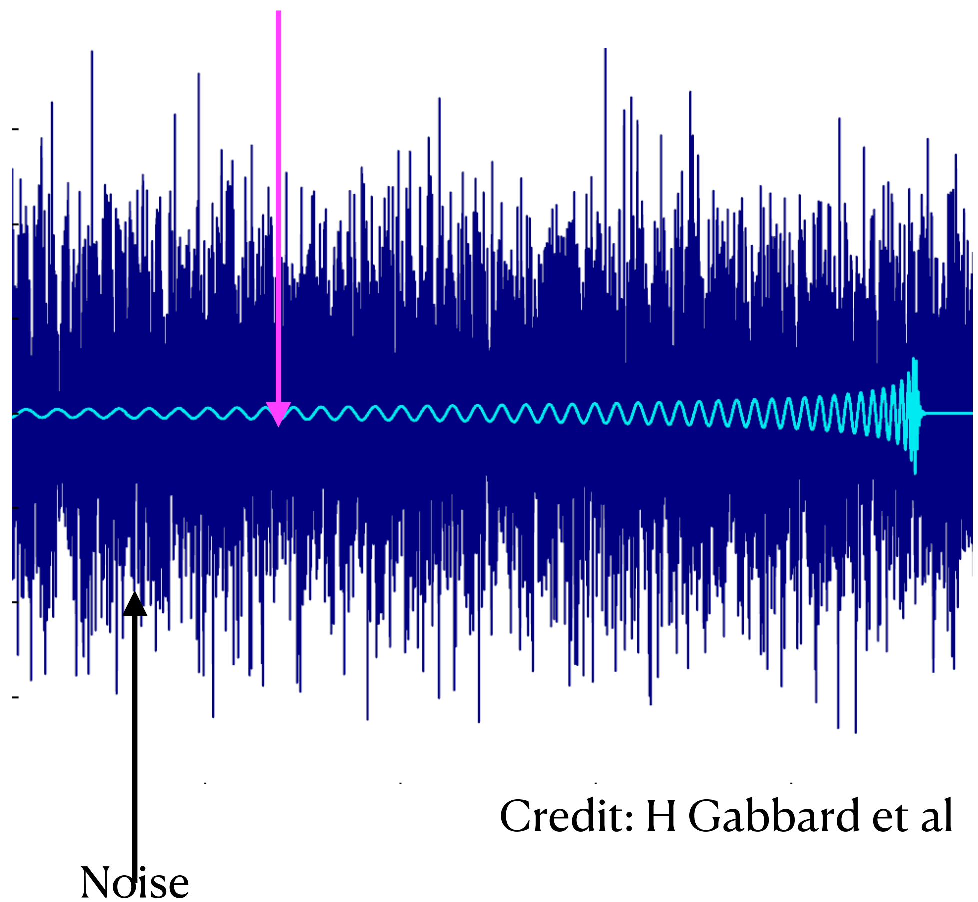
Credit: LSC



Probability distribution for component masses of the binary

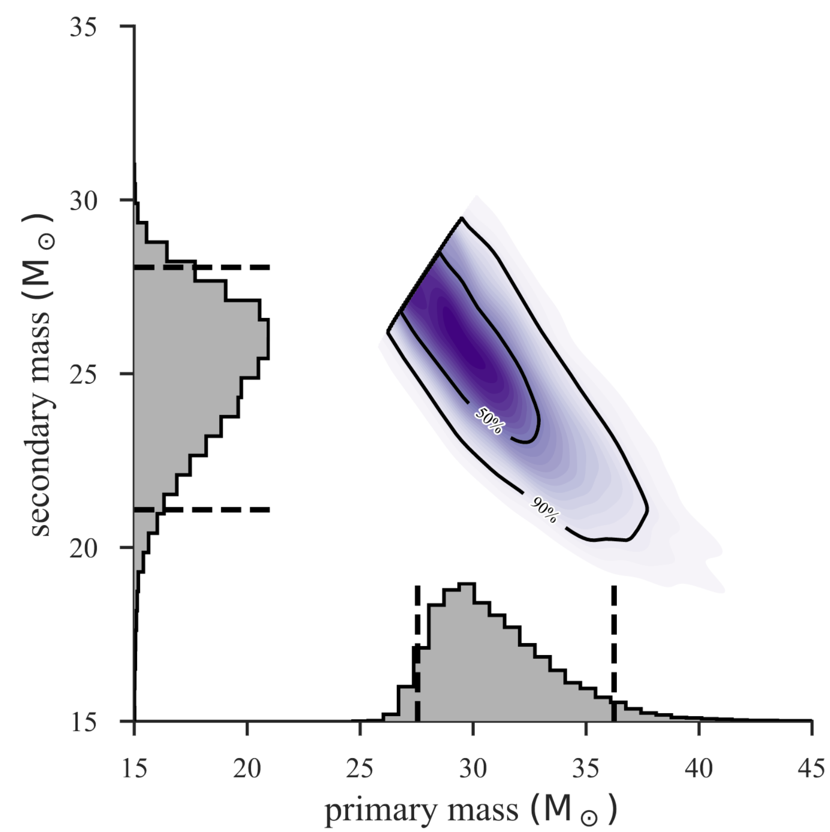
Problem in hand

Credit: LSC



Credit: H Gabbard et al

Parameter Estimation



Probability distribution for component masses of the binary

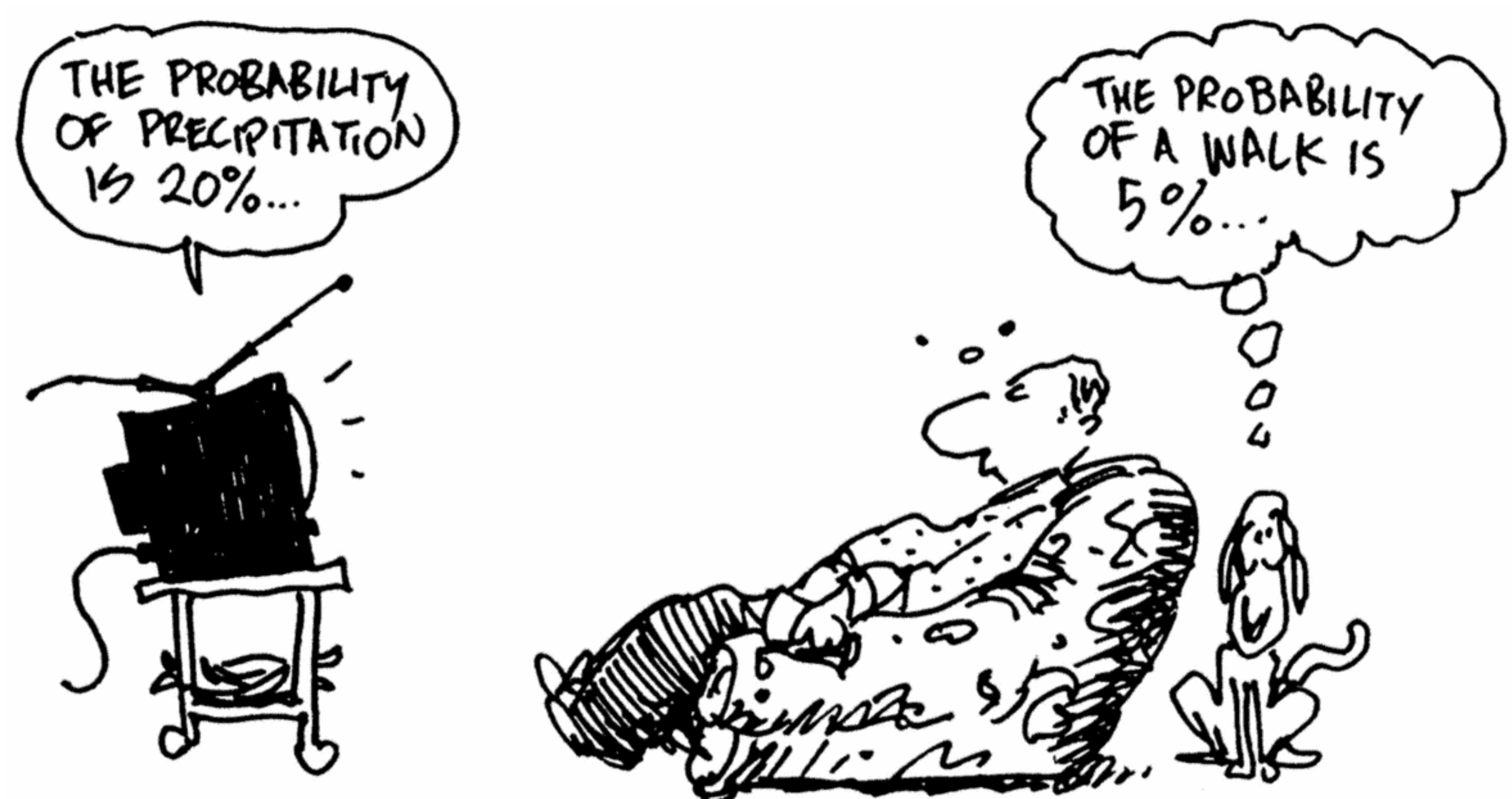
- How can we infer the source parameters of an unknown signal buried in noise and with what accuracy?

