Introduction to Nested Sampling

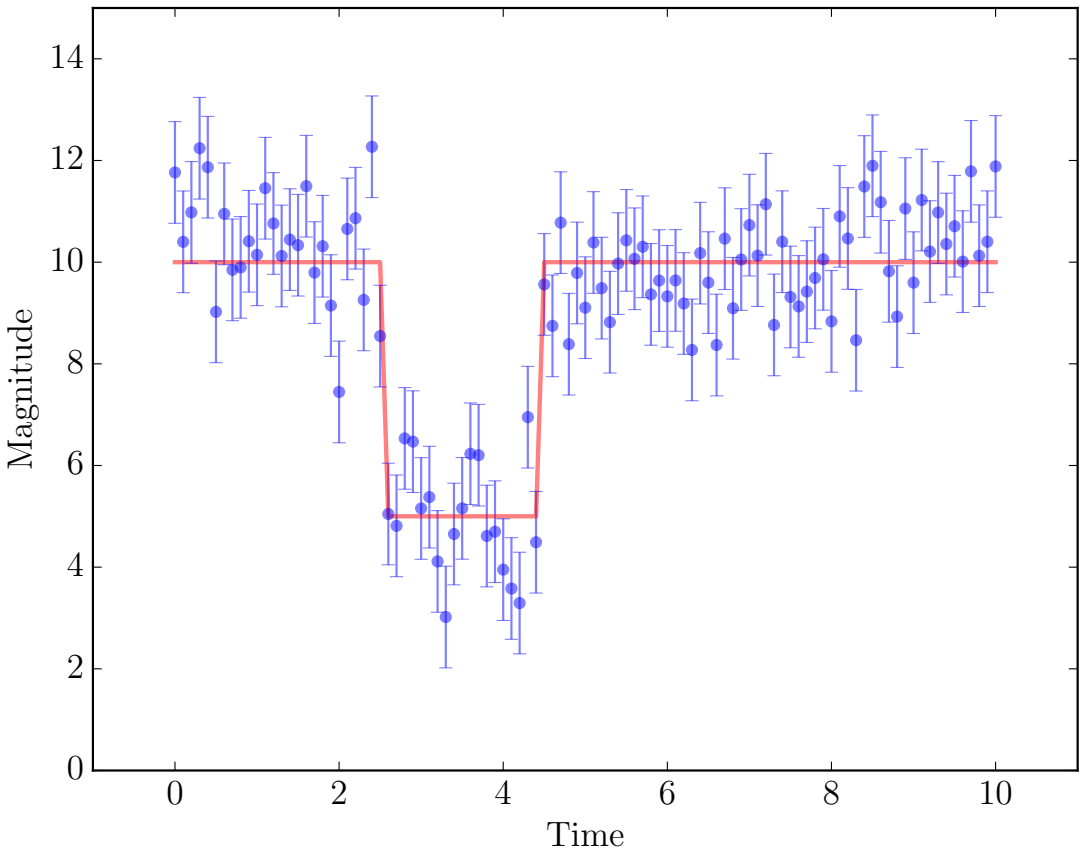
Brendon J. Brewer

Department of Statistics, The University of Auckland

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

@brendonbrewer

Transit Example



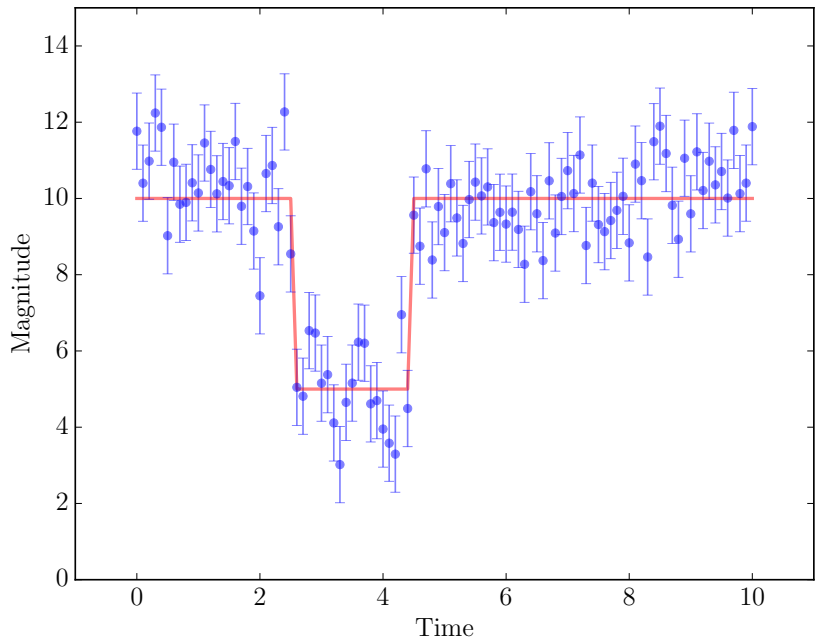
Related to the transit example...

- Realistic exoplanet transits
- Finding emission/absorption lines in spectra
- Finding stars/galaxies in an image
- All "curve fitting"

Transit Example: The Truth

The red curve was:

$$\mu(t) = \begin{cases} 10, & 2.5 \leq t \leq 4.5 \\ 5, & \text{otherwise.} \end{cases}$$



Transit Example: The Truth

The red curve was:

$$\mu(t) = \begin{cases} 10, & 2.5 \leq t \leq 4.5 \\ 5, & \text{otherwise.} \end{cases}$$

and the noise was added like this:

```
# Add noise
sig = 1.
y += sig*rng.randn(y.size)
```

Transit Example: Inference

Let's fit the data with this model:

$$\mu(t) = \begin{cases} A, & (t_c - w/2) \leq t \leq (t_c + w/2) \\ A - b, & \text{otherwise.} \end{cases}$$

