Nested sampling for evidence estimation

1 Inference and model comparison

When performing the analysis of some data d according to some model M parameterized by some finite number of parameters θ, we employ Bayes' theorem, which is derived by expressing the joint probability $\text{prob}(d, \theta|M)$ in two different ways:

$$\text{prob}(d, \theta|M) = \mathcal{L}(d|\theta, M)\pi(\theta|M) = p(\theta|d, M)\mathcal{Z}(d|M)\,.$$

The entries in this equation are:

- the *likelihood* $\mathcal{L}$ of the data given the model and a specific choice of its parameters - it is a probability distribution over the data (i.e. $\int \mathcal{L}(d|\theta, M)\mathrm{d}d = 1$), but we interpret it as a function of the parameters θ while the data are fixed;
- the *prior* π - a representation of our beliefs about the model's parameters before seeing the data;
- the *posterior* p for the model's parameters - a representation of our updated beliefs about them after seeing the data;
- the *evidence* $\mathcal{Z}$ for the model - a normalization constant for the posterior, or the likelihood for the data marginalized over all the model's parameters.

The likelihood and prior are part of the problem definition, and can typically be readily computed at any given value of the parameters. What we refer to as "computing the posterior" means obtaining a useful representation for it. If the parameter space is high-dimensional, evaluating the likelihood on a grid quickly becomes unpractical. A common solution to this issue is to use stochastic methods such as nested sampling Section 2, which allow us to obtain a set of samples θ_i distributed according to the posterior. This way of representing a posterior is convenient, since we can easily compute summary statistics from it, such as the expectation value of any function $f(\theta)$:

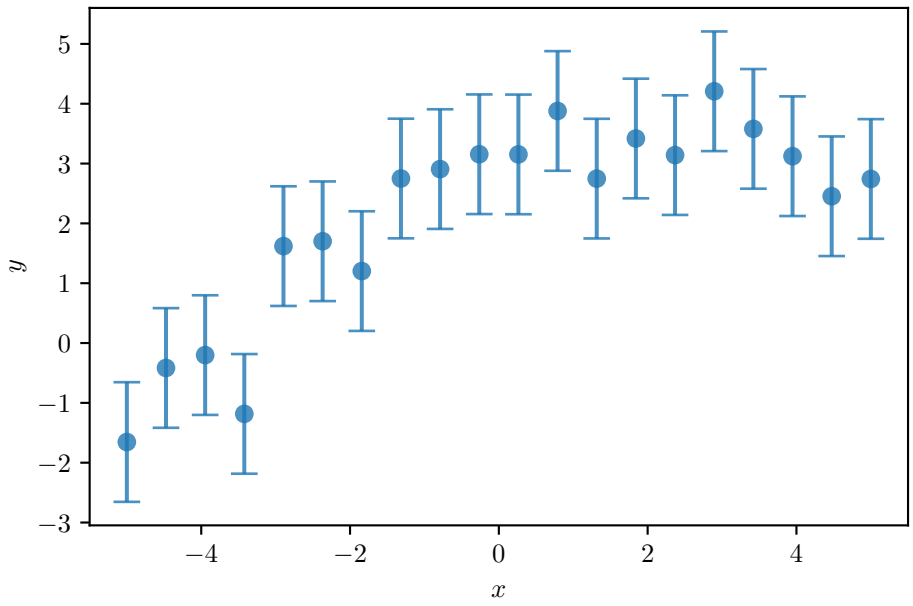
$$\langle f \rangle \approx \frac{1}{N}\sum_{i=1}^{N} f(\theta_i)\,.$$

1.1 Model comparison

In order to illustrate the use of the evidence $\mathcal{Z}$ for model comparison, we define a simple problem: suppose we have a set of data points (x_i, y_i) with uncertainties σ_y on the y values.

We will compare two simple *nested* models, namely a parabolic one with $y = ax^2 + bx + c$ and a linear one with $y = bx + c$. Model comparison can also be performed for non-nested models, but the nested case has some nice properties we will showcase later.

Let us generate the data and plot it: it will come from the parabolic model.



We can then analyze this data, using a Gaussian likelihood and setting (somewhat arbitrarily) uniform priors for the parameters a, b and c in the range $[-10, 10]$.

Since these analyses were done with Nested Sampling (see Section 2), we have estimates for the posterior distributions of the parameters as well as the evidences. This allows us to compute a Bayes Factor:

$$\log \mathrm{BF}^{\mathrm{parabola}}_{\mathrm{line}} = \log Z_{\mathrm{parabola}} - \log Z_{\mathrm{line}} \,,$$

which appears in the equation for the posterior odds for one model against the other:

$$\underbrace{\frac{p(M_{\mathrm{parabola}}|d)}{P(M_{\mathrm{line}}|d)}}_{\text{posterior odds}} = \mathrm{BF}^{\mathrm{parabola}}_{\mathrm{line}} \underbrace{\frac{p(M_{\mathrm{parabola}})}{p(M_{\mathrm{line}})}}_{\text{prior odds}} \,.$$

In this case, the curvature of the data is strong enough for the Bayes Factor to favor the parabolic model (although not by much).

```
parabola: Z = -36.67 +- 0.09
line: Z = -38.58 +- 0.07
log Bayes factor = 1.90 +- 0.11 nats for the parabolic model
```

Since the models we are considering were nested, we have a comparison point for our Bayes Factor: the Savage-Dickey density ratio. Our models being nested means that the linear one is a special case of the parabolic one (with $a = 0$); the priors for the other parameters are exactly the same in both cases. Then, it can be shown that the Bayes Factor is equal to the ratio of the posterior to the prior for the parabolic model, computed at the value for which the parabolic model reduces to the linear one:

$$\mathrm{BF}^{\mathrm{parabola}}_{\mathrm{line}} = \frac{p(a = 0|d, M_{\mathrm{parabola}})}{\pi(a = 0|M_{\mathrm{parabola}})}\,.$$

One might then think that, at least in this type of scenario, going through the evidence computation with a powerful tool such as Nested Sampling was unnecessary: the posterior distribution can be computed for cheaper, after all. However, as Figure 1 shows, actually computing the value for the posterior density distribution is not as easy. We can approximate it with tools like a histogram or a kernel density estimate, but any such method will suffer from low statistics if we need to explore the edges of the distribution, which is what we see here: we would need an enormous amount of posterior samples to reach the same accuracy on the ratio of the distributions we get with Nested Sampling.