Statistical Reasoning: Frequentist of Bayesian?

- Frequentist
  - Probabilities as an outcome of a repeatable experiment (e.g. coin tossing)
  - No prior information required.
  - Hypotheses/theories are not the outcome of a repeatable experiment.
- Bayesian
  - Can talk about the probability of a hypothesis, or of a theory, or the probability that a parameter within a theory takes a given value.
  - No need of a repeatable experiment.
  - Assumes prior information: “degree of belief” (updated in the light of data)
- Question: Which method will be applicable for GW data analysis?
- Answer: **Bayesian Inference**

Conditional Probability

- $p(A \mid B)$: conditional probability i.e. the probability of A given B.



Credit: Gonick & Smith

Conditional Probability: Example

- **Question**: What's the probability of drawing one red-suit card from a pack of 52 cards?
- **Answer**: 1/2 if the deck of cards is a regular one.
- **Question:** How do you express it mathematically?
- **Answer:**

$$p(\text{red} - \text{card} \,|\, \text{I}) = \frac{1}{2}$$

where I = "normal 52-card deck" is the background information. The vertical line is read as "given".

Conditional Probability: Example

- **Question**: What's the probability of drawing two red-suit cards from a pack of 52 cards?
- **Answer**: Let me draw a diagram here

$$
\begin{tikzcd}[arrows=-]
 & & & 0 \arrow[dl] \arrow[dr] & & & \\
 & & R0 \arrow[dl] \arrow[dr] & & B0 \arrow[dl] \arrow[dr] & & \\
 & R1 & & B1 \quad R1 & & B1 &
\end{tikzcd}
$$

So...

$$R_0 = \frac{26}{52} = \frac{1}{2} \text{ and } \mathrm{R}_1 = \frac{26-1}{52-1} = \frac{25}{51}$$

So the probability of occurring both is

$$R_0 R_1 = \frac{25}{102}$$

Conditional Probability: Example

- **Question**: How do you express it mathematically?
- **Answer**: We need to calculate $p(R_0, R_1 \,|\, I)$! Using simple rule of *conditional probability*

$$p(R_0, R_1 \,|\, I) = p(R_1 \,|\, R_0, I)\,p(R_0 \,|\, I)$$

With $p(R_0 \,|\, I) = \dfrac{1}{2}$ and $p(R_1 \,|\, R_0, I) = \dfrac{25}{51}$

So finally

$$p(R_0, R_1 \,|\, I) = \frac{1}{2} \times \frac{25}{51}$$

Bayes Theorem

- Conditional probability

$$p(A, B \mid I) = p(A \mid B, I)p(B \mid I)$$

- We can also write the same thing as

$$p(A, B \mid I) = p(B \mid A, I)p(A \mid I)$$

- Rearranging the terms in above two equations

$$p(A \mid B, I) = \frac{p(B \mid A, I)p(A \mid I)}{p(B \mid I)}$$

Bayesian Inference

- Inference means figuring something out from the data (d).
- Inference can be made two ways:

1. **Parameter estimation:** finding out the parameters (θ) of a model (M_A) that best fits the data (d).
2. **Model selection:** finding out which model, M_A or M_B, fits the data more effectively.

Parameter estimation

- Figuring out the model parameters θ, given the data d and the model M_A

$$\underbrace{p(\theta\,|\,d, M_A)}_{Posterior} = \frac{\overbrace{p(d\,|\,\theta, M_A)}^{Likelihood}\ \overbrace{p(\theta\,|\,M_A)}^{Prior}}{\underbrace{p(d\,|\,M_A)}_{Evidence}}$$

Used when comparing models

- Evidence is just a normalisation. So..

$$\underbrace{p(\theta\,|\,d, M_A)}_{Posterior} \propto \underbrace{p(d\,|\,\theta, M_A)}_{Likelihood}\ \ \underbrace{p(\theta\,|\,M_A)}_{Prior}$$

The degree of belief
After the experiment

The degree of belief
before the experiment

One Dimensional Example

- Given some specific observation y at time t, the data d

$$d(t) = y(t) + n \implies n = d(t) - y(t)$$

where n is random noise drawn from a gaussian distribution $\mathcal{N}(0,\sigma)$

- If the model is given by $M_A : y_A(t) = \sin(\omega t)$ with only parameter ω, the likelihood

$$\mathcal{L}(d \,|\, \omega, M_A) = p(d \,|\, \omega, y_A) \equiv p(n) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(d - y_A(t,\omega))^2}{2\sigma^2}\right]$$

- And, the posterior

$$p(\omega \,|\, d, y_A) \propto p(d \,|\, \omega, y_A) p(\omega \,|\, y_A)$$

- A good idea to work with log-likelihood for the sake of stability

$$\ln \mathcal{L} = -\frac{1}{2}\left(\frac{(d(t) - y_A(t,\omega))^2}{\sigma^2} + \ln(2\pi\sigma^2)\right)$$

One Dimensional Example

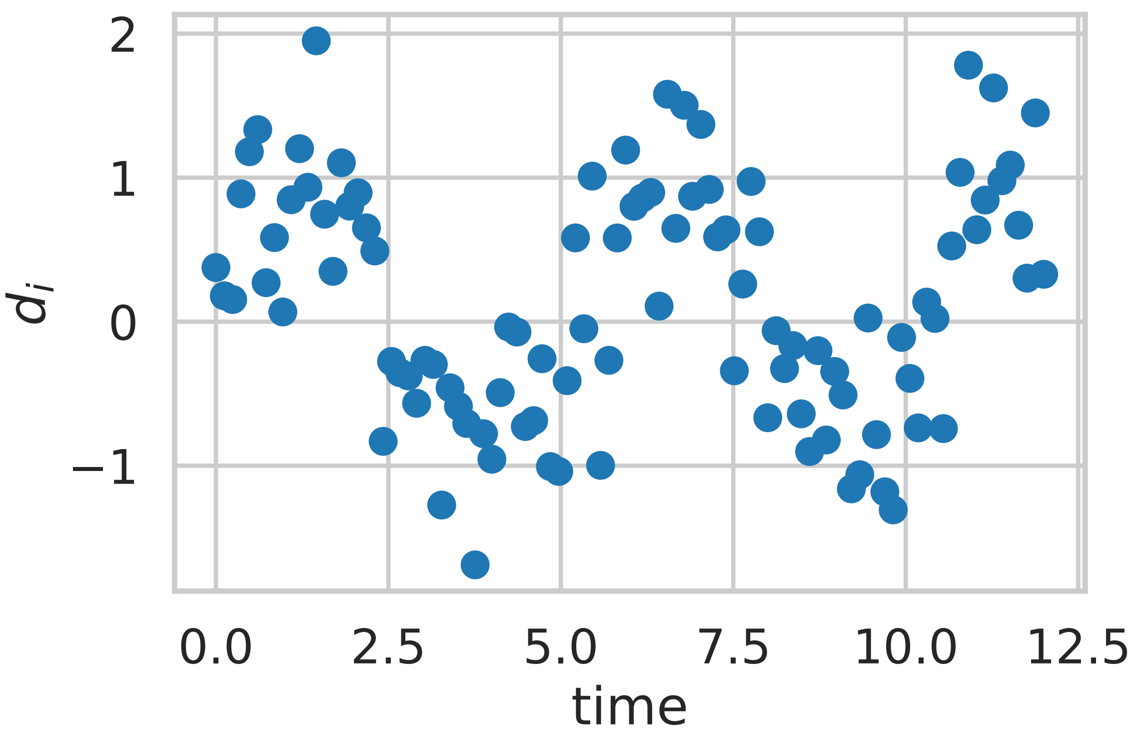
- For multiple observations,

$$\mathscr{L}(\mathbf{d}\,|\,\omega, M_A) = \prod_i \mathscr{L}(\mathbf{d_i}\,|\,\omega, M_A)$$

Or,

$$\ln \mathscr{L} = -\frac{1}{2}\Sigma_i \left(\frac{(d_i - y_A(t_i, \omega))^2}{\sigma^2} + \ln(2\pi\sigma^2) \right)$$

- Choose, say $\sigma = 0.1$ and $\omega_{\text{true}} = 1.2$, to simulate the data $d_i = y(t_i, \omega_{\text{true}}) + n_i$.



