Nested sampling cross-checks using order statistics

Andrew Fowlie, Will Handley, and Liangliang Su (June 2020). In: arXiv: 2006.03371 [stat.CO]

Andrew Fowlie

July 15, 2020

Nanjing Normal University

Table of contents

1. Model selection

2. Nested sampling

3. A new cross-check

Model selection

Model selection

Throughout science, we have the following problem:

I have data and some models. What is the status of my models in light of the data?

Bayesian updates

Compute the change in plausibility of a model in light of data relative to another model or set of models.

We just apply probability theory to the problem. All models treated equally.

Simple in theory; in practice there are difficulties.

Bayes factors

Let's pursue the Bayesian approach (Jeffreys 1939).

The Bayes factor (Kass and Raftery 1995) relates the relative plausibility of two models after data to their relative plausibility before data;

$$\textit{Posterior odds} = \textit{Bayes factor} \times \textit{Prior odds}$$

where

$$\textit{Bayes factor} = \frac{p\left(\textit{Observed data} \mid \textit{Model } a\right)}{p\left(\textit{Observed data} \mid \textit{Model } b\right)}$$

A nice result — by applying laws of probability, we see that models should be compared by nothing other than their ability to predict the observed data.

Bayesian evidence

The factors in the ratio are Bayesian evidences

$$\mathcal{Z} \equiv p\left(D \,|\, M\right) = \int_{\Omega_{\boldsymbol{\Theta}}} \mathcal{L}(\boldsymbol{\Theta})\, \pi(\boldsymbol{\Theta})\, \mathrm{d}\boldsymbol{\Theta},$$

where D is the observed data, $\mathcal{L}(\boldsymbol{\Theta}) = p\left(D \,|\, \boldsymbol{\Theta}, M\right)$ is the likelihood and $\pi(\boldsymbol{\Theta}) = P\left(\boldsymbol{\Theta} \,|\, M\right)$ is our prior, and $\boldsymbol{\Theta}$ are the model's parameters.

Bayesian evidence

The evidence is often the single most important number in the problem and I think every effort should be devoted to calculating it

Mackay (2003)

The single most important number in inference? Let's think about how to compute it!

It's a difficult integral

Multi-dimensional: Our models of physics might have many parameters. Even simple models contain $\mathcal{O}(10)$ parameters

Multi-modal: We don't live in Gaussian land. In physics, the likelihoods can feature degeneracies and multiple modes

Fat-tailed: Large variance if you try Monte Carlo integration

Nested sampling

Algorithm

Skilling's idea (Skilling 2004; Skilling 2006). We can write

$$\mathcal{Z} = \int \mathcal{L}(X)\, \mathrm{d}X$$

where the volume variable

$$\begin{aligned} X(\mathcal{L}^\star) &= \text{Fraction of prior volume with } \mathcal{L}(\boldsymbol{\Theta}) \geq \mathcal{L}^\star \\ &= \int_{\mathcal{L}(\boldsymbol{\Theta}) \geq \mathcal{L}^\star} \pi(\boldsymbol{\Theta})\, \mathrm{d}\boldsymbol{\Theta} \end{aligned}$$

and $\mathcal{L}(X(\lambda)) = \lambda$. This is a one-dimensional integral. We can approximate it by a Riemann sum

$$\mathcal{Z} \approx \sum \mathcal{L}(X) \Delta X$$

Compression

We haven't achieved much yet. The trick is how to estimate X?

Compression

0. Draw n_{live} samples from the prior — the live points
1. Denote the smallest likelihood amongst the live points by $\mathcal{L}^\star$
2. Replace that live point by one drawn from the constrained prior

$$\pi^\star(\boldsymbol{\Theta}) \propto \begin{cases} \pi(\boldsymbol{\Theta}) & \mathcal{L}(\boldsymbol{\Theta}) \geq \mathcal{L}^\star \\ 0 & \text{otherwise} \end{cases}$$

3. Make a statistical estimate of $X(\mathcal{L}^\star)$ from this procedure
4. Increment estimate of evidence, $\mathcal{Z} \to \mathcal{Z} + \mathcal{L}^\star \Delta X$
5. If we have completed evidence sum to given tolerance, stop. Otherwise go to 1.

So we evolve a set of n_{live} live points to higher and higher likelihoods, replacing one live point at a time.

