Bayesian Inference How-To: Part 2

Brendon J. Brewer

The University of Auckland

www.stat.auckland.ac.nz/~brewer

Twitter: @brendonbrewer

Topics for Today

We will look at:

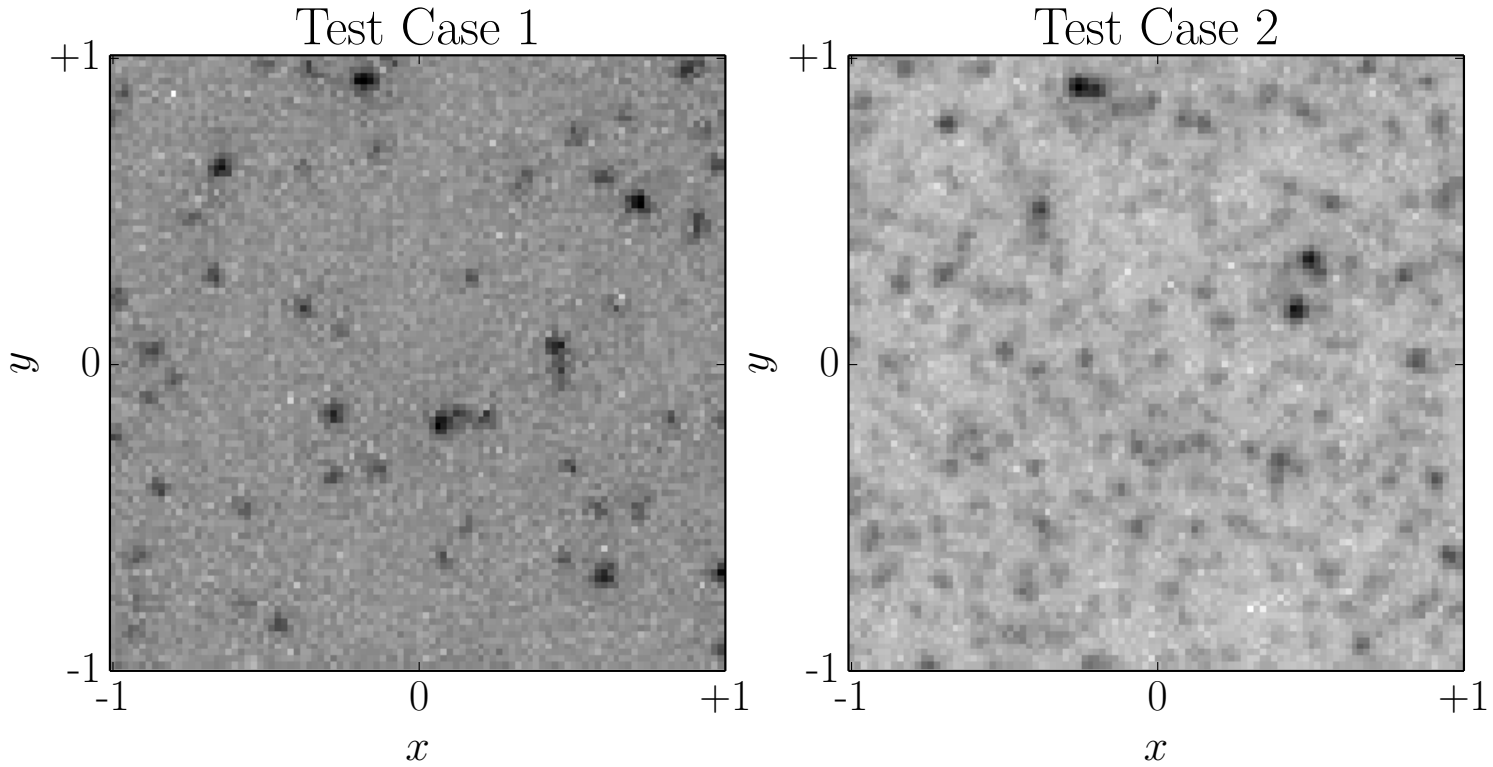
- Trans-dimensional MCMC
- Nested Sampling

Trans-dimensional MCMC

Trans-dimensional MCMC is useful when the model dimension is unknown. This arises quite frequently in astrophysics.

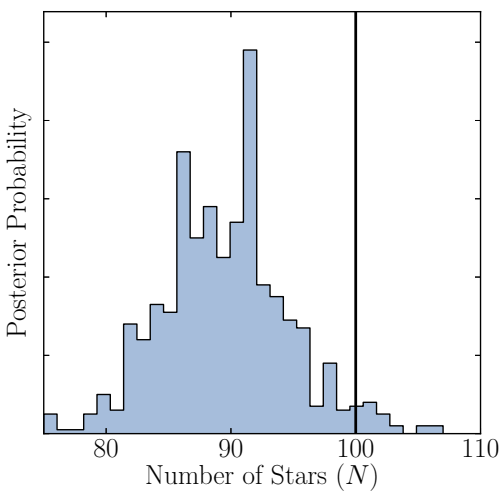
Trans-dimensional examples

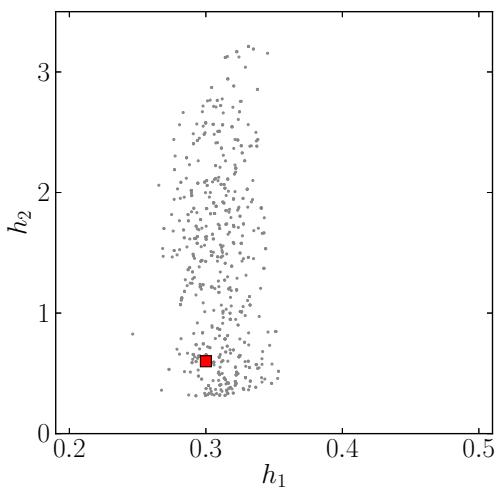
Some examples from my own work: How many stars are in these images (and what are their positions and fluxes)?