One Dimensional Example

- For multiple observations,

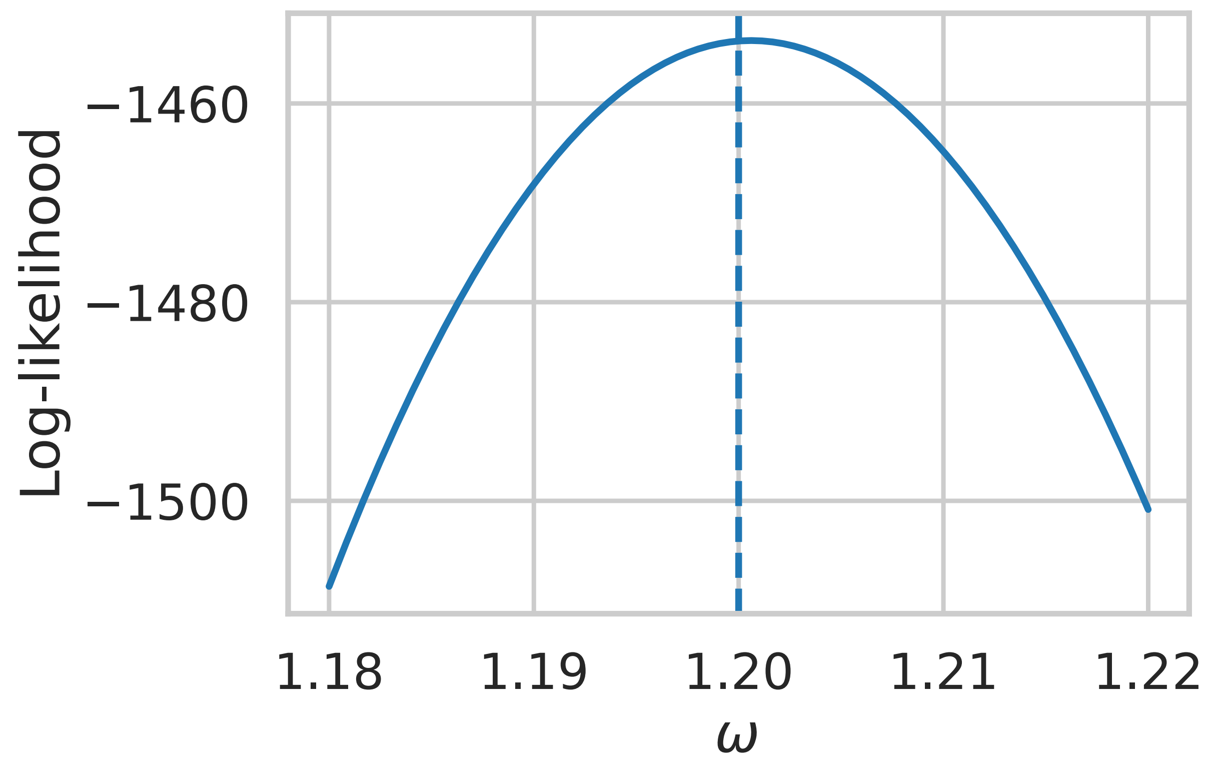
$$\mathscr{L}(\mathbf{d} \,|\, \omega, M_A) = \prod_i \mathscr{L}(\mathbf{d_i} \,|\, \omega, M_A)$$

Or,

$$\ln \mathscr{L} = -\frac{1}{2}\Sigma_i \left(\frac{(d_i - y_A(t_i, \omega))^2}{\sigma^2} + \ln(2\pi\sigma^2) \right)$$

- Choose, say $\sigma = 0.1$ and $\omega_{\text{true}} = 1.2$, to simulate the data $d_i = y(t_i, \omega_{\text{true}}) + n_i$.

- Compute the likelihood on grid of ω.



One Dimensional Example

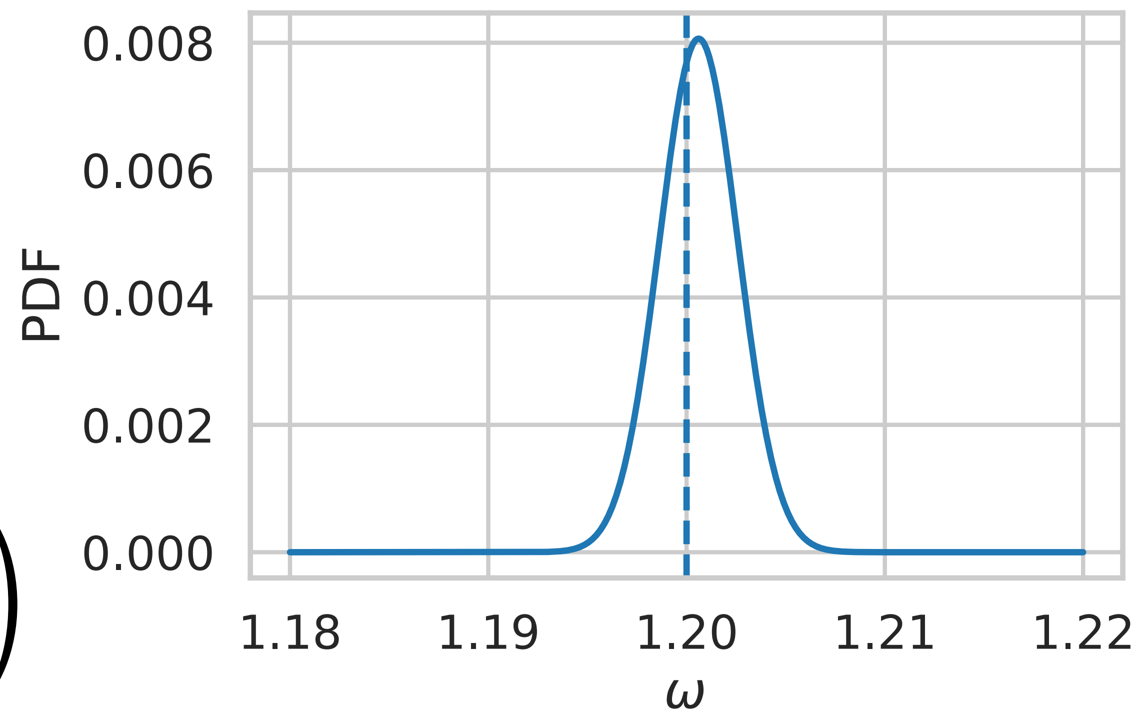
- For multiple observations,

$$\mathscr{L}(\mathbf{d}\,|\,\omega, M_A) = \prod_i \mathscr{L}(\mathbf{d_i}\,|\,\omega, M_A)$$

Or,

$$\ln \mathscr{L} = -\frac{1}{2}\Sigma_i\left(\frac{(d_i - y_A(t_i,\omega))^2}{\sigma^2} + \ln(2\pi\sigma^2)\right)$$

- Choose, say $\sigma = 0.1$ and $\omega_{\text{true}} = 1.2$, to simulate the data $d_i = y(t_i, \omega_{\text{true}}) + n_i$.



Posterior probability distribution

- Compute the likelihood on grid of ω. The posterior $\propto$ likelihood if prior is constant.

