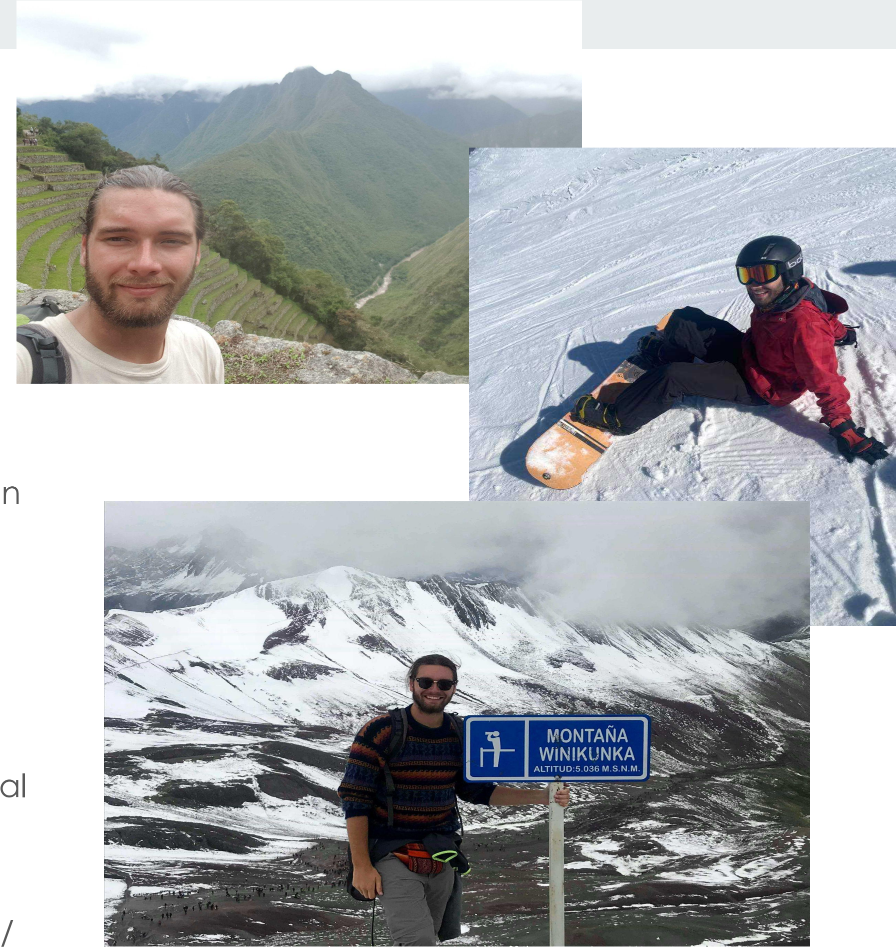
How I stopped worrying and learnt to love Bayesian Statistics

Matt Miles / Daniel Reardon

(Heavily influenced by lectures given by: Steve Taylor / Neil Cornish / Lindley Lentati)

About me

- My name is Matt!
- Postdoctoral researcher at Swinburne University in Melbourne
  - Submitted a thesis 3 weeks ago so I'm a post-doc without the doc
- Love all things millisecond pulsars, pulsar timing arrays and gravitational waves
- More than anything, pulsar noise and gravitational waves is what really excites me
- I also enjoy doing a lot of things very badly
  - Hiking / travelling / snowboarding / surfing / running / music

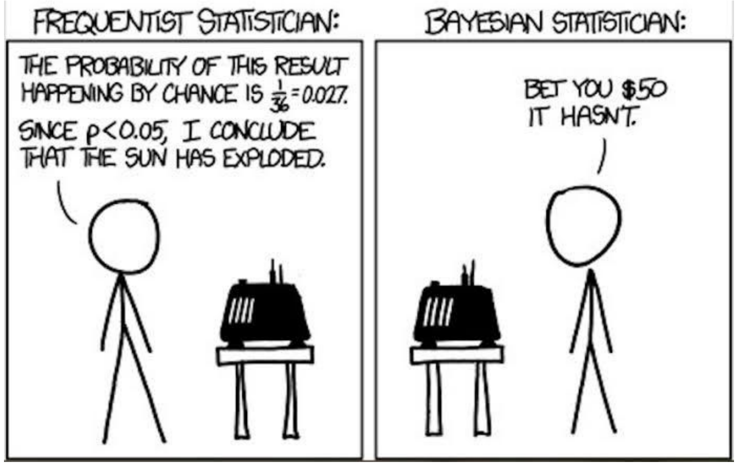


What, how, why?

Statistics are separated into two broad camps:

- Frequentist: **p(*d* | *h*)**
  - The likelihood of observing the data ***d*** given the model ***h***
- Bayesian: **p(*h* | *d*,*H*)**
  - The likelihood of observing the model ***h*** given data ***d*** under hypothesis ***H***

Both can give reasonable results, but in pulsar timing we prefer to use Bayesian statistics.


FREQUENTIST STATISTICIAN:
THE PROBABILITY OF THIS RESULT HAPPENING BY CHANCE IS $\frac{1}{36}$=0.027. SINCE $p<0.05$, I CONCLUDE THAT THE SUN HAS EXPLODED.
BAYESIAN STATISTICIAN:
BET YOU $50 IT HASN'T.

Why Bayes?

In the frequentist landscape, the signal is fixed but the data is considered random.

In the Bayesian landscape, the data is recorded (not random) but the signals are unknown and need to be ***inferred.***

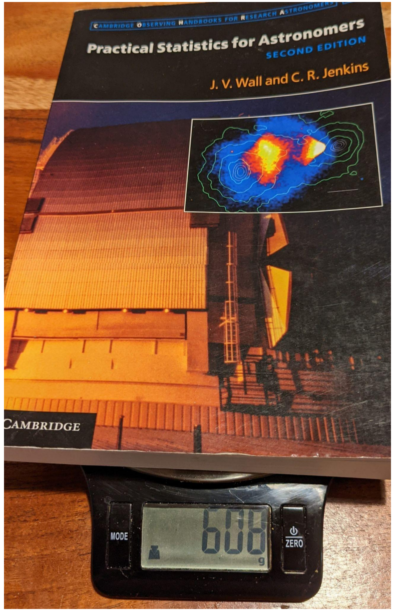
In pulsar timing, we are largely unaware of what signals are present in the data, leading us to require a Bayesian approach (with some frequentist techniques sprinkled in).

Measurements as a probability distribution

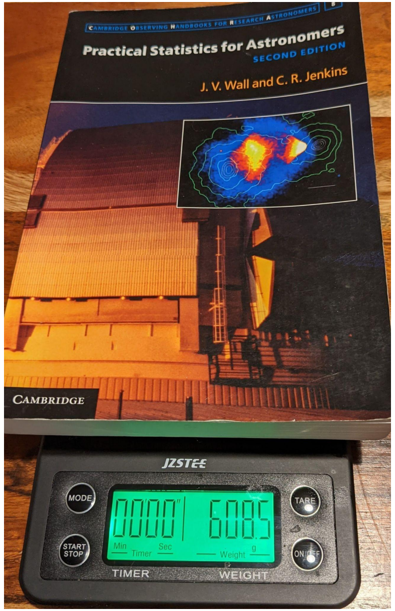
Measurements always come with uncertainty

- Measuring the mass of “*Practical statistics for Astronomers*”
  - Measurement precision is 1g or 0.1g depending on the scale.
  - Uncertainty is related to this precision.
  - Measurement uncertainty depends on the ***sampling distribution***

Error and uncertainty, relate to accuracy and precision respectively


Cambridge Observing Handbooks for Research Astronomers
Practical Statistics for Astronomers
SECOND EDITION
J. V. Wall and C. R. Jenkins
CAMBRIDGE
MODE
608
g
ZERO

