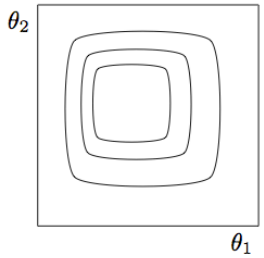


Bayesian Reasoning:
Nested Sampling

Recap: Model Comparison

As you have learnt in previous lectures, Bayes theorem provides us with a quantitative method of **selecting between competing hypotheses**

In order to answer this question we need to calculate the following ('odds') ratio of probabilities:

$$\frac{\Pr(\mathcal{M}_1|\boldsymbol{d})}{\Pr(\mathcal{M}_2|\boldsymbol{d})} = \frac{\Pr(\boldsymbol{d}|\mathcal{M}_1)}{\Pr(\boldsymbol{d}|\mathcal{M}_2)} \frac{\Pr(\mathcal{M}_1)}{\Pr(\mathcal{M}_2)} = \frac{E_1}{E_2} \frac{\Pr(\mathcal{M}_1)}{\Pr(\mathcal{M}_2)}$$

Evidence **Prior**

Recap: Model Comparison

$$E \equiv \Pr(\boldsymbol{d}|\mathcal{M})$$

$$= \int \Pr(\boldsymbol{d}|\boldsymbol{\theta}, \mathcal{M}) \Pr(\boldsymbol{\theta}|\mathcal{M}) \mathrm{d}\boldsymbol{\theta}$$

Likelihood (L) Prior (π)

where ***d*** are the data, ***ϑ*** are the model parameters and *M* is the model hypothesis we wish to test

In practice calculation of the **marginal likelihood** or **"Evidence"** *E* typically requires numerical integration

Recap: Model Comparison

Often the numerical integration is implemented using some form of Monte Carlo or ‘random sampling’ integration

For complex, multi-dimensional problems how do we find and sample the interesting regions of high likelihood value in the prior volume and so obtain an accurate estimate of the evidence for a given hypothesis?

Recap: MCMC methods

Specific forms of **Monte Carlo** integration include Metropolis-Hastings and Simulated Annealing

In these cases a chain of points converges to sampling the joint posterior distribution

However these techniques primarily focus on **parameter estimation**, and **neglect efficient model comparison**

Introducing Nested Sampling

The key premise of the **nested sampling** algorithm is to provide an **efficient** and **robust** numerical means of calculating the marginal likelihood for **model comparison**

Parameter estimation is a useful by-product of the algorithm

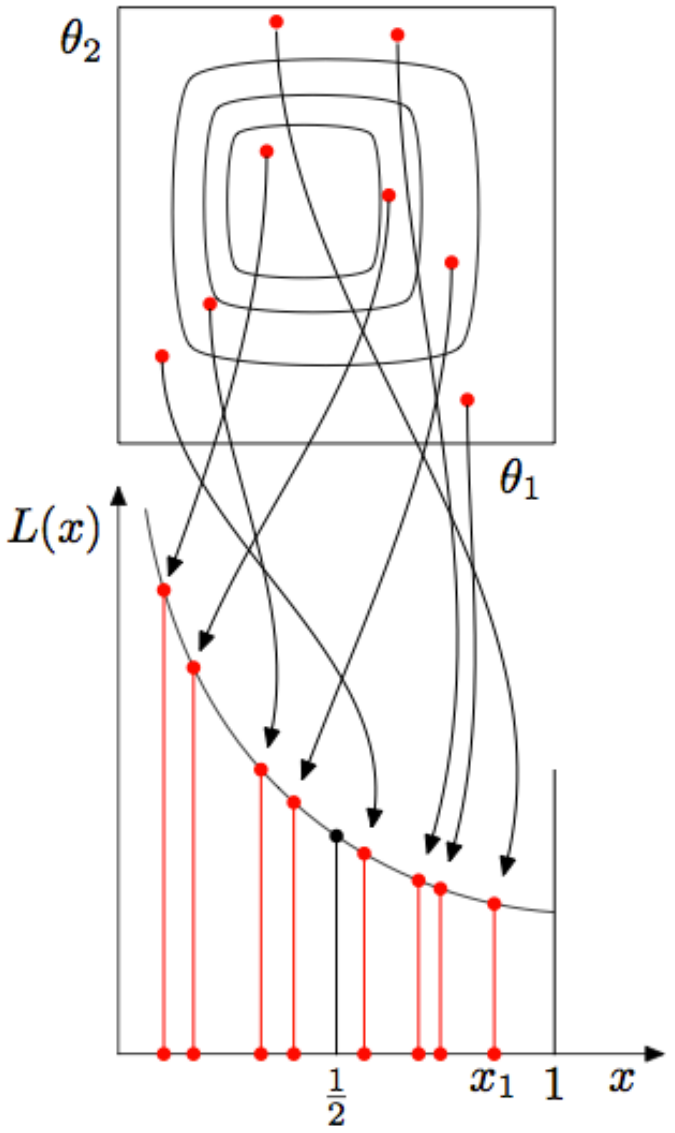
Nested sampling is statistical and therefore of the Monte Carlo class of algorithms

Basics of the algorithm

Initialization:

- *N* uniform random samples (‘**active sample**’) are drawn from the prior parameter space

