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 20, 2020

Nanjing Normal University

Table of contents

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?

Error theoretic

Construct a rule so that you'd wrongly reject the null hypothesis at a pre-specified rate in the long-run in an ensemble of experiments.

E.g., we reject at null hypothesis 95% confidence level.

We can't treat all models on equal footing — must specify a null — and doesn't consider only about the evidence from this experiment — we have to think about an ensemble of repeats.

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.

p-value

Compute a p-value

$$P\left(\text{data more or as extreme as that observed} \,\middle|\, \text{null hypothesis}\right)$$

Very popular in particle physics and elsewhere.

- Use it as a proxy for the plausibility of H_0. Small p-value $\Rightarrow$ H_0 implausible.
- Use it control error rate: if we reject null when p-value ≤ 0.05, for example, becomes error theoretic approach with error rate 0.05.

Bayesian evidence

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

$$\textit{Bayes factor} = \frac{p\,(\textit{Observed data}\,|\,\textit{Model a})}{p\,(\textit{Observed data}\,|\,\textit{Model b})}$$

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

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

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

Priors

Many consider the dependence of the Bayes factor on the priors to be a major problem.

No priors, no predictions

I need to compare your model's predicts with data. If you don't tell the plausible parameters, how am I to know what it predicts?

Sensitive to arbitrary choices

If the inference changes dramatically within a class of reasonable priors, we can't draw reliable conclusions.

Science is hard; it's hard to get reliable knowledge about the world. We often disagree about the consequences of experimental data.

How could it be any other way?

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

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

Compression

0. Draw n_{live} samples from the prior — the live points
1. Denote the smallest likelihood amongst the live points by $\mathcal{L}^\star$

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

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

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

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

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

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

3. Make a statistical estimate of $X(\mathcal{L}^\star)$ from this procedure

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

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, $\mathcal{Z} \to \mathcal{Z} + \mathcal{L}^\star \Delta X$

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

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

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

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 largest one, $t \equiv \max f_i$, 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

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

Compression

Consider the number of steps to reach the bulk of the posterior mass (typical set)

$$e^{-i/n_{\text{live}}} \approx e^{-H}$$

where H is the relative entropy from prior to the posterior.

This ultimately allows us to estimate the error in our estimate of $\log \mathcal{Z}$

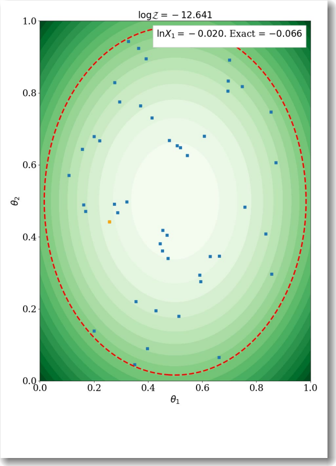
$$\Delta \log \mathcal{Z} \approx \sqrt{\frac{H}{n_{live}}}$$

So multi-dimensionality not fundamental problem; the problem is significant compression.

And precision goes like $1/\sqrt{\text{computational effort}}$ as usual.

2d Gaussian example

We take a two-dimensional Gaussian centered at $(0.5, 0.5)$. The analytic $\log \mathcal{Z} = 0$.



Exploration

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

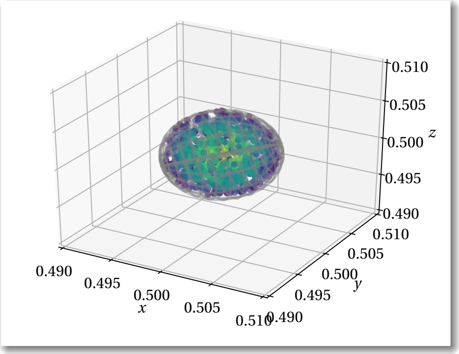
This step in nested sampling was needed for our estimates of the volume, X.

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

This requires an exploration strategy.