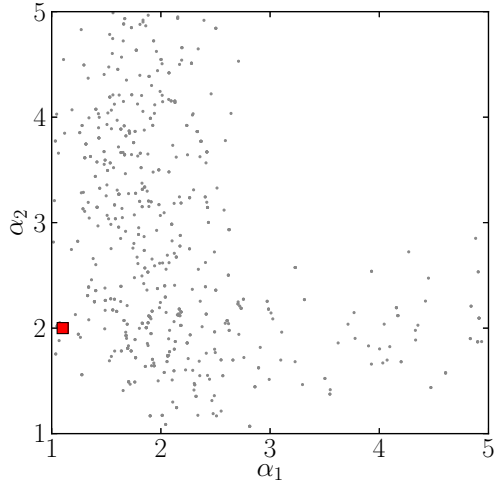


Trans-dimensional examples

How many stars were there?

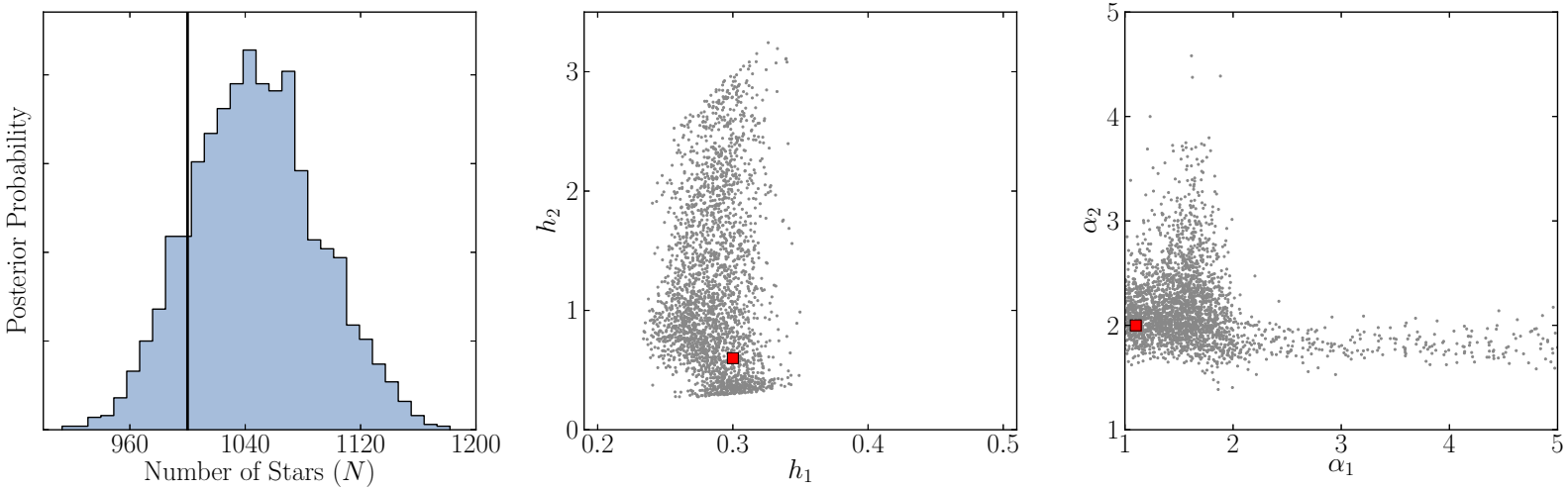






Trans-dimensional examples

How many stars were there?



Trans-dimensional examples

Perhaps you can see how to use MCMC to estimate the parameters of the stars (x and y position, and a brightness, for each), if we knew the number of stars.

But we want to know the number of stars!