- These **active samples** are re-ordered in ascending likelihood



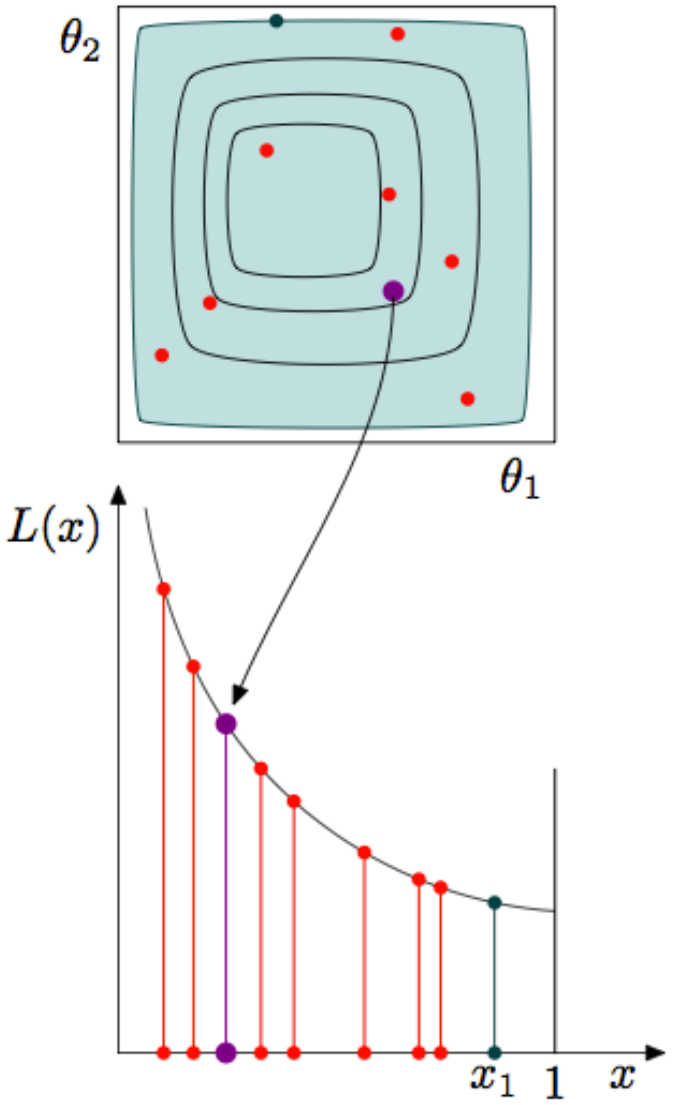
Basics of the algorithm

Iterative step:

- We remove the active sample with the lowest likelihood from the active sample and store it in the chain

- A new sample from the prior is added to the group, subject to the hard constraint that its likelihood is greater than the recently removed sample

- Continue this process iteratively



Basics of the algorithm

Each successive iteration defines an outer likelihood contour nested within the previous one, hence “Nested Sampling”