Exploration

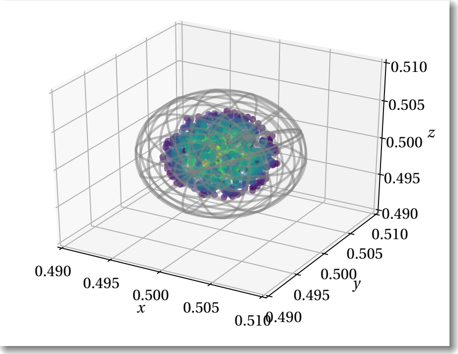
MultiNest (Feroz and Hobson 2008; Feroz, Hobson, and Bridges 2009; Feroz et al. 2013) — bound live points by ellipsoids. Use them to approximate iso-likelihood contour. Sample from the ellipsoids.



Two-dimensional Gaussian.

Exploration

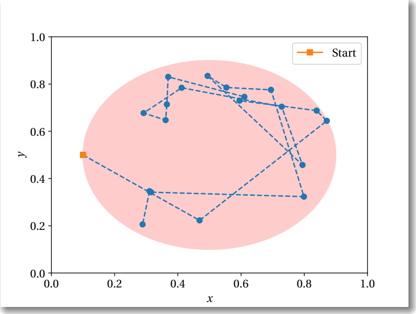
MultiNest (Feroz and Hobson 2008; Feroz, Hobson, and Bridges 2009; Feroz et al. 2013) — bound live points by ellipsoids. Use them to approximate iso-likelihood contour. Sample from the ellipsoids.



Expand to be safe — at expense of sampling efficiency.

Exploration

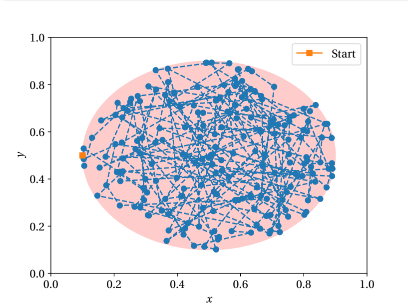
PolyChord (Handley, Hobson, and Lasenby 2015a; Handley, Hobson, and Lasenby 2015b) — slice sampling walk, starting from a randomly chosen live point.



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

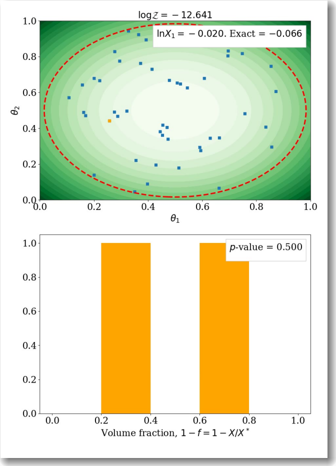
$$f_i = \frac{X(\mathcal{L}_i)}{X(\mathcal{L}^\star)}$$

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).

Histogram of the fractions f

Let's run the same nested sampling run, but this time monitor the fractions.



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.

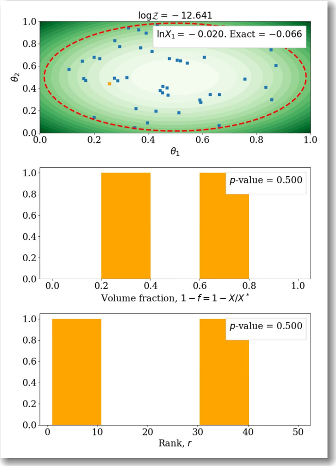
Sorry to sully a Bayesian algorithm with a p-value.

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.

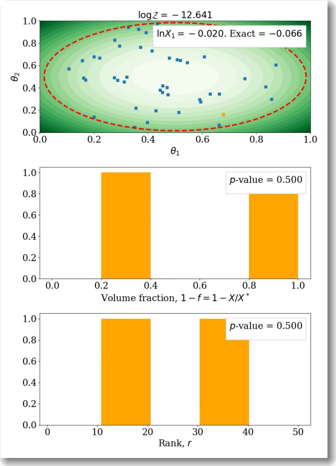
Histogram of ranks r

Let's run the same two-dimensional Gaussian nested sampling run but this time 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 a couple of them.