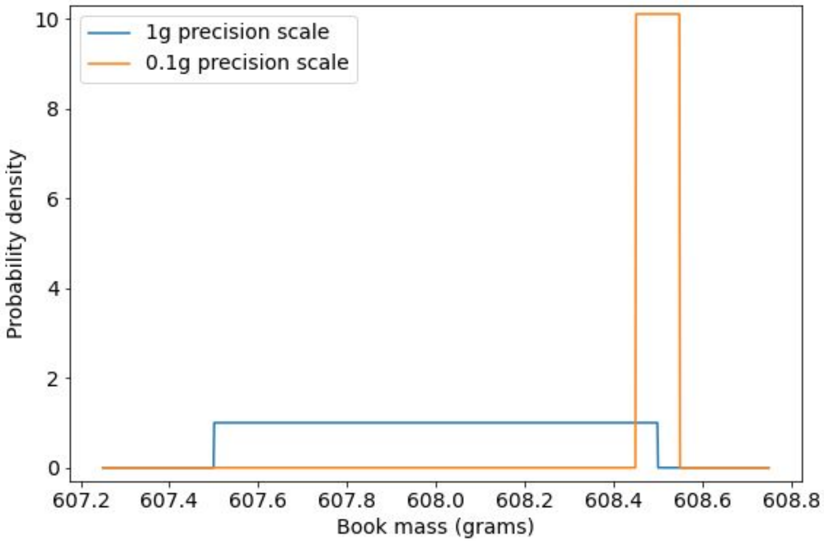
608g measurement


Cambridge Observing Handbooks for Research Astronomers
Practical Statistics for Astronomers
SECOND EDITION
J. V. Wall and C. R. Jenkins
CAMBRIDGE
JZSTEE
MODE
TARE
START STOP
ON/OFF
0000
608.5
Min
Sec
Timer
Weight
g
TIMER
WEIGHT

608.5g measurement

Uncertainty and probability density functions (PDFs)

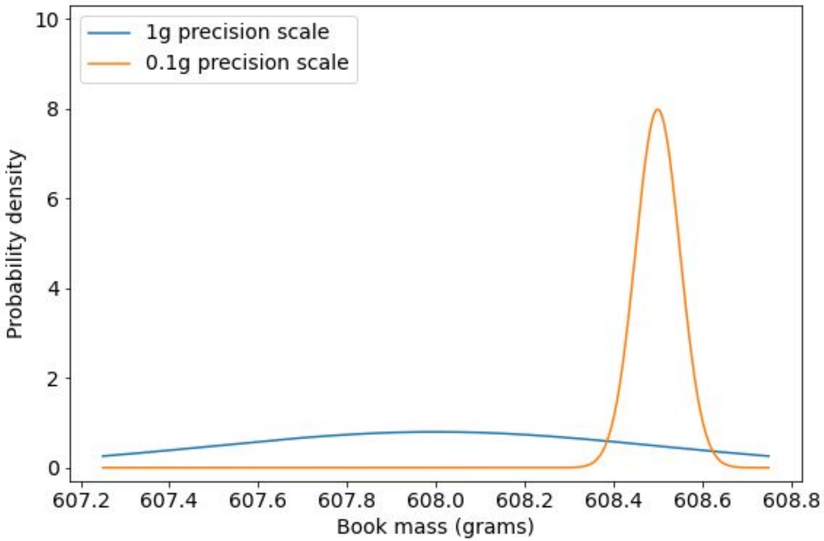
- Every measurement of a random variable can be described by a probability distribution that describes where the true value is likely to be.
- A model of this distribution, with a normalised area, is a *probability density function* (PDF)



Uncertainty and probability density functions (PDFs)

- Every measurement of a random variable can be described by a probability distribution that describes where the true value is likely to be.
- A model of this distribution, with a normalised area, is a *probability density function* (PDF)

- A common PDF is the normal distribution
  - Also called Gaussian or “Bell Curve”



Understanding the normal distribution (notebook)

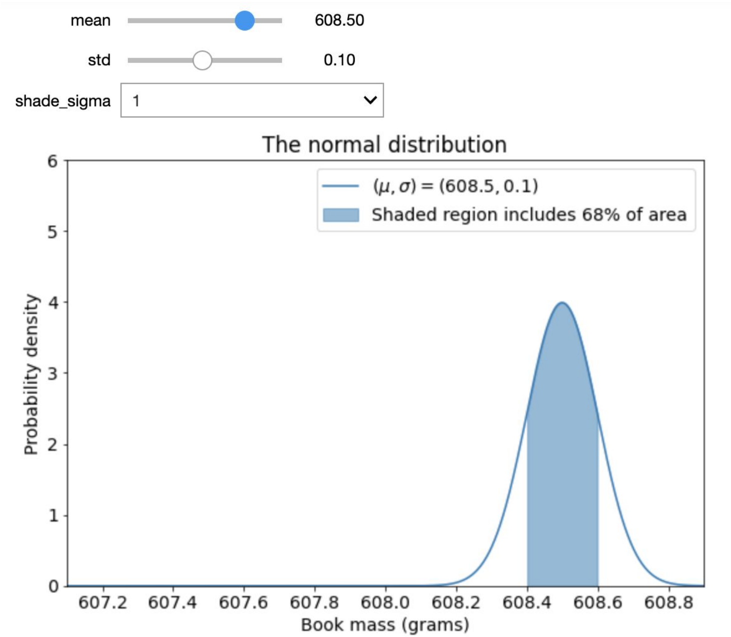
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$

An uncertainty (standard error), and a model PDF (e.g. normal), determines the confidence interval of measurements

- The "68-95-99.7" rule
  - 68% of the area is within 1σ from μ
  - 95% of the area is within 2σ from μ
  - 99.7% of the area is within 3σ from μ

In VM at:
~/exercices/lec4_data_analysis/data_analysis_model_fitting.ipynb

Run command "jupyter-notebook"



mean 608.50

std 0.10

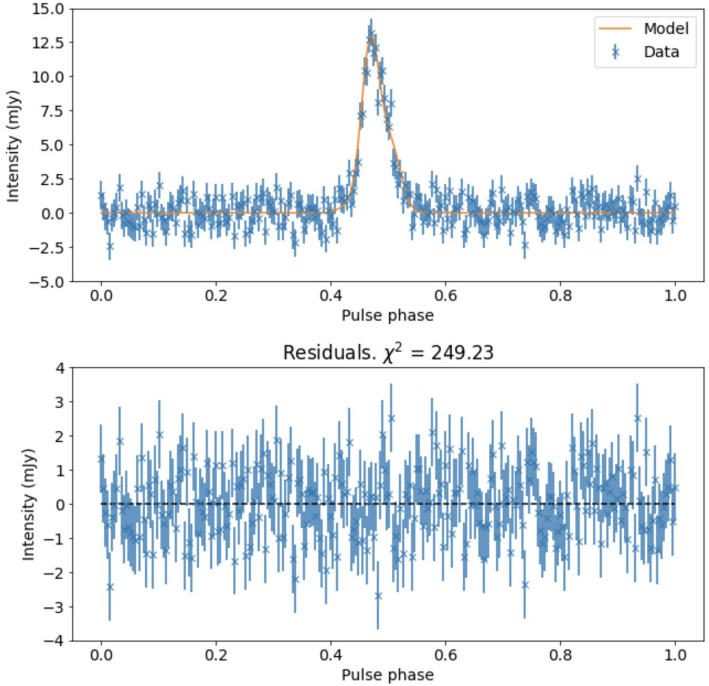
shade_sigma 1

Least squares fitting (e.g. in tempo2)

- Residual = Data - Model
- Least-squares fitting minimises the weighted, squared residual
- The chi-squared value is a weighted sum of the squared deviations:

$$\chi^2 = \sum_i \frac{(x_i - m_i)^2}{\sigma_i^2}$$

- Measurements x_i
- Model predictions m_i
- Uncertainties σ_i



A Gaussian likelihood function

Transitioning to Bayesian inference, the data are described by a likelihood. This is where the PDF of the data is important

