UNIVERSITY OF CAMBRIDGE

Lecture 13
Nested Sampling and MCMC in Astronomy

Lecturer: **Dr Dominic Anstey** (da401)

Overview

The importance in astronomy and cosmology

MCMC and Nested Sampling

Use of Bayesian data products

An example case – 21cm cosmology

Summary

Importance in Astronomy

- Applicable to science in general

- Measuring individual, unchanging properties

- Often making individual measurements with individual instruments

UNIVERSITY OF CAMBRIDGE

Bayes Theorem

$$\mathrm{P}\left(\theta_{\mathcal{M}}|\mathcal{D},\mathcal{M}\right) = \frac{\mathrm{P}\left(\mathcal{D}|\theta_{\mathcal{M}},\mathcal{M}\right)\mathrm{P}\left(\theta_{\mathcal{M}}|\mathcal{M}\right)}{\mathrm{P}\left(\mathcal{D}|\mathcal{M}\right)}$$

Likelihood

Posterior

Prior

$$\mathcal{P} = \frac{\mathcal{L}\Pi}{\mathcal{Z}}$$

Evidence

UNIVERSITY OF CAMBRIDGE

Bayesian Data Products

$$\mathcal{L}\Pi = \mathcal{P}\mathcal{Z}$$

Inputs:

- Model
- Prior
- Likelihood

Products:

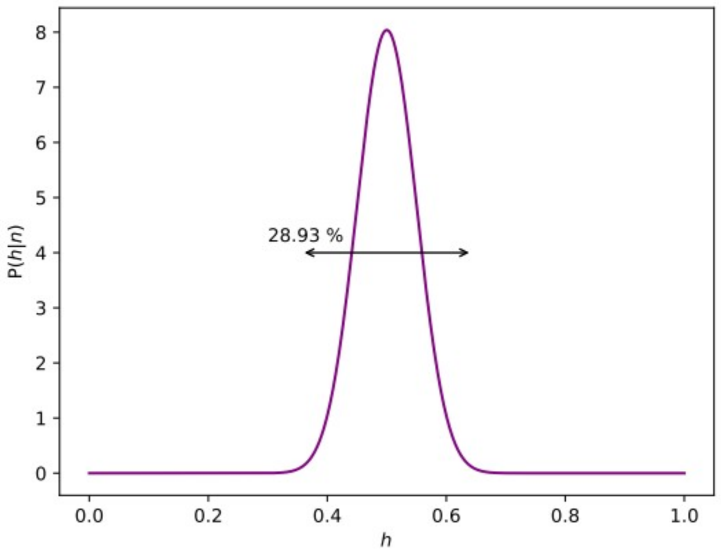
- Posterior
- Evidence

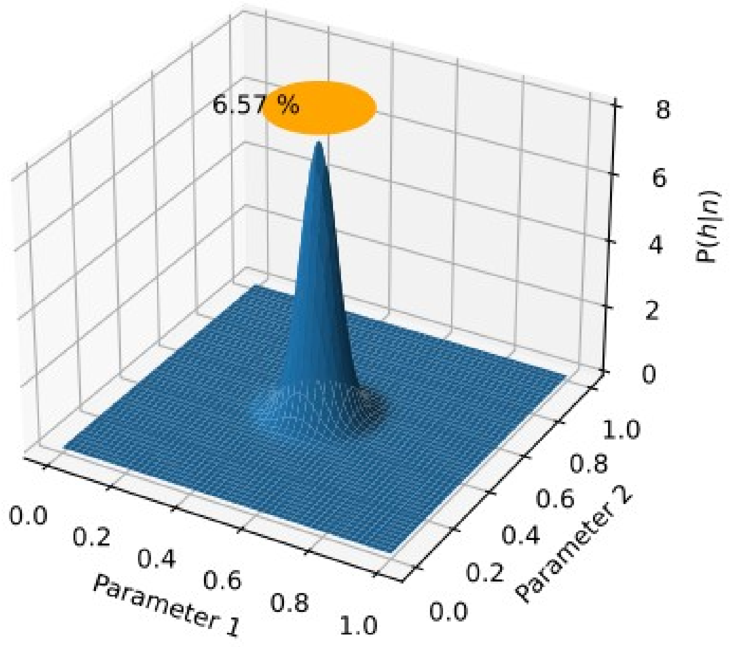
UNIVERSITY OF CAMBRIDGE

Posteriors and Parameter Estimation

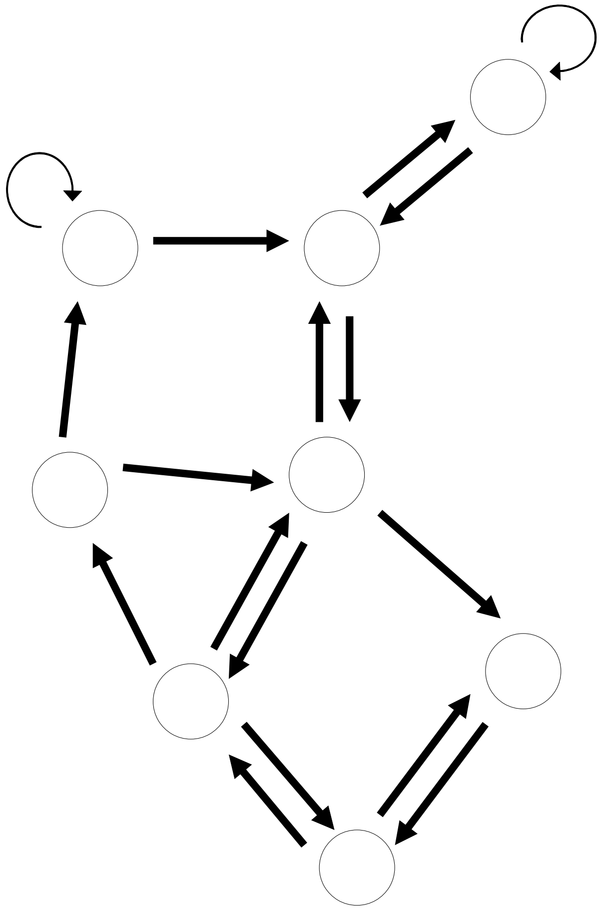
Find the distribution of most probable parameter values given the data

Requires a mechanism for efficiently exploring the parameter space





MCMC



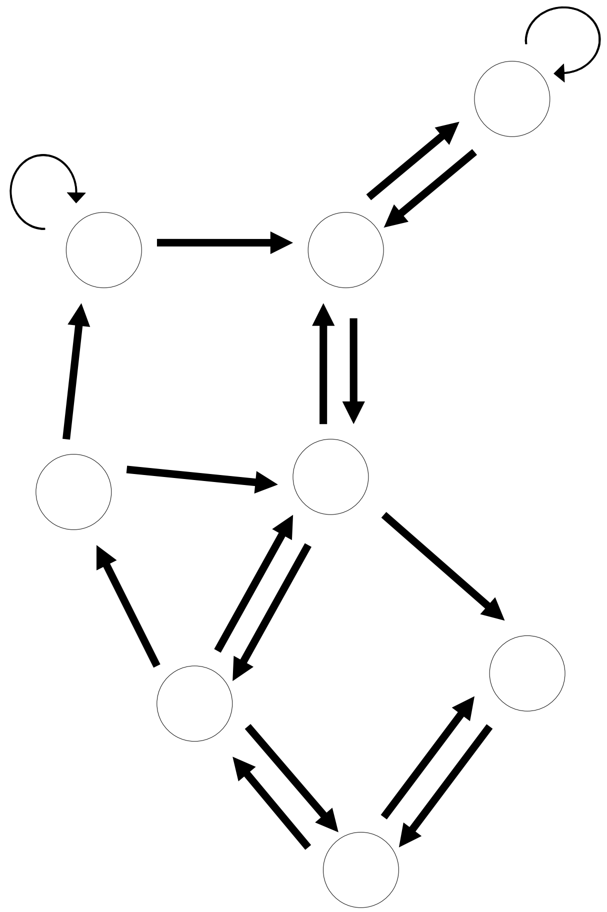
Markov Chain:
Network of states with defined probabilities of moving between those states.