Compression

We know that $X_0 = 1$. How much do we expect X to contract when we replace the worst point?

Drawing from the constrained prior means live points are distributed uniformly in X from 0 to $X(\mathcal{L}^\star)$.

In other words, the

$$f_i = \frac{X(\mathcal{L}_i)}{X(\mathcal{L}^\star)}$$

are uniformly distributed from 0 to 1.

Compression

We know that $X_0 = 1$. How much do we expect X to contract when we replace the worst point?

The smallest one, $t \equiv \min f_i$, gives us the compression. We can write

$$p(t) = \binom{n_{\text{live}}}{1} \cdot t^{n_{\text{live}}-1} \cdot 1 = n_{\text{live}} t^{n_{\text{live}}-1}$$

where the factors are combinatorial, the probability of $n_{\text{live}} - 1$ samples less than t, and lastly the probability density of a point at t.

Compression

We know that $X_0 = 1$. How much do we expect X to contract when we replace the worst point?

We find the expected compression:

$$\langle \log t \rangle = n_{\text{live}} \int_0^1 t^{n_{\text{live}}-1} \log t \, \mathrm{d}t = -\frac{1}{n_{\text{live}}}$$

Thus we may estimate that at iteration i

$$X_i \equiv X(\mathcal{L}_i^\star) \approx e^{-i/n_{live}}$$

Exploration

How can we find an independent sample from the constrained prior?

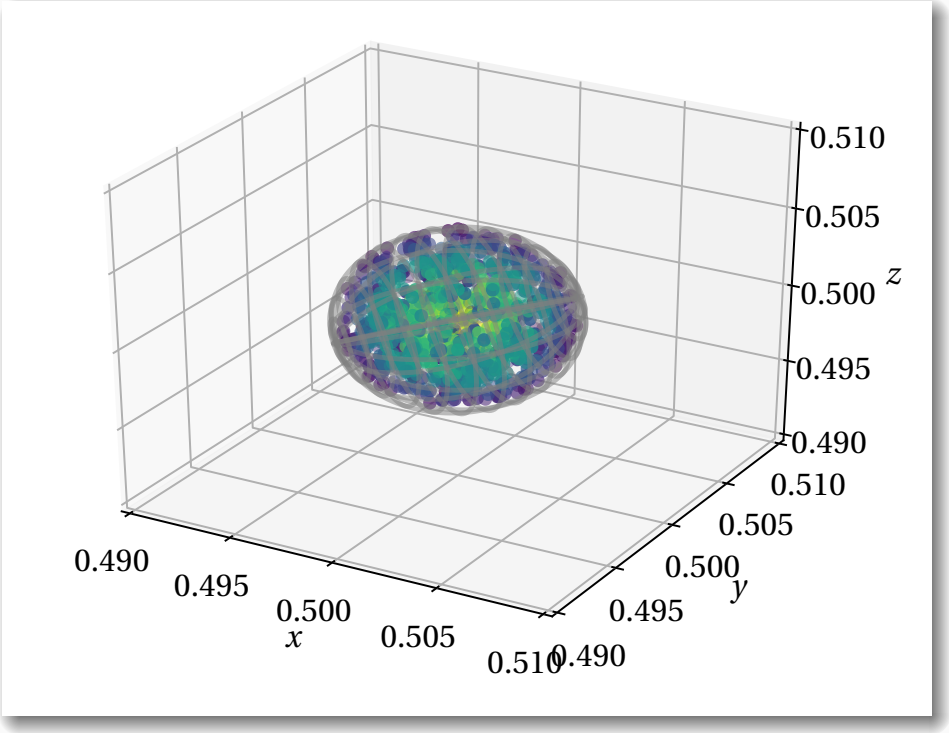
This step in nested sampling was needed for our estimates of the volume, X.

Failure to correctly sample from the constrained prior leads to faulty estimates of the evidence.

This requires an exploration strategy.