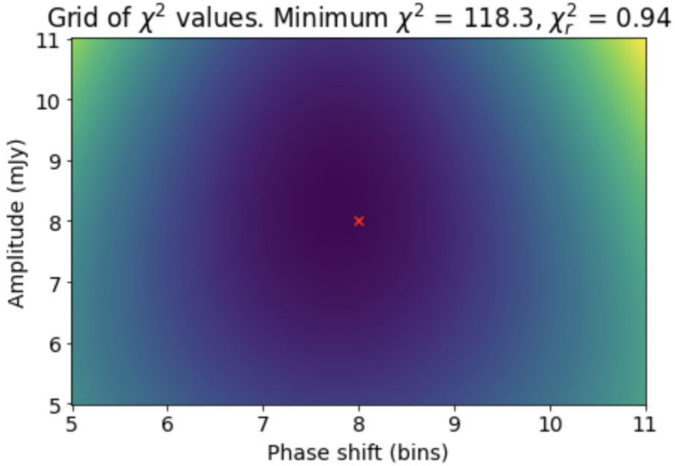
- Gaussian (normal) likelihood

$$P(X|M) \propto e^{-\frac{1}{2}\chi^2}$$

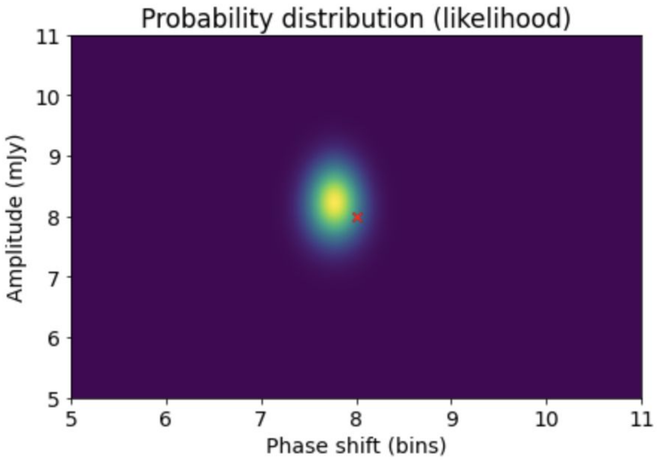
- In full, this comes from the equation for a normal PDF.
- Assume each measurement uncertainty describes a normal distribution

$$P(X|M) = \prod_i \frac{1}{\sigma_i\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x_i-m_i)^2}{\sigma_i^2}}$$

$$P(X|M) \propto e^{-\frac{1}{2}\sum_i \frac{(x_i-m_i)^2}{\sigma_i^2}}$$



Grid of χ^2 values. Minimum χ^2 = 118.3, χ_r^2 = 0.94



Probability distribution (likelihood)

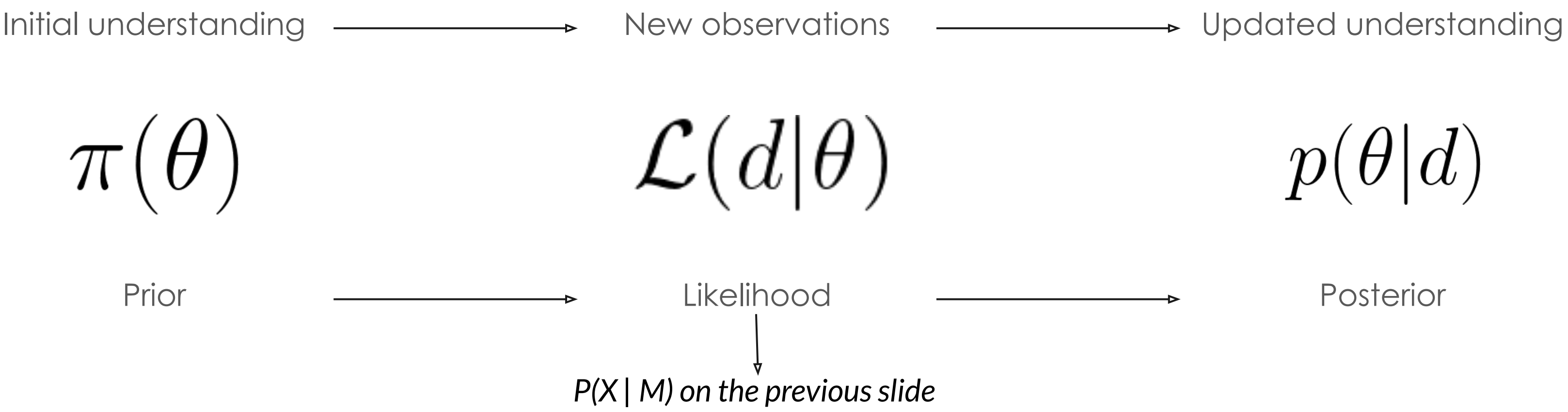
$$\theta = \text{model parameters}$$

$$d = \text{data}$$

Bayes Theorem

Credit: Neil Cornish

The primary aim of modern Bayesian inference is to construct a posterior distribution (Thrane and Talbot 2018)



Initial understanding → New observations → Updated understanding

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

Prior → Likelihood → Posterior

P(X | M) on the previous slide

$$\theta = \text{model parameters}$$

$$d = \text{data}$$

Bayes Theorem

Credit: Neil Cornish

The primary aim of modern Bayesian inference is to construct a posterior distribution (Thrane and Talbot 2018)

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

Prior → Likelihood → Posterior

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

These are related through a normalisation factor (evidence): $\mathcal{Z} = \int d\theta \mathcal{L}(d|\theta)\pi(\theta)$

"All models are wrong, but some are useful."
- George Box

Bayesian Inference

- The goal is to find the model that the data favours the most.
- We can find this by comparing the evidence of the models, giving us a Bayes factor.

For two models ***A*** and ***B*** we can compare their evidence such that:

Model A evidence

Bayes Factor

$$\mathcal{B}^{A}_{B} = \frac{\mathcal{Z}_{A}}{\mathcal{Z}_{B}}$$

Model B evidence

Bayesian Inference

Note: The true way to do this is to use the "Odds ratio", the product of the Bayes factor and the Prior Odds Ratio.

In PTA analyses, the prior odds are usually of unity, so the Bayes factor is appropriate.

If the Bayes factor is greater than unity, the data prefers model ***A.***

Otherwise, it prefers model ***B.***

Model A evidence

Bayes Factor

$$\mathcal{B}^{A}_{B} = \frac{\mathcal{Z}_A}{\mathcal{Z}_B}$$

Model B evidence

How is this done?

Product space sampling techniques are used to construct posterior distributions of the models given the data. Commonly these are:

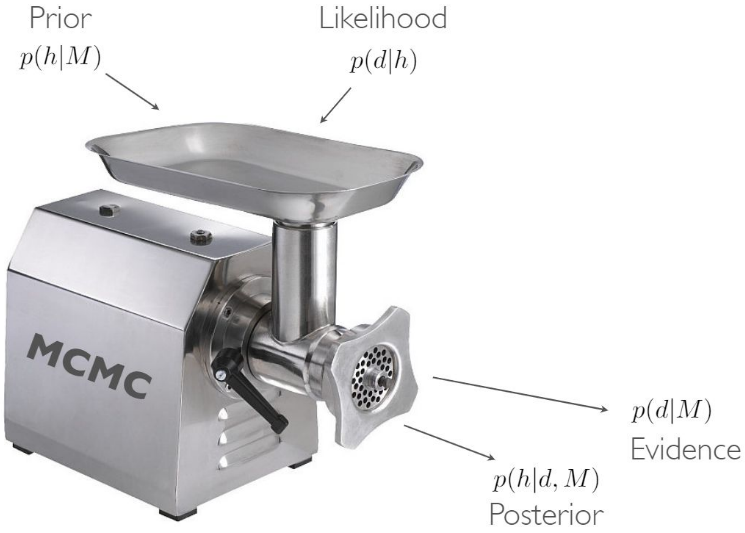
- Markov chain Monte Carlo (MCMC) techniques
  - Direct model comparison is possible so evidence calculation is not required
  - Very fast, but limited in model comparison and exploring multi-modal distributions
  - Attractive because it is trivially parallelisable on computing clusters

- Nested sampling techniques
  - Natively calculates the evidence of the model
  - Very useful for large scale model comparison
  - Can protect against unsampled prior spaces where complicated models are considered

How is this done?

MCMC is the most common in the PTA field and is what we use in this workshop

- If you’re interested in setting up nested sampling for PTA methods in the future, let me know and I can help you down the track!

