

# Lecture 13: Nested Sampling for Bayesian Inference

D. Jason Koskinen  
[koskinen@nbi.ku.dk](mailto:koskinen@nbi.ku.dk)

*Advanced Methods in Applied Statistics*  
*Feb - Apr 2020*

Photo by Howard Jackman  
University of Copenhagen

Niels Bohr Institute

# Comments

- For the following nested sampling lecture, I have included more references at the end of the slides as well as on the course webpage

# Bayes' Theorem (from Lecture 5)

- One can solve the respective conditional probability equations for  $P(A \text{ and } B)$  and  $P(B \text{ and } A)$ , setting them equal to give Bayes' theorem:

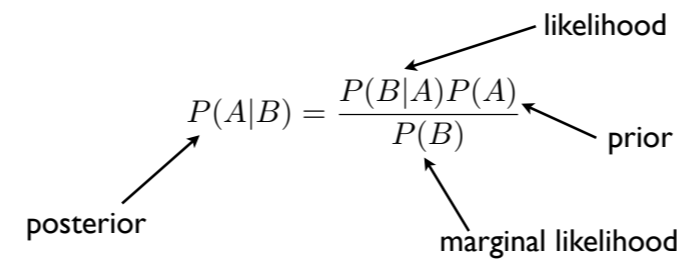
$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$


Diagram illustrating the components of Bayes' theorem:

- posterior:  $P(A|B)$
- likelihood:  $P(B|A)$
- prior:  $P(A)$
- marginal likelihood:  $P(B)$

$$\text{posterior} \propto \text{prior} \times \text{likelihood}$$

- The theorem applies to both frequentist and Bayesian methods. Differences stem from how the theorem is applied and, in particular, whether one extends probability to include some degree of belief.

# Slight Notation Shift

- Previously, we have focused on the posterior distribution  $P(\Theta|D,H)$  which is critical for parameter estimation and we used Markov Chain Monte Carlo for calculating the marginal likelihood  $P(D|H)$
- For model selection — versus parameter estimation — the marginal likelihood is important in its own right. The problem is that many MCMC methods are slow (simulated annealing).

$$P(\Theta|D, H) = \frac{P(D|\Theta, H) P(\Theta|H)}{P(D|H)}$$

$D$  are data

$\Theta$  are parameters

$H$  is hypothesis or model

–  $P(D|H)$  is the bayesian evidence

–  $P(D|H)$  is the same as the likelihood  $P(D|H,\theta)$ , but with  $\theta$  integrated out. This is why it's also known as the marginal likelihood.

# New Task

- If model selection is important then comparing models can be done via the respective posterior distributions

$$\frac{P(H_1|D)}{P(H_0|D)} = \frac{P(D|H_1)P(H_1)}{P(D|H_0)P(H_0)} = \frac{Z_1P(H_1)}{Z_0P(H_0)}$$

- The “marginal likelihood” is now rebranded as the “Bayesian evidence” and noted as  $Z$
- Reversing the traditional MCMC approach, the ‘evidence’ is now the primary target, and the posterior is a by-product
- Note: we won’t be doing model selection explicitly in this lecture, but it is the motivation for much of the following material

– Mention that  $Z_1/Z_0$  is the Bayes Factor.

# Nested Sampling

- In 2004, John Skilling came up with a new Monte Carlo sampling technique, known as nested sampling, to more efficiently evaluate the bayesian evidence ( $Z$ )

$$Z = \int \mathcal{L}(\Theta)\pi(\Theta)d\Theta$$

$\mathcal{L}$  is the likelihood

$\pi$  is the prior

- For higher dimensions of  $\Theta$  the integral for the bayesian evidence becomes challenging

- Could raster scan across  $\theta$ , but after a few dimensions this becomes impractical

# Nested Sampling

- If numerical integration in higher dimensions is troublesome, then we can transform the multi-dimensional integral to a one-dimensional integral, via

$$dX = \pi(\Theta)d\Theta$$
$$X(\lambda) = \int_{\mathcal{L}(\Theta) > \lambda} \pi(\Theta)d\Theta$$

- The new prior  $X$  is defined such that

$$Z = \int_0^1 \mathcal{L}(X)dX$$

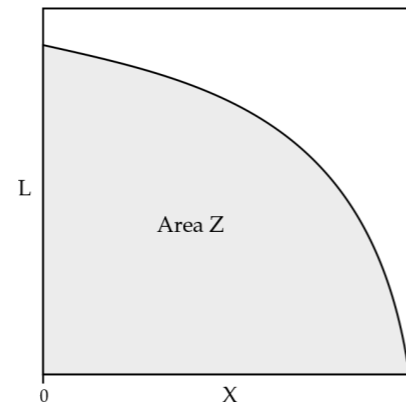
\*For more justification,  
see the original paper  
by J. Skilling

- Note that  $X$  is a probability (mass) function and can only be in the range from 0 to 1
- $\mathcal{L}(X)$  is also now a monotonically decreasing function
- A clever approx. to get  $X$  will be covered in later slides

– for high values of likelihood ( $\lambda$ ), the prior space is reduced and goes towards zero. As  $\lambda$  increases the enclosed mass  $X$  decreases from 1 to 0.