Exploration

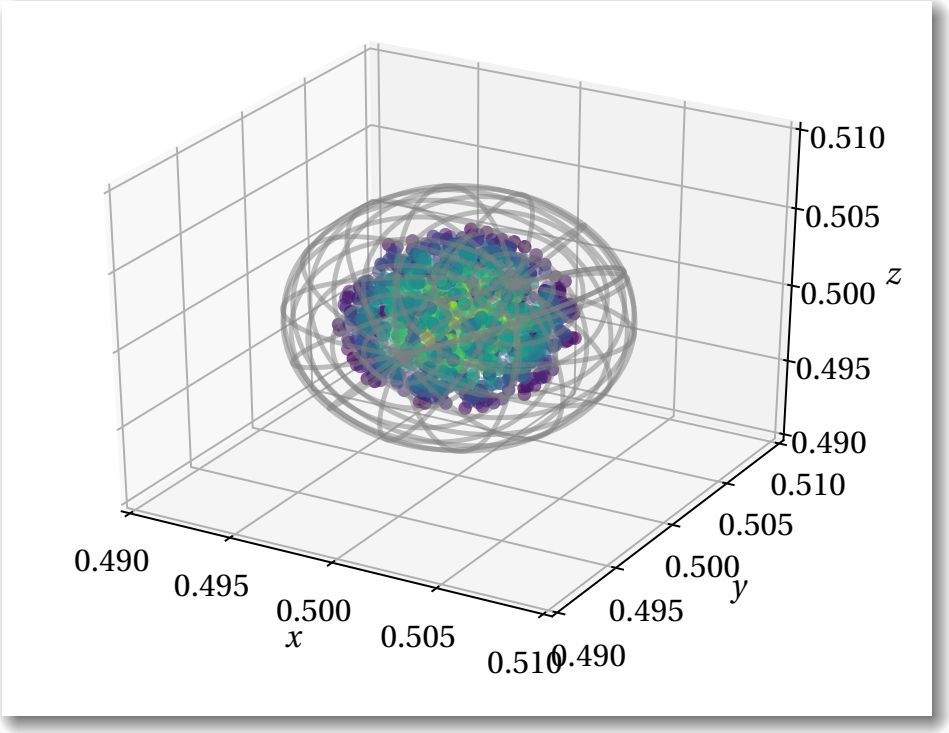
MultiNest (Feroz and Hobson 2008; Feroz, Hobson, and Bridges 2009; Feroz et al. 2013) — bound live points by ellipsoids. Use them to approximate iso-likelihood contour. Sample from the ellipsoids.



Two-dimensional Gaussian.

Exploration

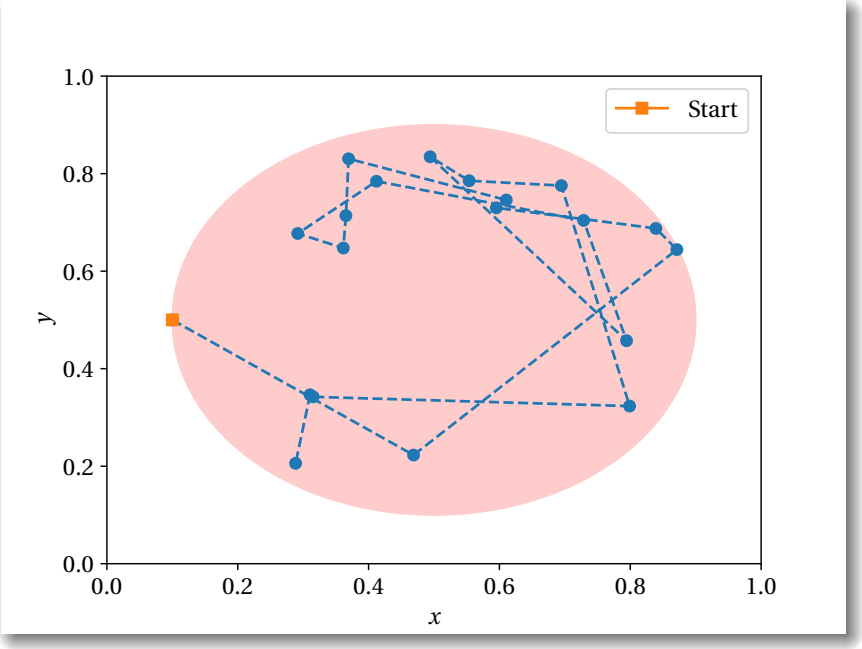
MultiNest (Feroz and Hobson 2008; Feroz, Hobson, and Bridges 2009; Feroz et al. 2013) — bound live points by ellipsoids. Use them to approximate iso-likelihood contour. Sample from the ellipsoids.



Expand to be safe — at expense of sampling efficiency.