Two key features:
No part of the network is isolated from any other

Memoryless – the probabilities of moving to the next state depend only on the current state and not the steps taken to reach that state

UNIVERSITY OF CAMBRIDGE

MCMC



Sampling a posterior through Markov Chain Monte Carlo Methods:

Define a Markov Chain with an underlying probability distribution matching the posterior probability distribution you want to evaluate

Perform one or more random walks (a Monte Carlo process) along the chain

After a burn-in period, the distributions of the walkers will approximate the underlying probability distribution

UNIVERSITY OF CAMBRIDGE

Evidence

$$\mathcal{Z} = \int \mathrm{P}\left(\mathcal{D}|\theta_{\mathcal{M}}, \mathcal{M}\right) \mathrm{P}\left(\theta_{\mathcal{M}}|\mathcal{M}\right) \mathrm{d}\theta_{\mathcal{M}} = \int \mathcal{L}\Pi \mathrm{d}\theta_{\mathcal{M}}$$

$$\mathrm{P}\left(\mathcal{M}|\mathcal{D}\right) = \frac{\mathrm{P}\left(\mathcal{D}|\mathcal{M}\right)\mathrm{P}\left(\mathcal{M}\right)}{\mathrm{P}\left(\mathcal{D}\right)} = \mathcal{Z}\frac{\mathrm{P}\left(\mathcal{M}\right)}{\mathrm{P}\left(\mathcal{D}\right)}$$

The Bayesian Evidence allows the relative probabilities of different models to be compared

$$\frac{\mathrm{P}\left(\mathcal{M}_1|\mathcal{D}\right)}{\mathrm{P}\left(\mathcal{M}_2|\mathcal{D}\right)} = \frac{\mathcal{Z}_1\mathrm{P}\left(\mathcal{M}_1\right)}{\mathcal{Z}_2\mathrm{P}\left(\mathcal{M}_2\right)}$$

UNIVERSITY OF CAMBRIDGE

Occam Penalty

$$\mathcal{Z} = \int \mathrm{P}\left(\mathcal{D}|\theta_{\mathcal{M}}, \mathcal{M}\right) \mathrm{P}\left(\theta_{\mathcal{M}}|\mathcal{M}\right) \mathrm{d}\theta_{\mathcal{M}} = \int \mathcal{L}\Pi \mathrm{d}\theta_{\mathcal{M}}$$

$$\mathcal{P} = \frac{\mathcal{L}\Pi}{\mathcal{Z}}$$

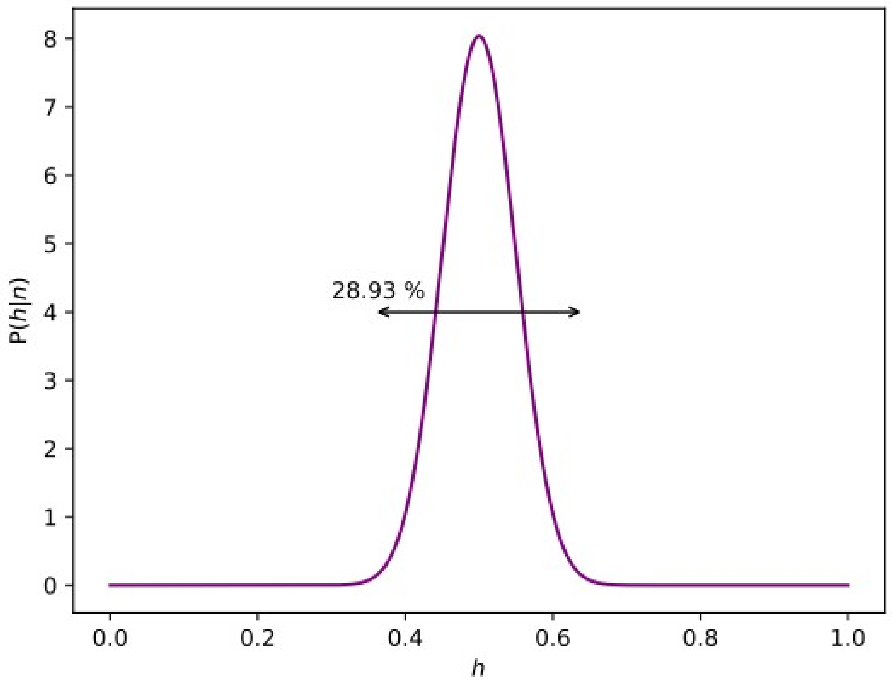
$$\log(\mathcal{Z}) = \int \mathcal{P} \log(\mathcal{L}) \mathrm{d}\theta_{\mathcal{M}} - \int \mathcal{P} \log\left(\frac{\mathcal{P}}{\Pi}\right) \mathrm{d}\theta_{\mathcal{M}}$$

Expectation of log likelihood over posterior distribution – quantifies the quality of the model fit

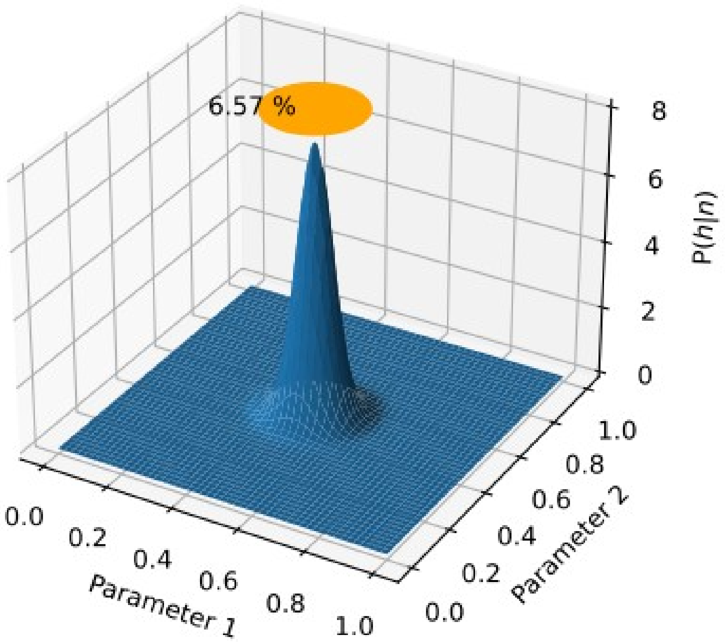
Ratio of posterior to prior – increases as finite prior probability is spread over a wider area

For two models that produce equally good fits, the one with more parameters will give a lower evidence: Bayesian evidence comparison naturally implements Occam’s Razor

The Curse of Dimensionality



1000 likelihood evaluations



1,000,000 likelihood evaluations

UNIVERSITY OF CAMBRIDGE

Nested Sampling

Aims to calculate the Bayesian Evidence:

$$\mathcal{Z} = \int \mathcal{L}\Pi d\theta$$

Need to reduce dimensions

Define a new quantity – the fraction of the prior volume contained with in a contour of constant likelihood

$$X\left(\mathcal{L}_{\text{contour}}\right) = \int_{\mathcal{L}>\mathcal{L}_{\text{contour}}} \Pi\left(\theta\right) \mathrm{d}\theta$$

The prior is the gradient of this value with respect to the parameters

$$\Pi\left(\theta\right) = \frac{\mathrm{d}X}{\mathrm{d}\theta}$$

Substitute this definition of the prior into the evidence calculation. Reduces the evidence calculation to a 1D integral

$$\mathcal{Z} = \int_0^1 \mathcal{L}\left(X\right) \mathrm{d}X \approx \sum_i^{N_{\text{samples}}} \mathcal{L}_i w_i$$

UNIVERSITY OF CAMBRIDGE

Nested Sampling

Draw a number of samples from the prior and calculate the likelihood of all of them

Identify the lowest likelihood point

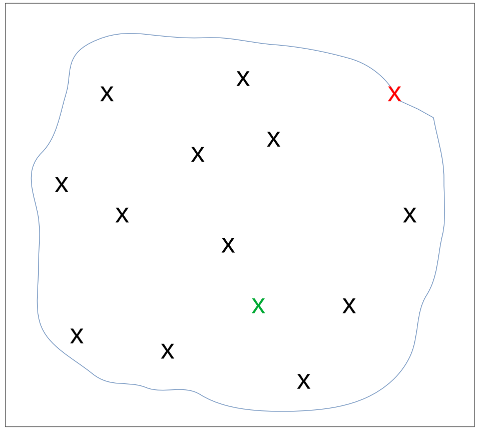
Draw a new point from the prior distribution, subject to the constraint that its likelihood is higher than the identified lowest point

