Computing the Bayesian evidence

From Bayes' theorem

$$P(x|d) = \frac{P(d|x,M)P(x|M)}{P(d|M)} \tag{1}$$

$$= \frac{\mathcal{L}(d|x)\pi(x)}{Z}, \tag{2}$$

where the normalising constant, known as the *Bayesian evidence*, is given by

$$Z = \int_{\mathcal{X}} \mathrm{d}x \; \mathcal{L}(x|d)\pi(x). \tag{3}$$

Computing the evidence, $Z = P(\text{data}|\text{model})$, is the key step in Bayesian model selection. This involves integrating over the full space of model parameters, $x \in \mathcal{X}$. The integral is usually impossible to perform analytically and difficult to evaluate with standard numerical integration methods (e.g. quadrature) for all but the simplest models. In this part of the course we will discuss various approaches to calculating the evidence.

0.1 Analytic computation

Analytically computing the Bayesian evidence is generally only possible when the posterior happens to be a simple, well-known probability distribution. This can happen in simple (usually low-dimensional) problems if we are able to find a *conjugate prior*.

0.2 Laplace's approximation

In general, finding the Bayesian evidence involves a complicated, high-dimensional integral over the model parameter space. In all but the simplest possible problems, these types of integrals are impossible to perform analytically. But it may be possible to approximate the evidence integral with another integral that can be evaluated analytically.

The *Laplace approximation* involves approximating the shape of the posterior using a multivariate normal distribution and then integrating this function instead.

If our problem involves parameters x, and we have prior and likelihood functions $\pi(x)$ and $\mathcal{L}(x)$ respectively, then the *unnormalised* posterior distribution $P^*(x|d)$ is given by

$$P^*(x|d) = \mathcal{L}(x|d)\pi(x)\,, \tag{4}$$

then the normalising evidence $Z \equiv P(D)$ is given by the following integral;

$$Z = \int_{\mathcal{X}} \mathrm{d}x \; P^*(x|d)\,. \tag{5}$$

Let us suppose we know that this distribution has a maximum value at the parameters $\hat{x}$. (It is usually relatively easy to find the parameters $\hat{x}$ numerically.) We now Taylor expand the quantity $\log P^*(x|d)$ about this peak location. In order to keep things simple, initially let's work in just 1-dimension where we have

$$\log P^*(x|d) = \log P^*(x|d) - \frac{c}{2}(x-\hat{x})^2 + \ldots\,, \tag{6}$$

where $c = -\mathrm{d}^2/\mathrm{d}x^2|_{x=\hat{x}} \log P^*(x|d)$. We assume that $c > 0$. There is no linear term because we have chosen to expand about the peak where the first derivative vanishes. Taking the exponential of this equation we find

$$\log P^*(x|d) \approx P^*(\hat{x}|d) \exp\left(-\frac{c}{2}(x-\hat{x})^2\right)\,, \tag{7}$$

the posterior is approximately Gaussian close to the peak, $\hat{x}$. If we assume that the parameter is free to take any value in the range $-\infty < x < \infty$, then we can integrate to find

$$Z \approx P^*(\hat{x}|d) \int_{-\infty}^{\infty} \mathrm{d}x \; \exp\left(-\frac{c}{2}(x-\hat{x})^2\right) = P^*(\hat{x}|d)\sqrt{\frac{2\pi}{c}}\,, \tag{8}$$

where we have used the standard result for the Gaussian integral.

The result in Eq. 8 is the Laplace approximation for the evidence in 1-dimension.

Notice that the Laplace approximation for the evidence in equation 8 is *not* invariant under a change of parameters. If we change variables $x \to x' = y(x)$, where $y(x)$ a suitable function of the original x, then the second derivative c transforms as $c \to c'$, where

$$c' = \frac{c}{\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2} \,. \tag{9}$$

If we instead use c' in equation 8 then we obtain a different approximation for Z. This is undesirable because the true evidence (defined in 5) is invariant under such a change of parameters. The accuracy of the Laplace approximation depends on which parameterisation we choose to use; this is expected, expressed in certain parameters the posterior may be well approximated as a Gaussian (in which case the Laplace approximation is expected to be accurate) but when expressed in terms of other parameters the posterior may be more complicated.

Exercise 0.1:
Derive the result for c' in equation 9.

We can repeat this calculation for a general multivariate problem. In this case the parameters vector $\vec{x}$ has components x_i where $i = 1, 2, \ldots, D$. We can approximate the logarithm of the unnormalised posterior distribution by the following multivariate Taylor series where we have expanded about the peak,

$$\log P^*(\vec{x}|d) \approx \log P^*(\hat{\vec{x}}|d) - \frac{1}{2}\sum_{i=1}^{D}\sum_{j=1}^{D} C_{ij}(x_i - \hat{x}_i)(x_i - \hat{x}_i) \,. \tag{10}$$

The constants C_{ij} form an $D \times D$ matrix and are given by the negative of the matrix of second derivatives (a.k.a. the *Hessian*) of the log-posterior;

$$C_{ij} = -\frac{\partial^2}{\partial x_i \partial x_j}\bigg|_{\vec{x}=\hat{\vec{x}}} \log P^*(\vec{x}|d) \,. \tag{11}$$

The derivatives may be evaluated either analytically or numerically. If we again assume that the parameter are free to take any value in the range $-\infty < x_i < \infty$, then we can integrate to find

$$Z \approx P^*(\hat{\vec{x}}|d)\sqrt{\frac{(2\pi)^D}{|\mathbf{C}|}}\,, \tag{12}$$

where $|\mathbf{C}|$ is the determinant of the matrix $\mathbf{C} = C_{ij}$ and where we have used the standard result for the multivariate Gaussian integral.