2 Nested sampling

Nested sampling, introduced by Skilling in 2006 (Skilling 2006), is an algorithm designed to compute the *evidence* $\mathcal{Z}$ as well as the posterior distribution on the parameters.

The evidence is computed through the integral

$$\mathcal{Z}(d|M) = \int \mathcal{L}(d|\theta, M)\pi(\theta|M)\mathrm{d}^n\theta\,.$$

This integral is in n dimensions (where n is the dimensionality of the parameter space) and therefore difficult to work with. The idea in nested sampling is to rewrite it in terms of the auxiliary variable X (employing the notation of Sivia and Skilling (2006)), defined as the prior volume contained within a likelihood constraint $\mathcal{L} \geq L$:

$$X(L) = \int_{\mathcal{L}(d|\theta,M)\geq L} \pi(\theta|M)\mathrm{d}^n\theta\,.$$

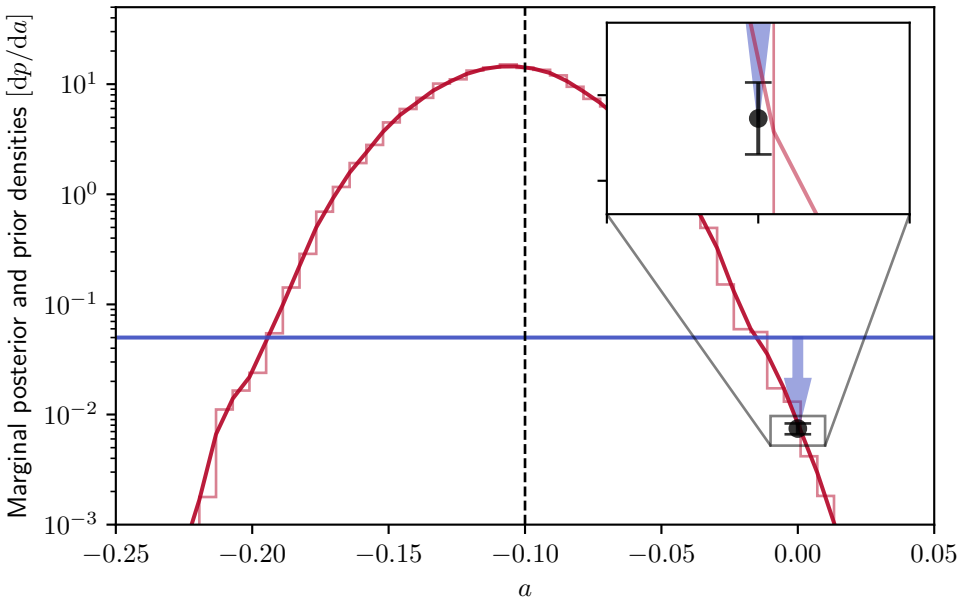


Figure 1: Consistency of the Savage-Dickey density ratio and Nested Sampling estimates for the Bayes Factor.

Since the prior is a normalized probability distribution and the likelihood ranges from 0 to $L_{\max}$, this variable will range from $X(0) = 1$ to $X(L_{\max}) = 0$, and it will always be decreasing.[1] We can define its inverse, $L(X)$, and rewrite the likelihood integral through integration by parts: the evidence is the expectation value of the likelihood, therefore

$$\mathcal{Z}(d|M) = \int_0^{L_{\max}} X(L)\mathrm{d}L = XL|_{L=0}^{L=L_{\max}} - \int_{X(L=0)}^{X(L=L_{\max})} L(X)\mathrm{d}X = \int_0^1 L(X)\mathrm{d}X\,. \qquad (1)$$

For a detailed discussion of the mathematical details, also see (Ashton et al. 2022).

2.1 Prior compression

Note that, by definition, X is directly related to probability mass: the prior probability of an interval $[X_0, X_1]$, i.e. the prior mass which has likelihood values $L(X_0) < \mathcal{L} < L(X_1)$, is

$$p([X_0, X_1]) = \int_{L(X_0)<\mathcal{L}(d|\theta,M)<L(X_1)} \pi(\theta|M)\mathrm{d}^n\theta = X_1 - X_0\,,$$

[1] It will not necessarily be *strictly* decreasing, but that is not a conceptual issue. One can introduce jitter in the likelihood, varying it at each point by an inconsequential amount, in order to ensure that the decreasing condition is verified precisely.

where we computed the integral on a likelihood "ring" by subtracting the inside volume from the outside one. This is useful to us, since it means that if we distribute points uniformly according to the prior their X values will also be uniformly distributed in $[0, 1]$.

If k points are uniformly distributed in an interval $[0, X^*]$, then the largest of them will have a X coordinate distributed as $X_{\max}/X^* \equiv t \sim \text{beta}(k, 1) = kt^{k-1}$,[2] where we defined the *compression ratio* t: when we discard the lowest-likelihood point with $X_{\max}$, we will have contracted the prior volume by a factor t. In terms of $\log t$, the distribution reads

$$\frac{\mathrm{d}p}{\mathrm{d}\log t} = t\frac{\mathrm{d}p}{\mathrm{d}t} = kt^k = ke^{k\log t}$$

This distribution is shown in Figure 2.