One Dimensional Example

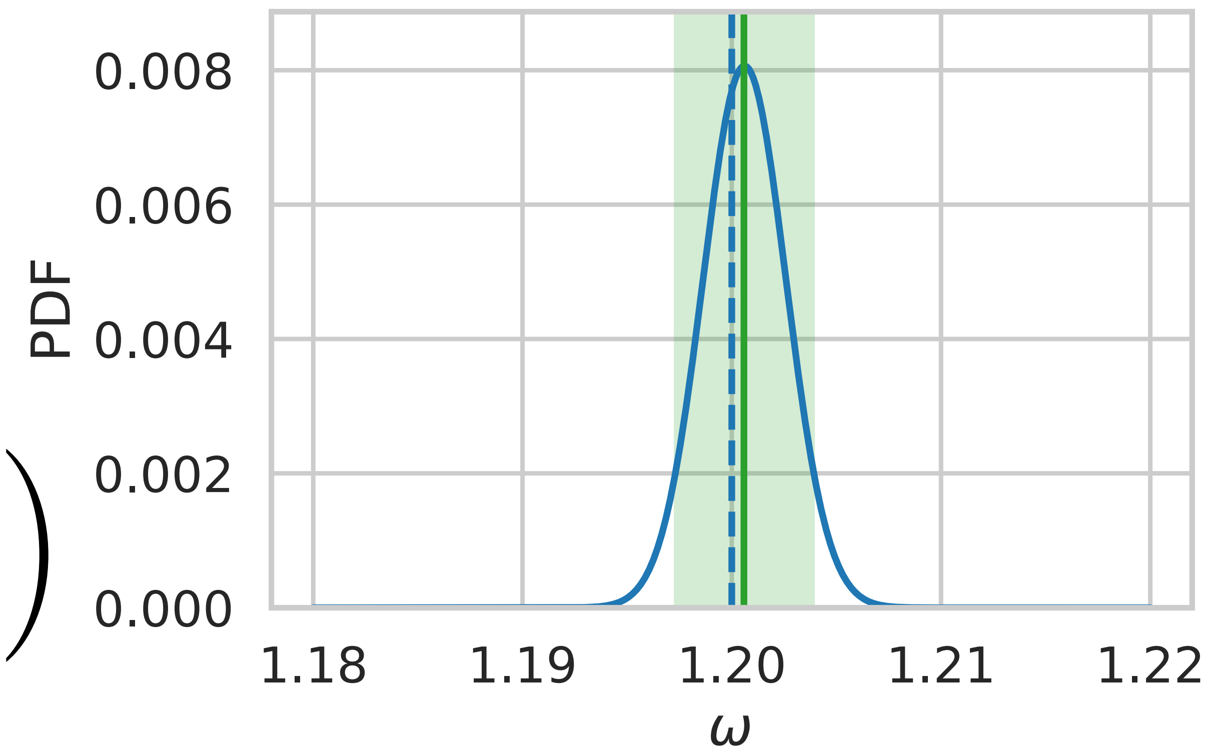
- For multiple observations,

$$\mathscr{L}(\mathbf{d} \,|\, \omega, M_A) = \prod_i \mathscr{L}(\mathbf{d_i} \,|\, \omega, M_A)$$

Or,

$$\ln \mathscr{L} = -\frac{1}{2}\Sigma_i \left(\frac{(d_i - y_A(t_i, \omega))^2}{\sigma^2} + \ln(2\pi\sigma^2) \right)$$

- Choose, say $\sigma = 0.1$ and $\omega_{\text{true}} = 1.2$, to simulate the data $d_i = y(t_i, \omega_{\text{true}}) + n_i$.



Posterior probability distribution

- Compute the likelihood on grid of ω. The posterior $\propto$ likelihood if prior is constant.
- Why is the peak (median) not consistent with the true value? Noise!
- How to quantify this? Bayesian answer is credible interval.
- We should always report inferences as ω has a median of XX and lies between YY and ZZ with 90% probability.

One Dimensional Example

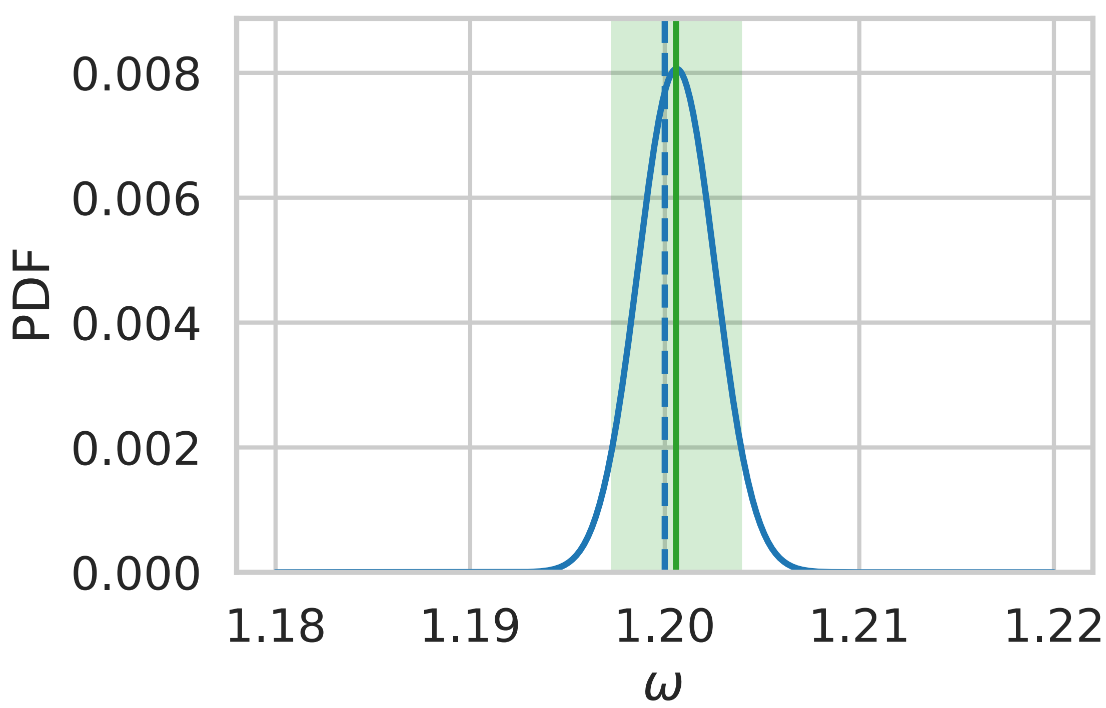
- For multiple observations,

$$\mathscr{L}(\mathbf{d}\,|\,\omega, M_A) = \prod_i \mathscr{L}(\mathbf{d_i}\,|\,\omega, M_A)$$

Or,

$$\ln \mathscr{L} = -\frac{1}{2}\Sigma_i\left(\frac{(d_i - y_A(t_i,\omega))^2}{\sigma^2} + \ln(2\pi\sigma^2)\right)$$

- Choose, say $\sigma = 0.1$ and $\omega_{\text{true}} = 1.2$, to simulate the data $d_i = y(t_i, \omega_{\text{true}}) + n_i$.



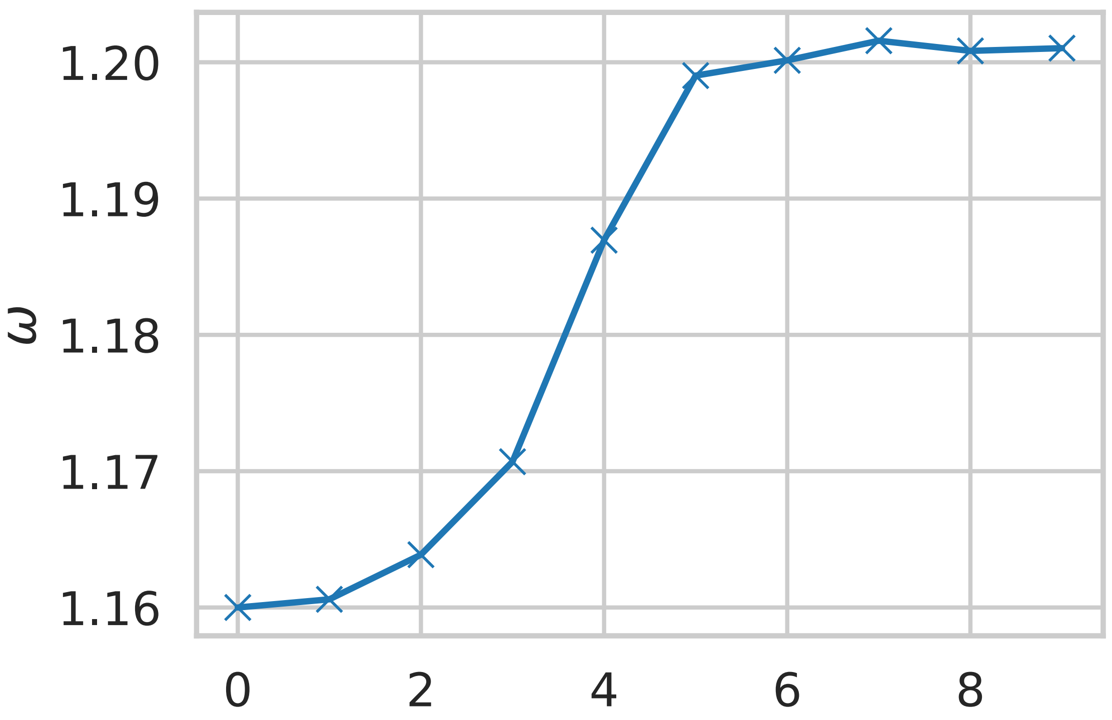
Posterior probability distribution

- Compute the likelihood on grid of ω. The posterior $\propto$ likelihood if prior is constant.
- Why is the peak (median) not consistent with the true value? Noise!
- How to quantify this? Bayesian answer is credible interval.
- What if the dimensionality (D) is reasonably high, then number of computations ~ $(\text{no . of grid points})^{\text{D}}$. Not feasible! Stochastic methods can come handy?!