Exploration

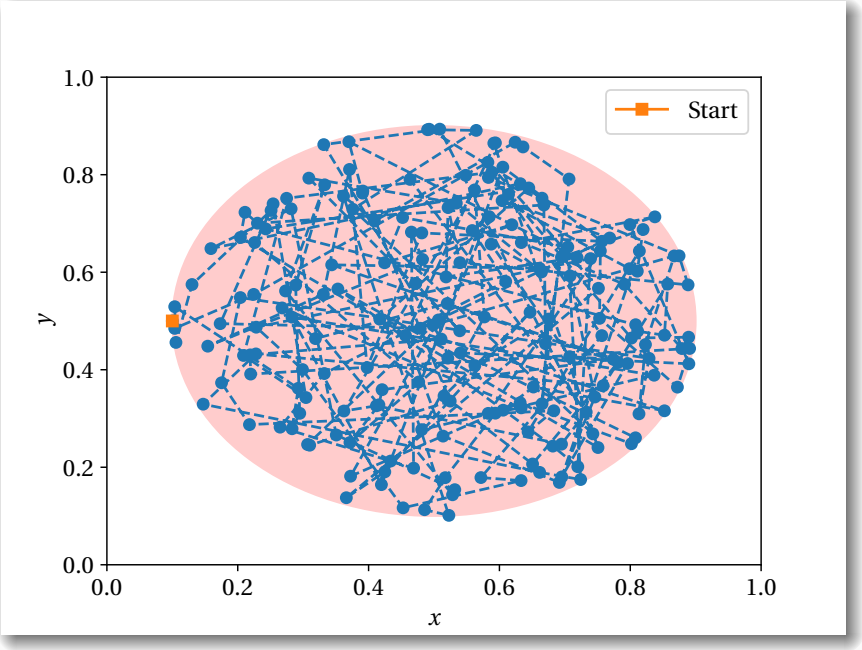
PolyChord (Handley, Hobson, and Lasenby 2015a; Handley, Hobson, and Lasenby 2015b) — slice sampling walk, starting from a randomly chosen live point.



Two-dimensional Gaussian. 20 steps.

Exploration

PolyChord (Handley, Hobson, and Lasenby 2015a; Handley, Hobson, and Lasenby 2015b) — slice sampling walk, starting from a randomly chosen live point.



200 steps. More steps to reduce correlation — at expense of sampling efficiency.

Things can go wrong...

- What if I don't expand the ellipsoids enough?
- What if I don't use enough steps?
- What if my exploration strategy isn't actually drawing independent samples from the constrained prior?

It would violate assumption and lead to faulty estimate of evidence.

But how would I know?

A new cross-check

What if we new the X of every sample?

Suppose we knew the X of every sample, $X(\mathcal{L}_i)$. We could look at

$$f_i = \frac{X(\mathcal{L}_i)}{X(\mathcal{L}^\star)}$$

it should be uniformly distributed from 0 to 1 as each new $X(\mathcal{L}_i)$ should be uniformly distributed from 0 to $X(\mathcal{L}^\star)$.

You could test whether the f indeed followed a uniform distribution (Buchner 2016).

What do we know?

We don't know that. We do know the likelihood of every new sample, $\mathcal{L}_i$, and that $X(\mathcal{L})$ is a monotonic function.

So we can rank the n_{live} points by $X(\mathcal{L}_i)$ by ranking them by $\mathcal{L}_i$.

The rank of every new sample, r, should be uniformly distributed from 1 to n_{live}.

It's just as likely to be the worst, second worst, …, second best, best likelihood.

We can test whether the r indeed follow a discrete uniform distribution.

KS test

To compare the samples with the uniform distribution, we compute a p-value form a Kolmogorov-Smirnoff test.