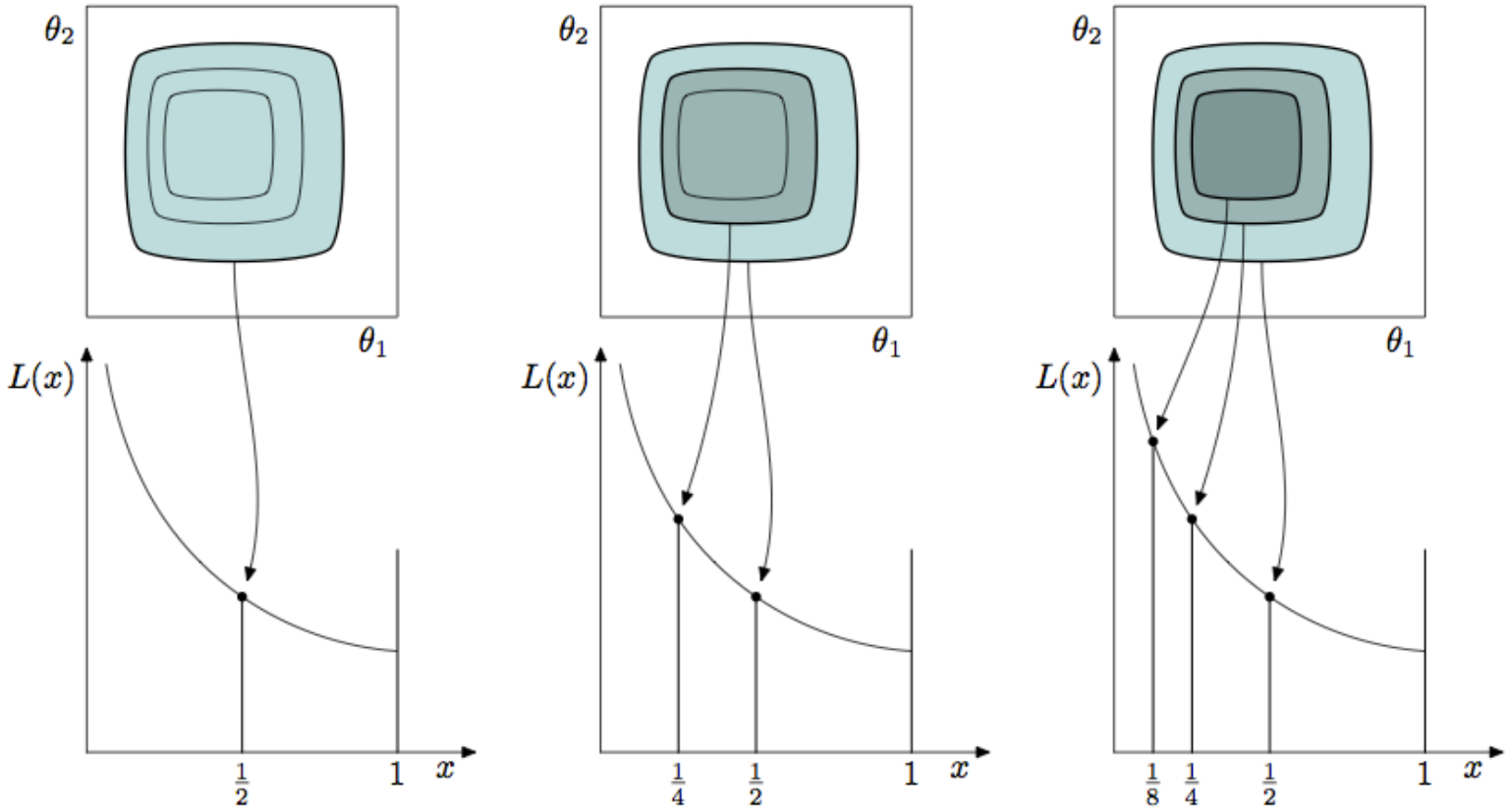


Image Credit: David Mackay (http://www.inference.phy.cam.ac.uk/bayesys/box/nested.pdf)

Basics of the algorithm

The **removed points** define a chain of increasing likelihood samples in *X*, where *X* is the fraction of the prior that contains likelihood larger than the sample

The marginal likelihood can then be calculated by integrating the likelihood over *X*

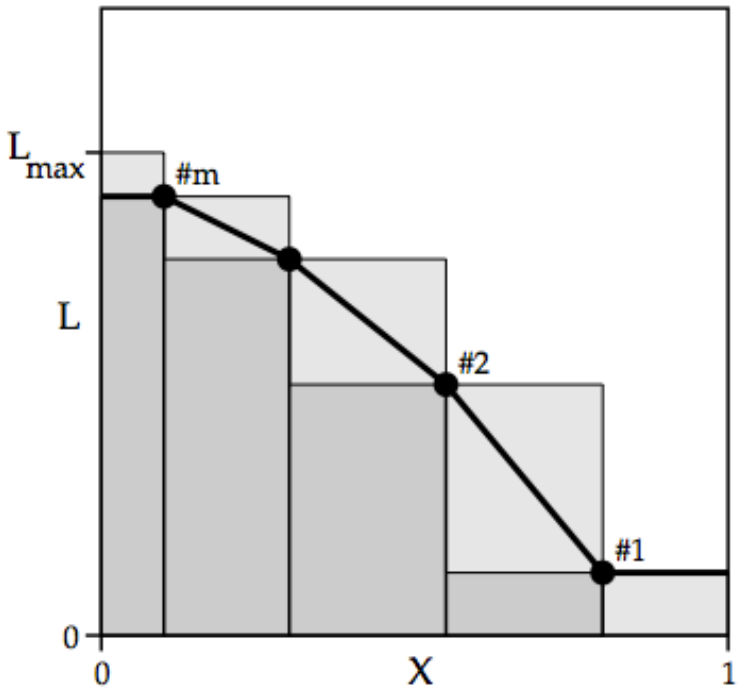
$$E = \int_0^1 L(X)\, dX$$

Basics of the algorithm

The integral can be **approximated by a sum** m samples of L_i, separated by widths w_i in X

$$E = \int_0^1 L(X)\, dX$$

$$E \leftarrow \sum_{i=1}^{m} L_i w_i$$



Basics of the algorithm

We can approximate the widths w_i using one of the following rules:

Rectangle Rule

$$w_i = X_{i-1} - X_i$$

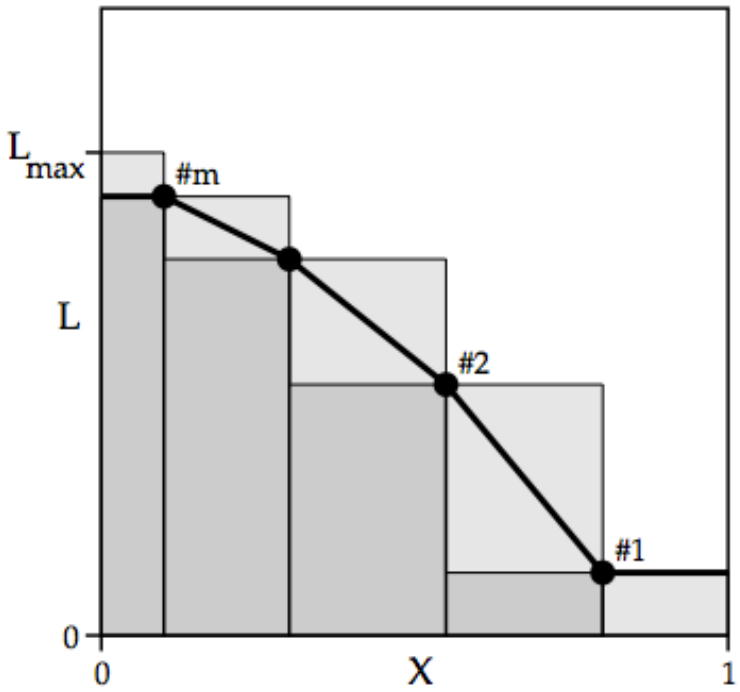
Error = $\mathcal{O}(1/m)$

Trapezoidal Rule

$$w_i = \frac{1}{2}(X_{i-1} - X_{i+1})$$

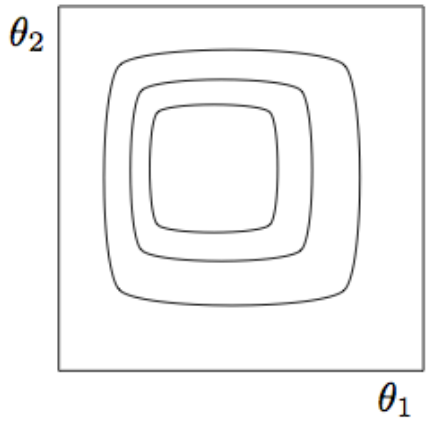
Error = $\mathcal{O}(1/m^2)$

Simpson's Rule etc…



Estimating the integral

The chain of rejected points define our L_i but what about the corresponding X_i required to calculate our widths?



The problem is for a general likelihood distribution we can't easily calculate the fraction of the prior volume X_i

Estimating the integral

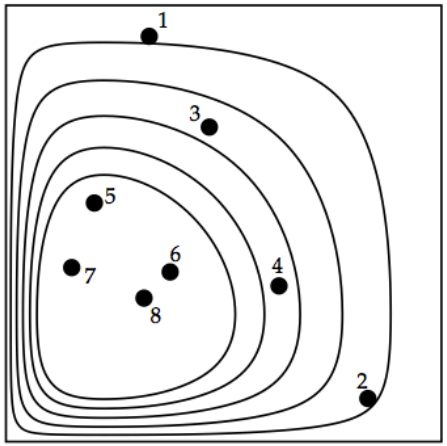
However it turns out that we can estimate X statistically

At each iteration the prior volume X_i occupied by the active samples shrinks by a value that has a well defined mean and variance

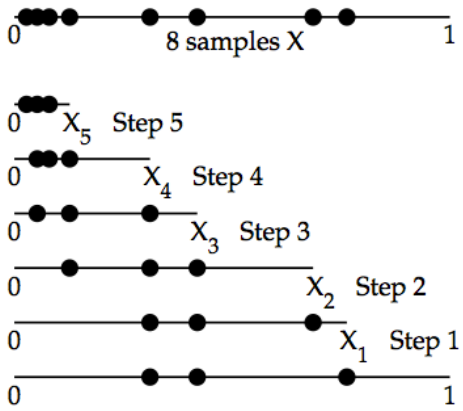
The uncertainty in this statistical estimate of X_i becomes less for large values of active samples N

Estimating the integral

X at each iteration is related to its previous value by a geometric sequence defined by the "shrinkage" factor *t*:



Parameter space



Enclosed prior mass X

$$X_i = t_i X_{i-1}$$

$$X_1 = t_1,$$
$$X_2 = t_1 t_2,$$
$$X_i = t_1 t_2 \ldots t_i,$$
$$X_m = t_1 t_2 \ldots t_m$$

Estimating the integral

We can use this behaviour to derive a statistical estimate of the amount by which the prior volume shrinks at each iteration

- $x_1, x_2 \ldots x_n$ (the positions of our active samples) are n independent uniform random variables between 0 and 1

- The probability $\Pr(x_i < x) = x$

Estimating the integral

- t is the prior volume contained by the active sample with lowest likelihood, selected to be added to the chain. So $t = \max\{x_1, x_2 \ldots x_n\}$

- $t < x$ if and only if $x_i < x$ for all i

$\Pr(t < x) = \Pr(x_1 < x) * \Pr(x_2 < x) * \ldots * \Pr(x_n < x) = x^n$

- The probability distribution of t values for a given iteration is equal to the differential of this cumulative probability

Estimating the integral

At each iteration, the amount by which the prior volume contained by the active samples shrinks is therefore distributed with probability

$$\Pr(t) = Nt^{N-1}$$

And the geometric mean (since we often operate in log space) and STD is given by

$$< \log t >= -\frac{1}{N} \qquad \sigma[\log t] = \frac{1}{N}$$

Estimating the integral

So at the ith iteration we can estimate the value of X_i based on the product of expected values of t (setting $X_0 = 1$):

$$\begin{aligned} \mathrm{X}_1 &= t_1, \\ \mathrm{X}_2 &= t_1 t_2, \\ \mathrm{X}_\mathrm{i} &= t_1 t_2 \ldots t_\mathrm{i}, \\ \mathrm{X}_\mathrm{m} &= t_1 t_2 \ldots t_\mathrm{m} \end{aligned}$$

$$X_\mathrm{i} \approx \exp\left(-i/N\right)$$

We have estimates of our widths
now for the integral of the likelihood!

$$w_\mathrm{i} = \frac{1}{2}(X_{\mathrm{i}-1} - X_{\mathrm{i}+1})$$

Estimating the integral

- We now have everything we need to estimate the marginal likelihood

1 $$X_{\mathrm{i}} \approx \exp\left(-i/N\right)$$

2 $$w_{\mathrm{i}} = \frac{1}{2}(X_{\mathrm{i}-1} - X_{\mathrm{i}+1})$$

3 $$E \leftarrow \sum_{\mathrm{i}=1}^{m} L_{\mathrm{i}} w_{\mathrm{i}}$$

Working in logarithmic space

In practice we sample the logarithm of the likelihood, rather than the absolute value of the likelihood itself