Stochastic Sampling

- First, let's try building a peak finding algorithm

    1. Choose a random value, say ω_0, and evaluate likelihood $\mathscr{L}_0$
    2. Find a new $\omega_1 = \omega_0 + \delta\omega$ (random shift) and calculate likelihood $\mathscr{L}_1$.
    3. Store the new ω and likelihood $\mathscr{L}_1$ if $\mathscr{L}_1 > \mathscr{L}_0$.
    4. Iterate step 2 ad 3 until the maximum.
    5. We're always walking uphill.



- Limitations:

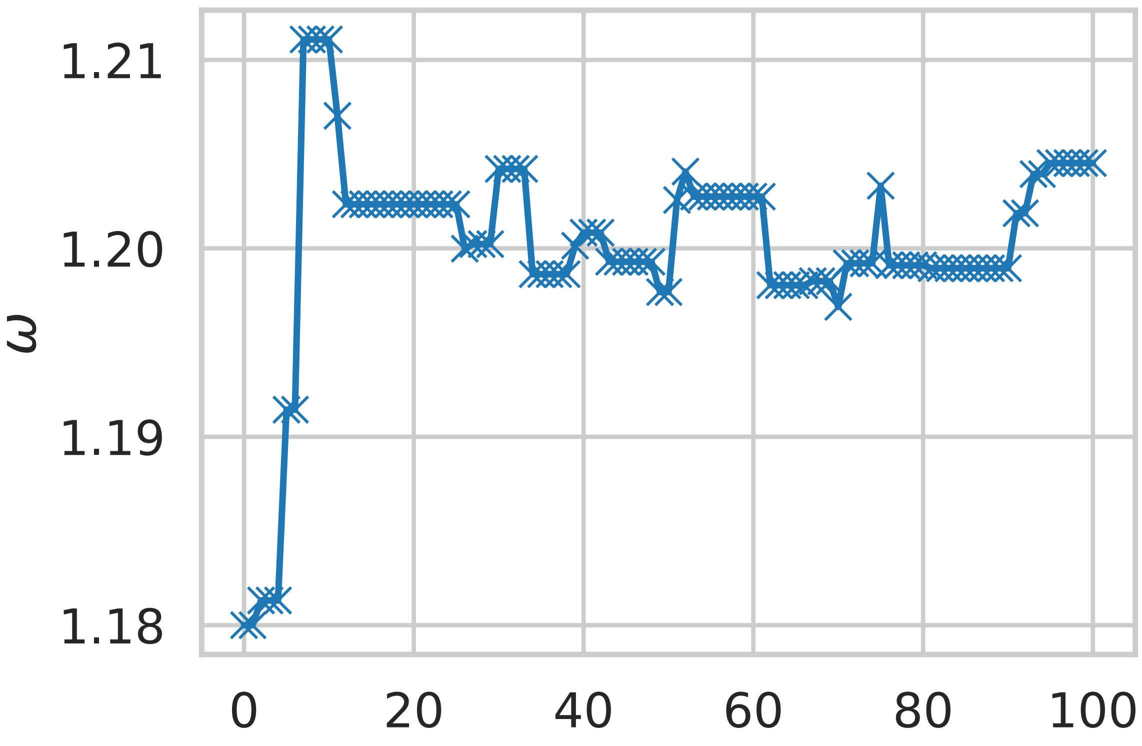
    (i) Does not tell us anything about the "structure" (shape of the distribution)

    (ii) bad if multimodal posteriors.

Metropolis-Hastings Sampler

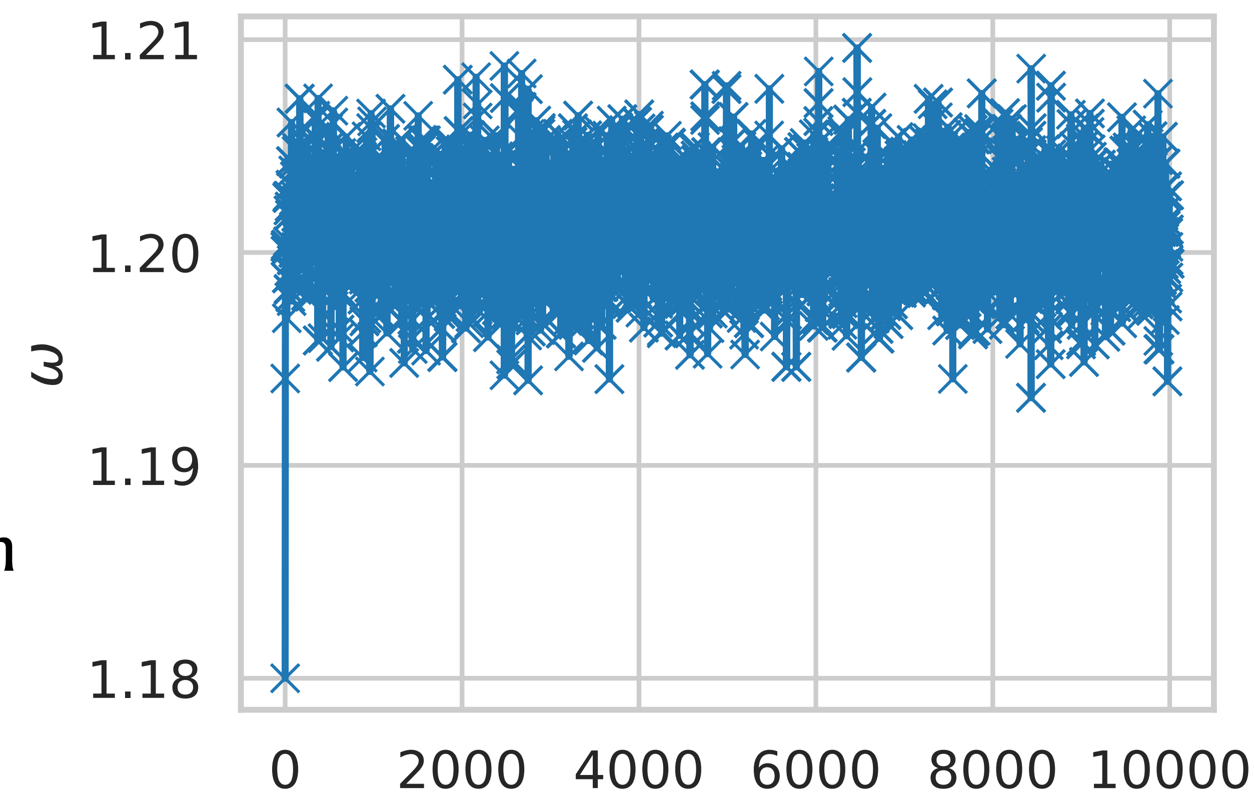
- Let us add some "randomness" in the steps