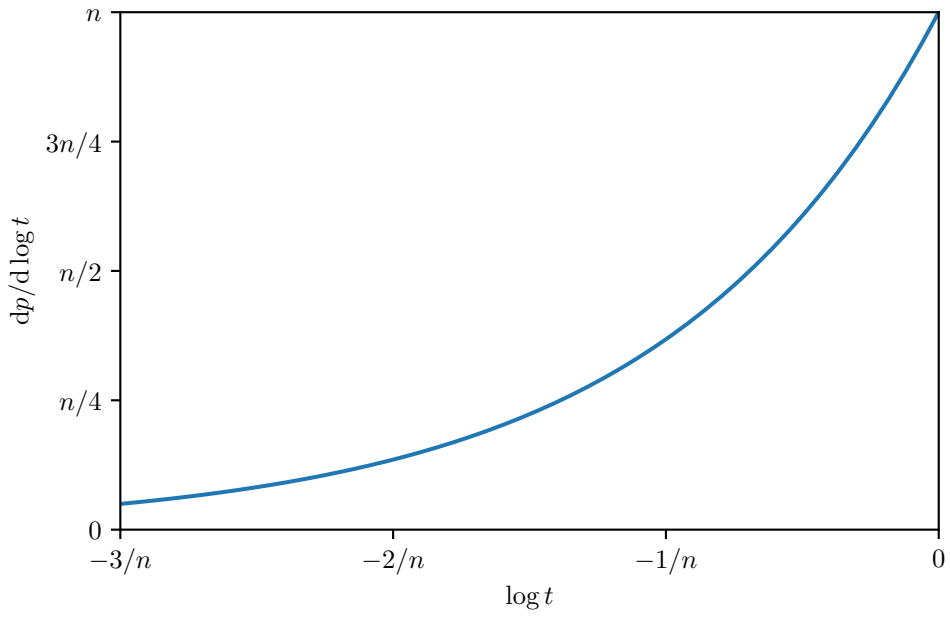


Figure 2: Volume compression

[2] In general, the n-th smallest out of k uniform random variates in $[0, 1]$ is distributed according to $\text{beta}(k, n-k+1) \propto x^{n-1}(1-x)^{k-n}$; here we are looking at the k-th smallest (i.e. the largest).

Its mean and standard deviation are:

$$\begin{aligned}\langle \log t \rangle &= \int_{-\infty}^{0} \log t \frac{\mathrm{d}p}{\mathrm{d}\log t} \mathrm{d}\log t \\ &= \int_{-\infty}^{0} \frac{x}{k} k e^x \frac{\mathrm{d}x}{k} \\ &= \frac{1}{k} \int_{-\infty}^{0} x e^x \mathrm{d}x \\ &= -\frac{1}{k}\end{aligned}$$

where $x = k \log t$, and

$$\begin{aligned}\left\langle (\log t)^2 \right\rangle - \frac{1}{k^2} &= \int_{-\infty}^{0} (\log t)^2 \frac{\mathrm{d}p}{\mathrm{d}\log t} \mathrm{d}\log t - \frac{1}{k^2} \\ &= \int_{-\infty}^{0} \frac{x^2}{k^2} k e^x \frac{\mathrm{d}x}{k} - \frac{1}{k^2} \\ &= \frac{1}{k^2} \int_{-\infty}^{0} x^2 e^x \mathrm{d}x - \frac{1}{k^2} \\ &= \frac{2-1}{k^2} = \frac{1}{k^2}\,.\end{aligned}$$

Compactly, we can then write $\log t = (-1 \pm 1)/k$.

2.2 Computing evidence and posterior

If we sample k points from the prior (with total volume X_0) and remove the one with the worst likelihood, the remaining prior volume will be $X_1 = t_1 X_0$, where $\log t_1$ has an expectation value of $-1/k$. This will be true at every iteration, so (Skilling 2006, sec. 5)

$$\log X_i = \sum_{j \le i} \log t_j \approx -\frac{i}{k} \pm \frac{\sqrt{i}}{k}$$

The prior is compressed exponentially, which is desirable since the typical set for the posterior is often several orders of magnitude smaller than the prior.

This allows us to construct a procedure to evaluate the integral in Equation 1, as long as we are able to do the following:

1. sample uniformly k points from the prior

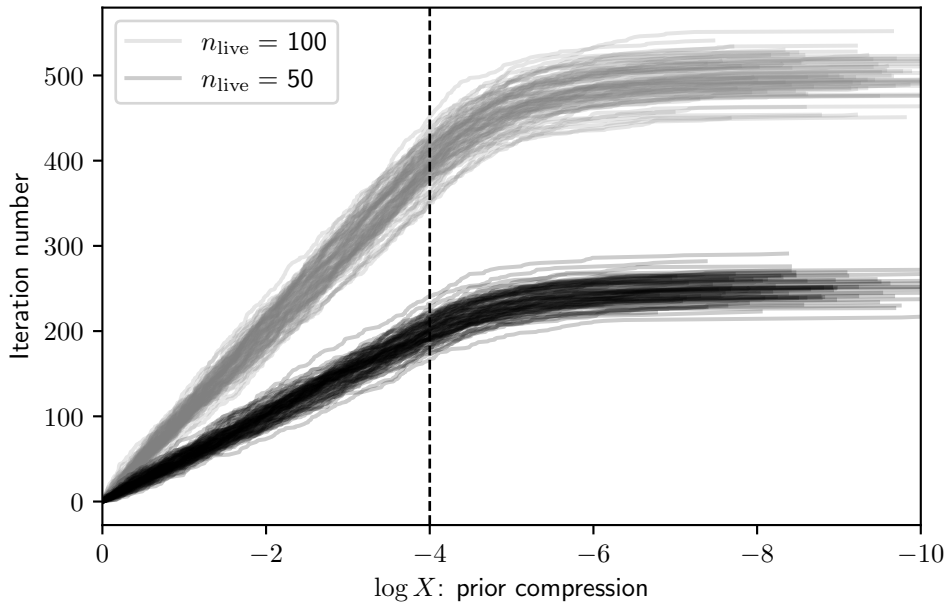


Figure 3: Volume compression by iteration. The number of live points is denoted here as $n_{\text{live}} = k$. Note how the horizontal slices are thinner when we have more live points: increasing them will decrease the uncertainty on the prior compression obtained at a fixed iteration number. After reaching a logarithmic volume compression of $-\log X = 4$ the runs are truncated, and the remaining points are shown: these are distributed uniformly and not geometrically in X, therefore they taper off in this logarithmic plot.