Replace the old lowest likelihood point with the new one

Repeat until the samples converge

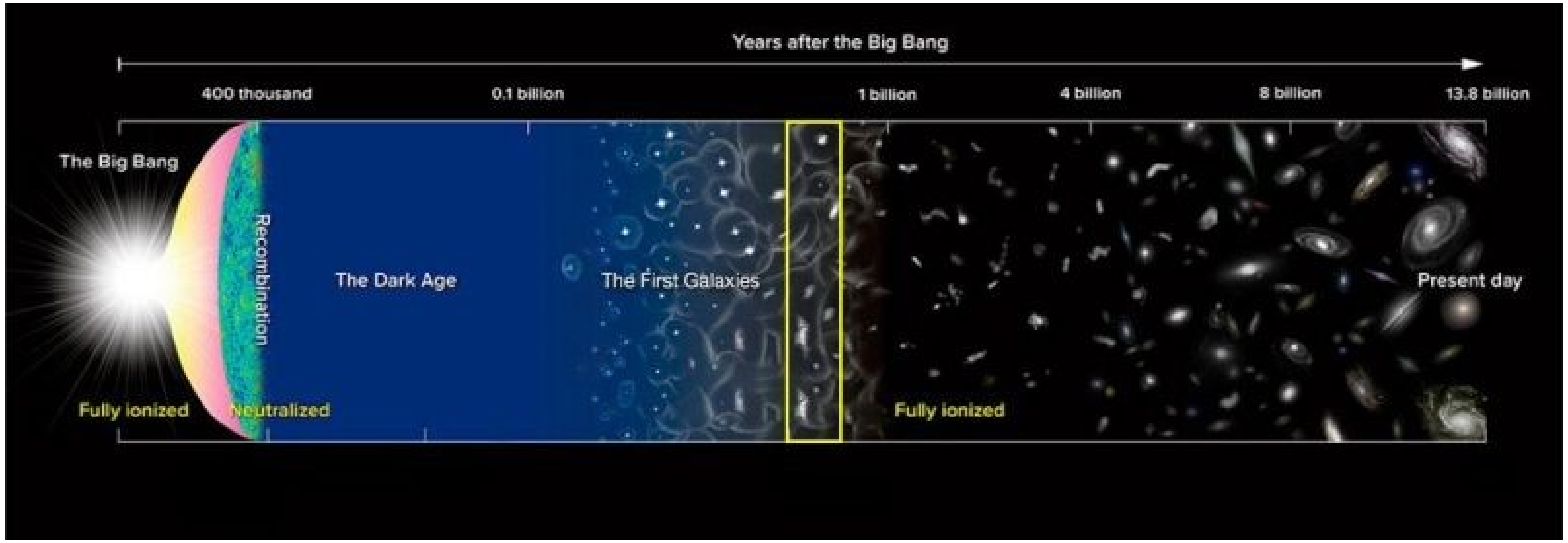
The discarded dead points are ordered in likelihood and uniformly spaced in log(X), as required

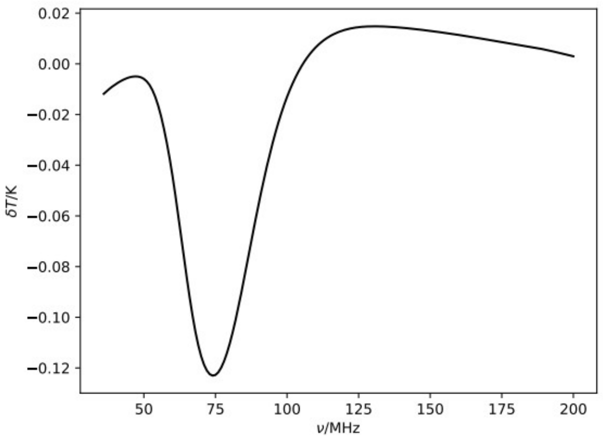
Final live point distribution gives the posterior as a side product



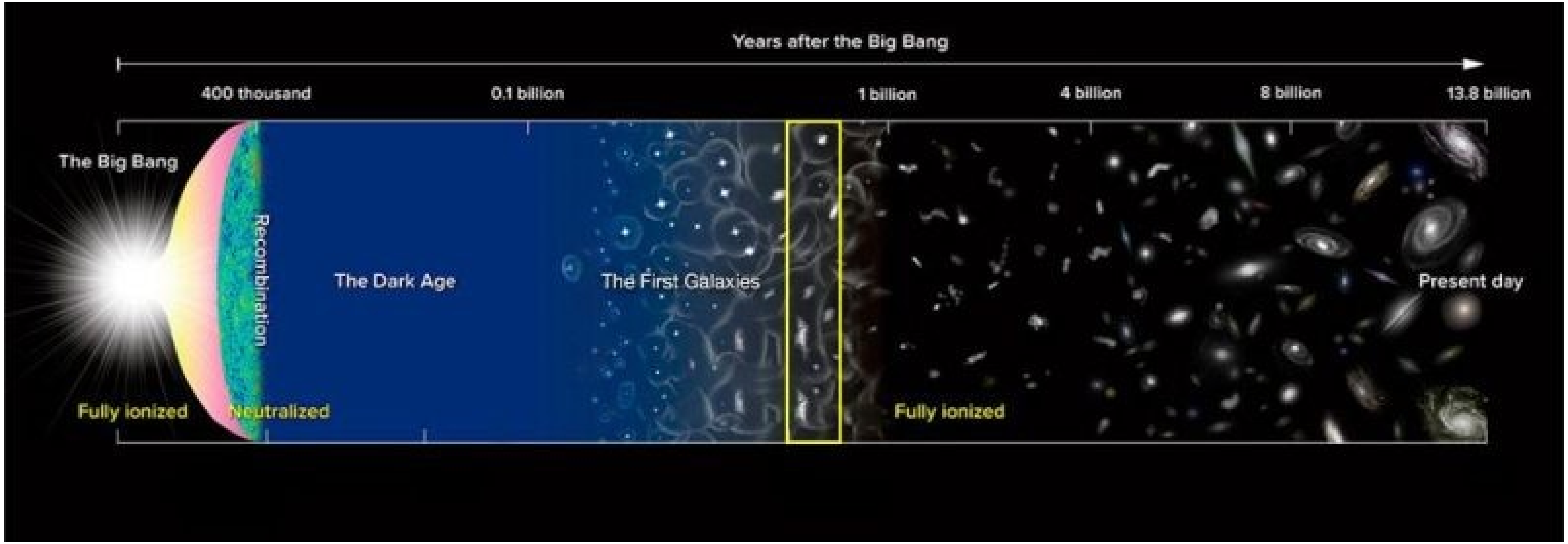
$$X_i \approx e^{\frac{-i}{n_{\rm live}}}$$

Use Case – 21cm Cosmology



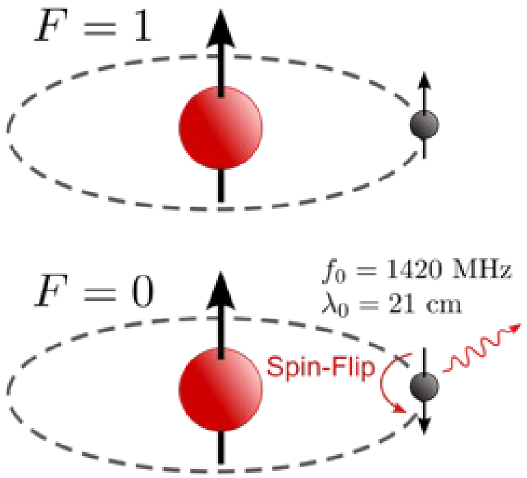


Dark Ages and Cosmic Dawn



UNIVERSITY OF CAMBRIDGE

Spin Temperature



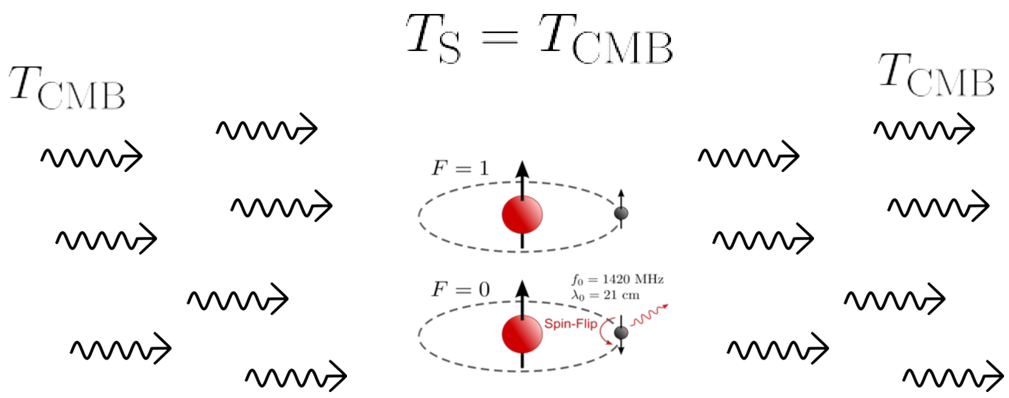
https://en.wikipedia.org/wiki/Hydrogen_line