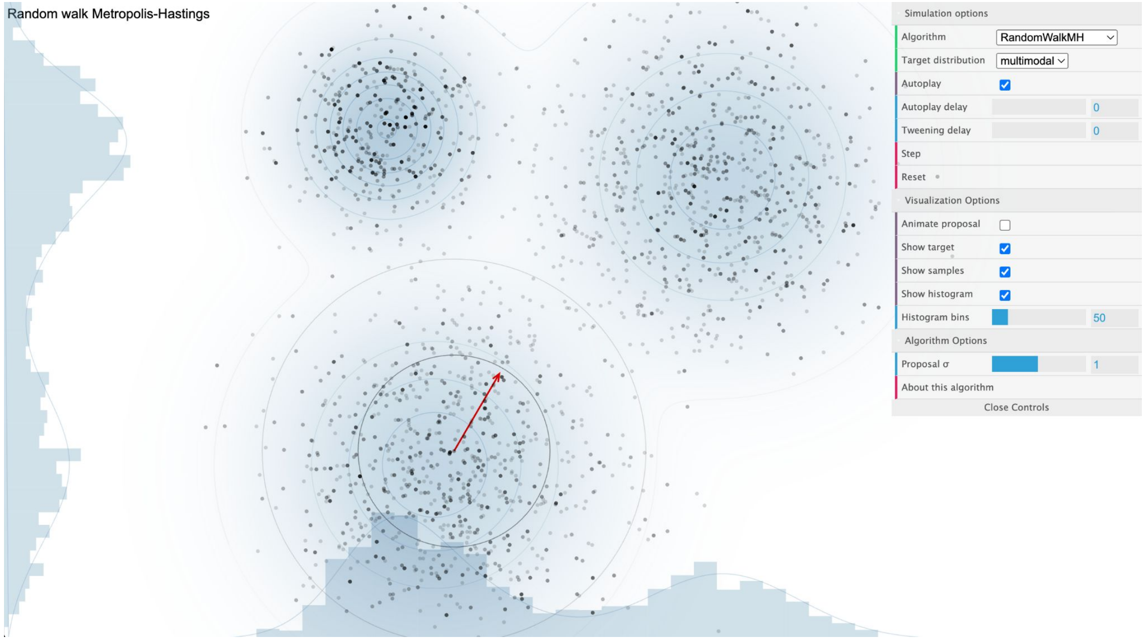


“The PTA analysis pipeline”

Credit: Neil Cornish

Visualising an MCMC sampler

https://chi-feng.github.io/mcmc-demo/app.html?algorithm=RandomWalkMH&target=multimodal



Simulation options

Algorithm RandomWalkMH

Target distribution multimodal

Autoplay ☑

Autoplay delay 0

Tweening delay 0

Step

Reset

Visualization Options

Animate proposal □

Show target ☑

Show samples ☑

Show histogram ☑

Histogram bins 50

Algorithm Options

Proposal σ 1

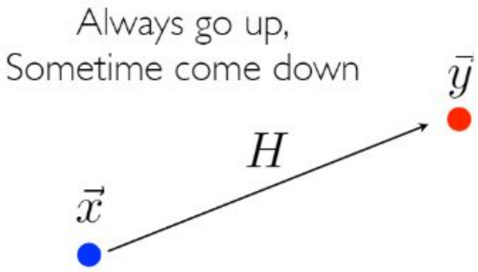
About this algorithm

Close Controls

How is this done?

What actually is MCMC?

- Random walk algorithm from x to y, with a selection criteria for moving forward
- Proposed by Metropolis, developed by Hastings
- The jump proposal (q) is important!
    - We want the MCMC to converge to the best solution
    - Bad jumps or complex distributions can stop this from happening



$$H = \min\left(1, \underbrace{\frac{p(\vec{y})}{p(\vec{x})}}_{\text{Prior}} \underbrace{\frac{p(d|\vec{y})}{p(d|\vec{x})}}_{\text{Likelihood}} \underbrace{\frac{q(\vec{x}|\vec{y})}{q(\vec{y}|\vec{x})}}_{\text{Proposal}}\right)$$

Steve Taylor's "The nanohertz gravitational wave astronomer": https://arxiv.org/pdf/2105.13270.pdf

How is this done?

In order to construct the posterior distributions, a prior range and a likelihood is needed.

The prior range can be selected based on physical expectations, but what about the likelihood?

$\mathbf{C} = \text{covariance matrix}$

$\overrightarrow{\delta t} = \text{timing residuals}$

$$\mathcal{L}(\overrightarrow{\delta t}|\overrightarrow{\eta}) \propto \frac{\exp\left(-\frac{1}{2}\overrightarrow{\delta t}^{T}\mathbf{C}^{-1}\overrightarrow{\delta t}\right)}{\sqrt{\det(2\pi\mathbf{C})}}$$

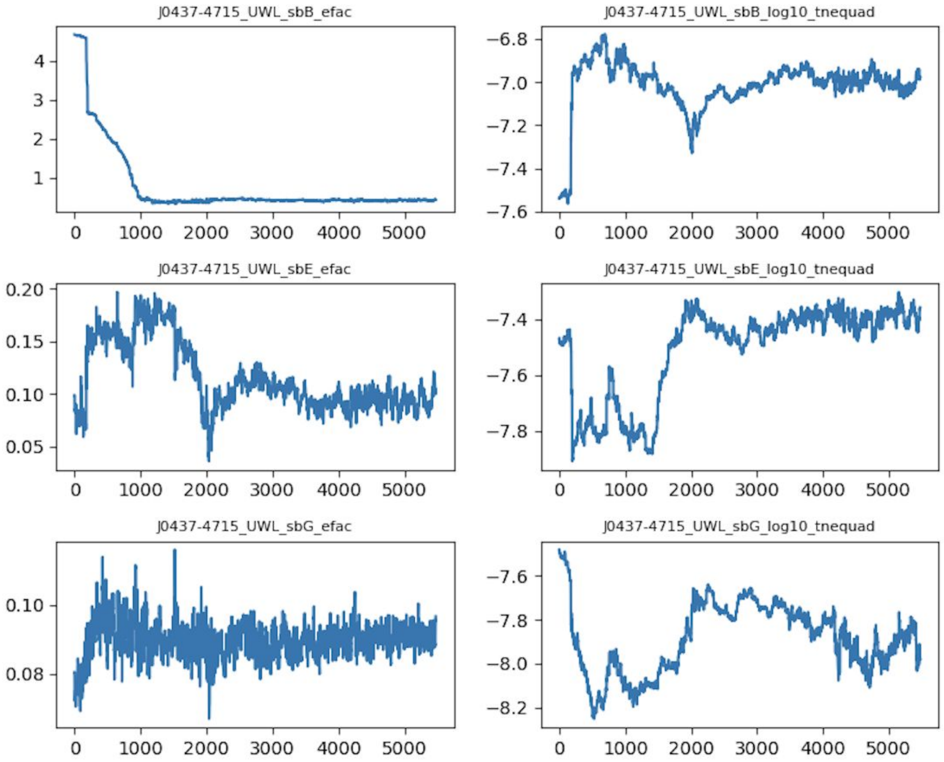
What does this give us?

The aim of the game is to construct the posterior distribution

Samples start at a random point, and progress towards higher likelihoods.

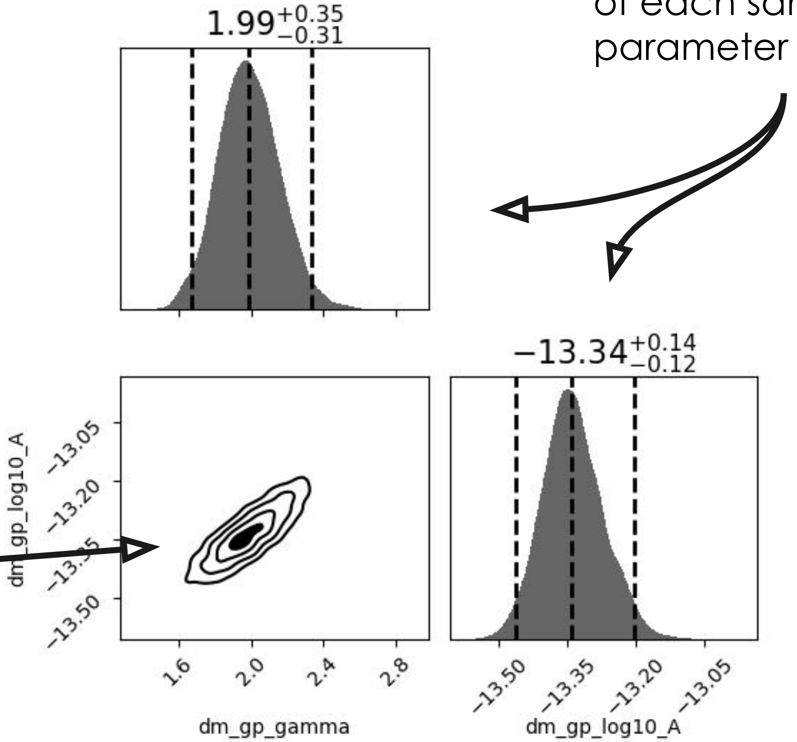
Eventually, the distribution of samples represents the posterior distribution. We then look at a "corner plot" of one and two dimensional histograms

A "trace" plot shows the progression of parameter samples



What does this give us?