2. find the lowest-likelihood point θ_1 and discard it: now the prior volume is approximately $X_1 = e^{-1/k}$; let us denote its likelihood as L_1
3. sample a new point θ_2 uniformly from the prior, constrained to $L_2 = L(\theta_2) > L_1$
4. repeat from point 2.

We are constructing a sequence of points θ_i with corresponding likelihoods L_i and approximate associated interior prior volumes $X_i \approx e^{-i/k}$. At this stage, we still need to discuss how step 3 is performed, as well as when the iteration should stop - for now, suppose that happens at some stage n_{iter}.

For the evidence, we can simply use a trapezoidal integration rule:

$$\mathcal{Z}(d|M) \approx \sum_{i=1}^{n_{\text{iter}}} w_i L_i = \sum_{i=1}^{n_{\text{iter}}} \frac{X_{i+1} - X_{i-1}}{2} L_i$$

It can be shown that this sum converges to the correct value in the $n_{\text{iter}} \to \infty$ and $n_{\text{live}} \to \infty$ limit.

These weights can also be used to estimate the posterior as follows: take the points θ_i and assign to each of them a weight

$$p_i = \frac{w_i L_i}{Z} \,.$$

It is readily seen that these weights add to 1, and heuristically they are computed as one might expect, by multiplying the prior mass and likelihood value for each of the regions we have subdivided the prior volume into.

Often one is be interested in a set of equally-weighted samples: these can be obtained, for example, through density estimation methods.

2.3 Commonly used plots

The quantities we discussed can be tracked through a Nested Sampling run as a useful diagnostic. Here we show some commonly used plots, based on a two-dimensional example run with a Gaussian covariance.

The assumptions are as follows:

- our prior is a 2-dimensional uniform distribution $\pi \sim \mathcal{U}([0,1]^2)$;
- our likelihood is a 2-dimensional normal distribution, with mean $\mu = [0.5, 0.5]$ and correlated covariance matrix

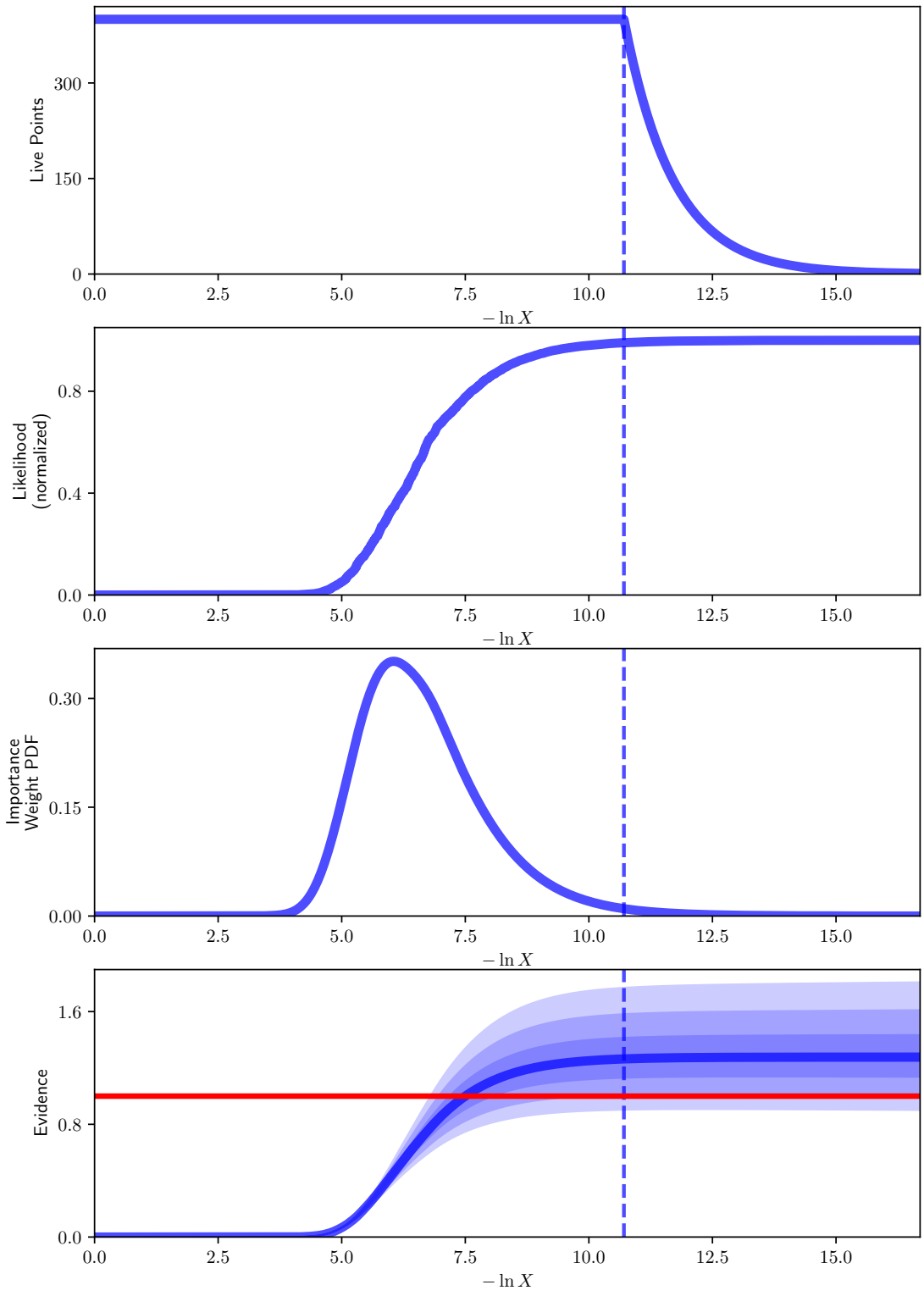
$$C = 3 \times 10^{-2} \begin{pmatrix} 1 & 0.95 \\ 0.95 & 1 \end{pmatrix}$$

We use 400 live points, and stop iterating when the contribution to the integral from a single iteration reaches $\Delta \log Z_i < 0.01$.