For each one, we will 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

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

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 $\boldsymbol{\Theta}$, 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	$8.0 \cdot 10^{-86}$	$1.4 \cdot 10^{-6}$

Gaussian shells

This multidimensional likelihood is

$$\mathcal{L}(\boldsymbol{\Theta}) = \mathrm{shell}(\boldsymbol{\Theta}; \boldsymbol{c}, r, w) + \mathrm{shell}(\boldsymbol{\Theta}; -\boldsymbol{c}, r, w)$$

where the shell function is a Gaussian

$$\mathrm{shell}(\boldsymbol{\Theta}; \boldsymbol{c}, r, w) = \frac{1}{\sqrt{2\pi} w} e^{-(|\boldsymbol{\Theta}-\boldsymbol{c}|-r)^2/(2w^2)}.$$

The highest likelihood region forms a shell of characteristic width w at the surface of a d-sphere of radius r.

With uniform priors between -6 and 6, the analytic evidence is approximately,

$$\mathcal{Z} = 2\langle |x|^{d-1} \rangle S_d / 12^d$$

where S_d is the surface area of an d-sphere and $\langle |x|^{d-1} \rangle$ is the $(d-1)$-th non-central moment of a Gaussian, $\mathcal{N}(r, w^2)$.

MultiNest. Gaussian shells, $\log \mathcal{Z} = -1.75, -14.59, -60.13, -112.42$

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	-1.75 ± 0.05	-0.06	-0.64	0.55	0.55
0.10	10	-14.59 ± 0.12	0.02	0.16	0.57	0.56
0.10	30	-59.61 ± 0.24	2.11	24.29	0.37	$7.3 \cdot 10^{-6}$
0.10	50	-110.15 ± 0.33	6.87	115.58	0.07	$3.7 \cdot 10^{-23}$
1	2	-1.71 ± 0.05	0.79	8.52	$4.7 \cdot 10^{-3}$	0.10
1	10	-13.92 ± 0.12	5.57	65.88	0.02	$1.1 \cdot 10^{-5}$
1	30	-57.57 ± 0.24	10.67	151.79	$7.7 \cdot 10^{-3}$	$1.4 \cdot 10^{-20}$
1	50	-107.97 ± 0.33	13.63	218.07	$3.5 \cdot 10^{-3}$	$3.6 \cdot 10^{-37}$
10	2	-1.73 ± 0.05	0.39	1.45	0.07	0.18
10	10	-11.73 ± 0.11	25.56	321.53	$6.8 \cdot 10^{-18}$	$1.7 \cdot 10^{-19}$
10	30	-55.41 ± 0.24	20.03	367.16	$3.0 \cdot 10^{-6}$	$9.3 \cdot 10^{-66}$
10	50	-105.82 ± 0.32	20.42	480.50	$9.3 \cdot 10^{-6}$	$2.2 \cdot 10^{-92}$

PolyChord. Gaussian shells, $\log \mathcal{Z} = -1.75, -14.59, -60.13, -112.42$

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	-1.74 ± 0.05	0.16	1.54	0.13	0.13
0.50	10	-14.59 ± 0.12	0.02	0.12	0.50	0.48
0.50	30	-60.12 ± 0.25	0.03	0.29	0.56	0.55
0.50	50	-112.33 ± 0.34	0.27	2.65	0.40	0.58
1	2	-1.75 ± 0.05	-0.02	-0.30	0.01	0.01
1	10	-14.59 ± 0.12	0.02	0.19	0.49	0.61
1	30	-60.46 ± 0.25	-1.36	-14.57	0.48	0.53
1	50	-113.52 ± 0.34	-3.26	-34.47	0.50	0.51
2	2	-1.74 ± 0.05	0.06	0.43	$6.1 \cdot 10^{-6}$	$2.1 \cdot 10^{-5}$
10	10	-14.05 ± 0.12	4.42	15.01	0.09	0.09
30	30	-38.78 ± 0.21	103.26	159.47	$3.5 \cdot 10^{-5}$	$5.2 \cdot 10^{-3}$
50	50	-64.20 ± 0.63	103.63	93.31	$5.2 \cdot 10^{-12}$	$3.8 \cdot 10^{-7}$