We don't know A, b, t_c, and w. But we do know the data D.

Transit Example: Parameters

We don't know A, b, t_c, and w. These are our unknown parameters. Let's find the posterior.

$$p(A, b, t_c, w|D) = \frac{p(A, b, t_c, w)p(D|A, b, t_c, w)}{p(D)}$$

and the marginal likelihood

$$p(D) = \int p(A, b, t_c, w)p(D|A, b, t_c, w)\, dA\, db\, dt_c\, dw$$

Uniform Priors

For all four parameters:

$$\begin{aligned} A &\sim U(-100, 100) \\ b &\sim U(0, 10) \\ t_c &\sim U(t_{\min}, t_{\max}) \\ w &\sim U(0, t_{\max} - t_{\min}) \end{aligned}$$

Where $t_{\min}$ and $t_{\max}$ give the time range of the dataset. Question: is this legitimate? Are we using the data to set our priors?

Sampling Distribution / Likelihood

Let's assume "gaussian noise":

$$p(y_i|A, b, t_c, w) = \prod_{i=1}^{N} \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma_i^2}\left(y_i - m(t_i; A, b, t_c, w)\right)^2\right].$$

or more concisely:

$$y_i|A, b, t_c, w \sim \mathcal{N}\left(m(t_i; A, b, t_c, w), \sigma_i^2\right).$$