Asteroseismology Example

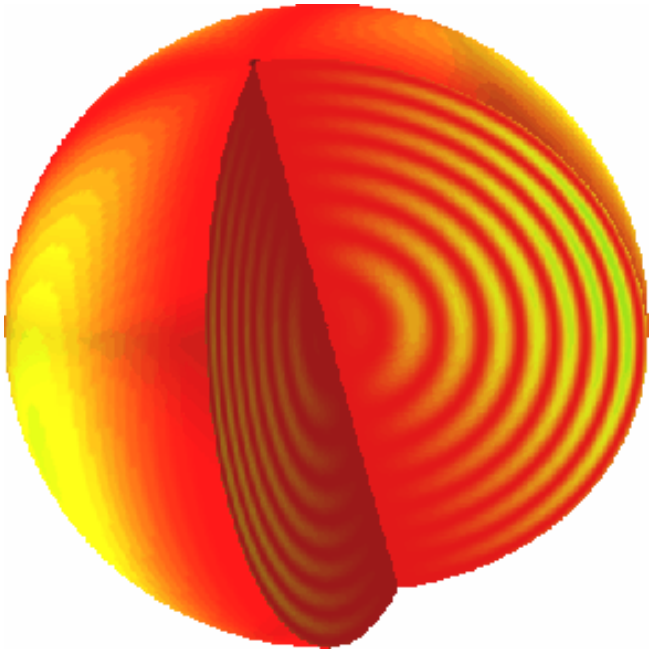


Image credit: Tim Bedding

Asteroseismology: Time Series Data

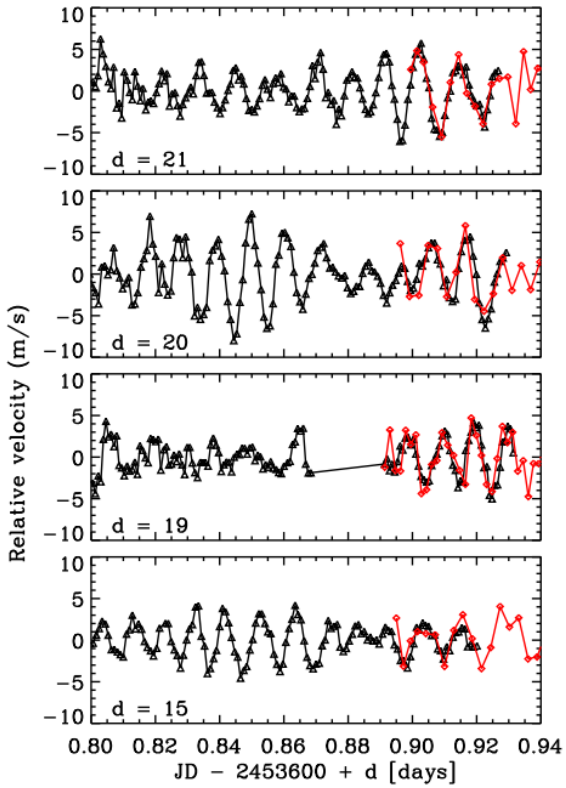