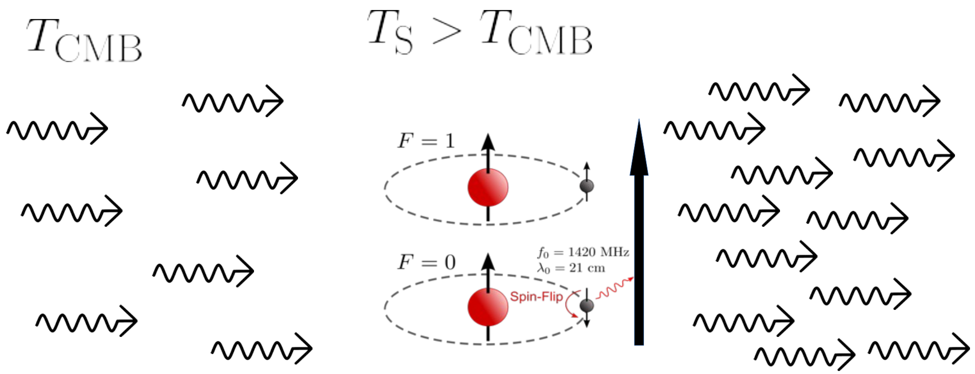
$$\frac{n_1}{n_0} = \frac{g_1}{g_0} e^{-\frac{T_*}{T_S}}$$

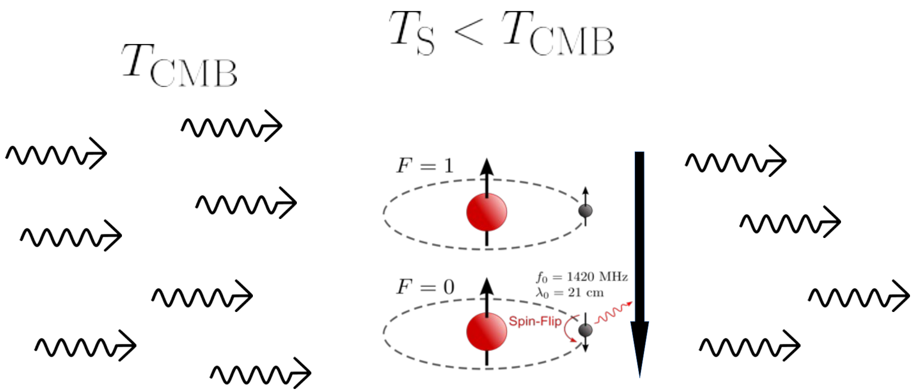
$$T_* = \frac{h f_o}{k_B}$$

UNIVERSITY OF CAMBRIDGE

21cm Signal







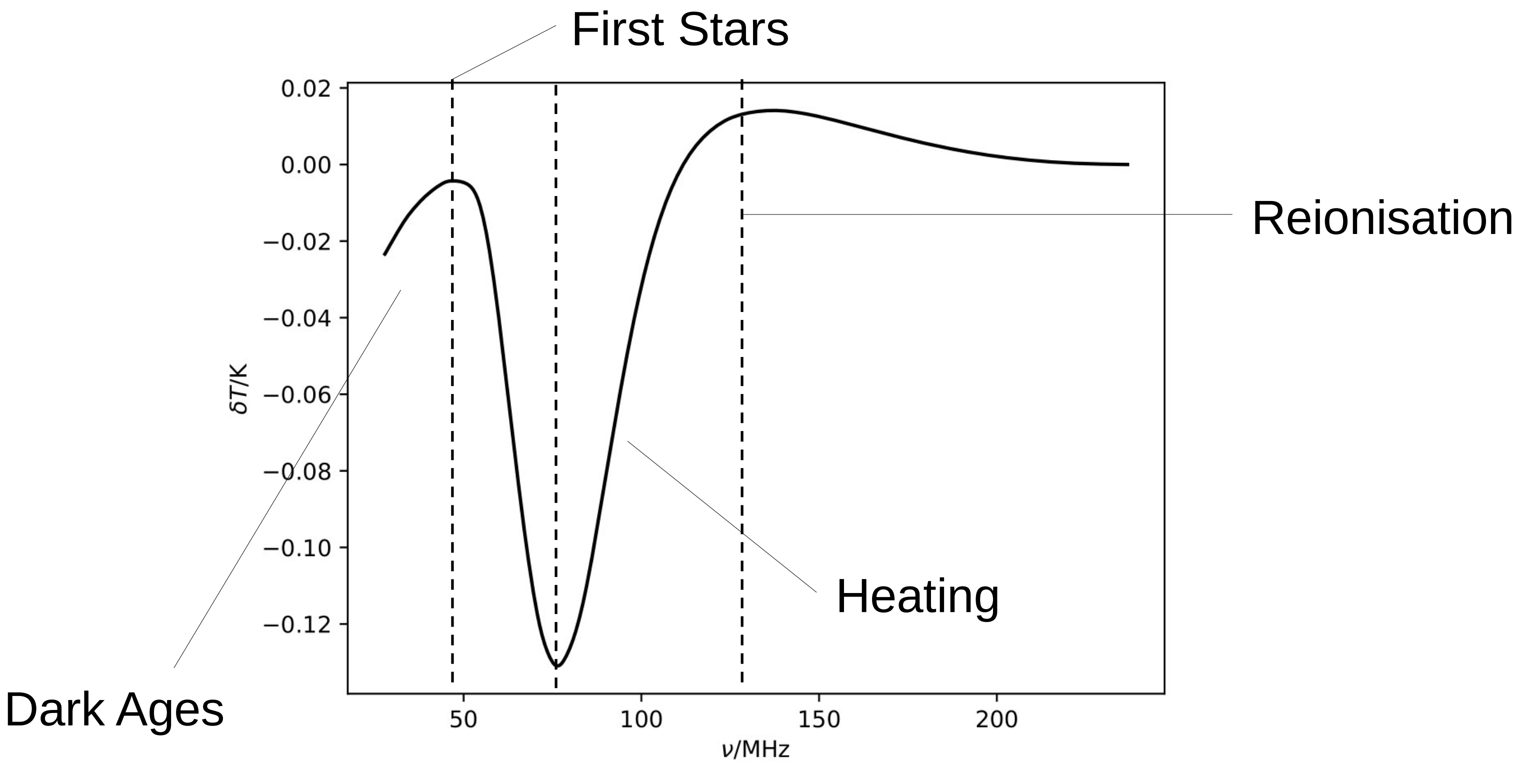
UNIVERSITY OF CAMBRIDGE

21cm Signal

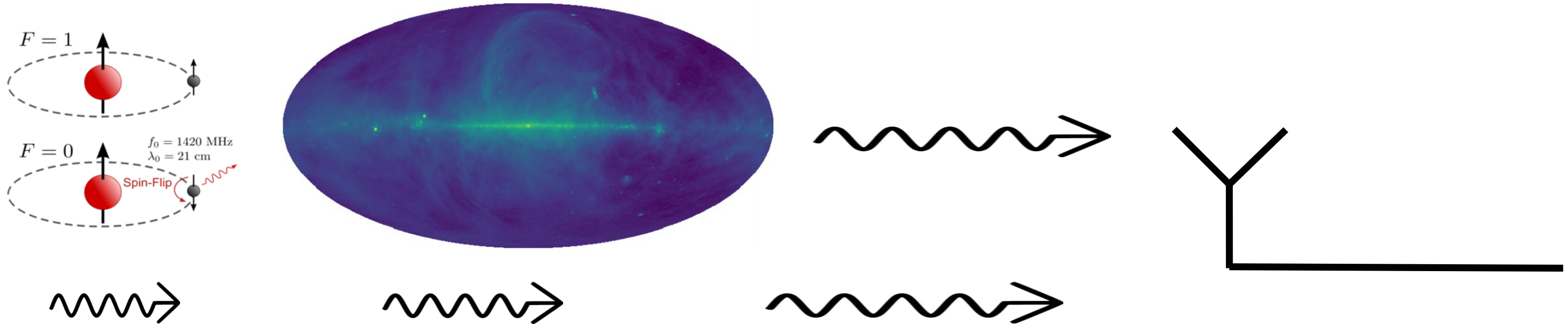
$$\delta T_{\rm b} \approx 27\,(1-\bar{x}_i)\left(\frac{T_{\rm S}-T_{\rm CMB}}{T_{\rm S}}\right)\left(\frac{1+z}{10}\right)^{\frac{1}{2}}\,\mathrm{mK}$$

Effects that alter Spin Temperature:

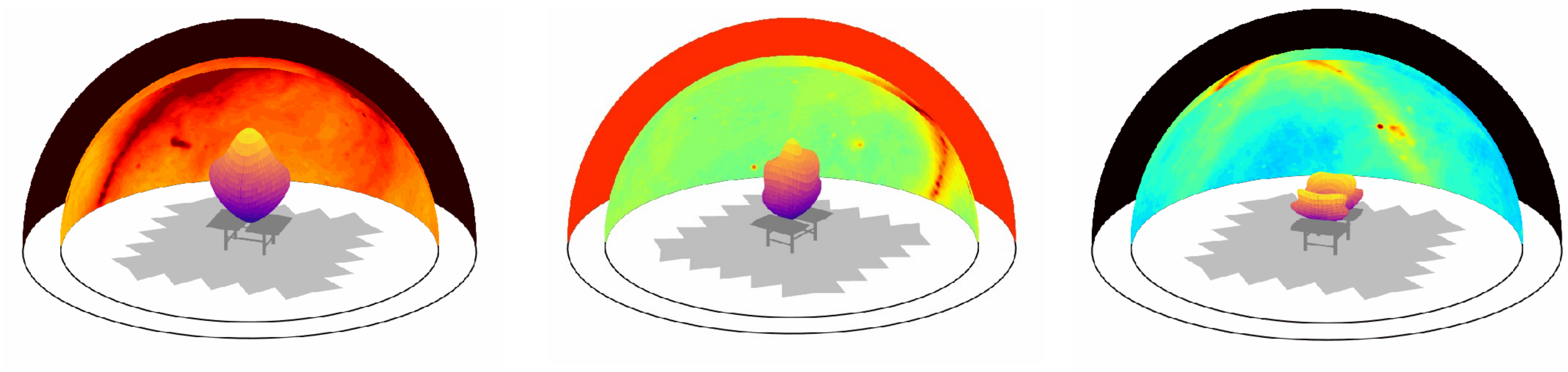
- Collisions
- Lyman-alpha photons
- Ionising UV photons



Foregrounds and Systematics



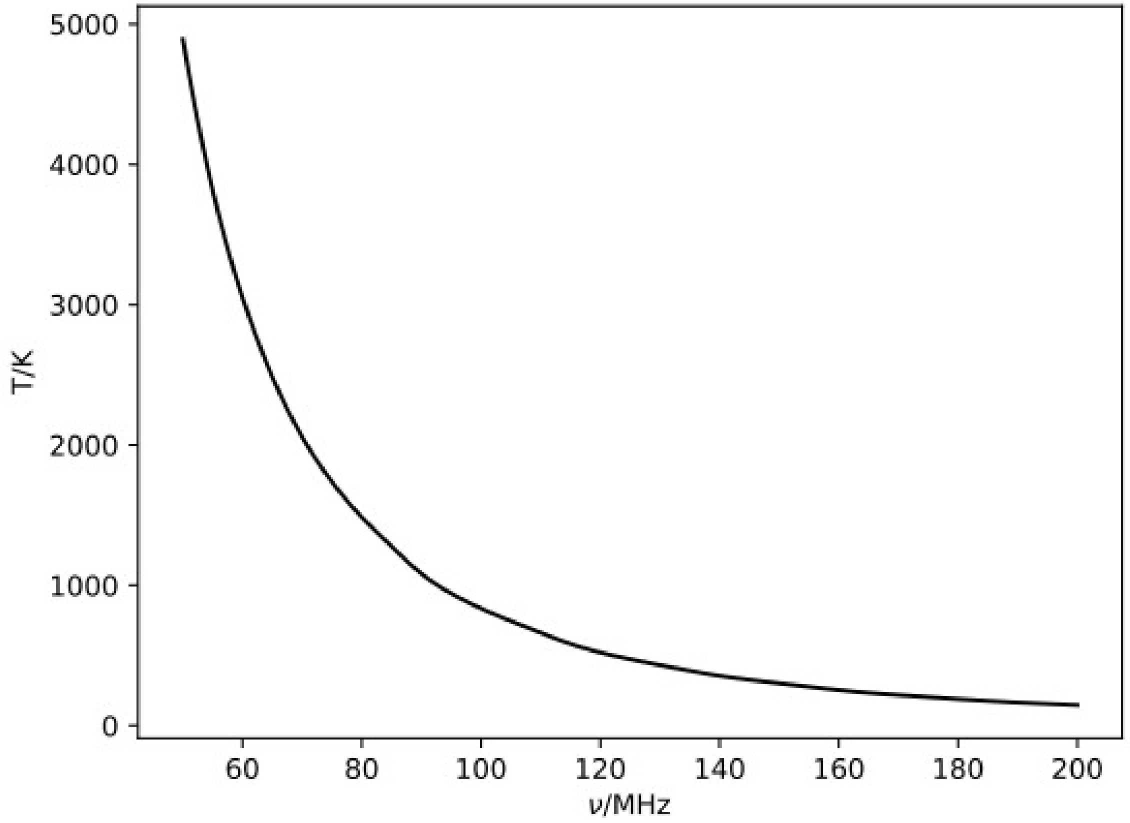
Galactic and extragalactic radio emissions



UNIVERSITY OF CAMBRIDGE

21cm Data

$$\mathcal{D} = \frac{1}{4\pi}\int D\left(\Omega,\nu\right)\left[T_{\mathrm{base}}\left(\Omega\right) - T_{\mathrm{CMB}}\right]\left(\frac{\nu}{\nu_{\mathrm{base}}}\right)^{-2.55}\mathrm{d}\Omega + T_{\mathrm{CMB}} + \widehat{\sigma_n}$$



UNIVERSITY OF CAMBRIDGE

Define a Model

Requirements for your model:

- Includes truth within parameter space
- Able to account for expected uncertainties
- Minimal degeneracies with other components

Analytical models

$$\mathcal{M} = a \sin\left(\frac{x}{b} - c\right)$$

- Fast to compute
- Well constrained

- Do not exist for all problems

Linear Models

$$\mathcal{M} = \sum_i a_i X_i$$

- Fast to compute
- Very general

- Can become very high dimensional
- Often not constrained

Forward Models

$$\mathcal{M} = F\left(\theta\right)$$

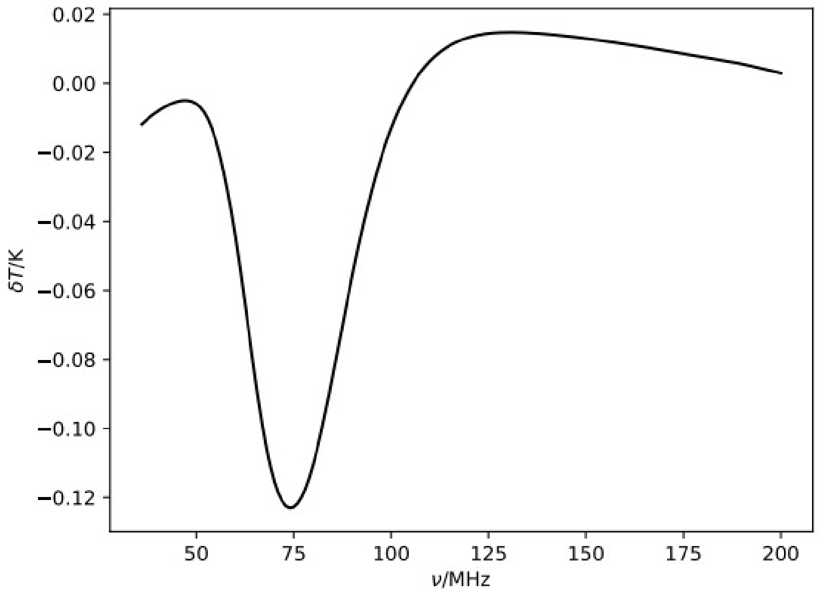
- Specific to problem
- Well constrained

- Slow to compute
    - Machine Learning

UNIVERSITY OF CAMBRIDGE

Define a Model

$$\mathcal{M} = T_{\mathrm{F}}\left(\nu, \theta_{\mathrm{F}}\right) + T_{\mathrm{S}}\left(\nu, \theta_{\mathrm{S}}\right)$$



$$T_{\mathrm{S}}\left(\nu, \theta_{\mathrm{S}}\right) = -Ae^{-\frac{1}{2}\left(\frac{\nu-\nu_0}{w}\right)^2}$$

Analytical Model

$$\theta_{\mathrm{S}} = \{A, \nu_0, w\}$$

UNIVERSITY OF CAMBRIDGE

Define a Model

$$\mathcal{M} = T_{\mathrm{F}}\left(\nu, \theta_{\mathrm{F}}\right) + T_{\mathrm{S}}\left(\nu, \theta_{\mathrm{S}}\right)$$

$$T_{\mathrm{F}}\left(\nu, \theta_{\mathrm{F}}\right) = \frac{1}{4\pi}\int D\left(\Omega, \nu\right)\left[T_{\mathrm{base}}\left(\Omega\right) - T_{\mathrm{CMB}}\right]\left(\frac{\nu}{\nu_{\mathrm{base}}}\right)^{-\beta} \mathrm{d}\Omega + T_{\mathrm{CMB}}$$