2.3.1 Run plot

The first diagnostic tool is the *run plot*, where we see the accumulation of the evidence integral as a function of the prior compression $\log X$

- the number of live points, which is constant up to the point where the nested sampling run is stopped: then, the remaining points are reused;
- the likelihood L_i, which is strictly increasing as expected;
- the posterior mass $w_i L_i / Z$, which increases and then decreases: first, the likelihood's increase dominates, but at some point its values start tapering, and the decrease in prior volume w_i dominates;
- the accumulation of evidence, with a corresponding error estimated as discussed in Section 2.4

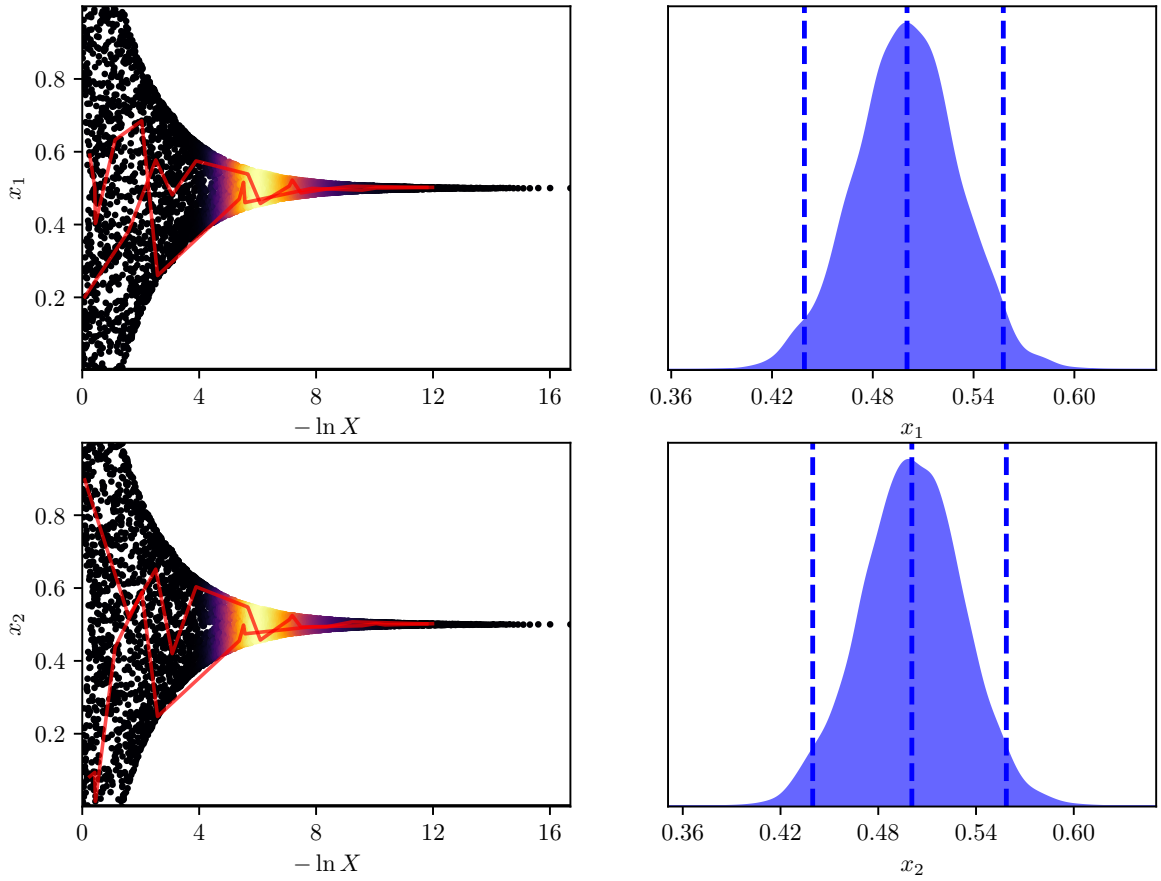

Live Points
300
150
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
$-\ln X$
Likelihood
(normalized)
0.8
0.4
0.0
$-\ln X$
Importance
Weight PDF
0.30
0.15
0.00
$-\ln X$
Evidence
1.6
0.8
0.0
$-\ln X$

2.3.2 Trace plot

The following plot is called a *trace plot*, which relates the evolution in prior compression to the actual values of the parameters. For each parameter, we get a scatter plot of its value as a function of the prior compression, and we can see its range shrink as we reach higher and higher values of the likelihood.