# New Likelihood in 1-D

- The bayesian evidence (Z) is now the 1-D integral of the re-parameterized likelihood (L(X)) integrated over the re-parameterized prior (X)
  - The shape of L(X) could be any shape, but it **is** monotonically decreasing from 0→1, and by construction is bounded at 0 and 1.



$$dX = \pi(\Theta)d\Theta$$

$$X(\lambda) = \int_{\mathcal{L}(\Theta) > \lambda} \pi(\Theta)d\Theta$$

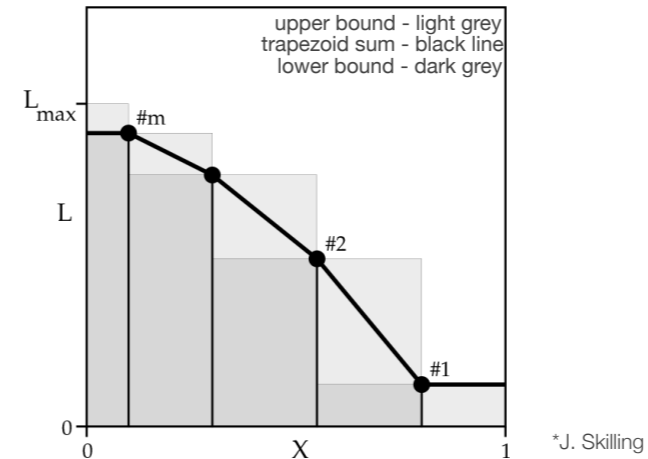
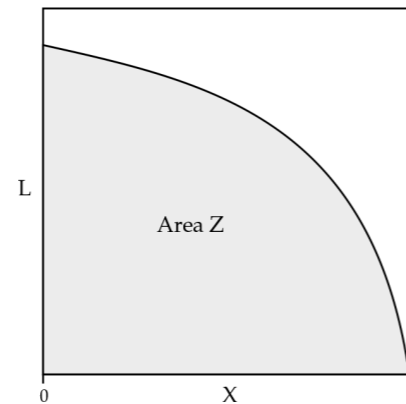
$$Z = \int_0^1 \mathcal{L}(X)dX$$

\*J. Skilling



# New Likelihood in 1-D

- The bayesian evidence is now the 1-D integral of the re-parameterized likelihood integrated over the re-parameterized prior
  - An analytic determination of the integral is not an option. If we could do it analytically, we wouldn't be using numerical integration.
  - Use points sampled in  $X$  to calculate the trapezoid sum
  - Diagram below (right) shows  $X$  and  $L$  for 4 sampled points



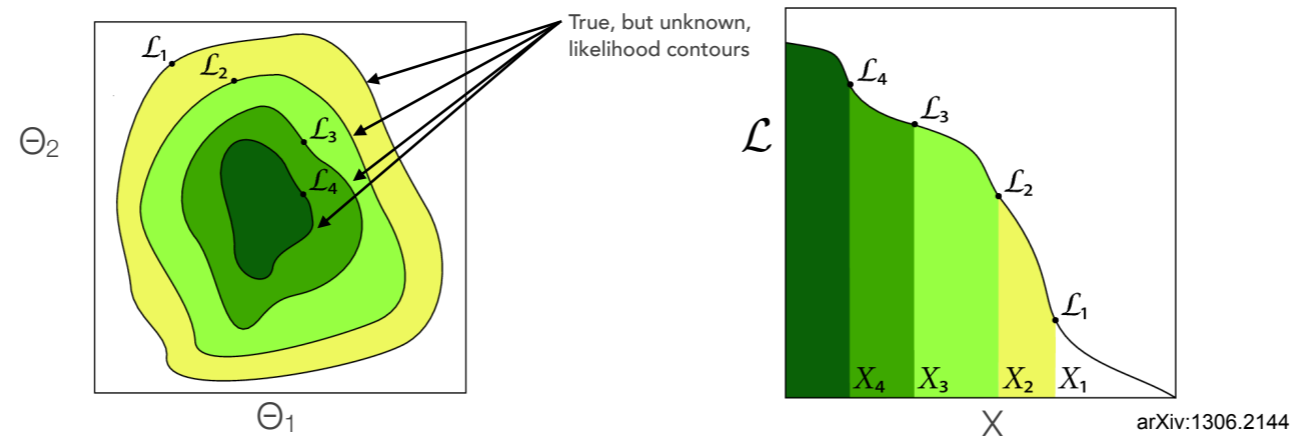
D. Jason Koskinen - Advanced Methods in Applied Statistics

9

– Each 'bin' has some given width  $\Delta X$ , and thus the area of that bin is  $\Delta X_i L_i$ . Because the bin width is determined as the distance between evaluated points, and the  $L_i$  is decreasing for increasing  $i$ , then the integral is the  $Z \geq \sum (X_i - X_{i+1}) L_i$ , or  $Z \leq \sum (X_{i-1} - X_i) L_i + X_{\max} L_{\max}$  depending on whether you're using the upper bound (light grey), lower bound (dark grey), or the trapezoidal sum (black).