Forward Model

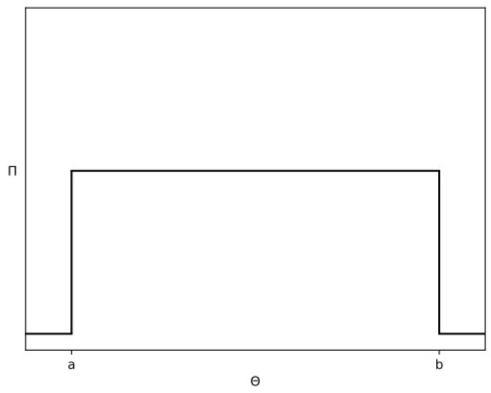
$$\theta_{\mathrm{F}} = \{\beta\}$$

Define a Prior

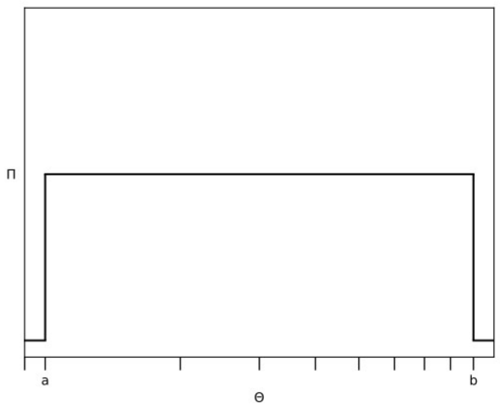
Requirements for your prior:

- Includes the truth within the parameter space
- Accurately reflects existing knowledge

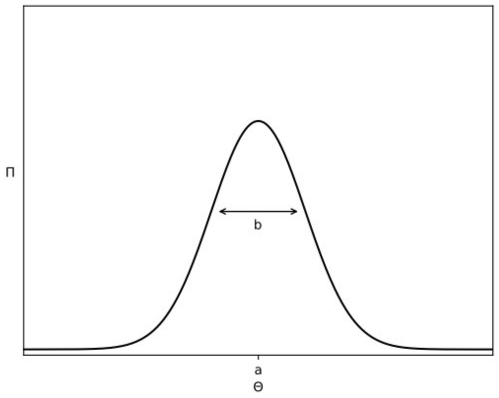
Uniform priors
- Known limits



Log-Uniform priors
- Known limits across orders of magnitude



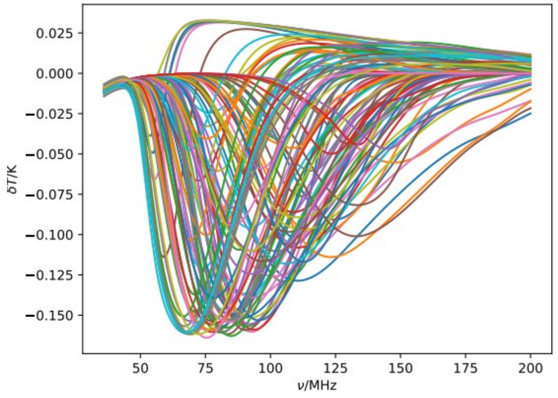
Gaussian priors
- Known expectation and uncertainty



Any other probability distribution
- Correlations
- Conditions
- Normalising flows
- etc.

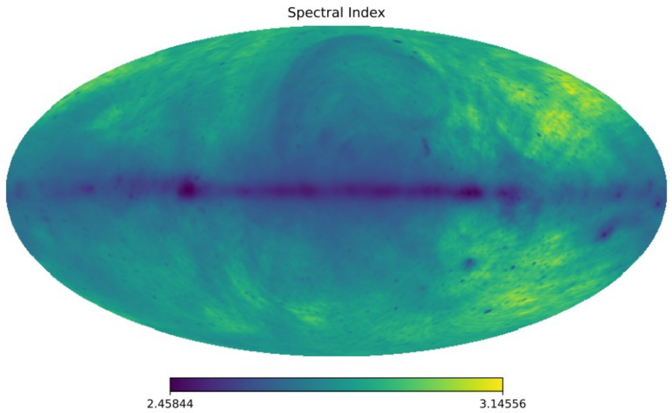
Define a Prior

$$\theta_{\mathrm{S}} = \{A, \nu_0, w\}$$



$$A = \{0, -0.17\}$$
$$\nu_0 = \{50, 150\}$$
$$w = \{10, 20\}$$

$$\theta_{\mathrm{F}} = \{\beta\}$$



$$\beta = \{2.45, 3.15\}$$

Define a Likelihood

Probability of observing data, given a model

Usually describes noise structure

$$\log \mathcal{L} = \sum_i -\frac{1}{2}\log\left(2\pi\sigma_{\mathrm{n}}^2\right) - \frac{1}{2}\left(\frac{\mathcal{D}_i - \mathcal{M}_i}{\sigma_{\mathrm{n}}}\right)^2$$

$$\log \mathcal{L} = -\frac{1}{2}\log\left((2\pi)^n|\underline{C}|\right) - \frac{1}{2}\left(\underline{\mathcal{D}} - \underline{\mathcal{M}}\right)^{\mathrm{T}}\underline{C}^{-1}\left(\underline{\mathcal{D}} - \underline{\mathcal{M}}\right)$$

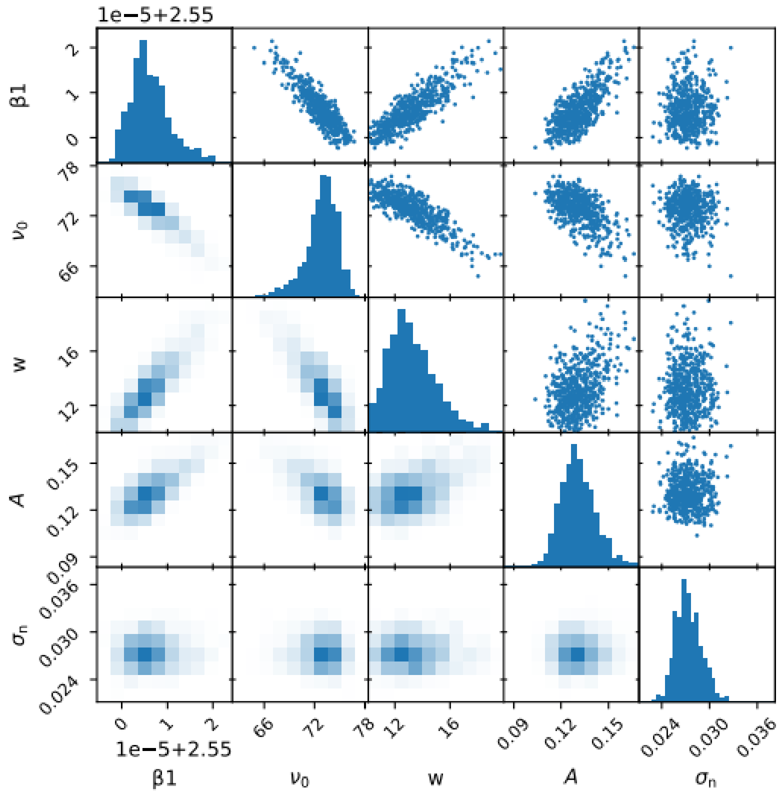
$$\sigma_n = \log\{10^{-4}, 10^{-1}\}$$

See lecture 16 for what to do if you cannot define a likelihood

UNIVERSITY OF CAMBRIDGE

Data Products

Nested Sampling using
PolyChord (Handley et al. 2015)



$$\log \mathcal{Z} = 309.1 \pm 0.4$$

UNIVERSITY OF CAMBRIDGE

Marginalising Nuisance Parameters

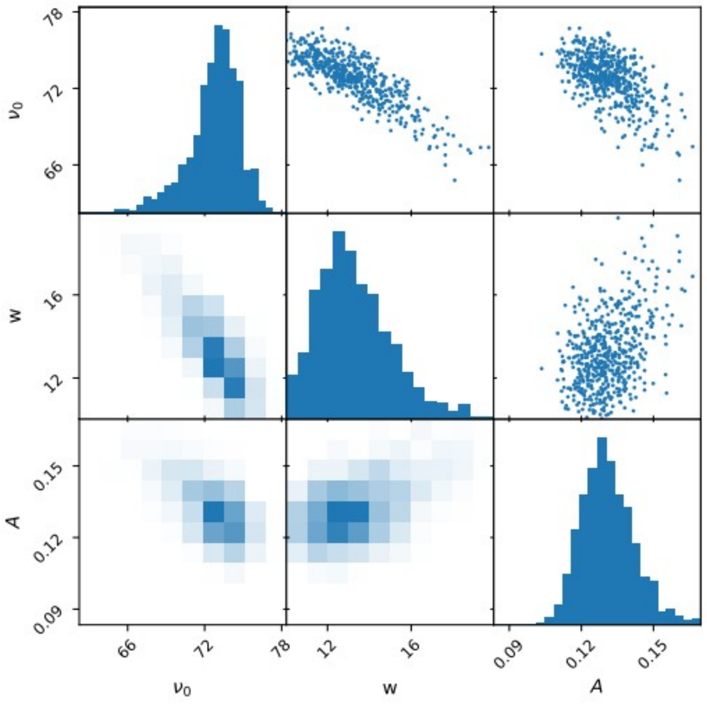
$$\mathrm{P}\left(\theta_{\mathcal{M}}|\mathcal{D},\mathcal{M}\right) = \mathrm{P}\left(\theta_{\mathcal{S}},\theta_{\mathcal{N}}|\mathcal{D},\mathcal{M}\right)$$