Image credit: Tim Bedding

Asteroseismology Example: Power Spectrum

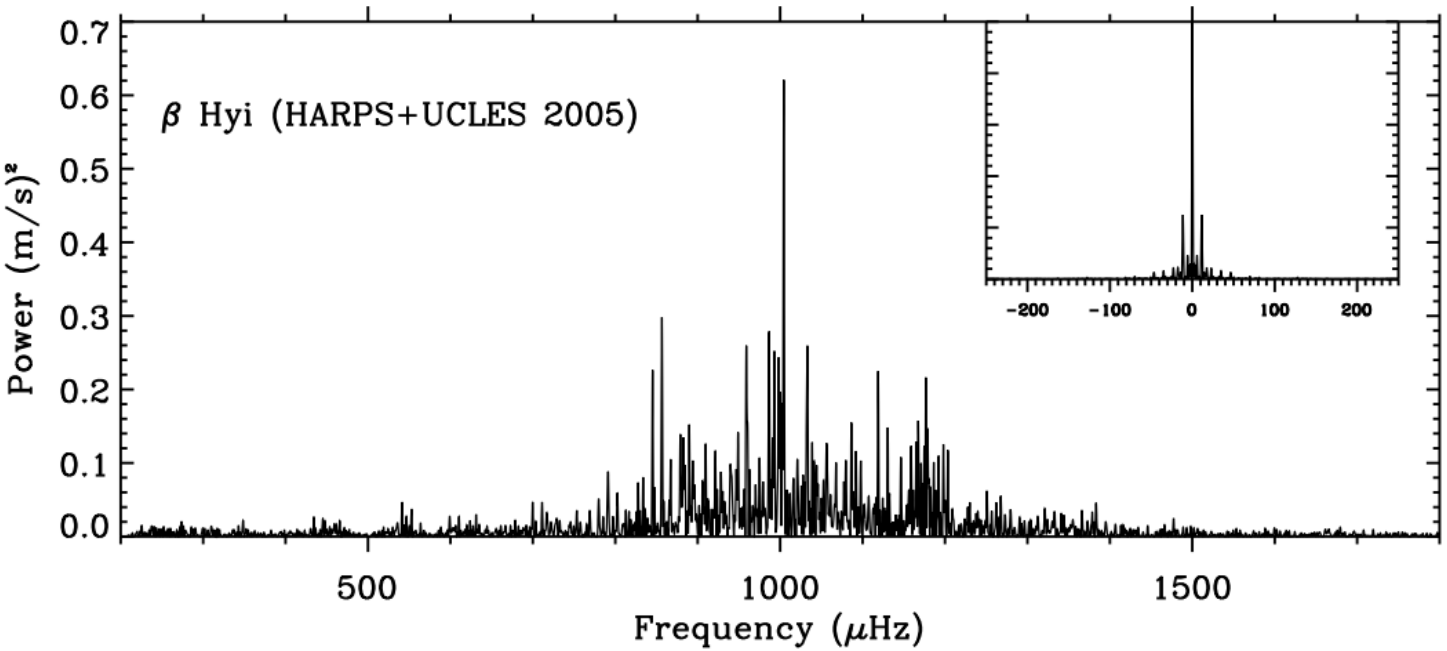
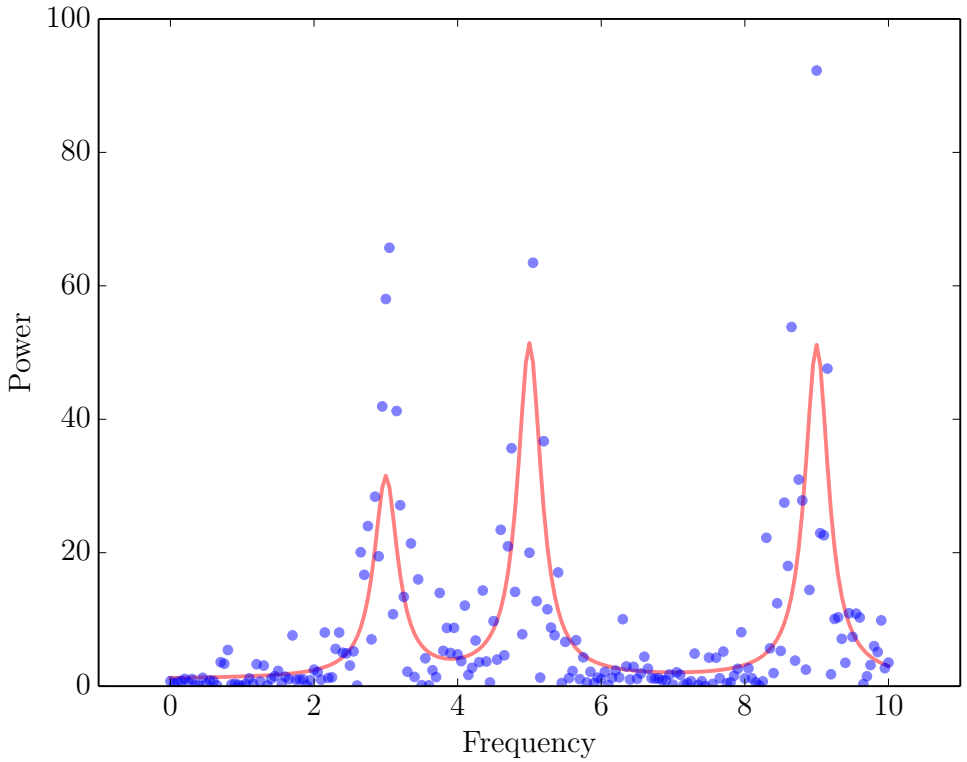
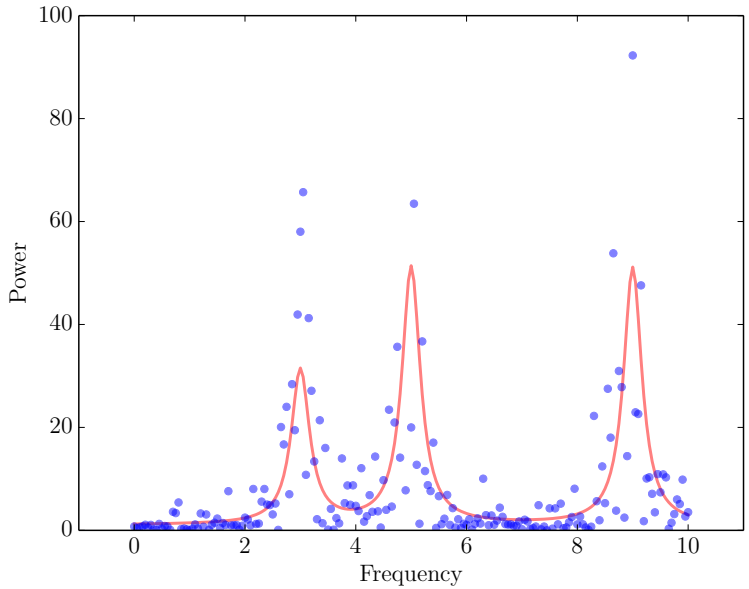