The aim of the game is to construct the posterior distribution



1D probability density distributions of each sampled parameter

2D probability density distribution of both the sampled parameters

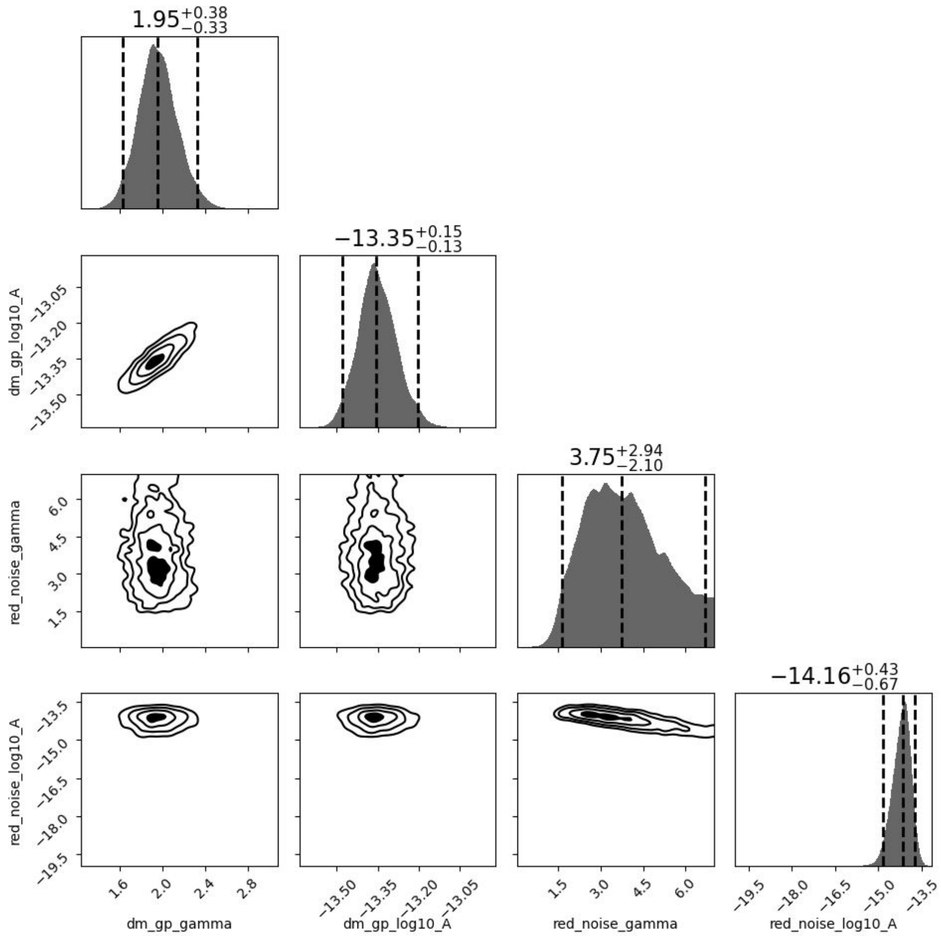
(Shows how the parameters relate to each other)

What does this give us?

The aim of the game is to construct the posterior distribution

- Started with 2 definitely well constrained parameters

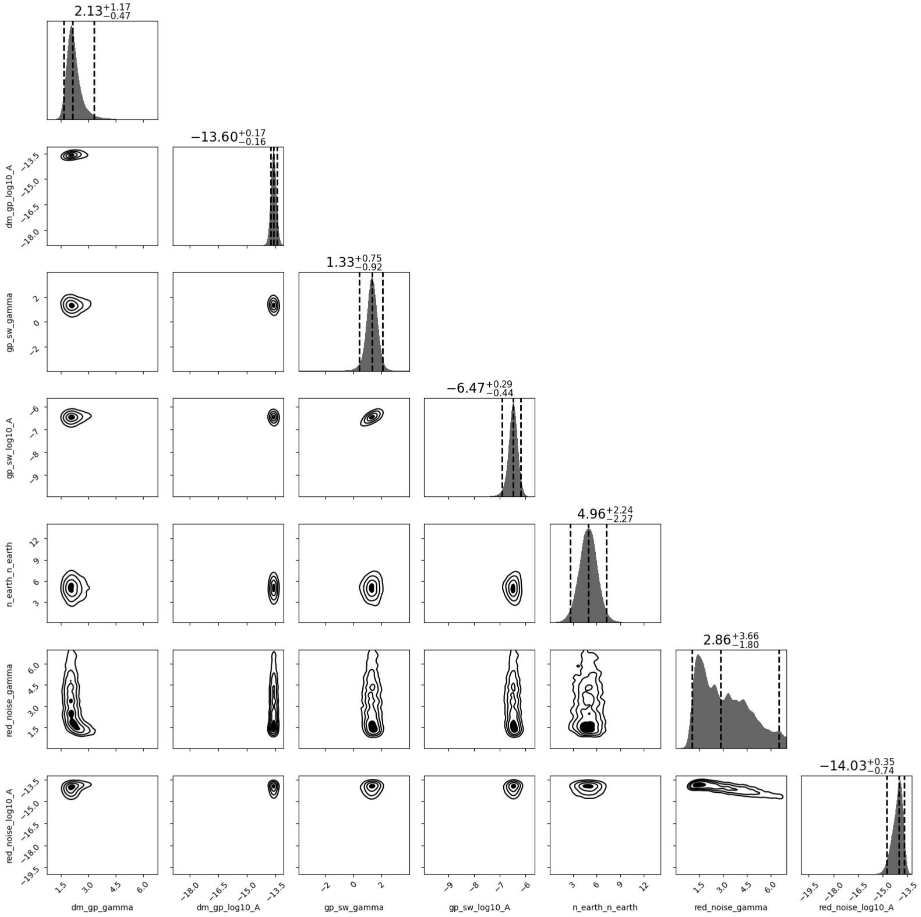
- So let's add more parameters and see if those processes are also constrained



What does this give us?

The aim of the game is to construct the posterior distribution

- Started with 2 definitely well constrained parameters
- So let’s add more parameters and see if those processes are also constrained
- And more…

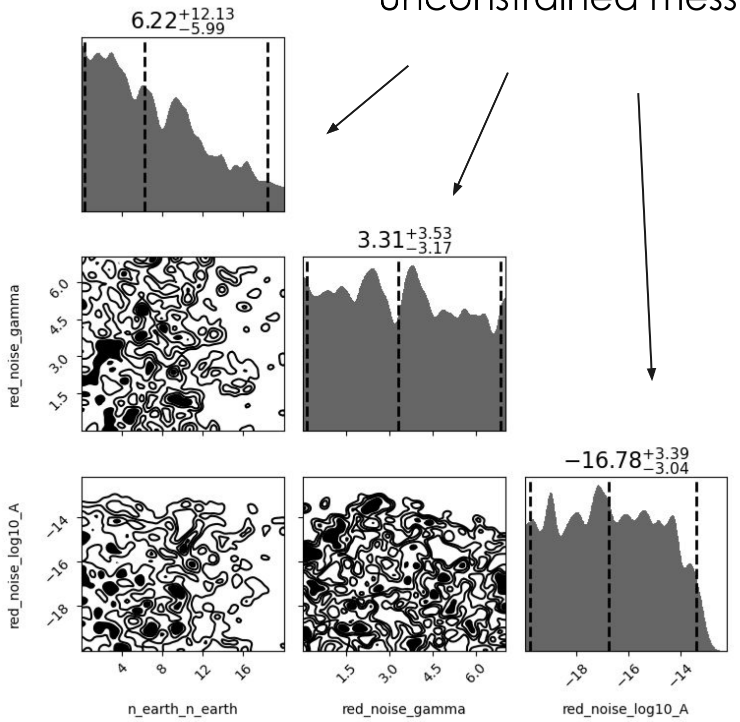


What does this give us?

The aim of the game is to construct the posterior distribution

- We can keep adding more but let's not
- The beauty of this modelling technique is it let's us check every possibility in a traceable way (if we want to)
- Importantly it lets us see where a model **isn't** favoured.