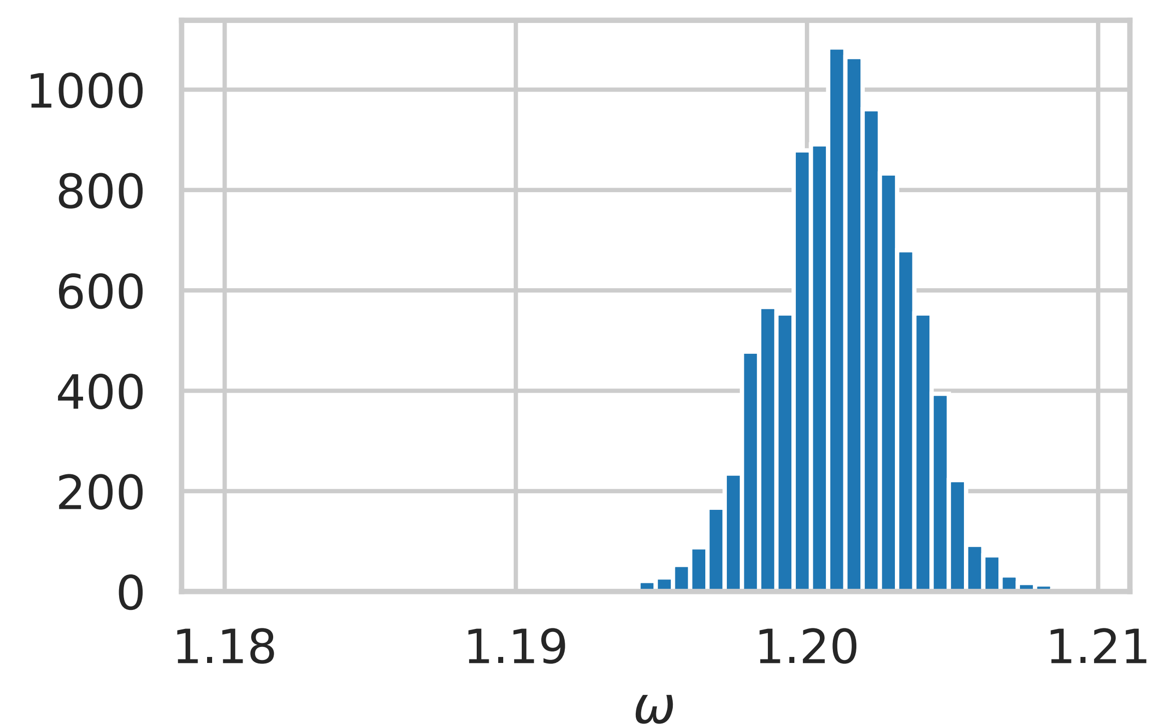
1. Choose a random value, say ω_0, and evaluate likelihood $\mathscr{L}_0$
2. Find a new $\omega_1 = \omega_0 + \delta\omega$ (random shift) and calculate likelihood $\mathscr{L}_1$.
3. Store the new ω and likelihood $\mathscr{L}_1$ if $\mathscr{L}_1 > \mathscr{L}_0\,\alpha$, where α is a random number $\in [0,1]$.
4. The samplers will walk both uphill and downhill.



Metropolis-Hastings Sampler

- Let us add some "randomness" in the steps
    1. Choose a random value, say ω_0, and evaluate likelihood $\mathscr{L}_0$
    2. Find a new $\omega_1 = \omega_0 + \delta\omega$ (random shift) and calculate likelihood $\mathscr{L}_1$.
    3. Store the new ω and likelihood $\mathscr{L}_1$ if $\mathscr{L}_1 > \mathscr{L}_0\,\alpha$, where α is a random number $\in [0,1]$.
    4. The samplers will walk both uphill and downhill.
- Limitations: (i) not efficient if multimodal posterior.



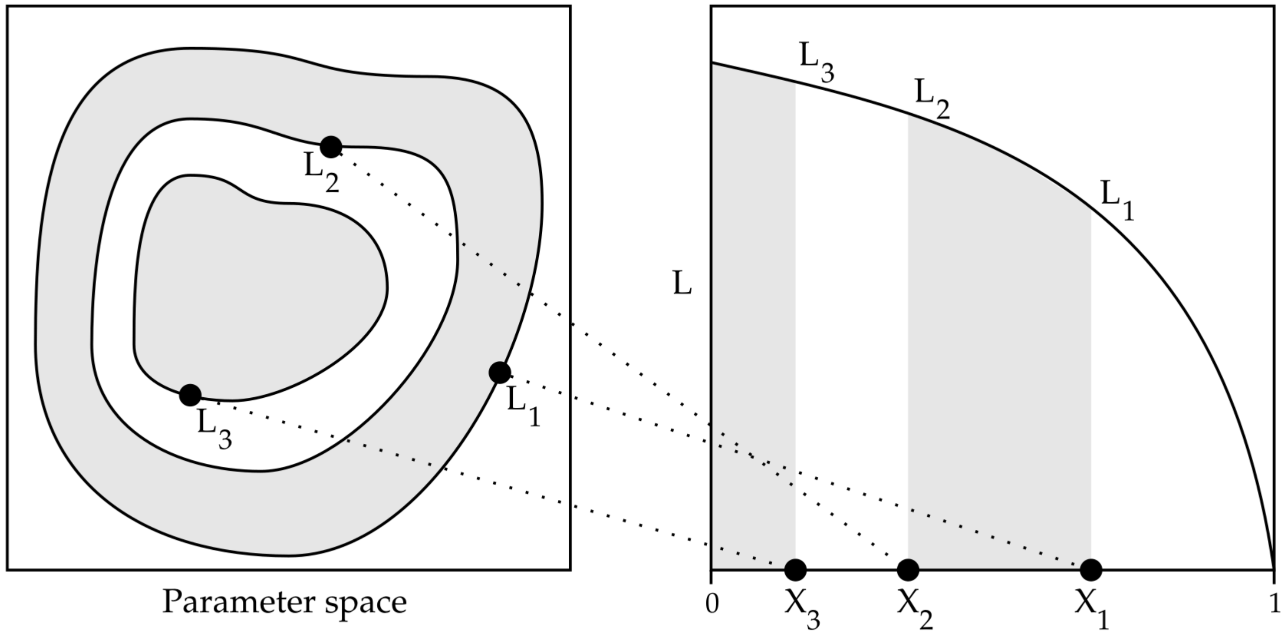


Nested Sampling

- Defining the prior volume as X such that $dX = p(\theta | M_A) d\theta$ where

$$X(\mathscr{L}) = \int_{p(d|\theta,M_A)>\mathscr{L}} d\theta\ p(\theta | M_A)$$

The total probability volume within a likelihood contour defined by $p(d | \theta, M_A) = \mathscr{L}$.



- The evidence,

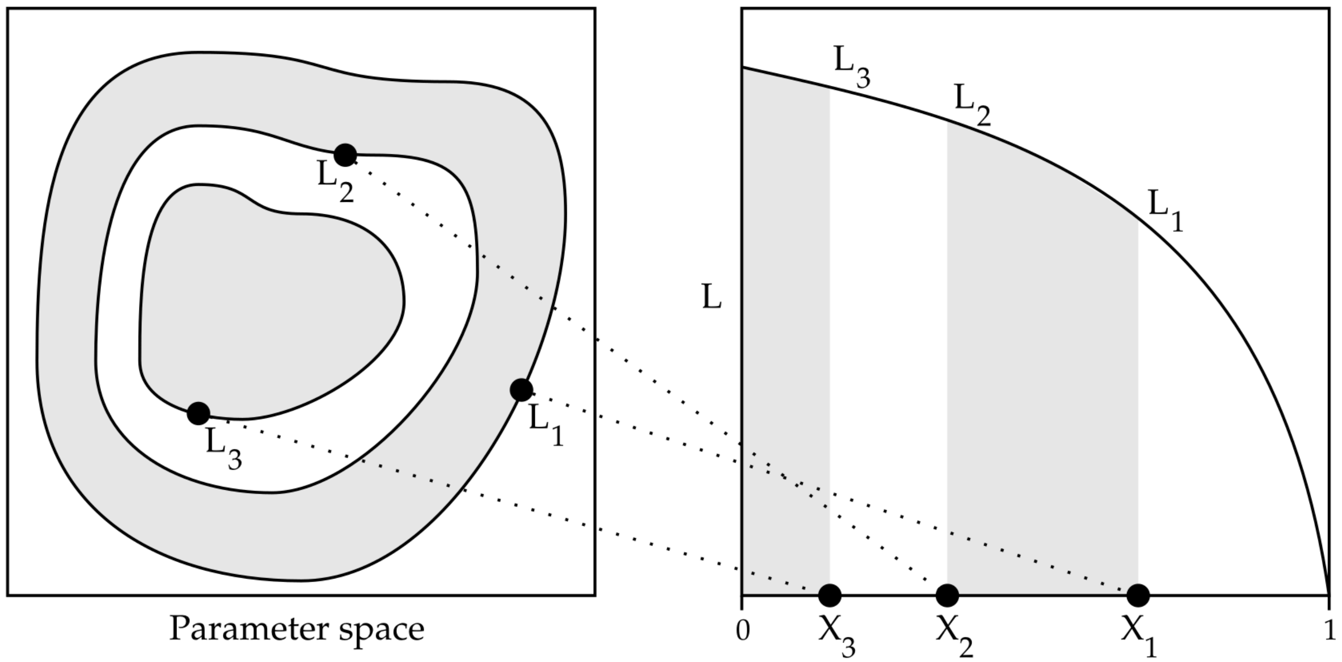
$$Z \equiv p(d | M_A) = \int_0^1 \mathscr{L}(X)\ dX$$