Image credit: Tim Bedding

Asteroseismology Example: Toy Dataset



Asteroseismology Example: Question



Given this data, how many peaks are there? And what are their parameters (position, height, width)?

Asteroseismology Example

Each of the peaks has a "Lorentzian" shape (same as the Cauchy distribution!):

$$m(x) = B + \sum_{i=1}^{N} \frac{A_i}{\left[1 + \left(\frac{x - c_i}{w_i}\right)^2\right]} \quad (1)$$

A_i = amplitude of ith component

c_i = center of ith component

w_i = width of ith component

Asteroseismology Example

The sampling distribution/likelihood is

$$y_i \sim \text{Exponential}(m(x_i;\theta)).$$

i.e.

$$p(\{y_i\}|\theta) \quad = \quad \prod_{i=1}^{n} \frac{1}{m(x_i;\theta)} \exp\left[-\frac{y_i}{m(x_i;\theta)}\right].$$

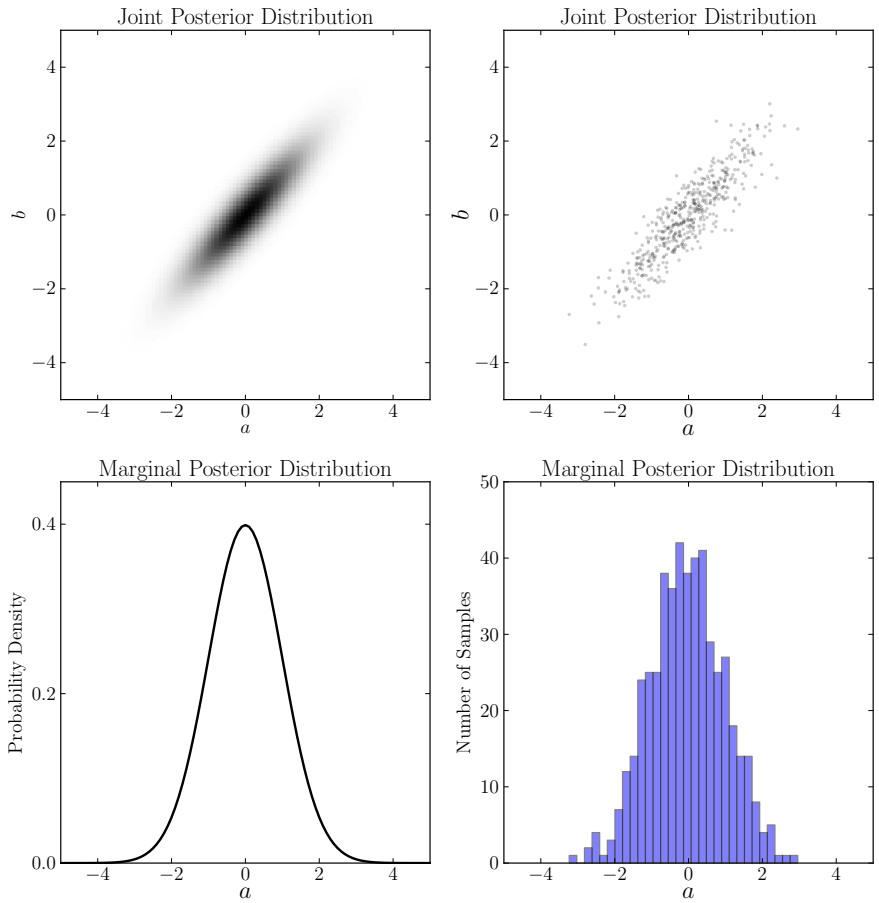
Asteroseismology Example

Some simple priors are:

$$
\begin{aligned}
N &\sim \text{Uniform}\left(\{0, 1, 2, ..., 9, 10\}\right) \\
\log(B) &\sim \text{Uniform}\left[\log(10^{-3}, \log(10^{3})\right] \\
c_i &\sim \text{Uniform}\left(x_{\min}, x_{\max}\right) \\
A_i &\sim \text{Exponential(mean=10)} \\
\log(w_i) &\sim \text{Uniform}\left[\log(0.01 x_{\text{range}}), \log(x_{\text{range}})\right]
\end{aligned}
$$

Recall: Monte Carlo

- **Marginalisation** becomes trivial
- We can quantify all uncertainties we might be interested in



Recall: The Metropolis Algorithm

- Start at some point θ in the hypothesis space.
- Loop

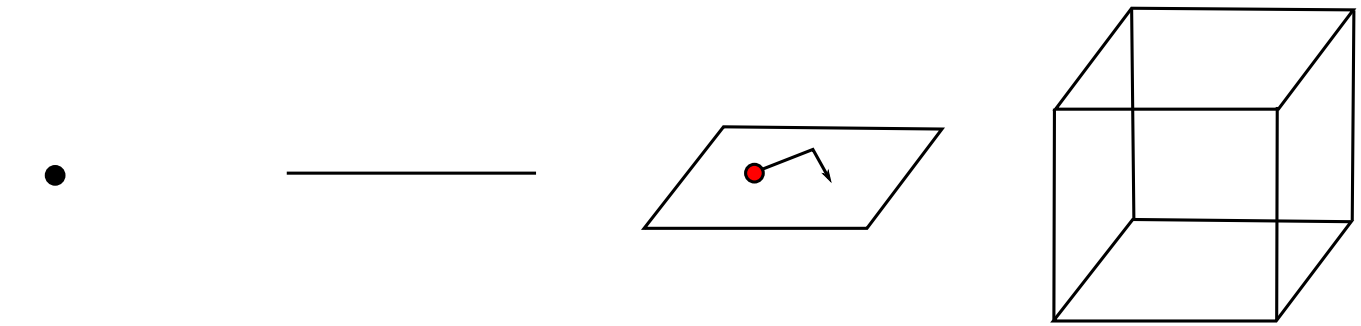
  $\{$

  - Generate **proposal** from some distribution $q(\theta'|\theta)$ (e.g. slightly perturb the current position).
  - With probability $\alpha = \min\left(1, \frac{p(\theta')p(D|\theta')}{p(\theta)p(D|\theta)}\right)$, accept the proposal (i.e. replace θ with θ').
  - Otherwise, stay in the same place.

  $\}$

Trans-Dimensional MCMC

For problems of unknown dimensionality, the hypothesis space is the union of several fixed-dimension hypothesis spaces. To do MCMC with these models, you need a way to move between models with different numbers of components.



Approaches to Trans-Dimensional MCMC

There are several approaches:

- Reversible Jump MCMC (Green, 1995)
- Birth and Death MCMC (Stephens, 2000)

Approaches to Trans-Dimensional MCMC

We will do our MCMC like this:

- Put 10 components in the model, and do MCMC as usual.
- Interpret the parameter N as the **number of components that are "switched on"**

Code for the priors

Let's take a look at the Python code for the priors. Remember, the prior appears in two places:

- The function `from_prior`, which we use to generate a starting point
- The function `log_prior`, which calculates the log of the prior density, which is used to determine the acceptance probability.

Code for the likelihood

Let's take a look at the Python code for the likelihood function.

Note how the calculation of the model curve $m(x)$ only sums over the first N model components, the ones that are **switched on**.

Label Switching Degeneracy

Imagine we found a solution with two peaks like this:

$$\begin{aligned} \text{Peak 1} : \{A, c, w\} &= \{5, 3, 2\} \\ \text{Peak 2} : \{A, c, w\} &= \{3, 7, 1\} \end{aligned}$$

Then the following solution is completely equivalent:

$$\begin{aligned} \text{Peak 1} : \{A, c, w\} &= \{3, 7, 1\} \\ \text{Peak 2} : \{A, c, w\} &= \{5, 3, 2\} \end{aligned}$$

Label Switching Degeneracy

When there are N peaks, the posterior will have $N!$ identical modes, corresponding to switching the order of the peaks.

We can add a proposal move that switches labels. Since the meaning of the models is the unchanged, this proposal will always be accepted.

Label Switching Degeneracy

The `shuffle` function chooses two switched-on peaks "at random" and swaps their parameter values.

Label Switching Degeneracy

When there are N peaks, the posterior will have $N!$ identical modes, corresponding to switching the order of the peaks.