Reasons for this include:

- Avoids very large and small numbers (which often occurs for many physical applications)

- In the case of nested sampling the sequence is geometric, and so the sampling is naturally in logarithmic space

- Uncertainties are symmetric in logarithmic space

So instead we estimate $\log(E)$ by calculating the logarithm of the sum of $L_i w_i$

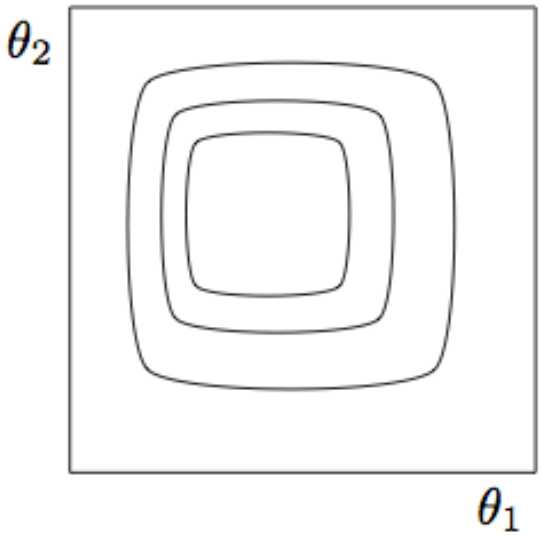
Choosing the replacement active sample

The hardest problem in Nested Sampling is sensibly choosing the replacement active point

Conditions are:

- New active point must be drawn from the prior
- New likelihood must be greater than L_i

Problem: As X_i tends to 0 the acceptance rate decreases steadily, reducing the efficiency of the algorithm



Choosing the replacement active sample

Solution:

Use Markov-Chain Monte Carlo to choose the new active point

1) Choose a surviving active point at random
- This ensures that the likelihood is sensible
- However we now need to loose knowledge of this point, otherwise we bias the sampling

2) Evolve this point by taking a small step in the prior (e.g. 0.1)

Choosing the replacement active sample

3) Accept this trial point subject to the constraint $L > L_i$ else reject

4) If number of acceptances greater than number of rejections then increase step size, else decrease

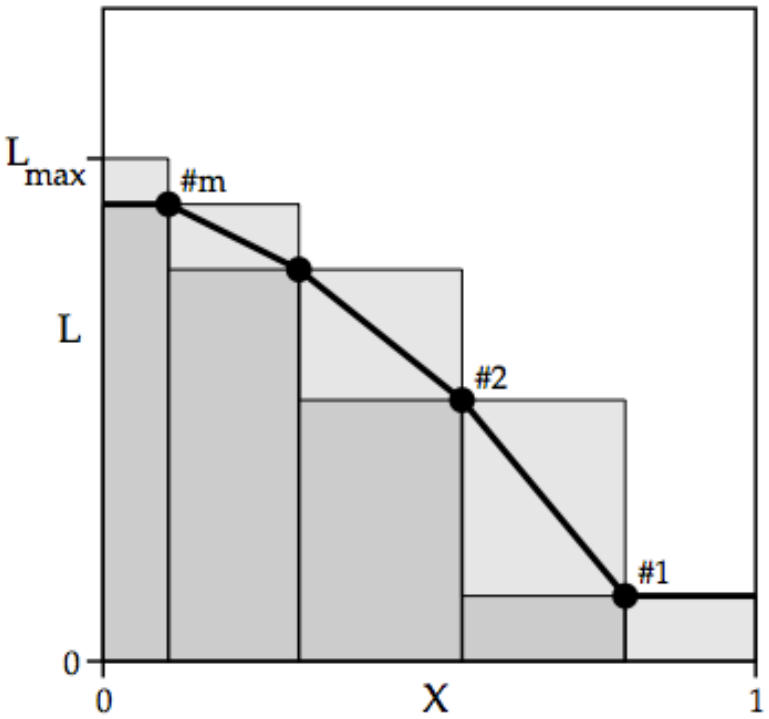
- This ensures that information is lost about the starting point and that the acceptance rate converges

5) Continue for approximately 20 or more steps

The termination condition

We can now estimate the marginal likelihood *E* for a large number of *N* active samples over *m* iterations

But how do we know when to stop? What is a good value of *m*?



The termination condition

At each iteration the largest possible contribution to the integral *E* from the remaining portion of the enclosed prior volume is given by:

$$\Delta E_{\mathrm{i,max}} = L_{\mathrm{max}} X_{\mathrm{i}}$$

L_{max} is the largest likelihood value found in the current sample of the *N* active samples

The iterations are stopped when the above maximum change in *E* is less than the required tolerance e.g. 0.1 in $\log(E_{\mathrm{i}})$

Uncertainty in the marginal likelihood

The power of Nested Sampling lies in being a simple method for estimating the marginal likelihood for model comparison

Unlike most MCMC methods, which require **multiple chains**, Nested Sampling can estimate the uncertainty in the marginal likelihood from **just a single run**

Uncertainty in the marginal likelihood

Contributions to the uncertainty in the marginal likelihood are:

- The discretization of the integral due to the width

e.g. rectangle rule $\mathcal{O}(1/m)$ or trapezoidal rule $\mathcal{O}(1/m^2)$

- The cumulative uncertainty in the statistically estimated value of *X* at each iteration

For an adequate number of iterations *m* the first error will be negligible and the second error dominates

Uncertainty in the marginal likelihood

The uncertainty in *E* is dominated by iterations that contribute the bulk of the likelihood

This region in prior has size approximately equal to

$$\log X = -H$$

where the quantity *H* is a measure of the "information" gained from including the data, over that from the prior, and is given by

$$H = \int \log\left(\frac{\mathrm{d}P}{\mathrm{d}X}\right) \mathrm{d}X \approx \sum_{\mathrm{i}=1}^{\mathrm{m}} \frac{L_{\mathrm{i}} w_{\mathrm{i}}}{E} \log\left(\frac{L_{\mathrm{i}}}{E}\right)$$

Uncertainty in the marginal likelihood

Recapping, the expected value and standard deviation in log X_i is given by

$$\log X_i \approx -(i \pm \sqrt{i})/N$$

If most of the contribution to *E* comes from the region

$$\log X = -H$$

Then the error in the estimate of the log(*X*) contributing most to *E* is $\sqrt{H/N}$

So $$\log(E) \approx \log\left(\sum_{i=1}^{m} L_i w_i\right) \pm \sqrt{\frac{H}{N}}$$

The model parameters

What about parameter estimation?

The resulting chain of rejected samples also provides an estimate of the posterior probability for the model parameters

$$\Pr(\boldsymbol{\theta}_\mathrm{i}|\boldsymbol{d},\mathcal{M}) = \frac{L_\mathrm{i} w_\mathrm{i}}{E}$$

Mean

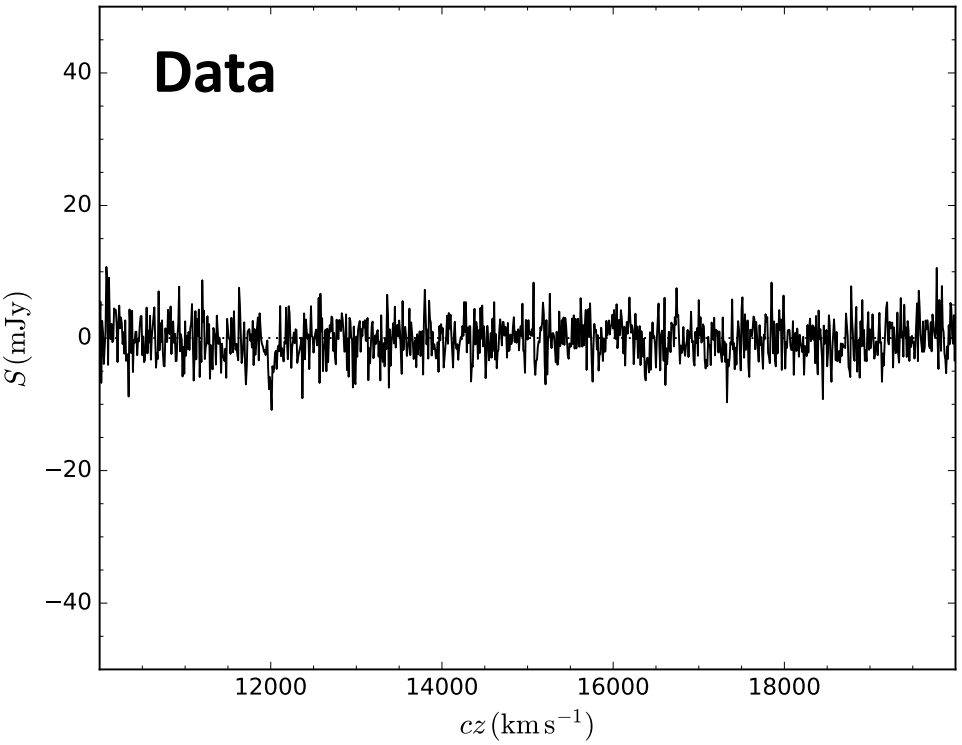
$$\langle Q \rangle = \sum_{\mathrm{i}=1}^{m} \frac{L_\mathrm{i} w_\mathrm{i}}{E} Q(\boldsymbol{\theta}_\mathrm{i})$$

Standard Deviation

$$\sigma[Q] = \left(\sum_{\mathrm{i}=1}^{m} \frac{L_\mathrm{i} w_\mathrm{i}}{E} Q(\boldsymbol{\theta}_\mathrm{i})^2 - \langle Q \rangle^2 \right)^{1/2}$$