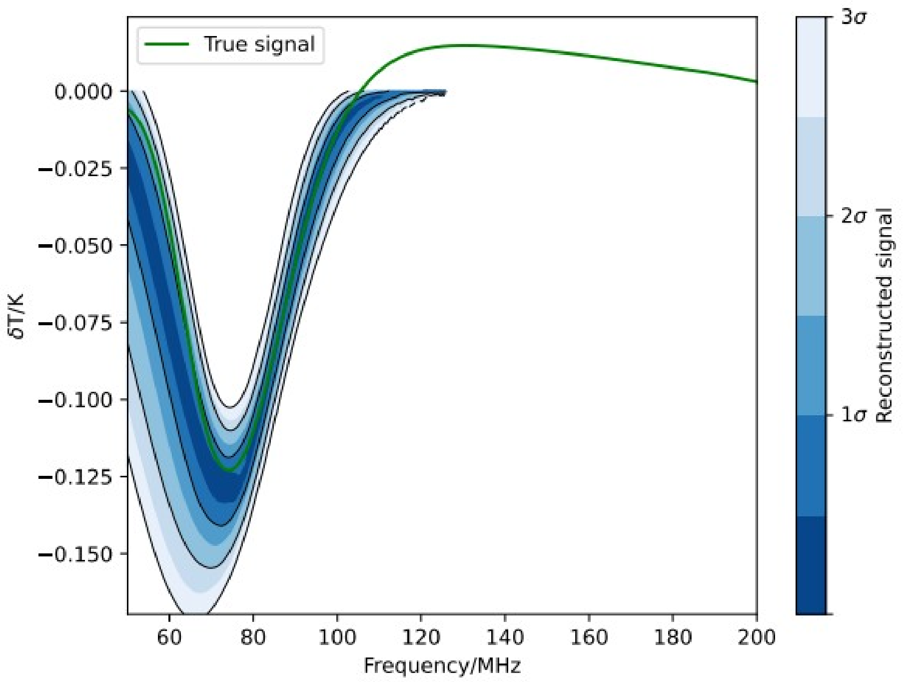
$$\mathrm{P}\left(A\right) = \sum_i \mathrm{P}\left(A, B_i\right)$$

$$\mathrm{P}\left(\theta_{\mathcal{S}}|\mathcal{D},\mathcal{M}\right) = \int \mathrm{P}\left(\theta_{\mathcal{S}},\theta_{\mathcal{N}}|\mathcal{D},\mathcal{M}\right)\mathrm{d}\theta_{\mathcal{N}}$$

Marginalisation removes posterior dependence on uninteresting parameters, accounting for all values they can take and how probable those values are

Marginalising Nuisance Parameters





UNIVERSITY OF CAMBRIDGE

Model Comparison

$$\frac{\mathrm{P}\left(\mathcal{M}_1|\mathcal{D}\right)}{\mathrm{P}\left(\mathcal{M}_2|\mathcal{D}\right)} = \frac{\mathcal{Z}_1\mathrm{P}\left(\mathcal{M}_1\right)}{\mathcal{Z}_2\mathrm{P}\left(\mathcal{M}_2\right)}$$

Preferential betting odds of model 1 in comparison to model 2 are

$$\frac{\mathcal{Z}_1}{\mathcal{Z}_2} : 1 \qquad\qquad e^{\log\mathcal{Z}_1 - \log\mathcal{Z}_2} : 1$$

Provided the same data set is used for both

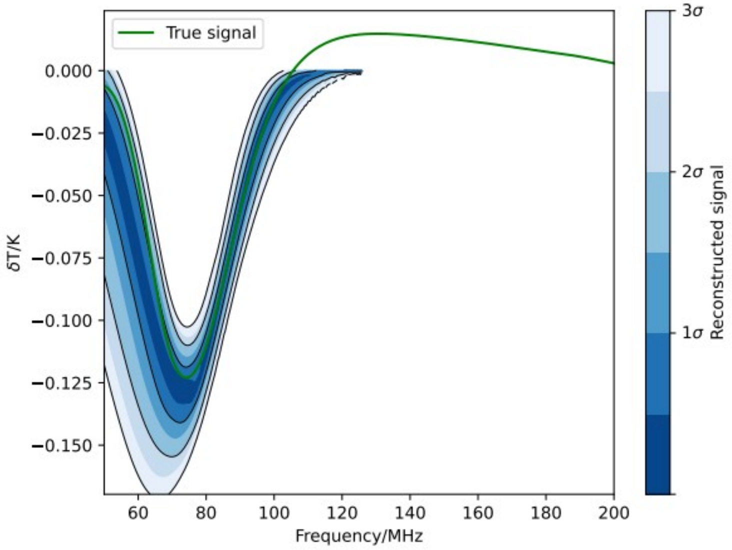
UNIVERSITY OF CAMBRIDGE

Model Comparison

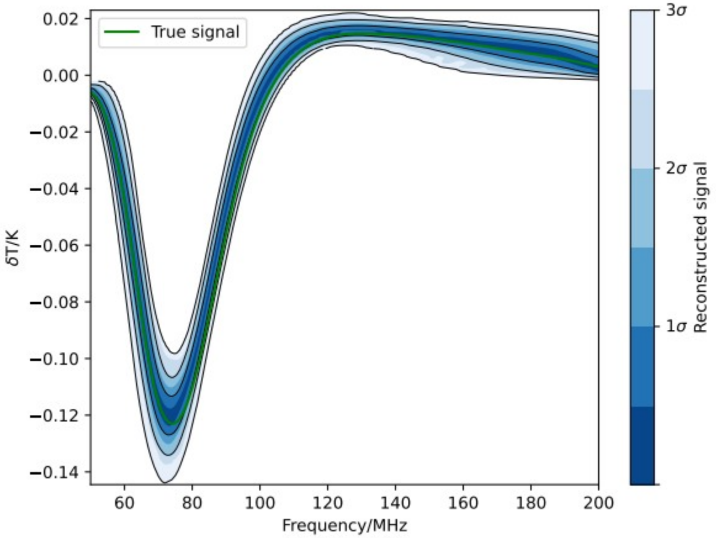
Consider, instead of analytical Gaussian signal model, we use a neural net train to simulate realistic signals from physical properties (lecture 15 will cover this more)

Fit both models and compare evidence

$$\log \mathcal{Z}_{\text{Gaussian Signal}} = 309.1 \pm 0.4$$



$$\log \mathcal{Z}_{\text{NN signal}} = 320.5 \pm 0.3$$



$$\text{NN} : \text{Gaussian} = e^{11.4} : 1 \approx 90000 : 1$$

UNIVERSITY OF CAMBRIDGE

Model Confidence

How confident can you be that you have detected the signal you are looking for?

$$\mathrm{P}\left(\mathrm{signal}\right) = \frac{\mathrm{P}\left(\mathcal{M}_{\mathrm{signal}}|\mathcal{D}\right)}{\mathrm{P}\left(\mathcal{M}_{\mathrm{no\,signal}}|\mathcal{D}\right)} = \frac{\mathcal{Z}_{\mathrm{signal}}}{\mathcal{Z}_{\mathrm{no\,signal}}}$$