We can add a proposal move that switches labels. Since the meaning of the models is the unchanged, this proposal will always be accepted.

Label Switching Degeneracy

The `shuffle` function chooses two peaks "at random" and swaps their parameter values.

Label Switching Degeneracy

Consider the marginal posterior distribution for x_1. It will be multimodal, because of the label-switching issue.

In models like this, we can plot a mixture of the posterior for x_1, x_2 (when it exists), and so on.

Part II: Nested Sampling

Nested Sampling is a Monte Carlo method (not necessarily MCMC) that was introduced by John Skilling in 2004.

It is very popular in astrophysics and has some unique strengths.

Marginal Likelihood

The **marginal likelihood** is useful for "model selection". Consider two models: M_1 with parameters θ_1, M_2 with parameters θ_2. The marginal likelihoods are:

$$\begin{aligned} p(D|M_1) &= \int p(\theta_1|M_1)p(D|\theta_1, M_1)\, d\theta_1 \\ p(D|M_2) &= \int p(\theta_2|M_2)p(D|\theta_2, M_2)\, d\theta_2 \end{aligned}$$

These are the normalising constants of the posteriors, within each model.

Bayesian Model Selection

If you have the marginal likelihoods, it's easy:

$$\frac{P(M_1|D)}{P(M_2|D)} = \frac{P(M_1)}{P(M_2)} \times \frac{P(D|M_1)}{P(D|M_2)}.$$

$$\text{(posterior odds)} = \text{(prior odds)} \times \text{(bayes factor)}$$

Challenging features

Another motivation: standard MCMC methods can get stuck in the following situations:

Nested Sampling

Nested Sampling was built to estimate the marginal likelihood.

But it can also be used to generate posterior samples, and it can potentially work on harder problems where standard MCMC methods get stuck.

Notation

When discussing Nested Sampling, we use different symbols:

$$p(D|M_1) \quad = \quad \int p(\theta_1|M_1)p(D|\theta_1, M_1)\, d\theta_1$$

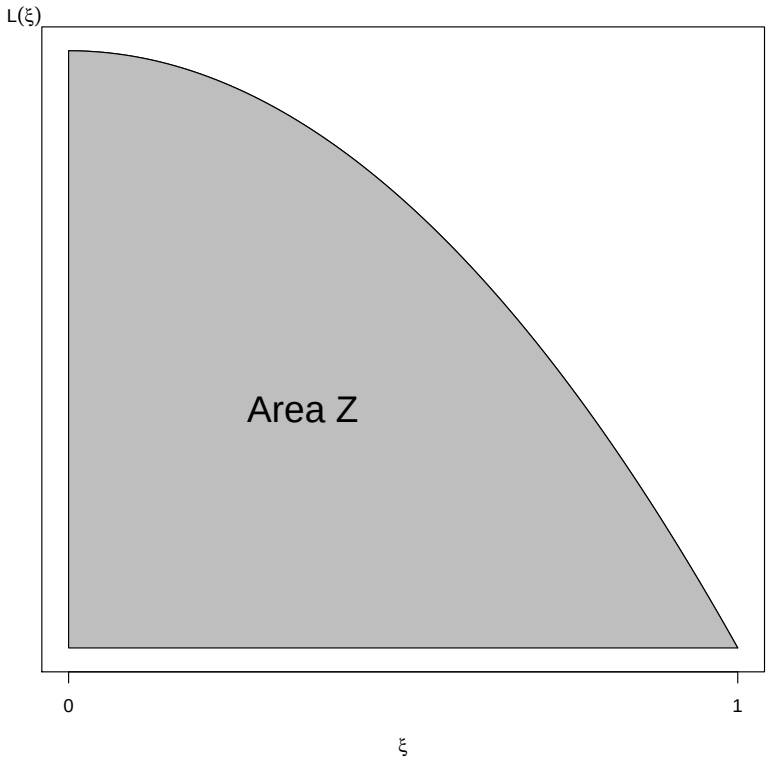
becomes

$$Z \quad = \quad \int \pi(\theta)L(\theta)\, d\theta.$$

Z = marginal likelihood, $L(\theta)$ = likelihood function, $\pi(\theta)$ = prior distribution.

Nested Sampling

Imagine we had an easy 1-D problem, with a Uniform(0, 1) prior, and a likelihood that was strictly decreasing.



Nested Sampling

The key idea of Nested Sampling: Our high dimensional problem can be mapped onto the easy 1-D problem. Figure from Skilling (2006):