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

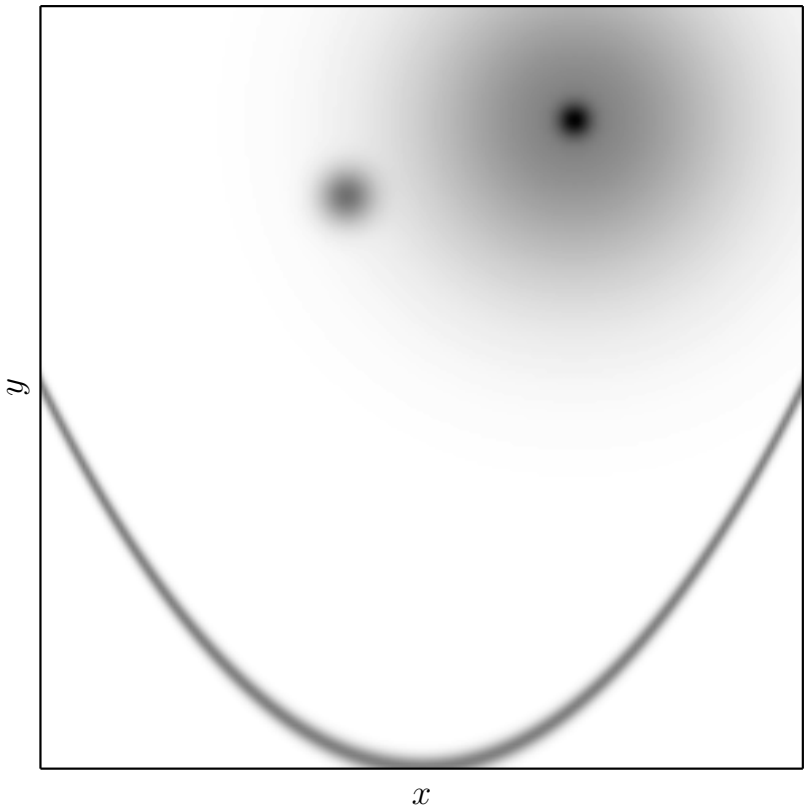
If you have the marginal likelihoods, it's easy:

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

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

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:

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

becomes

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

$Z =$ 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.

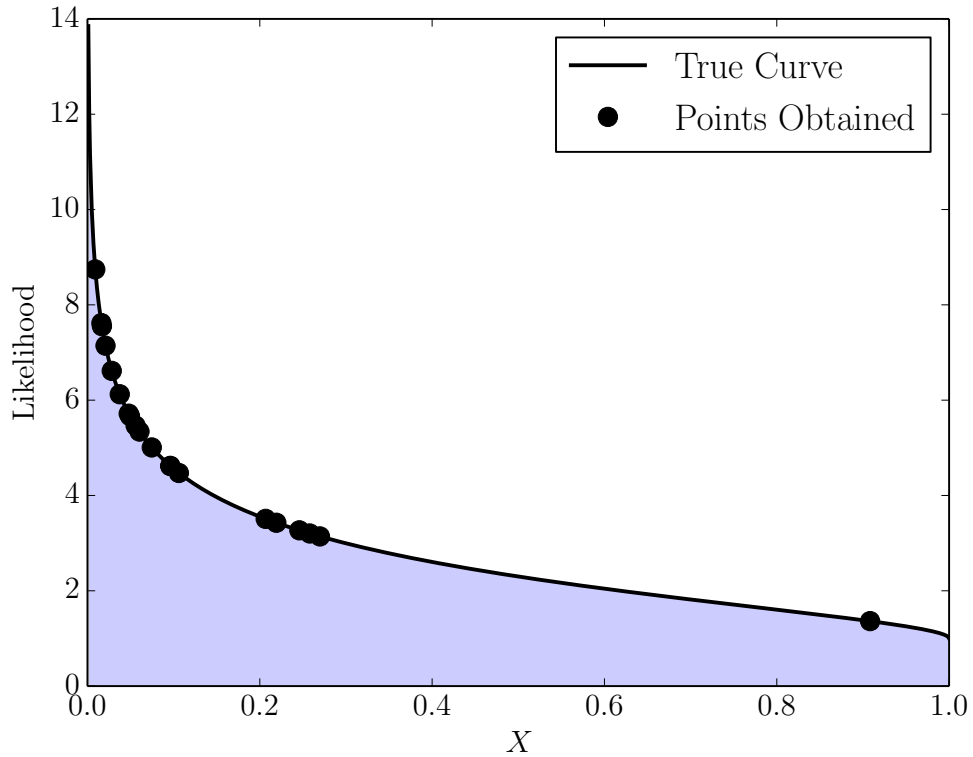


Figure: Likelihood function with area Z.