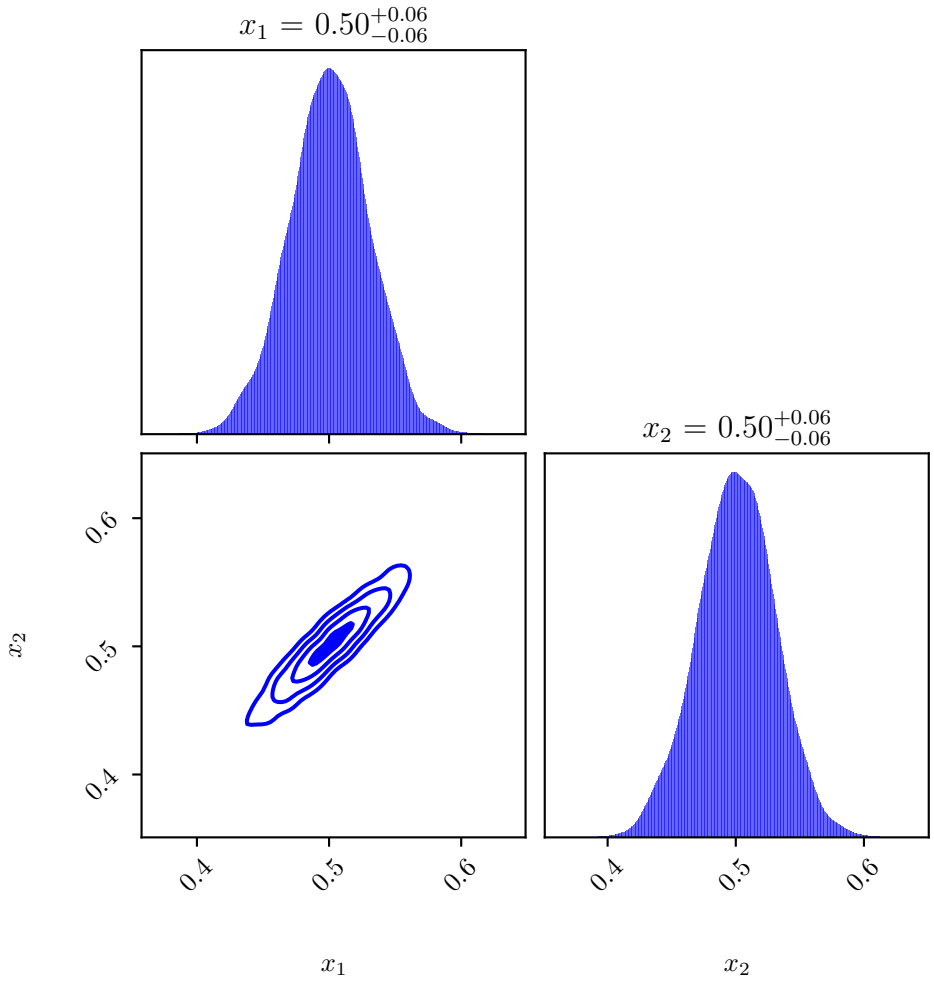
More specifically, this plot is showing the locations of the *dead points*, i.e. the lowest-likelihood points that get discarded at each iteration. When one of them gets replaced it will be assigned a new likelihood in the range $[L_i, L_{\max}]$, and after a certain number of iterations it will be replaced again. This is what the red lines are showing; they should be moving "randomly" across the parameter space, if we were to see them remain in a given region in the parameter space we would have an indication that our replacement algorithm is not performing properly.

The points are colored by their importance weights, $w_i L_i / Z$, and to the right we see a density estimate plot for the posterior distribution of each parameter together with a 95% confidence interval.



2.3.3 Corner plot

The *corner plot* is a great tool to extract physical understanding about our model. It shows the marginal density plot for all our parameters, as well as for each pair of parameters.



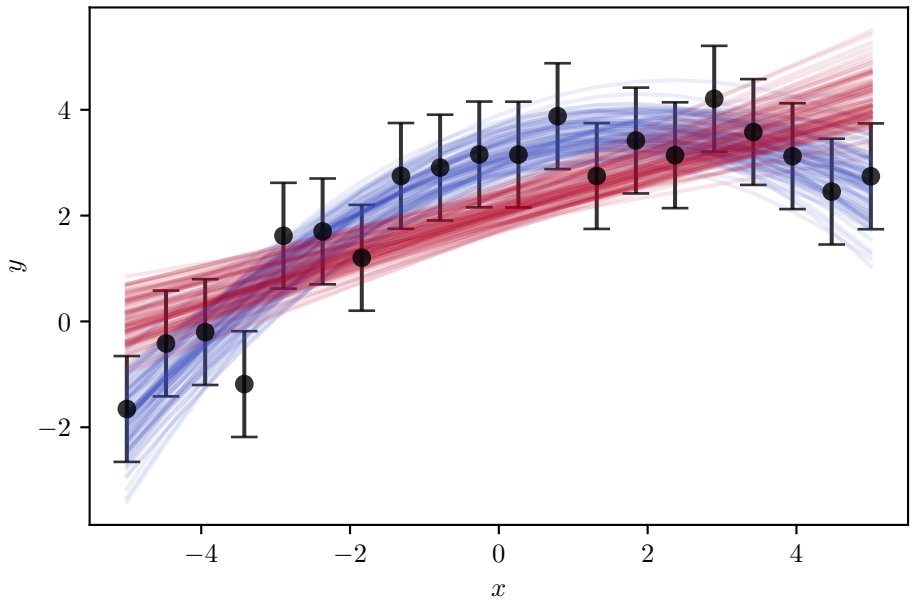
2.3.4 Posterior predictive density plot

For this example the previous example is not adequate, so we will use the problem from Section 1.1.

Is our model able to reproduce the data? A convenient visualization to gain insight into this question is the *posterior predictive distribution*: the distribution expected of new data, conditional on our model and the data we observed.

$$
\begin{aligned}
p(y_{\text{pred}}|y_{\text{obs}}, M) &= \int p(y_{\text{pred}}, \theta|y_{\text{obs}}, M)\mathrm{d}\theta \\
&= \int p(y_{\text{pred}}|\theta, y_{\text{obs}}, M)p(\theta|y_{\text{obs}}, M)\mathrm{d}\theta \\
&= \int \underbrace{p(y_{\text{pred}}|\theta, M)}_{\text{model}} \underbrace{p(\theta|y_{\text{obs}}, M)}_{\text{posterior}} \mathrm{d}\theta
\end{aligned}
$$

We can show this in different ways; an easy one to implement is by drawing samples from the posterior and plotting the model realization corresponding to each of them. The ensemble of curves will follow the posterior predictive distribution.