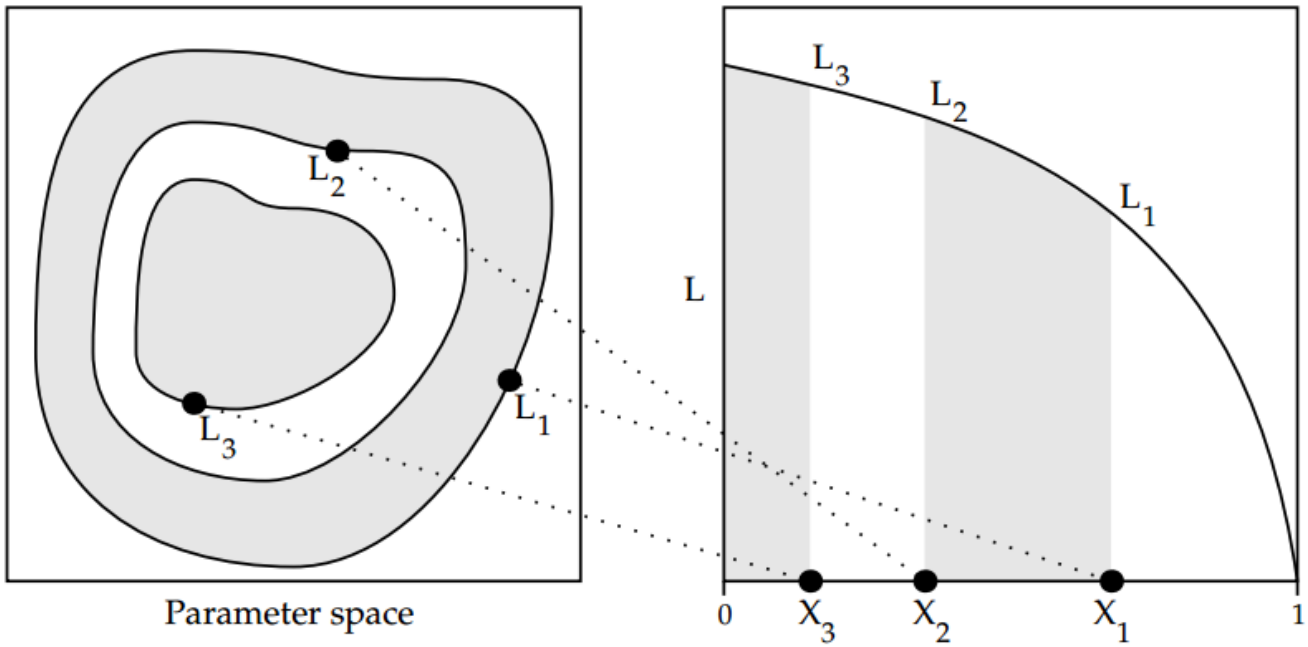


Figure 3: Nested likelihood contours are sorted to enclosed prior mass X.

Nested Sampling X

Define

$$X(L^*) = \int \pi(\theta) \mathbb{1}\left(L(\theta) > L^*\right) \, d\theta$$

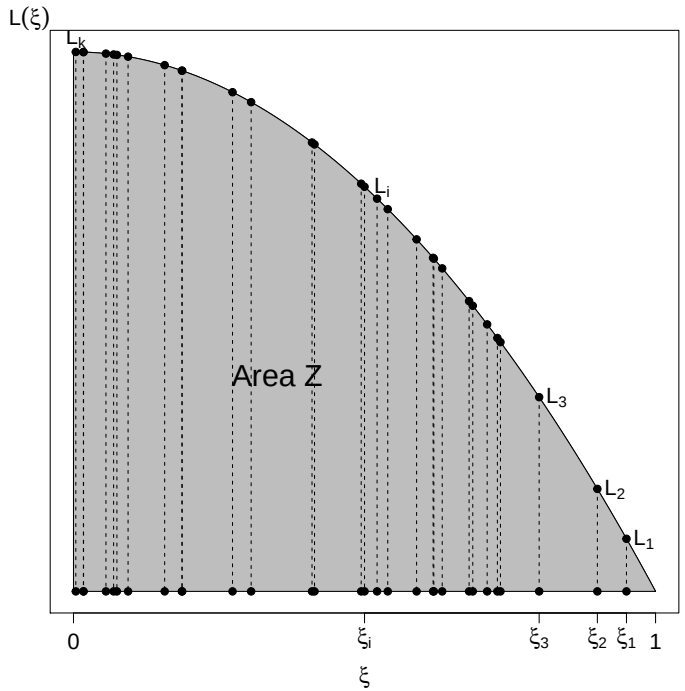
X is the **amount of prior probability** with likelihood greater than L^*.

Loosely, X is the **volume** with likelihood above L^*.

Higher $L^* \Leftrightarrow$ lower volume.

Numerical Integration

If we had some points with likelihoods L_i, and we knew the corresponding X-values, we could approximate the integral numerically, using the trapezoidal rule or something similar.



Nested Sampling Procedure

This procedure gives us the likelihood values.

- Sample $\theta = \{\theta_1, \ldots, \theta_N\}$ from the prior $\pi(\theta)$.
- Find the point θ_k with the worst likelihood, and let L^* be its likelihood.
- Replace θ_k with a new point from $\pi(\theta)$ but restricted to the region where $L(\theta) > L^*$.

Repeat the last two steps many times. The *discarded points* (the worst one at each iteration) are the output.

Generating the new point

We need a new point from $\pi(\theta)$ but restricted to the region where $L(\theta) > L^*$. The point being replaced has the worst likelihood, so **all the other points satisfy the constraint!**

So we can use one of the other points to initialise an MCMC run, trying to sample the prior, but rejecting any proposal with likelihood below L^*. See code.

Generating the new point

There are alternative versions of NS available, such as **MultiNest**, that use different methods (not MCMC) to generate the new point.

I also have a version of NS called **Diffusive Nested Sampling**, which is a better way of doing NS when using MCMC. I'm happy to discuss it offline.

Nested Sampling Procedure

Nested Sampling gives us a sequence of points with increasing likelihoods, but we need to somehow know their X-values!

Estimating the X values

Consider the simple one-dimensional problem with Uniform(0, 1) prior.

When we generate N points from the prior, the distribution for the X-value of the worst point is Beta$(N, 1)$. So we can use a draw from Beta$(N, 1)$ as a guess of the X value.

Estimating the X values

Each iteration, the worst point should reduce the volume by a factor that has a Beta$(N, 1)$ distribution. So we can do this:

$$\begin{aligned} X_1 &= t_1 \\ X_2 &= t_2 X_1 \\ X_3 &= t_3 X_2 \end{aligned}$$

and so on, where $t_i \sim$Beta$(N, 1)$. Alternatively, we can use a simple approximation.