- Evaluating the likelihoods $\mathscr{L}_i = \mathscr{L}(X_i)$ associated with monotonically decreasing sequence of prior volumes X_i: $0 < X_N < \ldots < X_2 < X_1 < X_0 = 1$

$$Z = \sum_{i=1}^{N} \frac{1}{2}(X_{i+1} - X_i)\ \mathscr{L}_i \implies p(\theta | d, M_A) = \frac{\frac{1}{2}(X_{i+1} - X_i)\ \mathscr{L}_i}{Z}$$

Nested Sampling

- Select a set of initial live points sampled from the prior.
- The point with the lowest likelihood is replaced with a new sample with higher likelihood.

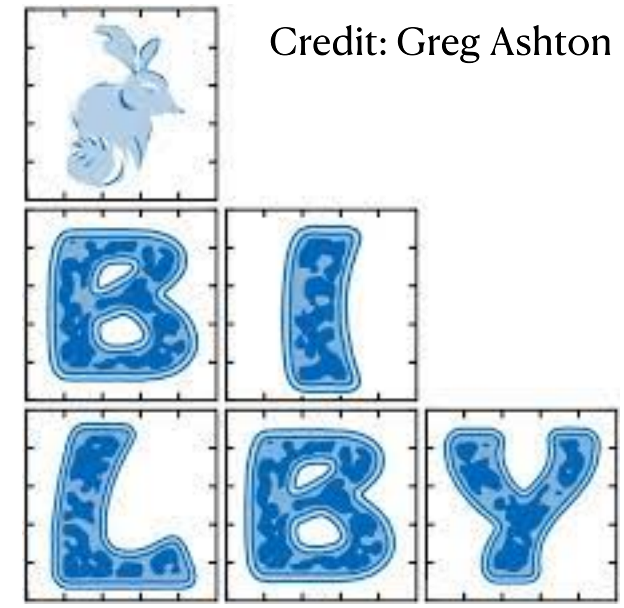


- Iterate this until reaching the stopping condition $\mathscr{L}_{\max} X_i / Z_i > e^{0.1}$ with $\mathscr{L}_{\max}$ is the maximum likelihood value.
- Checking whether the evidence estimate would change by more than a factor of ~0.1 if all the prior support were at the maximum likelihood.

Sampling Takeaways

- Stochastic samplers are about drawing random samples from a posterior distribution.
- MCM methods need tuning
  - Parallel tempering to recover the multi-modal posteriors
  - Does not calculate evidence by default. But not advised.
- Nested sampling is better than MCMC
  - It deals with multi-modal posteriors more effectively
  - Provides you evidences that can be useful for model selection.

GW Parameter Estimation: Bilby



Credit: Greg Ashton

- A generic **B**ayesian **I**nference **Lib**ra**ry**.
- Special support to gravitational-wave transients.
- Structure
  - Likelihood object

```
likelihood = bilby.gw.likelihood.base.GravitationalWaveTransient(
    interferometers, waveform_generator, priors, ….
)
```

  - Priors as python dictionaries
  - Samplers: dynasty, pymultinest, …, etc.

```
result = bilby.run_sampler(
    likelihood, prior, sampler="dynesty", outdir="outdir",
    label="GW150914",
    nlive=500, dlogz=0.1, …
)
```