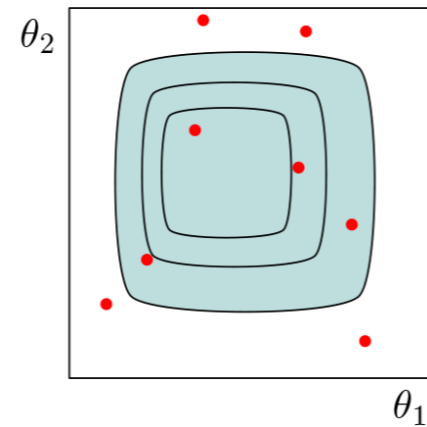
# Simple Cartoon

- For a simple 2-dimensional case, 4 'live' points are drawn at random. The likelihood for each point is calculated, and has an associated value of  $X$ .
  - Note that multiple points of  $\Theta_1$  and  $\Theta_2$  can have the same value of  $X$
  - This illustration nicely samples the space with only 4 points, which is uncommon and unrealistic



# Sampling

- Instead of relying on luck, it is better to sample the space sparsely where the new likelihood is low, and sample frequently in the space where the likelihood is high(er)



Shaded areas are the true underlying contours

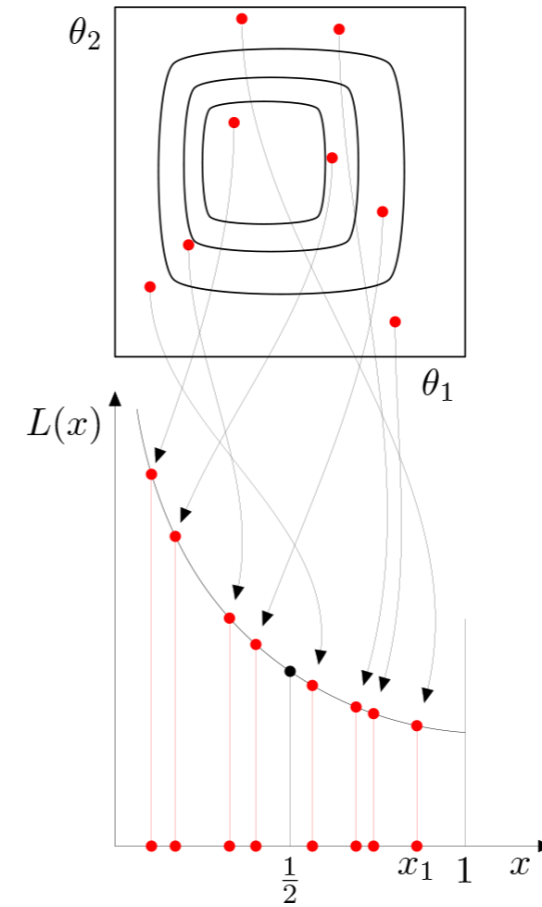
It is a flat prior in 2-D

Figure 51.3.  $N = 8$  points drawn uniformly from the prior.

<http://www.inference.phy.cam.ac.uk/bayesys/box/nested.pdf>

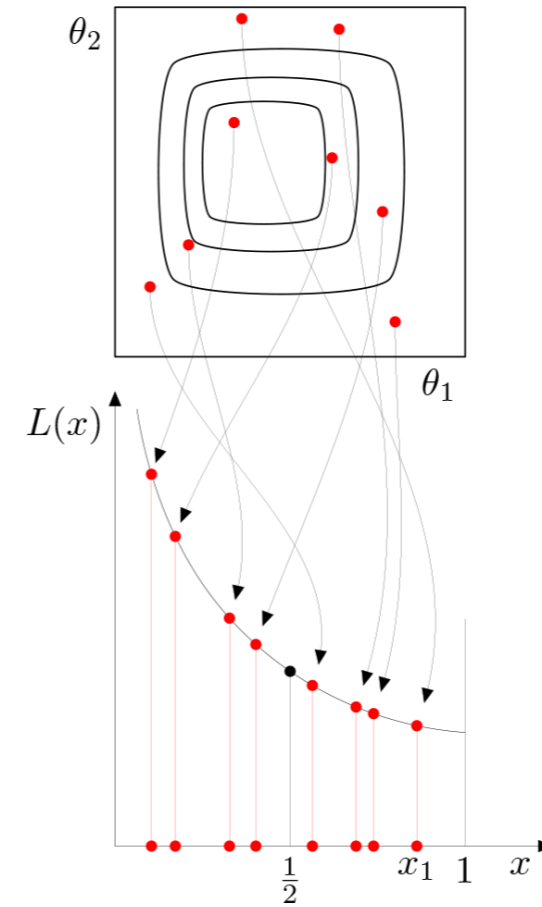
# Sampling Start

- Each of the 8 initial live points has a likelihood value  $L(x_i)$  that can be ordered:  $L(x_1) < L(x_2) < L(x_3) < L(x_4) < L(x_5) < L(x_6) < L(x_7) < L(x_8)$
- To get the  $x$ -values, 8 values are drawn from a uniform distribution in the range 0-1, and the largest  $x$ -value is defined as  $x_1$ 
  - Second largest  $x$ -value is  $x_2$ , third largest is  $x_3$ , etc.
- Can we use more than just the initial 8 points in some smart way?
  - **Absolutely!!**



# Sampling Start cont.

- In order to better sample where the likelihood is high, the point with the lowest  $L(x)$ , i.e.  $x_1$  in the diagram, is replaced by a new point  $x'$ 
  - A new point  $(\theta_1, \theta_2)$ , equivalently  $x'$ , is drawn from the prior which produced the initial points. Now in the range  $0 < x' < x_{\text{lowest}}$
  - $x'$  must satisfy that  $L(x') > L(x_{\text{lowest}})$
  - Remove the point  $x_{\text{lowest}}$ , but store it's values to calculate the likelihood integral, e.g. bayesian evidence
- Next slide covers other approx. for values of  $x$



# Pseudo-Code

Generate  $n$  points from the prior

Loop where  $i$  increments as  $i=1,2,3,\dots$

{

\* Find the point  $X_{\text{worst}}$  with the lowest likelihood,  $L_{\text{worst}}$ .  
Remove it from the population, but store it for  
results. Estimate the value of  $X_{\text{worst}}$  as  $((N-1)/N)^i$ , for  
 $N$  live points

\* Add a new livepoint generated from the prior.  
The new live point must satisfy that  $L(X_{\text{new}}) > L(X_{\text{worst}})$ .