Deterministic Approximation

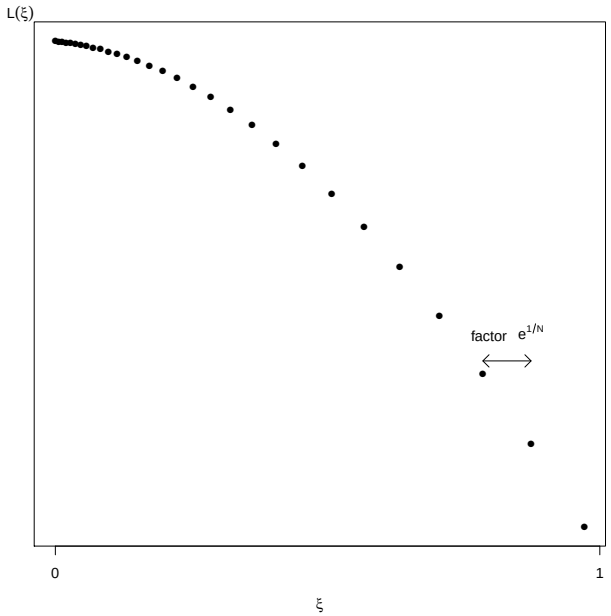


Figure: Deterministic approximation. Each iteration reduces the volume by a factor $\approx e^{-1/N}$. e.g. if $N = 5$, the worst likelihood accounts for about 1/5th of the remaining prior volume.

Posterior Distribution from Nested Sampling

The posterior sample can be obtained by assigning weights W_j to the discarded points:

$$W_j = \frac{L_j w_j}{Z}$$

where $w_j = X_{j-1} - X_{j+1}$ is the "prior weight/width" associated with the point. The "effective sample size" is given by

$$ESS = \exp\left(-\sum_{j=1}^{m} W_j \log W_j\right)$$

Information

NS can also calculate the **information**, also known as the Kullback-Liebler divergence from the prior to the posterior.

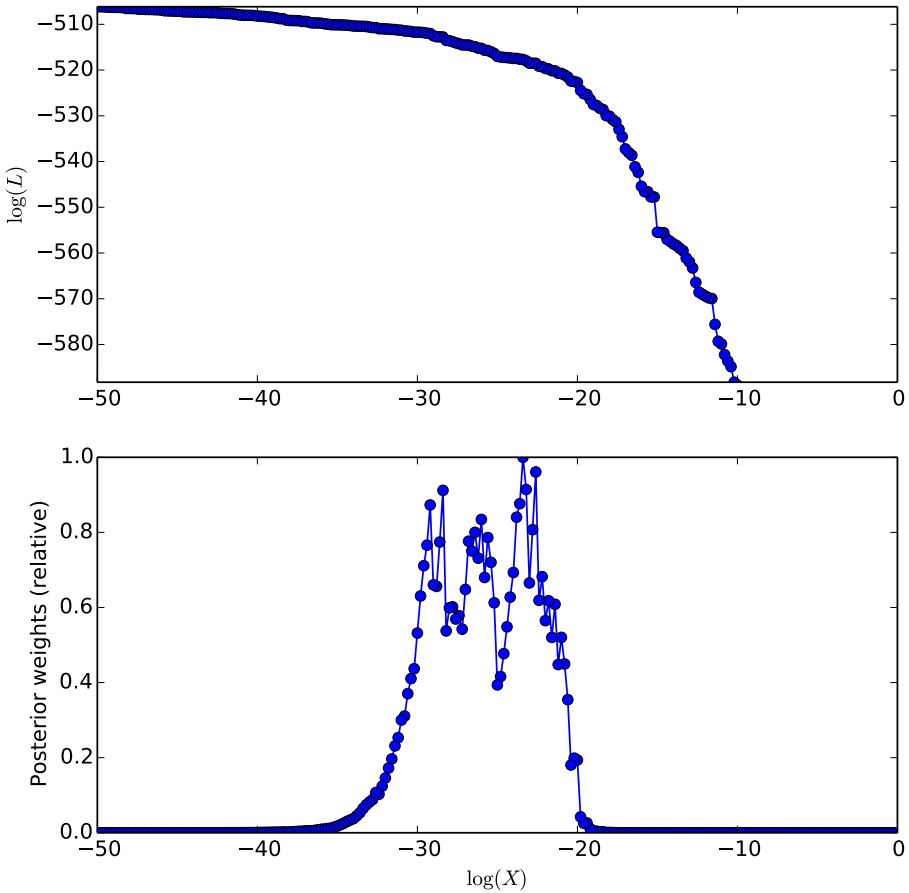
$$\begin{aligned} \mathcal{H} &= \int p(\theta|D) \log \left[\frac{p(\theta|D)}{p(\theta)} \right] d\theta \\ &\approx \log \left(\frac{\text{volume of prior}}{\text{volume of posterior}} \right) \end{aligned}$$

Nested Sampling Code

I have written a basic implementation of Nested Sampling in Python.

Let's use it on the transit problem and the asteroseismology problem.

Nested Sampling Plots



Nested Sampling Plots

A necessary but not sufficient condition for everything being okay is that you see the entire peak in the posterior weights.

If it's not there, you haven't done enough NS iterations. i.e. your parameter values have lower likelihoods than what is typical of the posterior distribution.

Nested Sampling Plots

The shape of the $\log(L)$ vs. $\log(X)$ plot is also informative: if it is straight for a long time, or concave up at some point, your problem contains a phase transition, and it's a good thing you used Nested Sampling!