Application to a science problem

Example: Detection of a spectral line in a low signal-to-noise spectrum



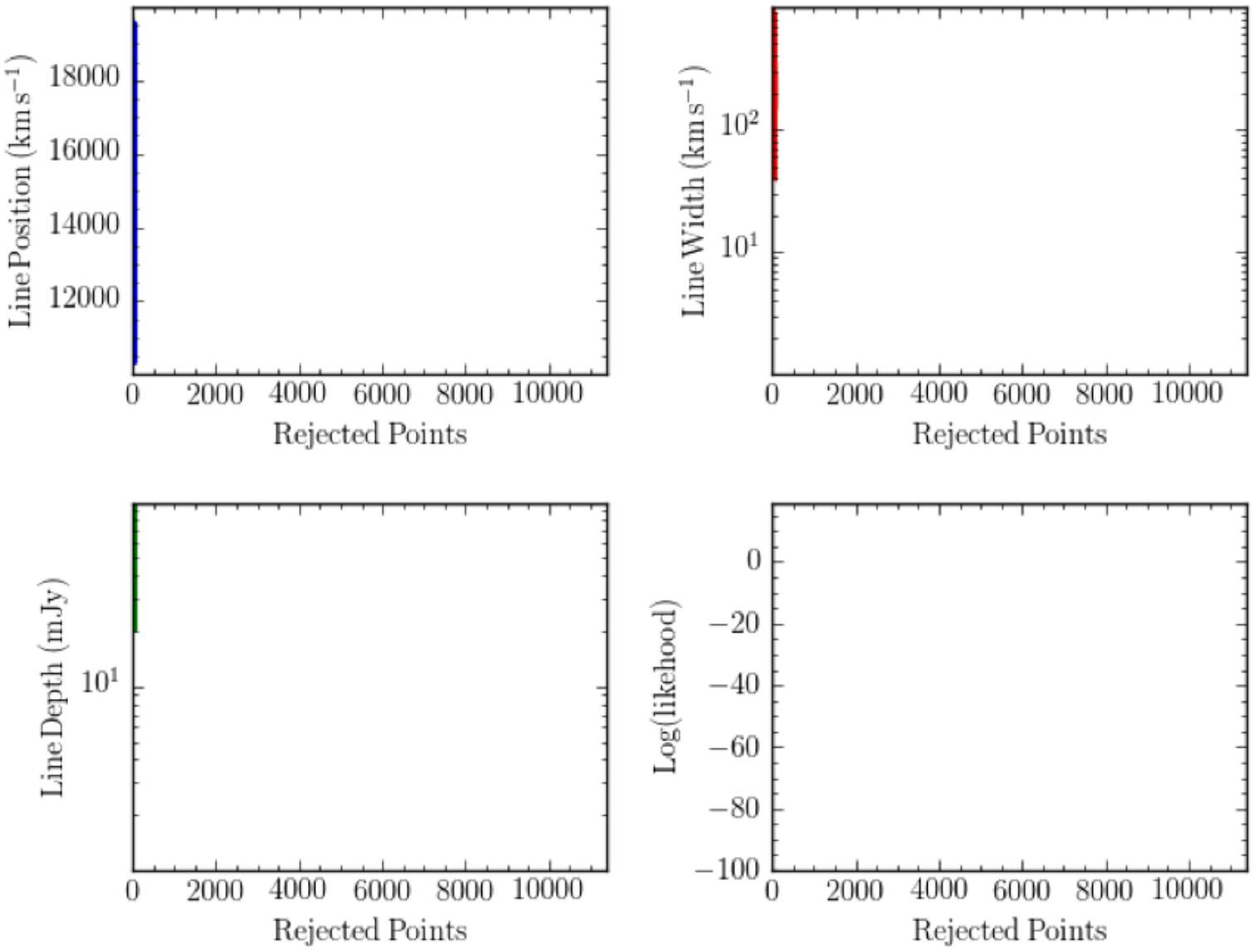
Likelihood function:

$$L = \frac{1}{\sigma^N \sqrt{(2\pi)^N}} \exp\left[-\frac{\sum_i (d_i - m_i)^2}{2\sigma^2}\right]$$

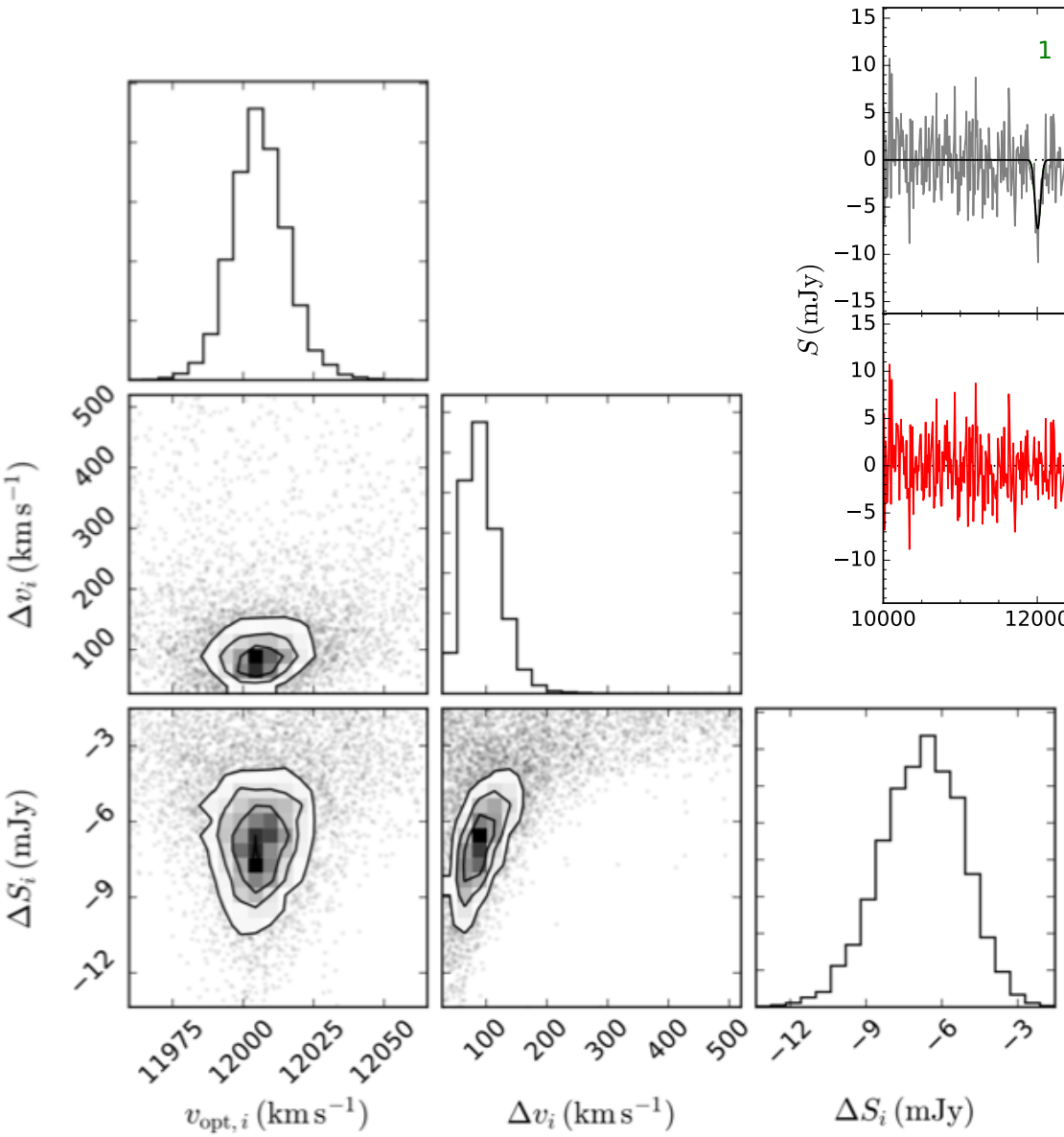
Hypothesis 1: No spectral line (null)