}

Other estimates of  $X$  can be

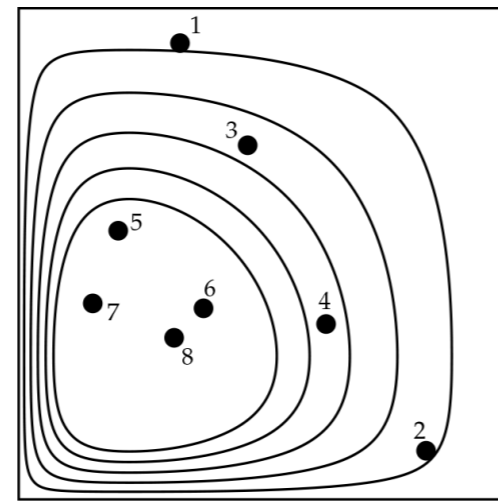
\*  $((N-1)/N)^i$

\*  $(N/(N+1))^i$

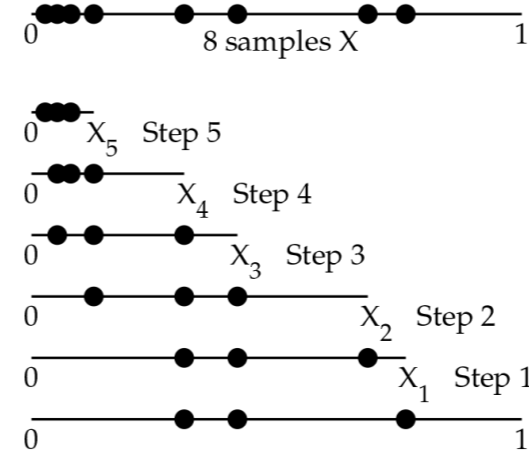
\*  $\exp(-i/N)$

\*G. F. Lewis

# Sampling more



Parameter space



Enclosed prior mass  $X$

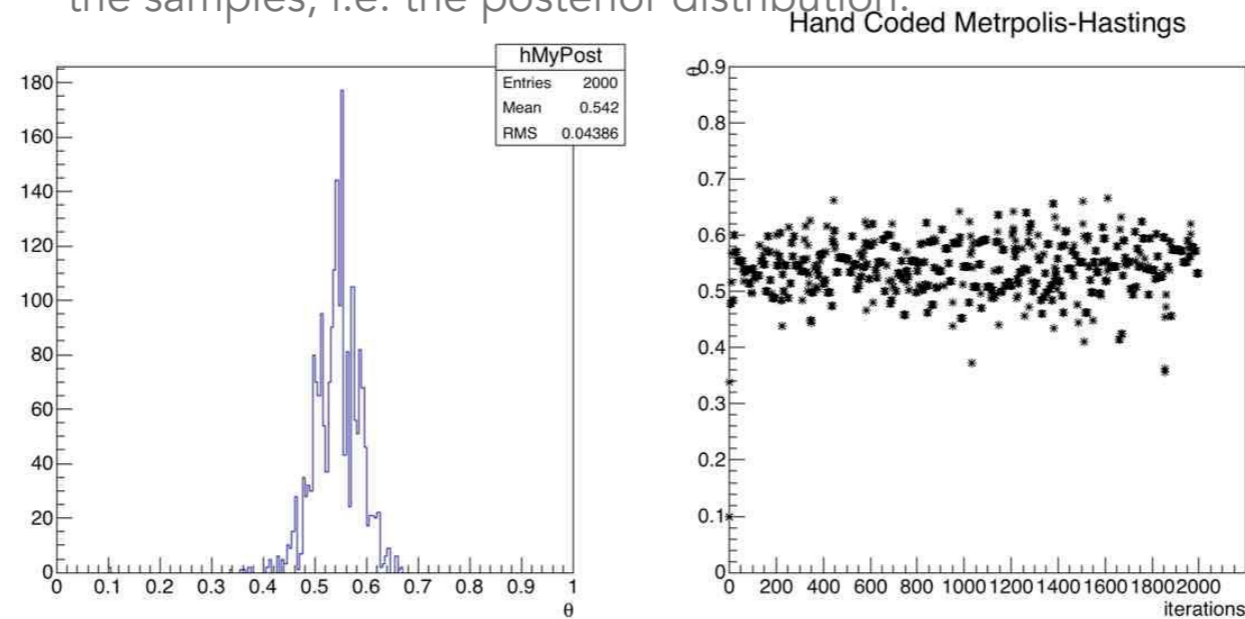
Figure 4: Nested sampling for five steps with a collection of three points. Likelihood contours shrink by factors  $\exp(-1/3)$  in area and are roughly followed by successive sample points.

\*J. Skilling 2006

## Exercise #3 (cont.)

\*Reminder from the lecture  
about Markov Chain Monte  
Carlo

- For 2000 iterations plot Markov Chain Monte Carlo samples as a function of iteration, as well as a histogram of the samples, i.e. the posterior distribution.