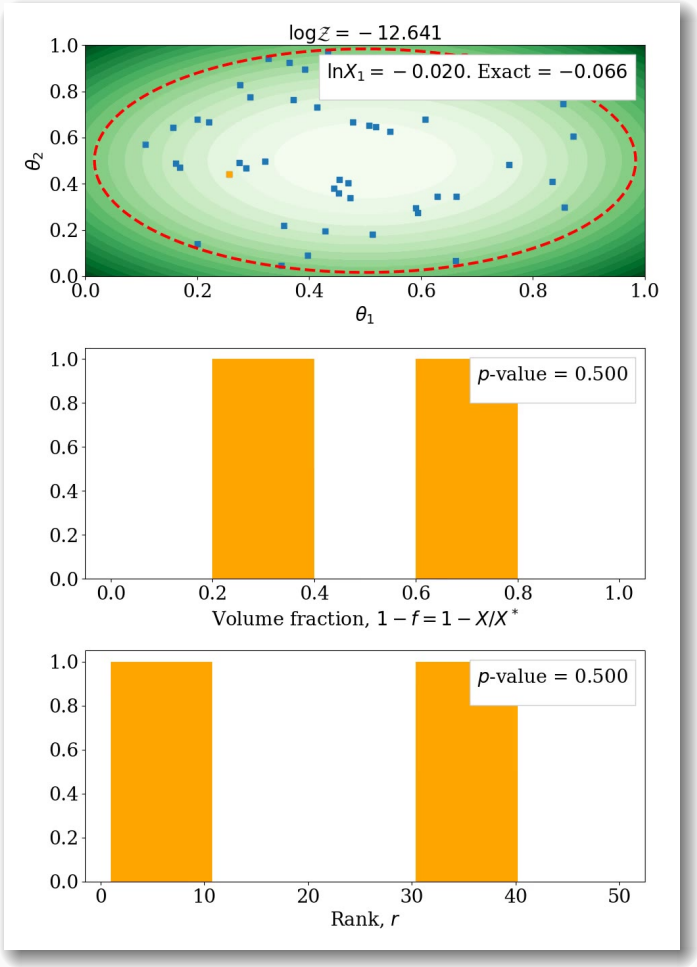
Sorry to sully a Bayesian algorithm with a p-value.

We use all the iterations and we test chunks of n_{live} iterations.

The latter stops biased periods in long runs being diluted by lots of unbiased iterations.

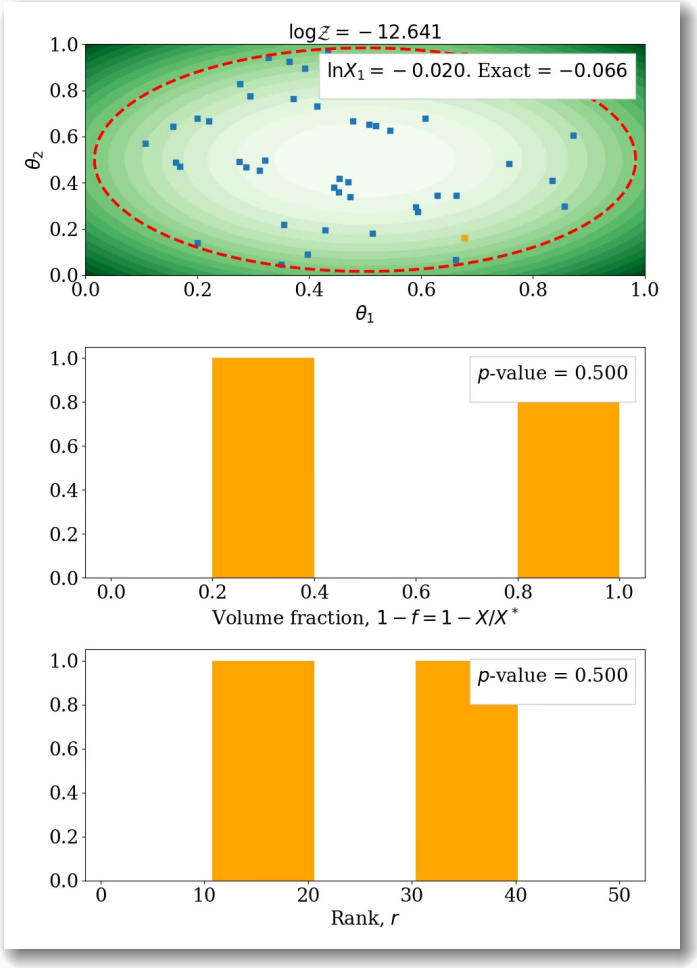
Histogram of ranks r

Let's run a two-dimensional Gaussian nested sampling run. We will monitor the fractions and the insertion ranks, r.



Detecting faults

This time, let's make the sampling biased by sampling from the wrong iso-likelihood contour — we find the correct one then contract it by a (random) factor 0.8 ± 0.1.



Detecting faults

We see a tiny p-value and a biased overestimate of $\mathcal{Z}$ — overestimated because the likelihoods that we draw are greater than they should be.