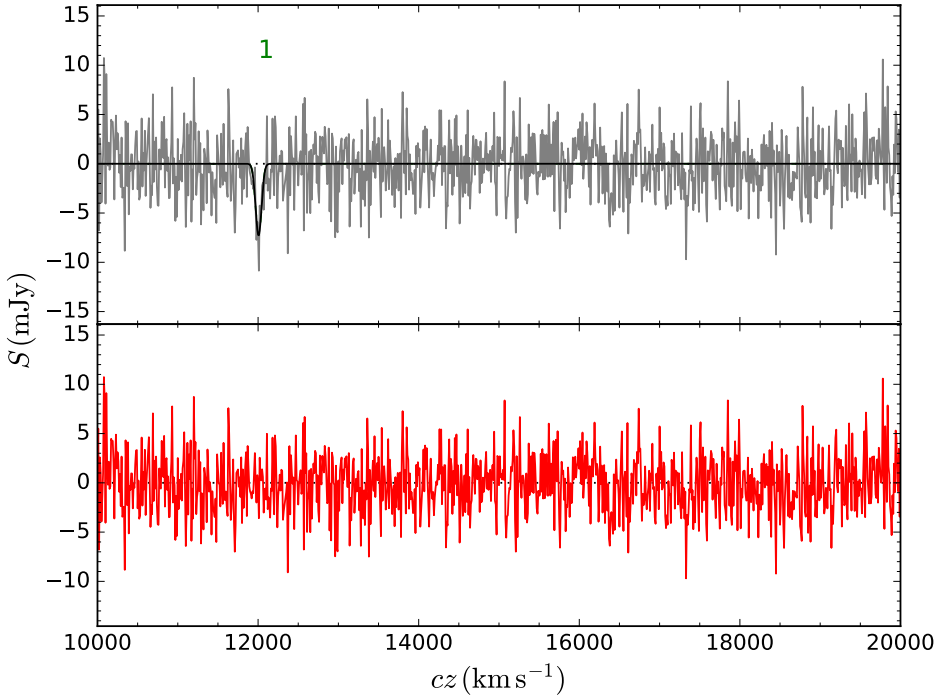
Hypothesis 2: Gaussian model of spectral line

Application to a science problem



Application to a science problem





Log(E) = 8.43 +/- 0.10

Pr(Line) = 99.98%

Further Reading

- "Data Analysis: A Bayesian Tutorial", 2nd Edition, D. S. Sivia & J. Skilling
- http://www.inference.phy.cam.ac.uk/bayesys/nest.pdf
- "Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses", F. Feroz & M. P. Hobson (2008), MNRAS

Python module to run multi-modal Nested sampling

- http://johannesbuchner.github.io/PyMultiNest/

Extra slides

Recap: The procedure

Steps:

1) Identify the prior and likelihood function for the problem
2) Initialize $X_0 = 1$ and $E = 0$
3) Take N random points ("active points") sampled from the prior distribution
4) Re-order the N samples by their likelihood values
5) Select the lowest likelihood sample point L_i
6) Set $X_i = \exp(-i/N)$ and $w_i = (X_{i-1} - X_{i+1})/2$
7) Increment E by $L_i w_i$
8) Check maximum change in E for termination condition (true = go to 11)
9) Replace with a new sample drawn from prior and condition $L > L_i$
10) Start again from 5
11) Increment E by final contribution from central prior region

$$N^{-1}(L(\theta_1) + \ldots + L(\theta_N))X_m$$