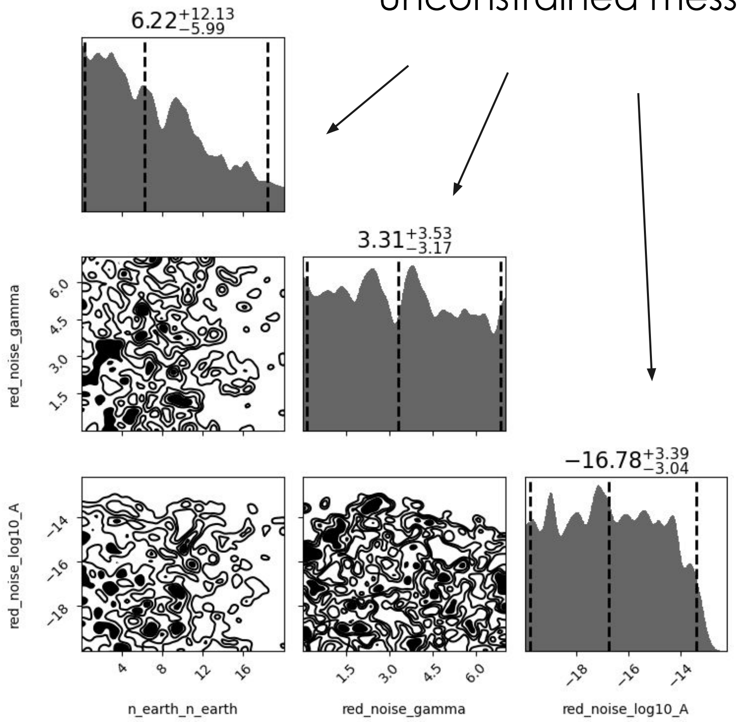
Terrible unconstrained mess:



What does this give us?

The aim of the game is to construct the posterior distribution

- The evidence comparisons and Bayes Factors that we mentioned before are your friend
- The MPTA has 85 pulsars and each could possibly have ~75 models
- As much as I love pulsars, I don't want to look at 6375 possible models
    - Neither do you



Terrible unconstrained mess:

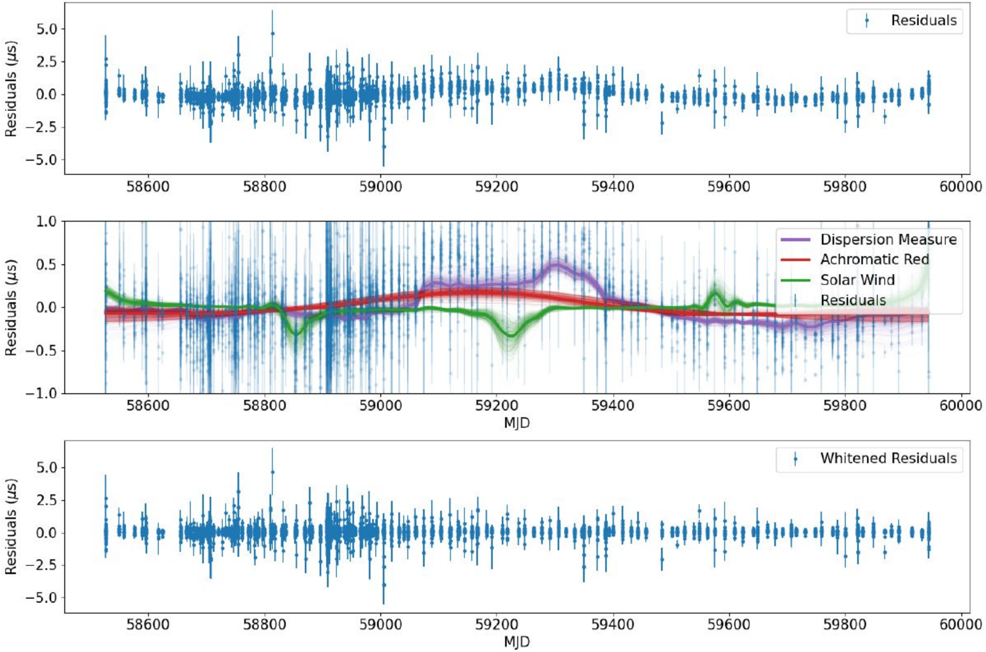
Hierarchical analysis

- Start simple, build to complexity
- Assumptions will lead to incorrect models

The timing residuals of J1909-3744 are on the right

I’ve identified three signals in the data which are coloured in the middle panel

With the data I have, I’m as correct as I can be - but it may not be the truth



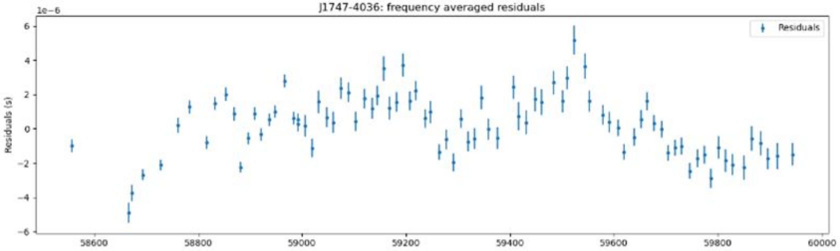
Hierarchical analysis

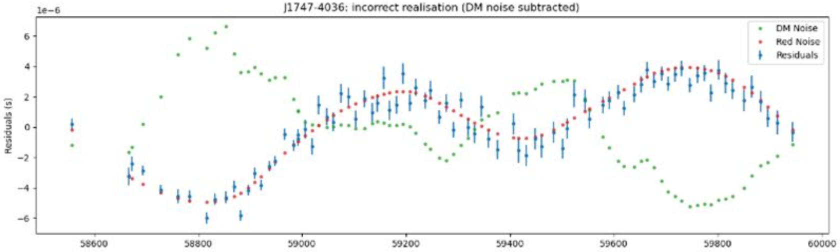
J1747-4036 is a prime example of why hierarchical analysis is important

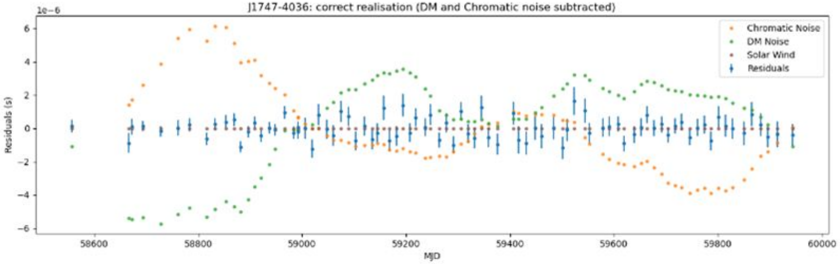
- Frequency averaged residuals with no noise reduction

- Noise processes reported in MPTA 2.5YR, chromatic noise removed
  - Strong achromatic noise!

- Advanced (hierarchical) noise modelling revealed achromatic noise was not favoured

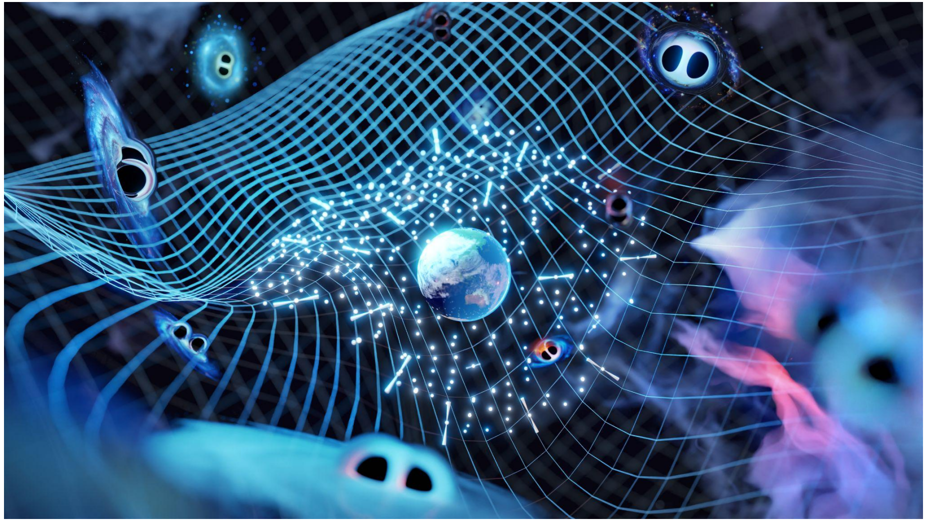

J1747-4036: frequency averaged residuals


J1747-4036: incorrect realisation (DM noise subtracted)


J1747-4036: correct realisation (DM and Chromatic noise subtracted)

Where do gravitational waves come into all of this?

- So far we've thought about modelling signals in a single pulsar
- We (me) care about modelling an entire ensemble of many pulsars
- We want to do this to find the signal of gravitational waves
- How does that work?