Correlations

The ranks r are not, however, independent — the distribution of the live points only changes by one point every iteration.

If live points a are clustered together in X, insertion indexes in that region are unlikely.

We ignore this complication. However, if anything, correlations make the insertion ranks repel each other,

Toy problems

Toy problems

In our paper, we introduce 4 toy problems. Here we discuss only one of them.

We compute the evidence using MultiNest and PolyChord, and p-values from our test.

We do 100 repeats. And good efr $\ll 1$ and bad efr $\gg 1$ exploration settings.

Gaussian

Multi-dimensional Gaussian likelihood

$$\mathcal{L}(\mathbf{\Theta}) \propto e^{-\frac{\sum(\mathbf{\Theta}-\mu)^2}{2\sigma^2}}$$

We pick a uniform prior from 0 to 1 for each dimension.

The analytic evidence is always $\log \mathcal{Z} = 0$ since the likelihood is a pdf in $\mathbf{\Theta}$, modulo small errors as the infinite domain is truncated by the prior.

We pick $\mu = 0.5$ and a diagonal covariance matrix with $\sigma = 0.001$ for each dimension.

MultiNest. Gaussian, $\log \mathcal{Z} = 0$

Tiny p-values and biased results shown in red.

Smaller efr $\Leftrightarrow$ stricter run

efr	d	$\log \mathcal{Z}$	Inaccuracy	Bias	p-value	Rolling
0.10	2	-0.00 ± 0.10	-0.04	-0.47	0.50	0.49
0.10	10	0.01 ± 0.23	0.04	0.48	0.59	0.60
0.10	30	0.38 ± 0.41	0.93	10.56	0.52	$2.7 \cdot 10^{-4}$
0.10	50	2.08 ± 0.52	3.98	41.25	0.38	$4.5 \cdot 10^{-24}$
1	2	-0.00 ± 0.10	-0.04	-0.46	0.52	0.49
1	10	0.57 ± 0.23	2.43	26.07	0.21	$1.2 \cdot 10^{-4}$
1	30	2.35 ± 0.40	5.83	63.82	0.23	$2.2 \cdot 10^{-23}$
1	50	4.06 ± 0.52	7.81	92.99	0.30	$1.3 \cdot 10^{-34}$
10	2	-64.75 ± 0.11	-532.44	-6.95	$7.7 \cdot 10^{-3}$	0.06
10	10	2.81 ± 0.23	12.30	150.55	$2.1 \cdot 10^{-6}$	$1.7 \cdot 10^{-19}$
10	30	4.30 ± 0.40	10.75	174.47	0.02	$3.1 \cdot 10^{-68}$
10	50	6.04 ± 0.52	11.66	197.79	0.08	$1.1 \cdot 10^{-93}$

PolyChord. Gaussian, $\log \mathcal{Z} = 0$

Tiny p-values and biased results shown in red.

Smaller efr $\Leftrightarrow$ stricter run

efr	d	$\log \mathcal{Z}$	Inaccuracy	Bias	p-value	Rolling
0.50	2	0.01 ± 0.11	0.11	1.03	0.54	0.60
0.50	10	-0.00 ± 0.23	-0.01	-0.10	0.48	0.52
0.50	30	-0.06 ± 0.41	-0.15	-1.61	0.54	0.57
0.50	50	-0.05 ± 0.52	-0.10	-0.85	0.58	0.51
1	2	-0.02 ± 0.11	-0.19	-1.96	0.42	0.48
1	10	-0.04 ± 0.23	-0.17	-2.20	0.55	0.59
1	30	-0.83 ± 0.41	-2.06	-20.73	0.61	0.46
1	50	-2.48 ± 0.52	-4.73	-54.22	0.49	0.59
2	2	-0.01 ± 0.11	-0.12	-0.89	0.47	0.53
10	10	2.20 ± 0.23	9.50	30.29	0.13	0.22
30	30	48.37 ± 0.64	112.25	70.58	$8.2 \cdot 10^{-10}$	0.02
50	50	69.74 ± 3.05	23.31	106.51	$8.0 \cdot 10^{-86}$	$1.4 \cdot 10^{-6}$

Summary of toy problem

A lot of numbers…

- Less strict exploration settings or high number of dimensions
- … leads to a biased estimate of evidece
- … often detected by tiny p-value by our test

Example from cosmology

Cosmology

Handley considered Bayesian evidence for a spatially closed Universe (Handley 2019a). Evidences from combinations of four datasets were computed using PolyChord for a spatially flat Universe and a curved Universe.