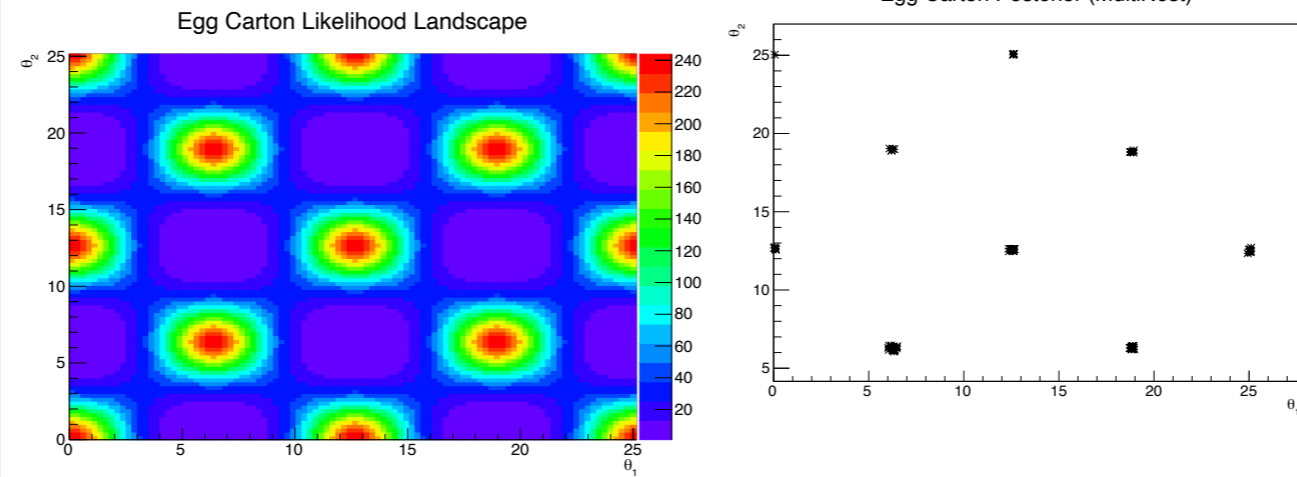


– We start to sample the posterior as the iterations progress



# Nested Sampling in Action

- The 'Egg Carton' likelihood landscape is a benchmark likelihood landscape for difficulty and stress testing of bayesian sampling techniques



# Nested Sampling Benefits

- Samples sparsely in low likelihood regions and samples densely where the likelihood is high
- Can handle irregular likelihood landscapes
- Many applications require nothing more than setting the range over which to generate 'live points'
  - Does not require lots of tuning
  - Most of the time the sampling prior is uniform, i.e. flat
- The true value of the maximum likelihood estimator is not essential to be known, it just needs to be within the region where the points are sampled
- Efficient when compared to other MCMC methods

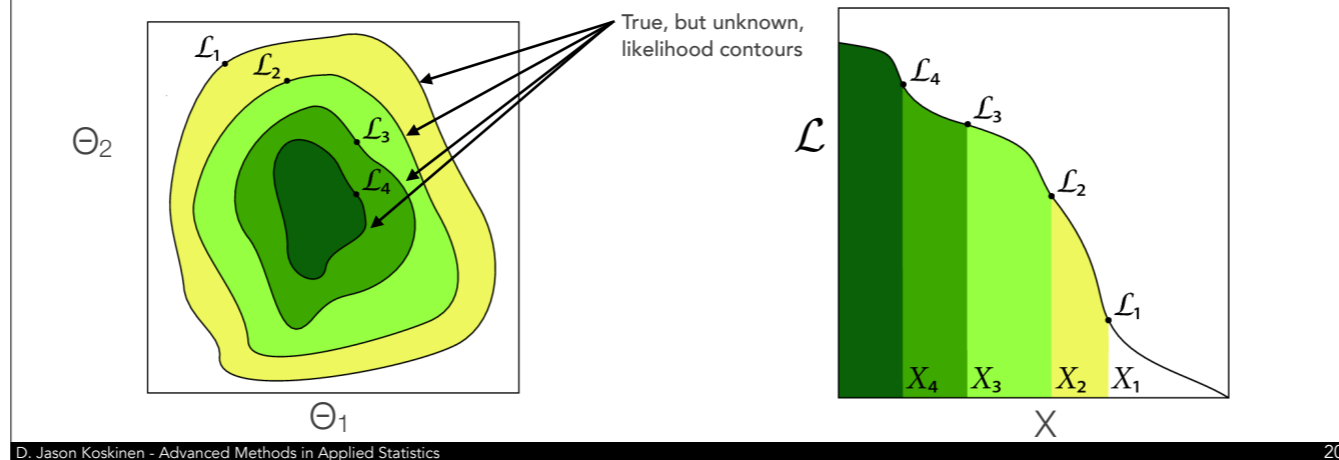
# Cons

- Similar to every other fitting technique, there is no guarantee that any best-fit values are global best-fit values
- No rigorous termination criterion
  - There is always the possibility that there exist some unsampled regions in  $X$  which have very large likelihood values which will contribute to the bayesian evidence value  $Z$
- Unlike other MCMC algorithms which sample near the current point, many nested sampling algorithms sample uniformly over the full parameter space
  - Higher dimensions can see slow-downs
- Trapezoidal summing will induce some uncertainty and *possibly* small bias

– A potential small bias is better than a large bias from getting trapped in a local minima or not converging to the maximum a posteriori (MAP)

# Big Issue

- How do we actually sample new nested points  $X'$  that are better than the current  $X_{\text{lowest}}$ , where  $X_{\text{lowest}}$  has the lowest likelihood?
- In  $n$ -dimensions and without knowing the true likelihood contours, this is problematic.