Fit the same data set with another model in which everything is identical except that the component of interest has been removed

$$\log \mathcal{Z}_{\mathrm{no\,signal}} = 239.5 \pm 0.6$$

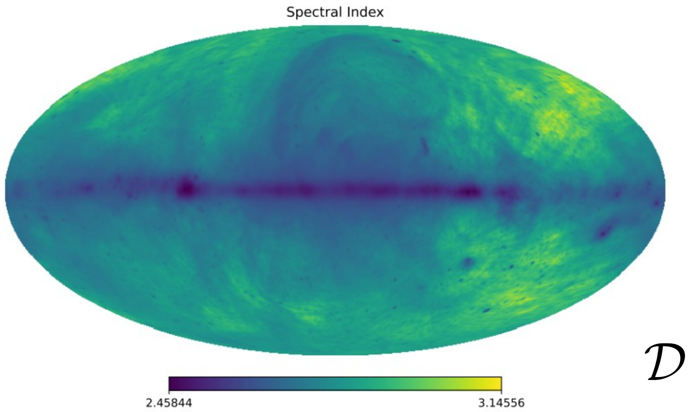
$$\boldsymbol{\log \mathcal{Z}_{\mathrm{Gaussian\ Signal}} = 309.1 \pm 0.4}$$

$$e^{69.6} : 1$$

UNIVERSITY OF CAMBRIDGE

Model Optimisation

Consider a case where our model cannot exactly match the data to within noise



$$\mathcal{D} = \frac{1}{4\pi}\int D\left(\Omega,\nu\right)\left[T_{\mathrm{base}}\left(\Omega\right) - T_{\mathrm{CMB}}\right]\left(\frac{\nu}{\nu_{\mathrm{base}}}\right)^{-2.55}\mathrm{d}\Omega + T_{\mathrm{CMB}} + \widehat{\sigma_n}$$

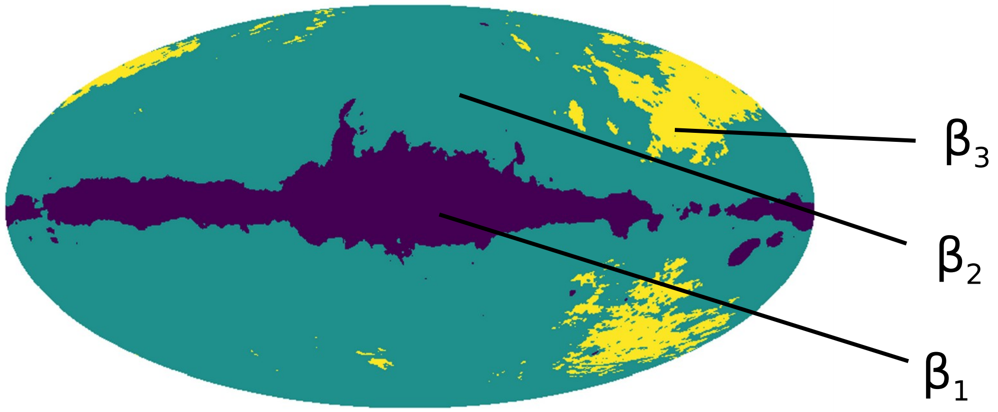
$$\mathcal{D} = \frac{1}{4\pi}\int D\left(\Omega,\nu\right)\left[T_{\mathrm{base}}\left(\Omega\right) - T_{\mathrm{CMB}}\right]\left(\frac{\nu}{\nu_{\mathrm{base}}}\right)^{-\beta(\Omega)}\mathrm{d}\Omega + T_{\mathrm{CMB}} + \widehat{\sigma_n}$$

UNIVERSITY OF CAMBRIDGE

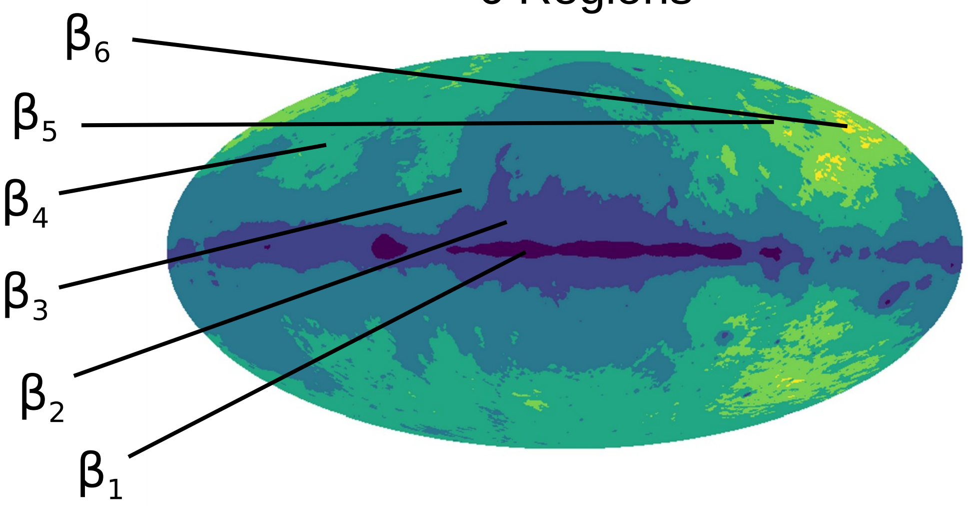
Model Optimisation

Parametrise by spectral index

3 Regions



6 Regions

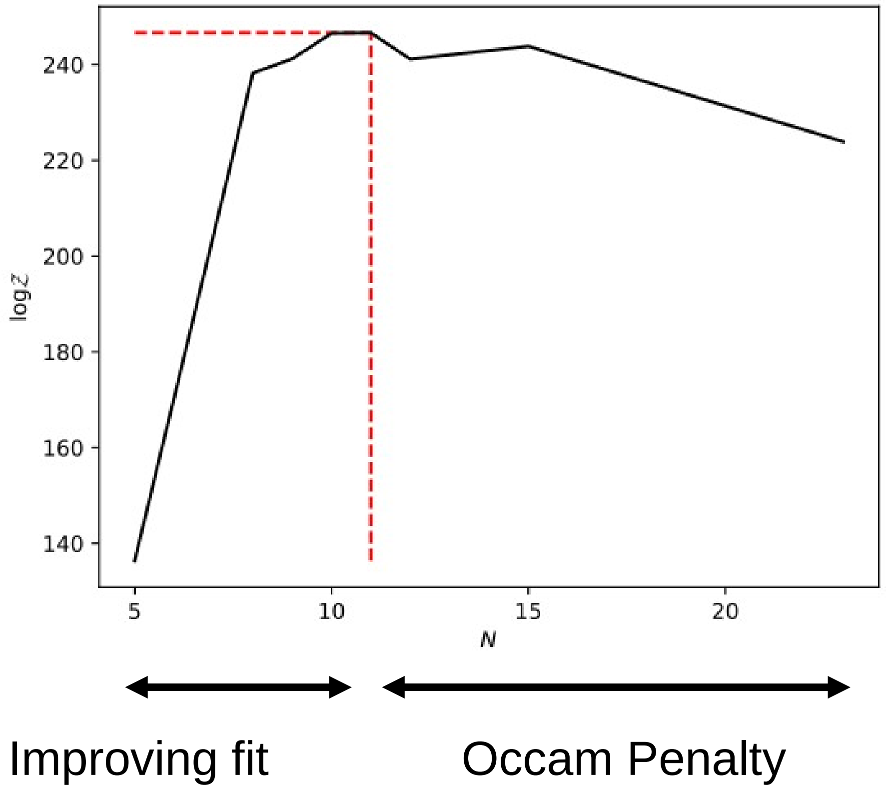


$$T_{\mathrm{F}}\left(\nu, \theta_{\mathrm{F}}\right) = \frac{1}{4\pi} \int D\left(\Omega, \nu\right) \left[\sum_{n=1}^{N} M_n\left(\Omega\right) \left(T_{\mathrm{base}}\left(\Omega\right) - T_{\mathrm{CMB}}\right) \left(\frac{\nu}{\nu_{\mathrm{base}}}\right)^{-\beta_n} \right] \mathrm{d}\Omega + T_{\mathrm{CMB}}$$

How many parameters/how complex a model should be used?

Model Optimisation

Models that give better fits to the data give a higher Bayesian evidence, but additional parameters in the model that do not improve the fit are penalised in the Bayesian evidence



$$\log \mathcal{Z}_{\max} = \log \mathcal{Z}_{N=11} = 246.6 \pm 0.4$$

UNIVERSITY OF CAMBRIDGE

Summary

Recaped MCMC and Nested Sampling and their respective uses

Learnt the basics of 21cm cosmology

Discussed how to define models, priors and likelihoods in practice

Covered how to used Bayesian data products to interpret results:

- Marginalising nuisance parameters
- Comparing different models
- Optimising inexact models
- Quantifying confidences in results