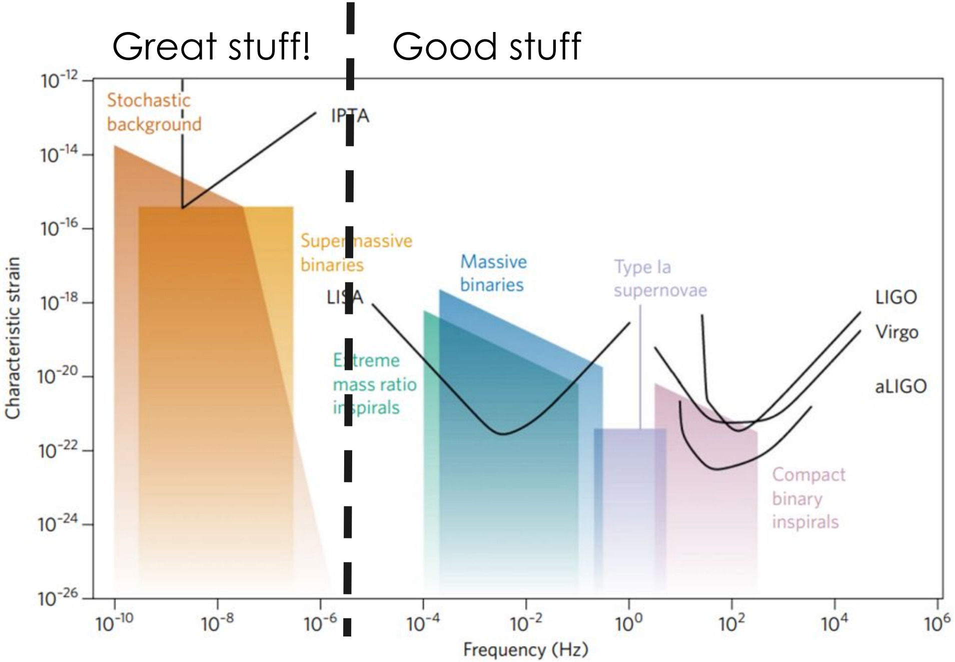
Credit: Carl Knox

Credit: Andrea N. Lommen (2017)

Lightning tour of GW signals in PTAs

What do we think we're looking for:

- Signals from supermassive black hole binaries
- Incoherent superposition of all these signals in the observable universe
- We call this a stochastic gravitational wave background (SGWB)



Lightning tour of GW signals in PTAs

This exists in PTA data in ***two*** distinct and important ways:

- Statistically identical noise
- Spatially correlated signals

Credit: Shannon & Cordes (2010)

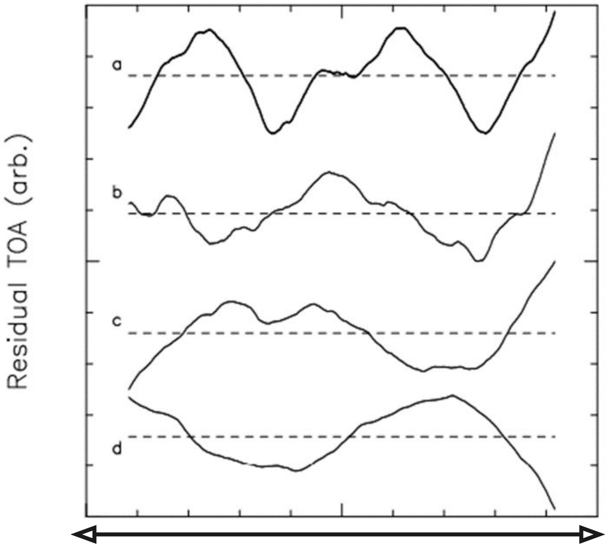
Lightning tour of GW signals in PTAs

This exists in PTA data in ***two*** distinct and important ways:

- **Statistically identical noise**
- Spatially correlated signals

STATISTICALLY IDENTICAL

INDEPENDENT REALISATIONS



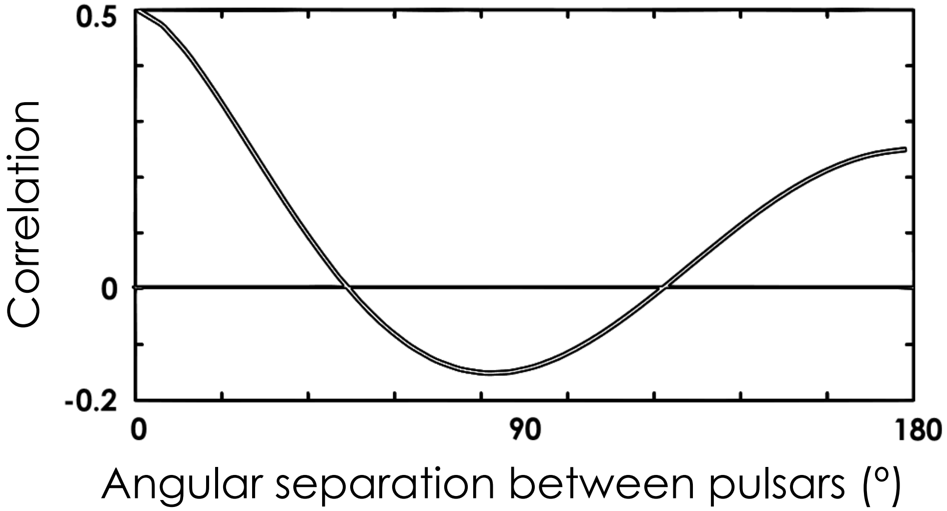
TIME

Lightning tour of GW signals in PTAs

This exists in PTA data in ***two*** distinct and important ways:

- Statistically identical noise
- **Spatially correlated signals**

“Smoking gun” of a SGWB

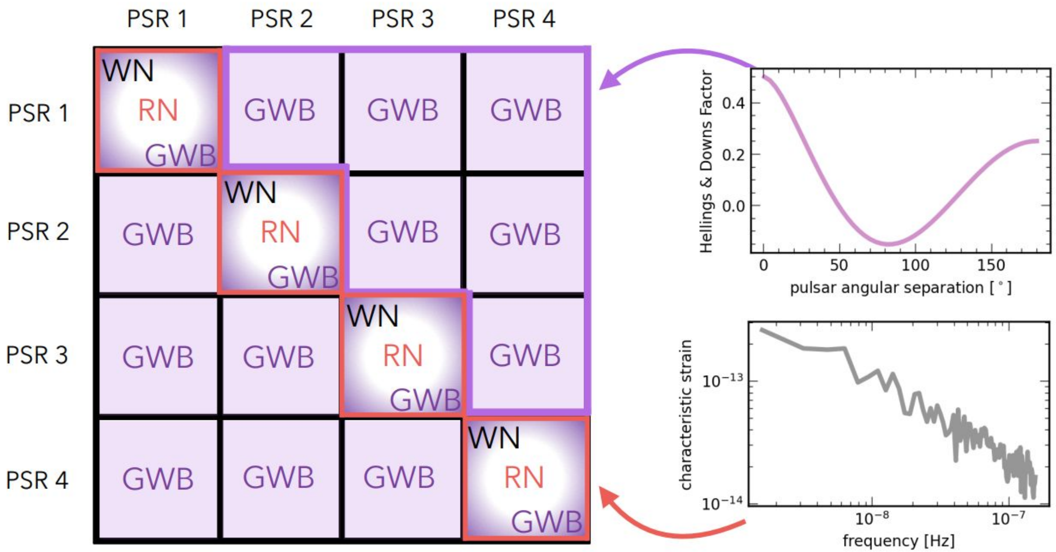


Credit: Taylor et al., 2022

Lightning tour of GW signals in PTAs

Does this change how we search for signals?