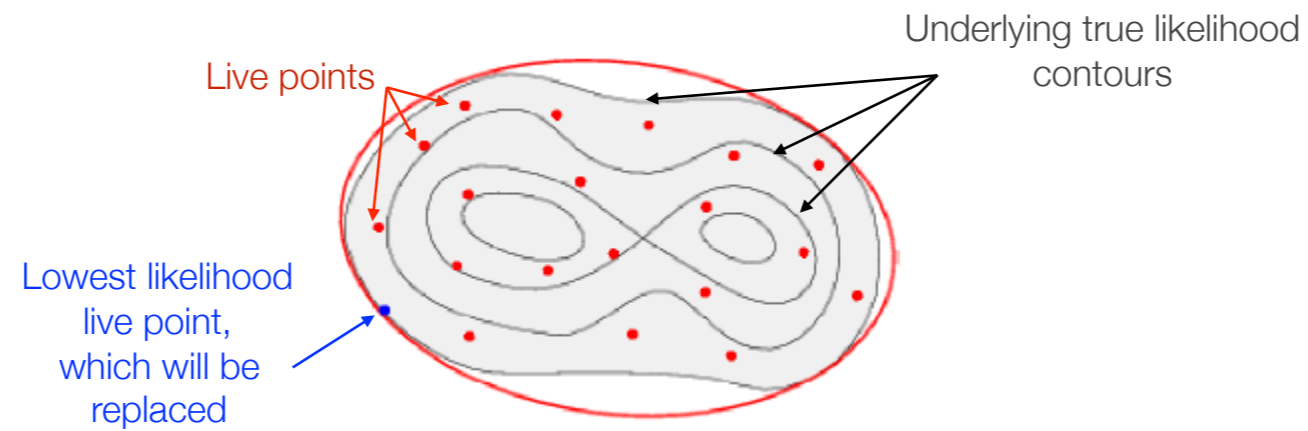
– Point out that when we have the underlying True distribution it's trivial, but the whole point is that we don't know the underlying true Bayesian Evidence or Likelihood

# MultiNest Application

- Crude nested sampling was somewhat inefficient when it came to multi-modal likelihood landscapes
  - But, much better than conventional maximum likelihood fitters when it comes to not getting stuck in local minima
- Instead of using a multi-dimensional uniform prior for each replacement point, use an n-dimensional ellipsoid for resampling
  - The hyper-ellipsoid is defined by the current iteration live points
  - The hyper-ellipsoid for re-sampling has a small enlargement margin as a safeguard

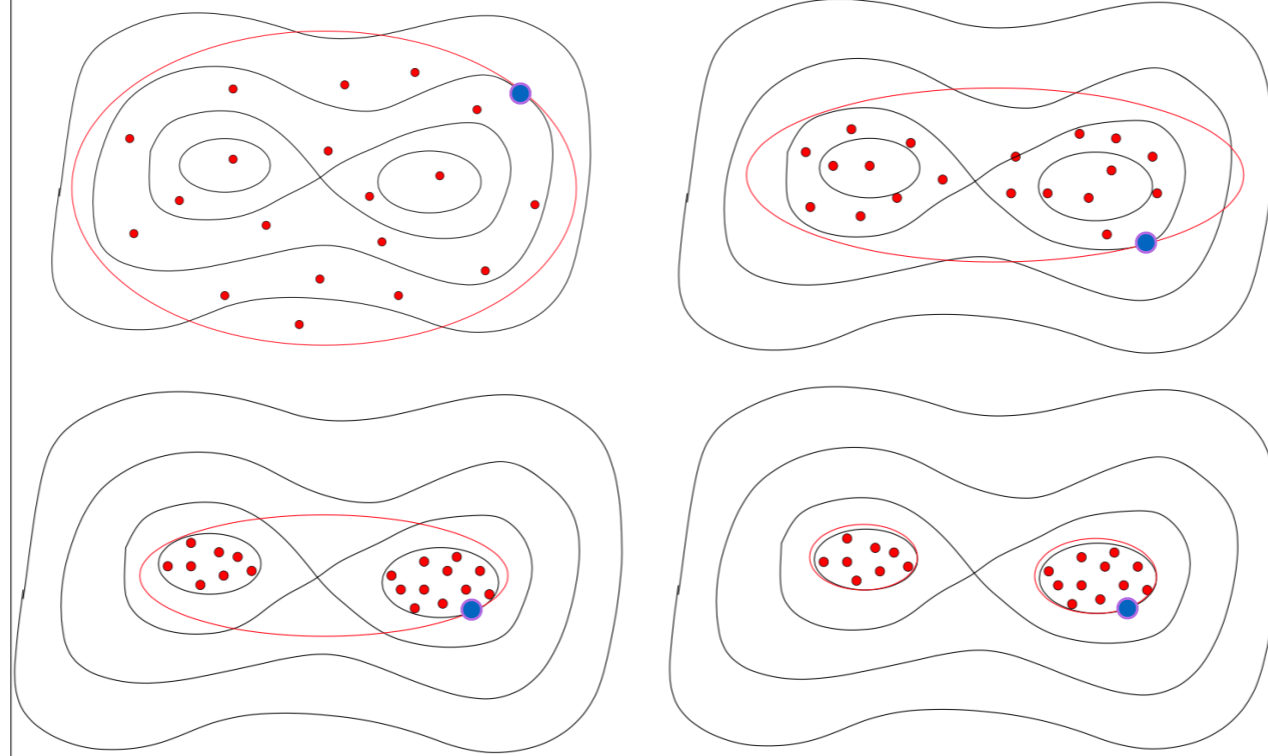
# MultiNest Ellipsoid Sampling

- Start with a sample of live points using a uniform prior in n-dimensional cube
- After a few iterations resampling within an ellipsoid we have:



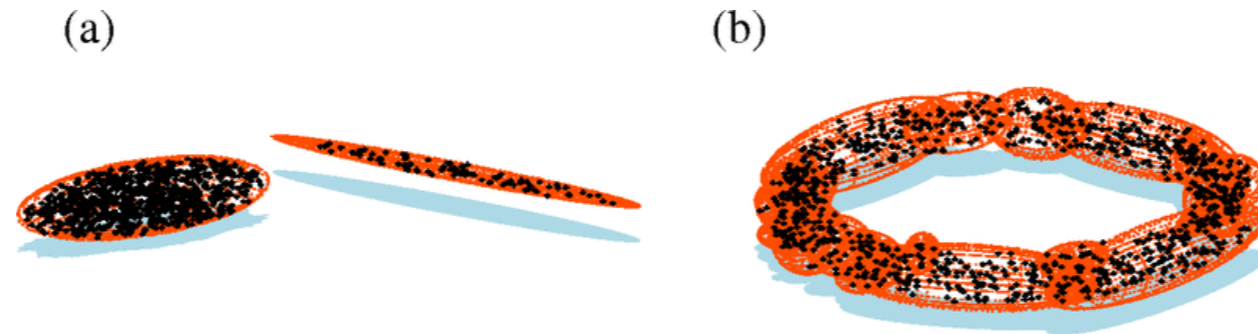
\*F. Feroz

# MultiNest Evolution



# MultiNest Pictures

- Dis-joint regions, as in fig. (a), as well as multi-dimensional multi-modal regions, as in figs. (a) and (b), can be found efficiently without continual resampling of the whole space





# Nested Sampling

- Can be an excellent method to map out a likelihood/probability landscape that is complicated
- MultiNest is very nice, but the base package requires Fortran, even though there are nice wrapper packages in other software languages

# Packages

- In Python there are a handful of nestling sampling packages
  - pymultinest (<https://johannesbuchner.github.io/PyMultiNest/>)
  - nestle (<http://kbarbary.github.io/nestle/>)
  - SuperBayeS (<http://www.ft.uam.es/personal/rruiz/superbayes/?page=main.html>)

# Exercise Egg Carton

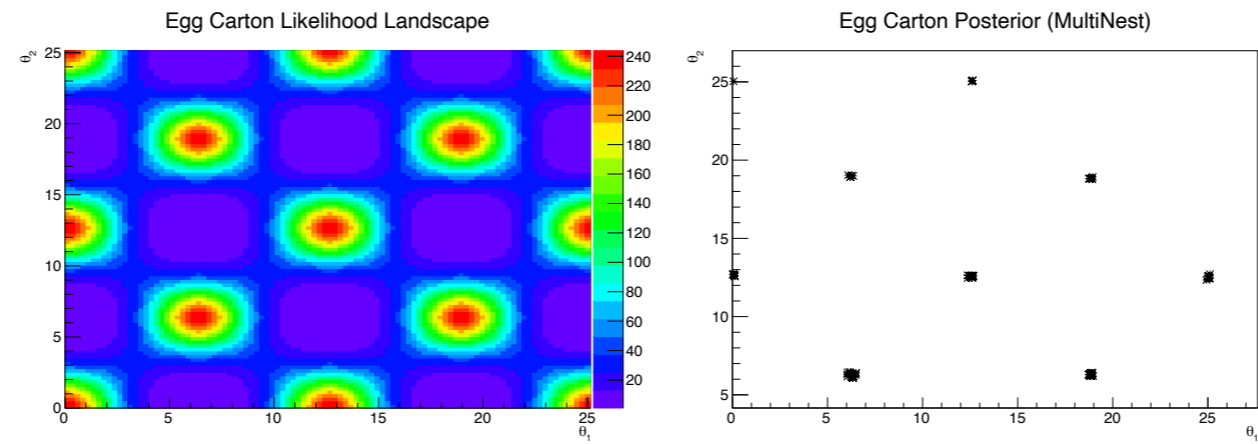
- The task is to produce a posterior distribution using a (hopefully) nested sampling algorithm for the classic 2-dimensional egg carton likelihood

$$\mathcal{L}(\theta_1, \theta_2) \propto \cos(\theta_1) \cos(\theta_2)$$

- First, make sure you have a nested sampling algorithm package installed
- Second, make a plot of the raster scan of the the 2-D likelihood for reference
- Third, make a plot of the posterior distribution from the sampling algorithm

# Exercise Egg Carton cont.

- The raster scan across  $\theta_1$  and  $\theta_2$  and the posterior distribution



# Exercise Gaussian Shell/Cylinder

- Another example is the 2- or 3-dimensional gaussian shell
  - The probability is highest, i.e. centered, on the surface of a sphere or cylinder, and has a gaussian width
  - Looking at 3D gaussian surfaces is tough, so we will do a projection into 2D for visualization

$$\mathcal{L}(\vec{\theta}) = \text{circ}(\vec{\theta}; \vec{c}, r, \sigma)$$

$$\text{circ}(\vec{\theta}; \vec{c}, r, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[ -\frac{(|\vec{\theta} - \vec{c}| - r)^2}{2\sigma^2} \right]$$