2.4 Information gain

The typical set of the posterior is often "much smaller" than the prior: how can this be quantified? A useful way to do so is the relative entropy, or Kullback-Leibler divergence between posterior and prior, often also called *information gain* or H:

$$
H = \mathrm{KL}(p \parallel \pi) = \int p(\theta) \log\left(\frac{p(\theta)}{\pi(\theta)}\right) \mathrm{d}\theta = \int p(X) \log p(X) \mathrm{d}X \,.
$$

The base of the logarithms defines the unit used for the information gain:

- if they are natural logarithms (base e) the values are in *nats*;

- if they are base 2 logarithms, the values are in *Shannons* or *bits*;
- if they are base 10 logarithms, the values are in *Hartleys* or *dex*.

One way to interpret this is related to the signal-to-noise ratio (Sivia and Skilling 2006, eq. 9.15):

$$H \sim \#\text{ active components in the data} \times \log(\text{signal-to-noise ratio})$$

It can also be seen as a "volumetric compression" from the prior to the posterior. If the distributions are uniform, this intuition is exact (Petrosyan and Handley 2022): suppose we have $\pi(\theta) \equiv [\theta \in V_\pi]/V_\pi$ and $p(\theta) \equiv [\theta \in V_p]/V_p$, where the two volumes are such that $V_p \subset V_\pi$.[3] Then,

$$\mathrm{KL}(p \parallel \pi) = \int_{V_\pi} \frac{[\theta \in V_p]}{V_p} \log \frac{V_\pi}{V_p} \mathrm{d}\theta = \log \frac{V_\pi}{V_p},$$

The integral can be restricted to the V_p volume, since the integrand is zero outside it. Another useful special case to develop an intuition is the one in which our prior is a Gaussian distribution with standard deviation σ, and our posterior is centered on the same value but its standard deviation is $\sigma' = \sigma/k$. Then, the information gain is (Buchner 2022):

$$\mathrm{KL}(\mathcal{N}(\mu, \sigma') \parallel \mathcal{N}(\mu, \sigma)) = \log k + \frac{1}{2}\left(\frac{1}{k^2} - 1\right).$$

We can compute the theoretical value for the entropy in the example above, since the computation is analytic in the Gaussian case: it is known that the differential entropy of a d-dimensional multivariate Gaussian random variable $f \sim \mathcal{N}(\mu, C)$ is (Cover and Thomas 2006, theorem 8.4.1):

$$-\int_{\mathbb{R}^d} f(\theta) \log f(\theta)\ \mathrm{d}^d\theta = \frac{d}{2}\log(2\pi e) + \log\det C\,.$$

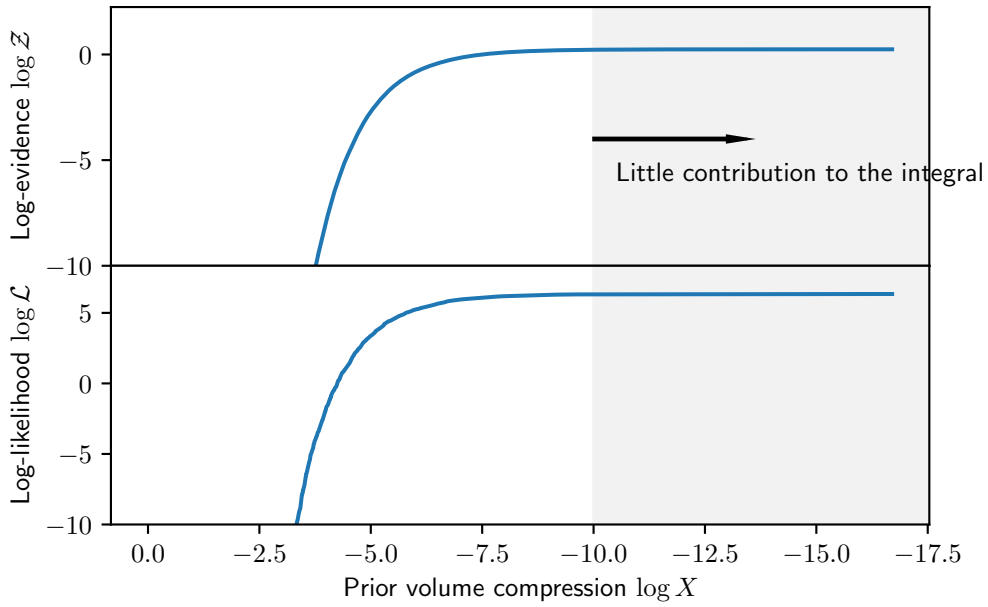
The integral we need to compute H is very similar. The sign is opposite, and instead of being over $\mathbb{R}^2$ it is only over $[0,1]^2$. The region outside of this box is, however, more than 10σ from the mean, therefore its contribution is negligible. The results are compatible:

```
H obtained from sampling = 5.082629561778999
Analytical estimate for H = 5.339190178719785
```

[3]The square brackets are the Iverson bracket, as defined by Knuth (1992).

2.4.1 Algorithm termination

As described, the nested sampling algorithm could continue in its compression forever, with ever-smaller compression $\log X$. However, if the likelihood function is bounded, after a certain number of iterations the prior volume will be concentrated around its maximum value,



2.4.2 Exploration time

The information gain provides an estimate of the expected volumetric compression: we expect to be sampling from the bulk of the posterior mass when $\log X \sim -H$, which will take approximately $n_{\text{live}}H \pm \sqrt{n_{\text{live}}H}$ iterations due to Poisson variability.

If $n_{\text{live}}H$ is large, we expect the dominant source of uncertainty to be the one on the compression reached at that stage - it will be much larger than the uncertainty on the compression in the $\mathcal{O}(\sqrt{n_{\text{live}}H})$ iterations required to traverse the posterior mass. This leads us to the estimate

$$\texttt{std}(\log Z) \approx \sqrt{\frac{H}{n_{\text{live}}}}\,.$$

So, for a fixed variance on the evidence, we need to have $n_{\text{live}} \propto H$; therefore, the number of iterations is approximately $n_{\text{iter}} \propto n_{\text{live}}H \propto H^2$.