- Yes - kind of
  - A representation of a PTA covariance matrix is on the right
- Single pulsar noise models (what we were looking at before) make up the cells on the diagonal
- The statistically identical signal is the ***GWB*** in the diagonal
- The spatially correlated signal is all the other ***GWB***'s



	PSR 1	PSR 2	PSR 3	PSR 4
PSR 1	WN RN GWB	GWB	GWB	GWB
PSR 2	GWB	WN RN GWB	GWB	GWB
PSR 3	GWB	GWB	WN RN GWB	GWB
PSR 4	GWB	GWB	GWB	WN RN GWB

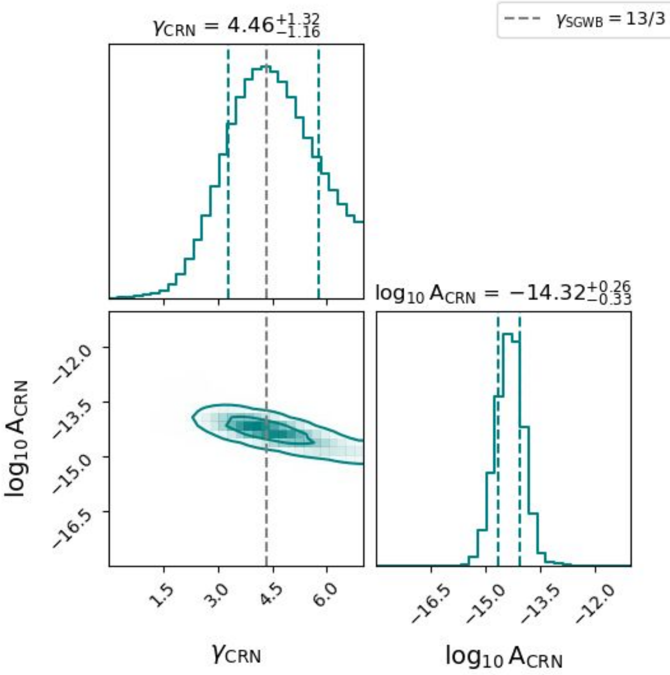
Lightning tour of GW signals in PTAs

What does this look like in PTA data?

- **Statistically identical noise**
- Spatially correlated signals

A statistically identical signal in the MPTA data

Maybe it's GWs, maybe not



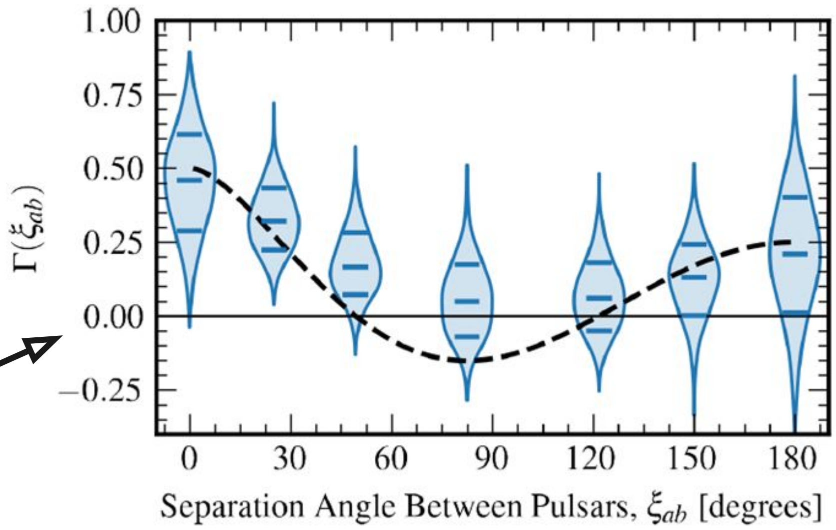
Lightning tour of GW signals in PTAs

What does this look like in PTA data?

- Statistically identical noise
- **Spatially correlated signals**

The results of the last NANOGrav search for a correlated signal

Best published result so far, but not definitive just yet

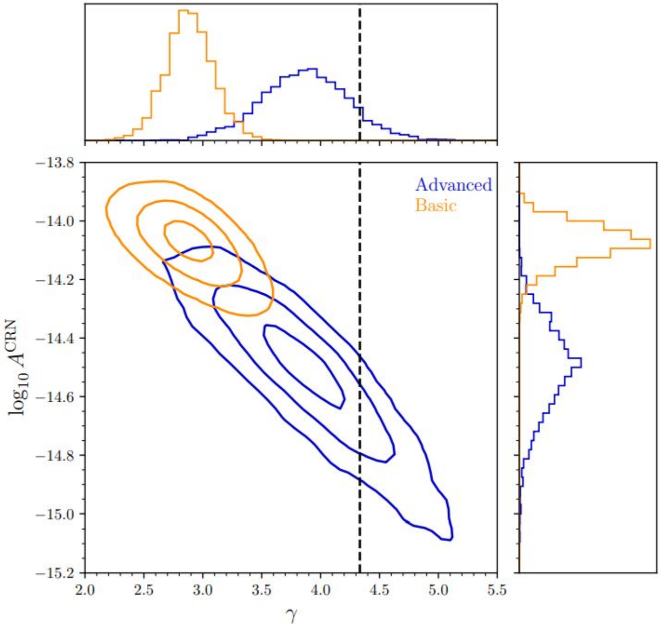


Credit: Agazie et al., 2023

Why is this actually important?

We know at least this so far:

- Bayesian inference is robust compared to frequentist
- Through hierarchical modelling we can get as close to the real signals as possible
- Not doing this appropriately creates **very** different inferred signals in the data
- This flows through to a PTA search for GWs

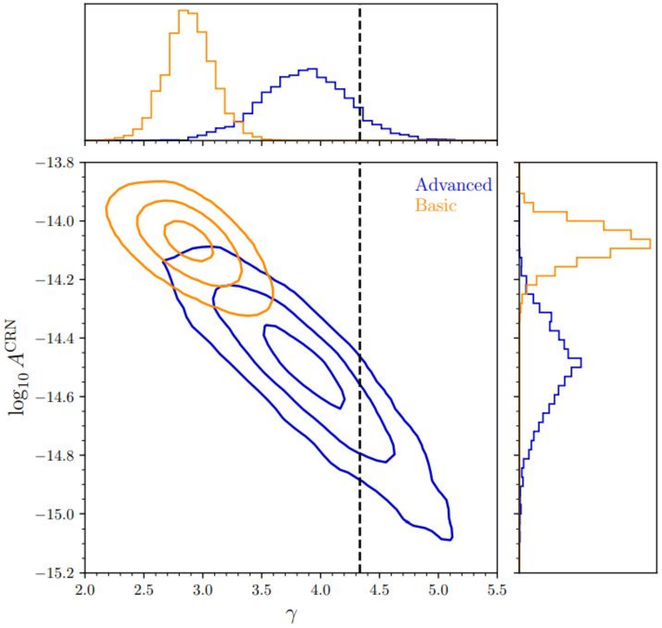


Credit: Reardon et al., 2023

Why is this actually important?

- A definitive detection is right around the corner

- It’s better to be safe than sorry when this happens, and hierarchical noise modelling is safer by far



Credit: Reardon et al., 2023

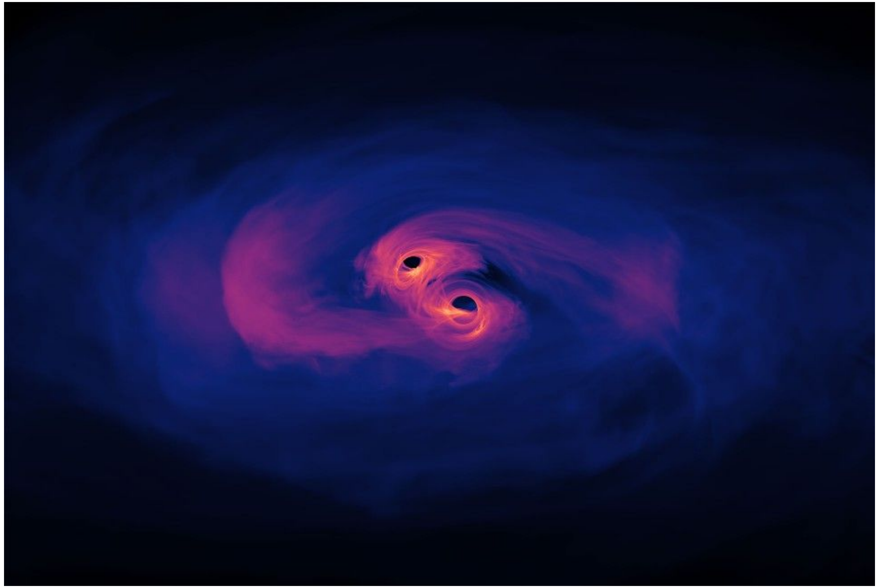
Summary

- Bayesian inference is your friend!
  - There are tools that have been developed to help you do this!

- Hierarchical modelling is a safe way forward

- The name of the game is finding GWs, modelling the noise correctly is important to do this (we think)

- There's so much we don't know and we need your help!



Credit: NASA