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):

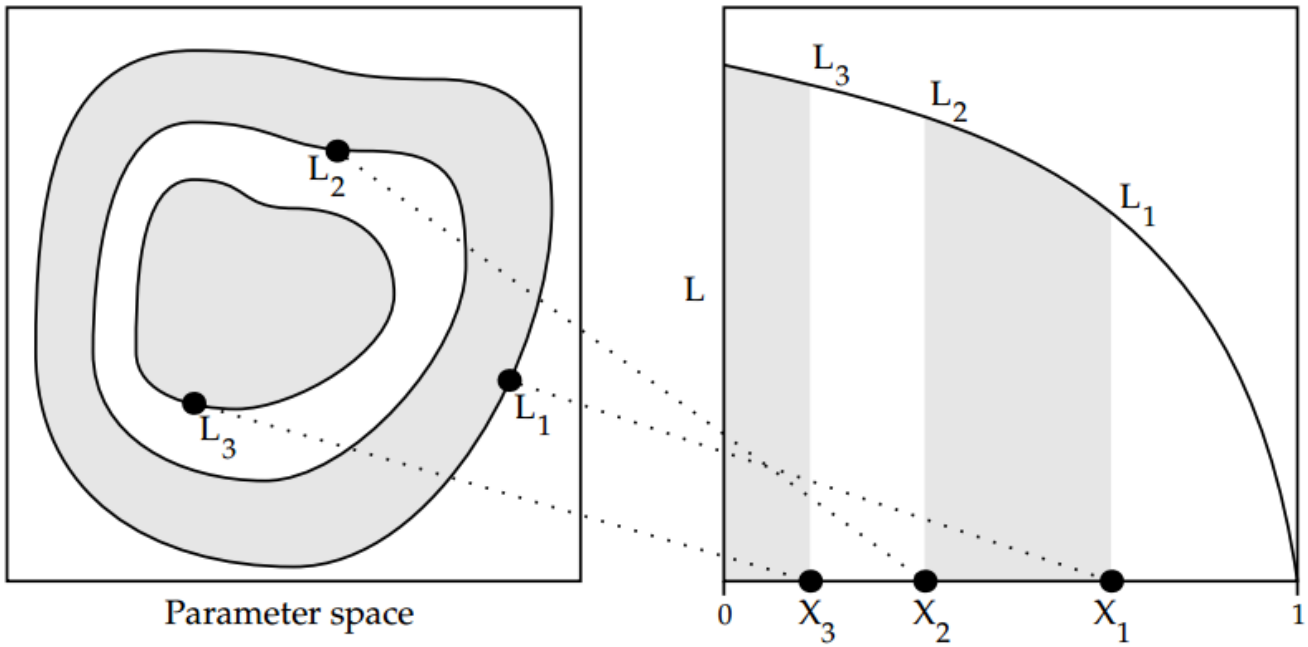


Figure 3: Nested likelihood contours are sorted to enclosed prior mass X.

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

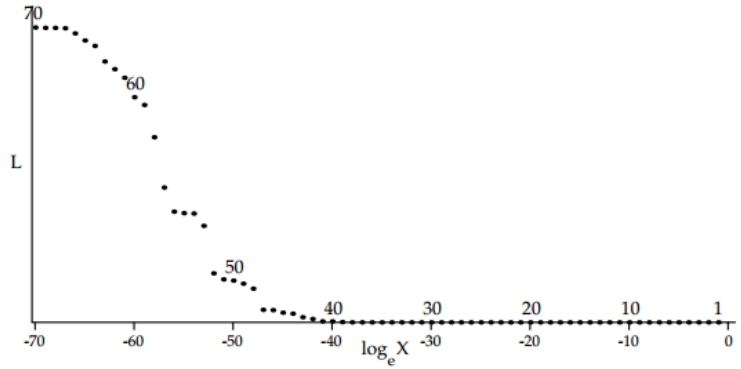


Figure 7: The same sequence of nested sampling points as in Fig. 6, shown at the crude central estimate of their X values (*i.e.* uniformly in $\log X$).

deviation.