2.4.3 Sampling options

Nested sampling is, in a way, a meta-algorithm: the sub-algorithm used to sample a new point from the prior subject to the constraint that its likelihood be higher than the last rejected point is a crucial determinant of the computational complexity and behavior of the implementation. This problem is called Likelihood-Restricted Prior Sampling.

Buchner (2021) gives a review of several possible options, which can be classified into:

- MCMC-based methods
- region-based methods
- hybrid methods
- ML-based methods

In all cases,

3 Nested sampling acceleration techniques

The time complexity for nested sampling can be estimated as

$$T \propto n_{\text{live}} \times \langle t_{\text{like}} \rangle \times \langle n_{\text{replace}} \rangle \times H$$

where

- n_{live} is the number of live points used;
- $\langle t_{\text{like}} \rangle$ is the average likelihood evaluation time;
- $\langle n_{\text{replace}} \rangle$ is the average number of likelihood evaluations required to find a replacement for each dead point;
- H is the information gain between posterior and prior.

Based on this equation, we can think of several ways to accelerate sampling (Roulet and Venumadhav 2024).

3.1 Varying the number of live points

The original nested sampling algorithm evolves a fixed number of live points throughout the run. It has since been proposed that it can be more efficient to vary it (Speagle 2020), depending on what is the quantity we are more interested in evaluating.

A nested sampling run gives us estimates for both the evidence and the posterior. If what we are most interested in is precisely computing the evidence, we should allocate more points in

the early stages of compression, even though the posterior mass is very low there: this will give us better statistics on the active volume when we reach the posterior bulk.

If we are more interested in the posterior, on the other hand, we can do the initial compression with fewer points and use more once we reach the posterior bulk.

It turns out (Higson et al. 2019) that, while both these objectives can be optimized independently, the standard choice of using a constant number of live points is not on the Pareto front for the uncertainties on evidence and posterior: a balanced dynamical allocation of points can improve both compared to the constant case.

3.2 Likelihood acceleration

A direct way to speed up any likelihood-based approach is to accelerate the evaluation of the likelihood, $\langle t_{\rm like} \rangle$. The way to do so is heavily dependent on the model being considered.

3.3 Replacement efficiency

As discussed in Section 2.4.3, there are many different ways to solve the likelihood-restricted sampling problem. While there is no panacea here, trying different parameters for the proposal method or switching to a different one altogether is a useful strategy when looking for acceleration.

3.4 Prior deformation

The compression H is a significant determinant of the sampling cost. Often, we want to use wide, uninformative priors, which significantly increases the computational cost, even though it is known that the posterior is contained within a relatively small region within them.

A useful observation ((Chen, Feroz, and Hobson 2022),(Petrosyan and Handley 2022)) is that when studying an inference problem defined by a likelihood $\mathcal{L}$ and prior π we can equivalently study a different one, defined by $\tilde{\mathcal{L}}$ and $\tilde{\pi}$: as long as at every point the condition

$$\tilde{\mathcal{L}}(\theta)\tilde{\pi}(\theta) = \mathcal{L}(\theta)\pi(\theta)$$

holds, the evidences, as well as the posteriors, will be the same. As long as we have some approximate knowledge about the shape of the posterior, then, we can apply a carefully-chosen shift to the prior which "zooms in" on the region of interest, reducing the compression required. If the error in $\log \mathcal{Z}$ is kept constant, reducing H results in a quadratic acceleration of the inference.

References

Ashton, Greg, Noam Bernstein, Johannes Buchner, Xi Chen, Gábor Csányi, Andrew Fowlie, Farhan Feroz, et al. 2022. "Nested Sampling for Physical Scientists." *Nature Reviews Methods Primers* 2 (1): 1–22. https://doi.org/10.1038/s43586-022-00121-x.

Buchner, Johannes. 2021. "Nested Sampling Methods." http://arxiv.org/abs/2101.09675.

———. 2022. "An Intuition for Physicists: Information Gain from Experiments." August 26, 2022. https://doi.org/10.48550/arXiv.2205.00009.

Chen, Xi, Farhan Feroz, and Michael Hobson. 2022. "Bayesian Posterior Repartitioning for Nested Sampling." July 4, 2022. https://doi.org/10.48550/arXiv.1908.04655.

Cover, Thomas M., and Joy A. Thomas. 2006. *Elements of Information Theory 2nd Edition*. 2nd edition. Hoboken, N.J: Wiley-Interscience.

Higson, Edward, Will Handley, Michael Hobson, and Anthony Lasenby. 2019. "Dynamic Nested Sampling: An Improved Algorithm for Parameter Estimation and Evidence Calculation." *Statistics and Computing* 29 (5): 891–913. https://doi.org/10.1007/s11222-018-9844-0.

Knuth, Donald E. 1992. "Two Notes on Notation." http://arxiv.org/abs/math/9205211.

Petrosyan, Aleksandr, and William James Handley. 2022. "SuperNest: Accelerated Nested Sampling Applied to Astrophysics and Cosmology." December 4, 2022. http://arxiv.org/abs/2212.01760.

Roulet, Javier, and Tejaswi Venumadhav. 2024. "Inferring Binary Properties from Gravitational Wave Signals." February 17, 2024. https://doi.org/10.1146/annurev-nucl-121423-100725.

Sivia, Devinderjit, and John Skilling. 2006. *Data Analysis: A Bayesian Tutorial*. Oxford University Press. https://books.google.com?id=lYMSDAAAQBAJ.

Skilling, John. 2006. "Nested Sampling for General Bayesian Computation." *Bayesian Analysis* 1 (4): 833–59. https://doi.org/10.1214/06-BA127.

Speagle, Joshua S. 2020. "DYNESTY: A Dynamic Nested Sampling Package for Estimating Bayesian Posteriors and Evidences." *Monthly Notices of the Royal Astronomical Society* 493 (April): 3132–58. https://doi.org/10.1093/mnras/staa278.