The Bayes factors showed that a closed Universe was favoured by odds of about 50/1 for a particular set of data.

There were 22 NS computations in total (Handley 2019b).

Cosmology

We ran our cross-check on each of the 22 NS runs finding p-values in the range 4% to 98%.

THis does not suggest problems with the NS runs. The p-value of 4% is not particularly alarming, especially considering we conducted 22 tests.

	Flat		Curved	
Data	p-value	Rolling p-value	p-value	Rolling p-value
BAO	0.89	0.82	0.07	0.05
lensing+BAO	0.72	0.54	0.19	0.43
lensing	0.26	0.14	0.04	0.64
lensing+SH_0ES	0.08	0.08	0.78	0.04
Planck+BAO	0.39	0.56	0.14	0.43
Planck+lensing+BAO	0.68	0.69	0.70	0.27
Planck+lensing	0.94	0.49	0.89	0.72
Planck+lensing+SH_0ES	0.92	0.92	0.33	0.82
Planck	0.81	0.69	0.84	0.88
Planck+SH_0ES	0.20	0.48	0.92	0.97
SH_0ES	0.59	0.59	0.98	0.98

Summary

- Nested sampling is a popular algorithm for computing Bayesian evidence
- We developed the first test of single nested sampling runs
- Appears to work nicely on toy and realistic problems
- Could become an important part of nested sampling analysis
- Could become a best practice to apply the check whenever using nested sampling

References

Buchner, Johannes (July 2016). "A statistical test for Nested Sampling algorithms." In: *Statistics and Computing* 26, pp. 383–392. arXiv: 1407.5459 [stat.CO].

Feroz, F., M. P. Hobson, and M. Bridges (2009). "MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics." In: *Mon. Not. Roy. Astron. Soc.* 398, pp. 1601–1614. arXiv: 0809.3437 [astro-ph].

Feroz, F. et al. (2013). "Importance Nested Sampling and the MultiNest Algorithm." In: *The Open Journal of Astrophysics*. arXiv: 1306.2144 [astro-ph.IM].

References ii

Feroz, Farhan and M. P. Hobson (2008). "Multimodal nested sampling: an efficient and robust alternative to MCMC methods for astronomical data analysis." In: *Mon. Not. Roy. Astron. Soc.* 384, p. 449. arXiv: 0704.3704 [astro-ph].

Fowlie, Andrew, Will Handley, and Liangliang Su (June 2020). "Nested sampling cross-checks using order statistics." In: arXiv: 2006.03371 [stat.CO].

Handley, W. J., M. P. Hobson, and A. N. Lasenby (2015a). "PolyChord: nested sampling for cosmology." In: *Mon. Not. Roy. Astron. Soc.* 450.1, pp. L61–L65. arXiv: 1502.01856 [astro-ph.CO].

– (Nov. 2015b). "PolyChord: next-generation nested sampling." In: *Mon. Not. Roy. Astron. Soc.* 453.4, pp. 4384–4398. arXiv: 1506.00171 [astro-ph.IM].

References iii

Handley, Will (Aug. 2019a). "Curvature tension: evidence for a closed universe." In: arXiv: 1908.09139 [astro-ph.CO].

– (Aug. 2019b). *Curvature tension: evidence for a closed universe (supplementary inference products)*. Version 1.0.0. Zenodo. URL: https://doi.org/10.5281/zenodo.3371152.

Jeffreys, Harold (1939). *The Theory of Probability*. Oxford Classic Texts in the Physical Sciences. Oxford University Press. ISBN: 978-0-19-850368-2, 978-0-19-853193-7.

Kass, Robert E. and Adrian E. Raftery (1995). "Bayes Factors." In: *J. Am. Statist. Assoc.* 90.430, pp. 773–795.

Skilling, John (Nov. 2004). "Nested Sampling." In: *American Institute of Physics Conference Series*. Ed. by Rainer Fischer, Roland Preuss, and Udo Von Toussaint. Vol. 735, pp. 395–405.

References iv

Skilling, John (2006). “Nested sampling for general Bayesian computation.” In: *Bayesian Analysis* 1.4, pp. 833–859.