$c$  is the center of the sphere/cylinder

$r$  is the radius

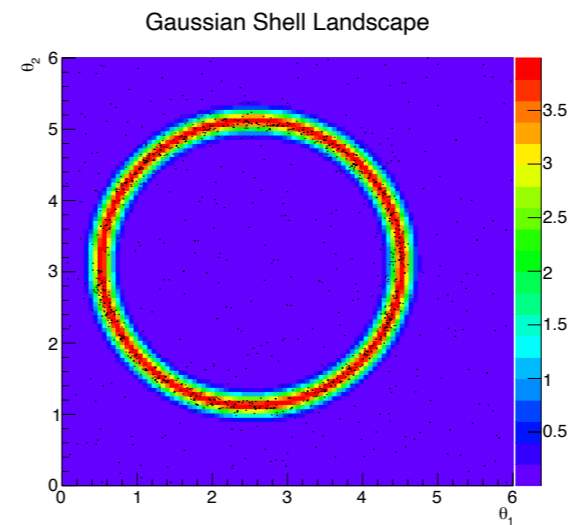
$\sigma$  is the gaussian width

$\theta$  is a/the sample point as a vector, e.g.  $(x,y,z,\dots)$  in cartesian coordinates

# Exercise Gaussian Shell/Cylinder

cont.

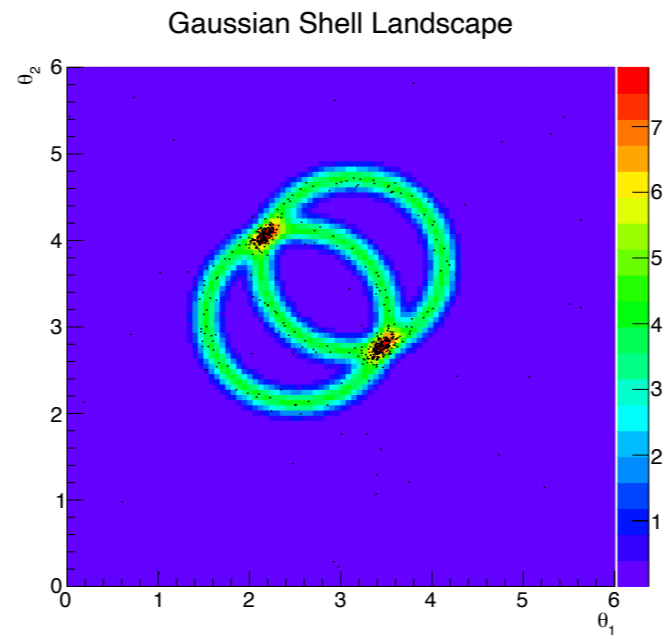
- Similar to the Egg Carton exercise generate the following plots:
  - For a single cylinder/sphere of  $r=2$ ,  $\sigma=0.1$ , centered at  $c=(2.5, 3.1)$
  - Plot the underlying probability/likelihood space
  - Plot the posterior sampling
- Note that there might be issues with the computer/machine precision when calculating  $\exp()$  or  $\ln()$  for negative, extremely large, or extremely small values related to the likelihood



# Exercise Gaussian Shell/Cylinder

cont.

- Repeat the previous task with two overlapping spheres/cylinders
  - For  $r=1$ ,  $\sigma=0.1$ , with one centered at  $c1=(2.5, 3.1)$  and the other at  $c2=(3.1, 3.7)$



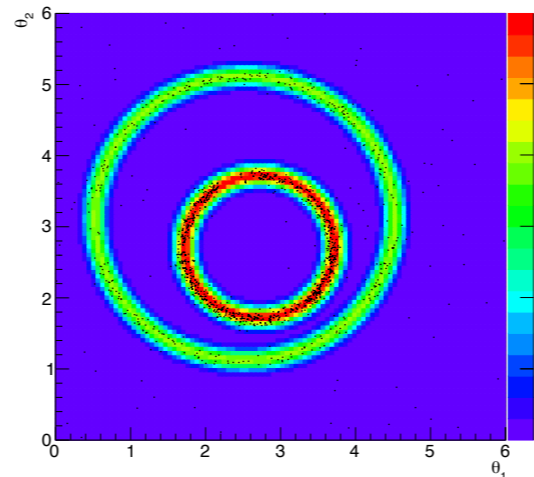
# Exercise Nested Cylinder

- Using the following likelihood for the two cylinders plot the underlying likelihood and posterior distribution:

$$\mathcal{L}(\vec{\theta}) = \text{circ}(\vec{\theta}; \vec{c}_1, r_1, \sigma_1) + 1.5 \text{circ}(\vec{\theta}; \vec{c}_2, r_2, \sigma_2)$$

- $\sigma_{1,2}=0.1$ ,  $c_1=(2.5, 3.1)$  and  $c_2=(2.7, 2.7)$  and  $r_1=2$  and  $r_2=1$

Gaussian Shell Landscape





# Extra

- Try higher dimensionality landscapes, e.g. 16-dimensions, and see if the sampler starts to slow down dramatically for the gaussian shell hyper-sphere likelihood

# References

- Excellent and readable paper by developer John Skilling
  - <http://projecteuclid.org/euclid.ba/1340370944>
- Nested sampling implementation in nice steps and detail slides ([http://www2.stat.duke.edu/~fab2/nested\\_sampling\\_talk.pdf](http://www2.stat.duke.edu/~fab2/nested_sampling_talk.pdf)). Also a sly reference to 'going down the rabbit hole'
- MultiNest
  - Slides by F. Feroz ([http://www.ics.forth.gr/ada5/pdf\\_files/Feroz\\_talk.pdf](http://www.ics.forth.gr/ada5/pdf_files/Feroz_talk.pdf))
  - Papers (<http://arxiv.org/abs/0809.3437>, <http://arxiv.org/abs/1306.2144>)