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 of examples

- Applied check to NS runs on several toy functions with known analytic results in 2 – 50 dimensions
- Detect problematic runs for MultiNest and PolyChord for many problems, settings and dimensions
- Easy to apply to realistic examples
- Problem detected by our cross-check usually corresponds to biased estimate of the evidence, though in a few cases the evidence estimate remains reasonable

Speculation

Birth index

During a run, we can make a sequence (of length n_{iter}) of birth indexes — the iterations at which the point that died was born.

The insertion ranks we have been discussing are actually a lossless compression of the sequence of birth indexes.

- $n_{\text{live}}^{n_{\text{iter}}}$ possible sequences of insertion ranks
- There are not $n_{\text{iter}}!$ possible sequences of birth indexes — points cannot die before they were born! At each iteration, there are only n_{live} possible birth indexes that could become the dead point $\Rightarrow n_{\text{live}}^{n_{\text{iter}}}$ possible sequences.

A new summation?

The classic nested sampling estimator of the evidence uses

- The $\mathcal{L}$ of every dead point
- The iteration at which each dead point died

A new summation?

The classic nested sampling estimator of the evidence uses

- The $\mathcal{L}$ of every dead point
- The iteration at which each dead point died

What if we used as well

- The iteration at which each dead point was born

A new summation?

The hypothetical improved estimator could

- Reduce variance
- Be more robust with respect to biases in sampling from the constrained prior

A new summation?

At the moment, we estimate $\mathcal{Z}$ from the statistical estimates of the volume, $\{X\} \equiv X_0, X_1, X_2, \ldots$

$$\overline{\mathcal{Z}} = \int \mathcal{Z}(\{X\}) p\left(\{X\}\right) \prod \mathrm{d}X \quad \text{where} \quad \mathcal{Z}(\{X\}) = \sum \mathcal{L}_i^\star (X_i - X_{i+1})$$

and we have $0 < X_j < X_{j-1}$,

$$p\left(X_j \,|\, X_{j-1}\right) = \frac{n_{\mathrm{live}}}{X_{j-1}} \left(\frac{X_j}{X_{j-1}}\right)^{n_{\mathrm{live}}-1}$$

A new summation?

At the moment, we estimate $\mathcal{Z}$ from the statistical estimates of the volume, $\{X\} \equiv X_0, X_1, X_2, \ldots$

$$\overline{\mathcal{Z}} = \int \mathcal{Z}(\{X\}) p\left(\{X\}\right) \prod \mathrm{d}X \quad \text{where} \quad \mathcal{Z}(\{X\}) = \sum \mathcal{L}_i^\star (X_i - X_{i+1})$$

and we have $0 < X_j < X_{j-1}$,

$$p\left(X_j \mid X_{j-1}\right) = \frac{n_{\text{live}}}{X_{j-1}} \left(\frac{X_j}{X_{j-1}}\right)^{n_{\text{live}}-1}$$

What if we condition on the insertion ranks?

$$\overline{\mathcal{Z}} = \int \mathcal{Z}(\{X\}) p\left(\{X\} \mid \textit{insertion ranks}\right) \prod dX$$

That is, what if we use $p\left(\{X\} \mid \textit{insertion ranks}\right)$ rather than just $p\left(\{X\}\right)$? Can we do it? It could be a kind of Rao-Blackwellization.

Not yet

Simulating from $p(\{X\})$ was easy.

Simulating from $p(\{X\} \mid \text{insertion ranks})$ isn't so straight-forward.

On the other hand, it doesn't require any likelihood evaluations.
So even if it is slightly involved, it probably won't affect the overall computation time for realistic problems.

If you want to solve this problem, let's talk!

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
- Hints towards a better estimate of the evidence

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].

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].

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.

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