Bilby Output

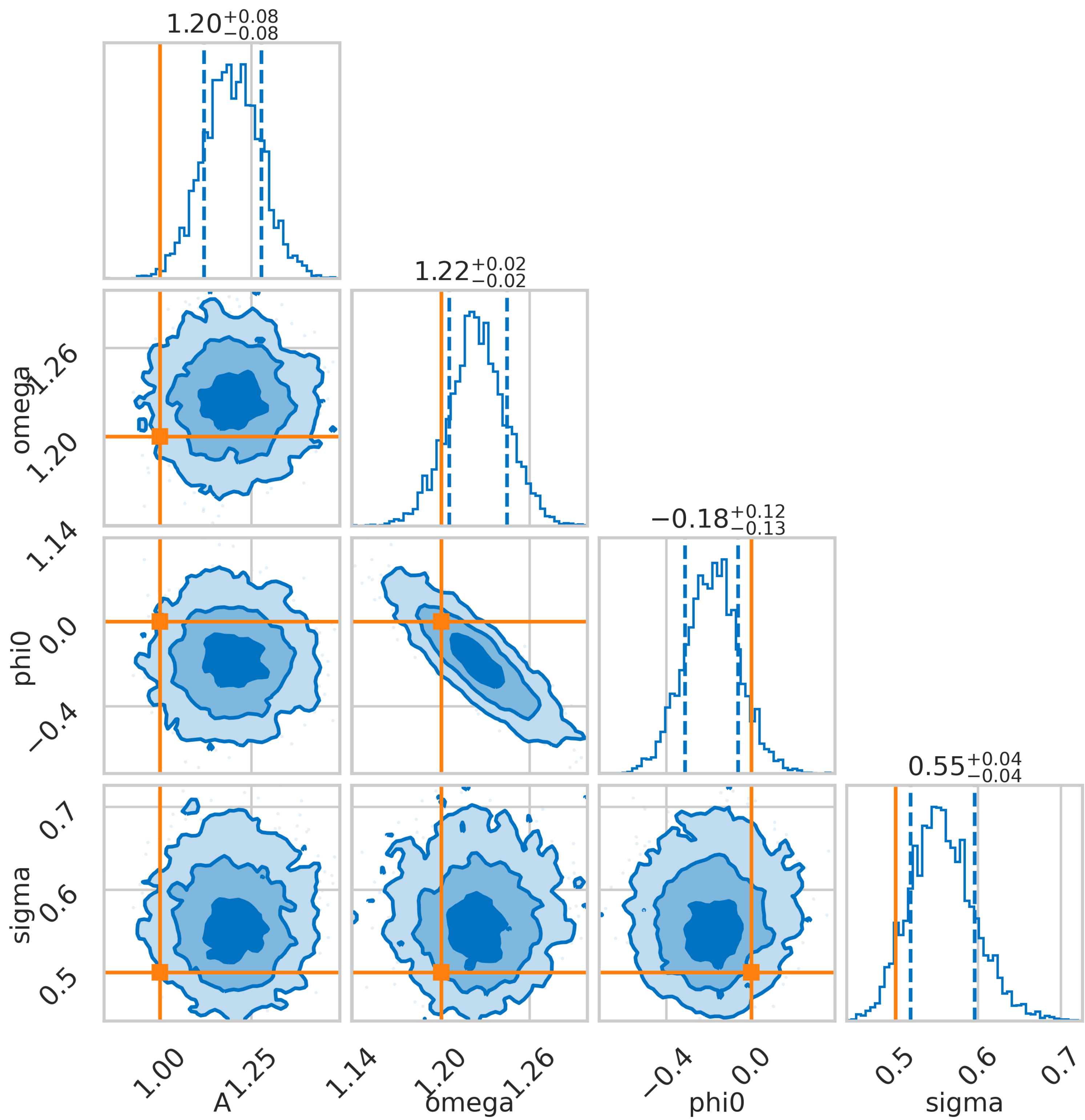
- Let the model M_A

  $$y(t) = A\sin(\omega t + \phi_0)$$

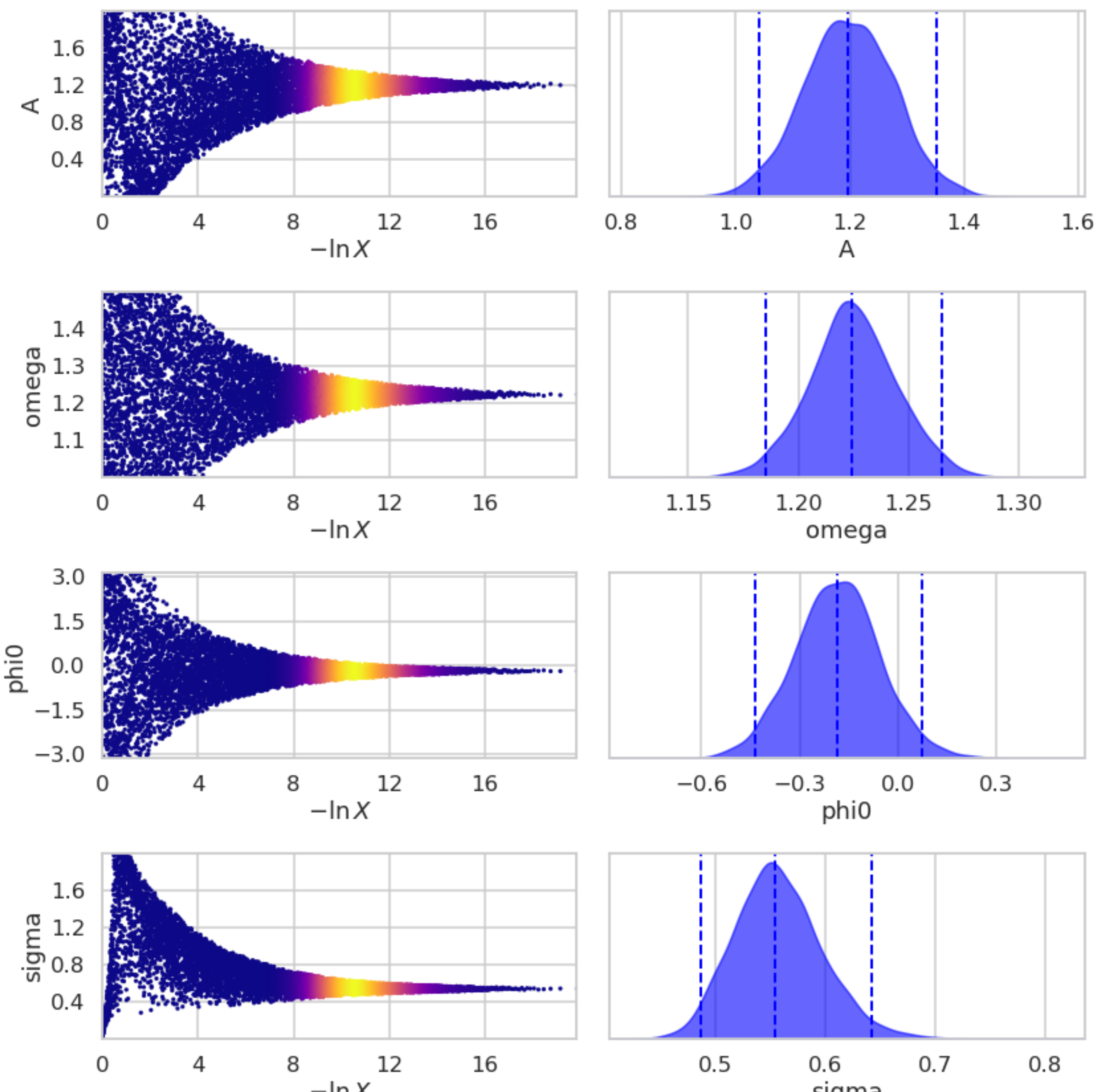
  parameters A, ω, and ϕ_0.

- Result object contains information about posteriors, priors, and likelihood, ..., etc.

- Just `result.plot_corner()` will give us



Bilby Trace Plots

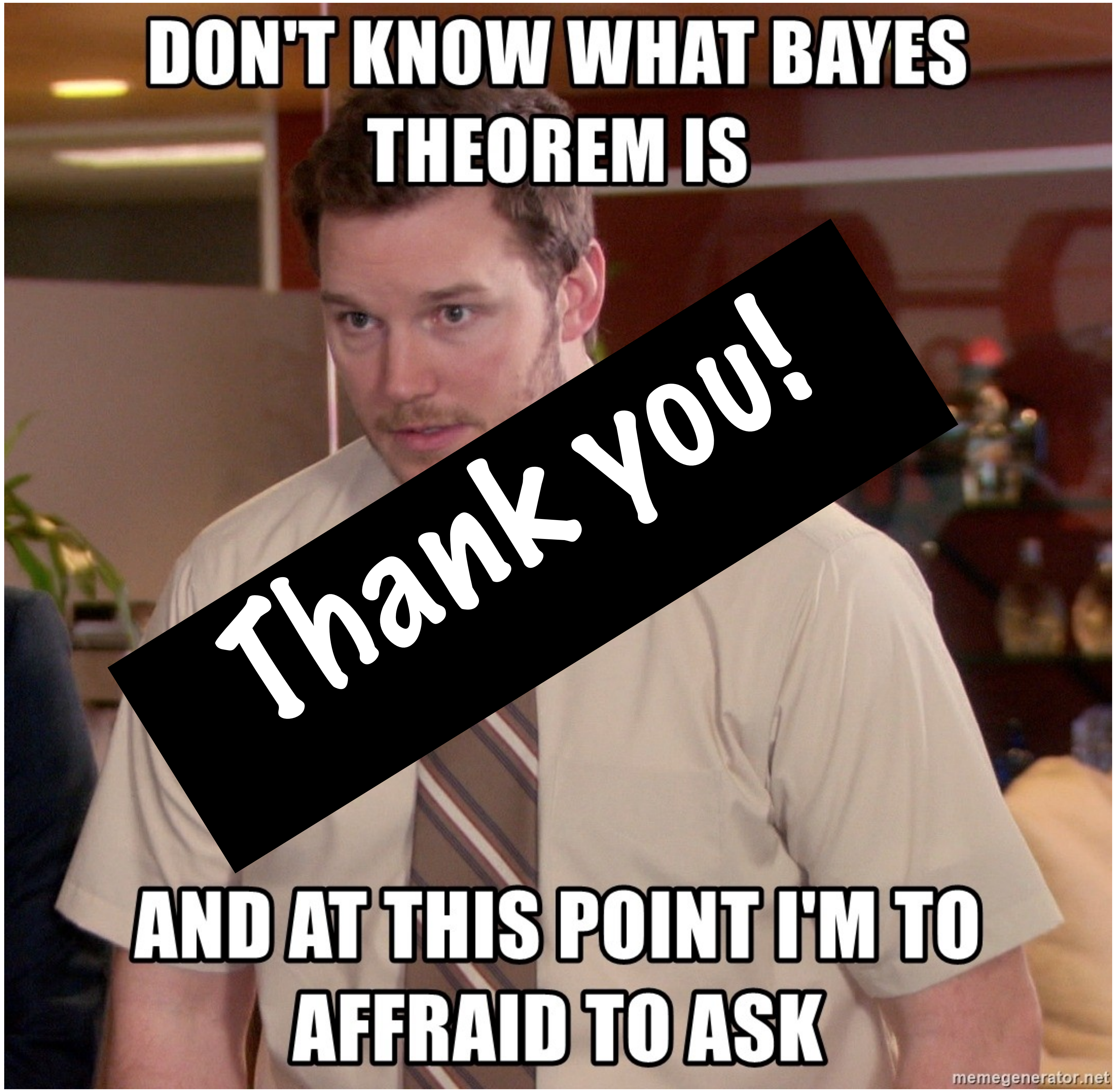


Conclusion

- Parameter estimation of a compact binary merger in GW is a high dimensionality problem.
- Need stochastic samplers to sample the likelihood in such case.
- Output is probability distributions of the parameters due to noise uncertainty in the data.
- Bayesian inference is key to the parameter estimation in GW sources, especially for compact binary mergers.
- Bilby is one such Bayesian Inference Library to perform parameter estimation.

References

- Bilby: Ashton et al 2018 (https://lscsoft.docs.ligo.org/bilby/)
- Data Analysis: A Bayesian Tutorial by D. S. Sivia & J. Skilling
- An Introduction to Bayesian Inference in GW Astronomy, Thrane & Talbot (2018)
- GWOSC: https://www.gw-openscience.org/
- GWpy: https://gwpy.github.io/
- PyCBC: https://pycbc.org/


DON'T KNOW WHAT BAYES THEOREM IS
Thank you!
AND AT THIS POINT I'M TO AFRAID TO ASK
memegenerator.net