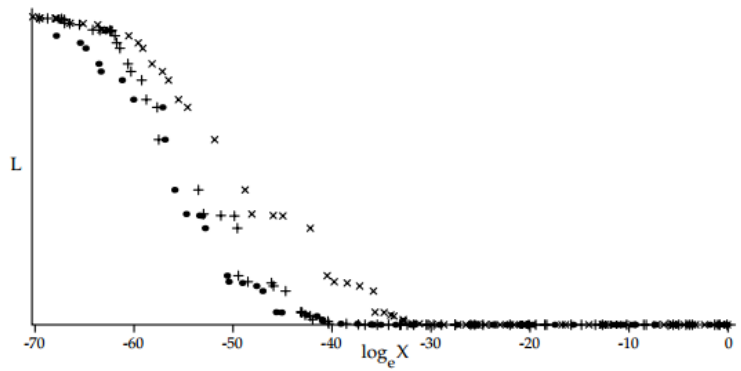
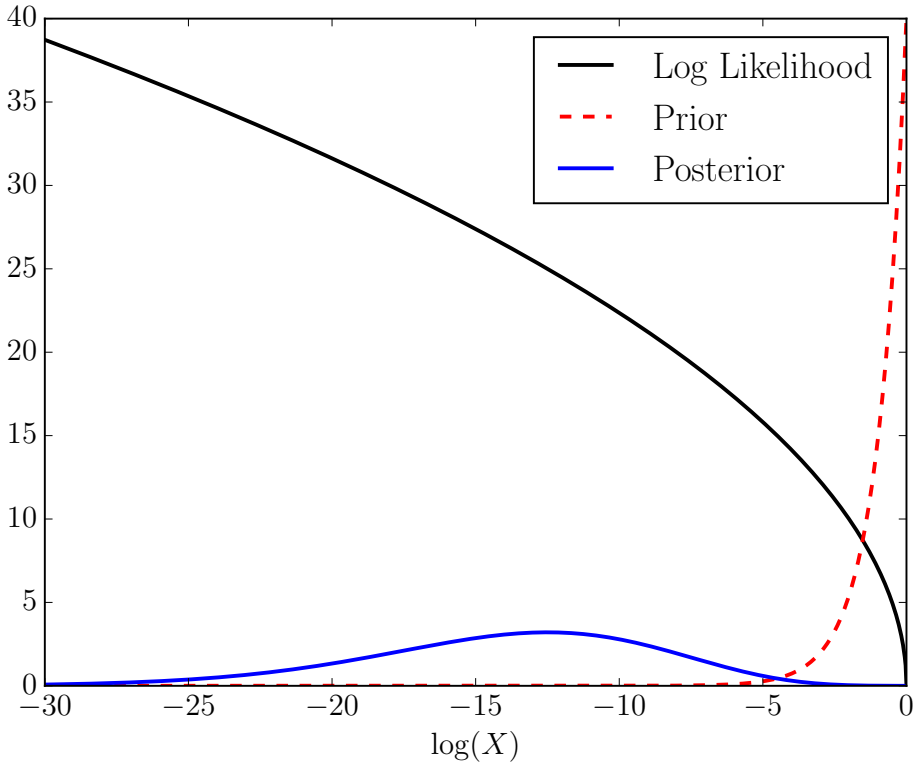


Figure 8: The same sequence of nested sampling points as in Fig. 6, showing three random assignments $(\bullet, +, \times)$ for X values taken from $\Pr(X)$.


Log Likelihood
Prior
Posterior
40
35
30
25
20
15
10
5
0
−30
−25
−20
−15
−10
−5
0
log(X)

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.

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!