The result in Eq. 12 is the full *Laplace approximation*. This is a useful semi-analytic method for approximating the Bayesian evidence. It is only semi-analytic because the location of the maximum $\hat{\vec{x}}$ (and possibly the derivatives C_{ij}) must still be found numerically; however, this is usually much easier than numerically evaluating a high dimensional integral.

Again, as with the 1-dimensional version, the Laplace approximation in equation 12 is *not* invariant under a change of parameters, $\vec{x} \to x' = x'(x)$.

The Laplace approximation is expected to work reasonably well if the posterior has a single, narrow region of significant support that is far from any boundaries in the parameter space and that is well approximated by a multivariate Gaussian. The accuracy of the Laplace approximation can be improved if it is possible to find a change of model parameters that makes the posterior more Gaussian And if the posterior contains multiple, widely-separated modes then Laplace's approximation can be applied separately to each peak.

Nevertheless, the Laplace approximation is really very crude. Among other shortcomings, there is no practical way to estimate the uncertainty in the result in Eq. 12. The usefulness of the Laplace approximation lies primarily in the fact that it easy (and computationally cheap) to evaluate.

0.3 Thermodynamic integration

Thermodynamic integration[1] is an MCMC-based method than can be use to estimate the Bayesian evidence integral in Eq. 3.

Thermodynamic integration introduces an *annealing*, or *inverse temperature* parameter β into the likelihood. The modified *annealed likelihood* is defined as $\mathcal{L}(d|x,\beta) = \mathcal{L}(d|x)^\beta$ where $0 < \beta \leq 1$. The notation here is chosen to be reminiscent of that in statistical

[1] Goggans & Chi (2004) "Using Thermodynamic Integration to Calculate the Posterior Probability in Bayesian Model Selection Problems", AIP Conf Proc, 707, 59 https://doi.org/10.1063/1.1751356.

mechanics where $\beta^{-1} = k_{\mathrm{B}}T$ commonly appears in, for example, the Boltzmann distribution. We then write down a modified version of Bayes' theorem for the model parameters x conditioned on a fixed value of β; this modified posterior is given by

$$P(x|d,\beta) = \frac{\mathcal{L}(d|x)^{\beta}\pi(x)}{Z(\beta)}, \tag{13}$$

$$\text{where} \quad Z(\beta) = \int_{\mathcal{X}} \mathrm{d}x \; \mathcal{L}(d|x)^{\beta}\pi(x). \tag{14}$$

The true, unmodified likelihood is recovered by setting $\beta = 1$ (this is called the low-temperature limit) and a flat likelihood equal to 1 everywhere is obtained by setting $\beta = 0$ (this is called the high-temperature limit). Therefore, $Z(\beta = 1) = Z$ (this follows from Eq. 3) and $Z(\beta = 0) = 1$ (this follows from the fact that prior is a normalised probability distribution).

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\beta}\log Z(\beta) &= \frac{1}{Z(\beta)}\frac{\mathrm{d}Z}{\mathrm{d}\beta} && (15)\\
&= \frac{1}{Z(\beta)}\int_{\mathcal{X}} \mathrm{d}x \; \pi(x)\mathcal{L}(d|x)^{\beta}\log\mathcal{L}(d|x) && (16)\\
&= \int_{\mathcal{X}} \mathrm{d}x \; P(x|d,\beta)\log\mathcal{L}(d|x) && (17)\\
&= \mathrm{E}_x[\log\mathcal{L}(d|x)|\beta] && (18)
\end{aligned}$$

On the last line we have written the integral using a notation that emphasises that this is the expectation of the quantity $\log\mathcal{L}(d|x)$ over realisations of the model parameters x drawn from the modified Bayesian posterior at a fixed value of the inverse temperature β. Using the fact that $Z(1) = Z$ and $Z(0) = 1$, we can integrate Eq. 15 to find

$$\log Z = \int_0^1 \mathrm{d}\beta \; \mathrm{E}_x[\log\mathcal{L}(d|x)|\beta]. \tag{19}$$

The method of thermodynamic integration relies on the ability of the MCMC methods described earlier in the course to approximate the expectation value of functions of the parameters.

We pick a series of increasing values for the inverse temperatures $0 \leq \beta_\mu \leq 1$ for $\mu = 0, 1, 2, 3, \ldots, M$ with $\beta_0 = 0$ and $\beta_M = 1$. The sequence $\beta_0^{-1}, \beta_1^{-1}, \beta_2^{-1}, \ldots, \beta_M^{-1}$ is called the *temperature ladder*. For each temperature a MCMC is run where the target distribution is the annealed posterior $P(x|d,\beta_\mu)$. These Markov chains are used to obtain N posterior samples $x_i^{(\beta_\mu)} \overset{\mathrm{iid}}{\sim} P(x|d,\beta_\mu)$, for $i = 0, 1, \ldots, N-1$. The expectation in Eq. 18 is then approximated by the Monte-Carlo sum,

$$\mathrm{E}_x[\log\mathcal{L}(d|x)|\beta_\mu] \approx \frac{1}{N}\sum_{i=0}^{N-1}\log\mathcal{L}(d|x_i^{(\beta_\mu)}). \tag{20}$$

Finally, the evidence can be obtained by evaluating the 1-dimensional integral in Eq. 19 using the trapezium rule;

$$\log Z \approx \frac{1}{2}\sum_{\mu=1}^{M}\big(\mathrm{E}_x[\log\mathcal{L}(d|x)|\beta_\mu] + \mathrm{E}_x[\log\mathcal{L}(d|x)|\beta_{\mu-1}]\big)\Delta\beta_\mu, \tag{21}$$

where $\Delta\beta_\mu = \beta_\mu - \beta_{\mu-1}$.

This procedure allows us to use MCMC methods to evaluate the evidence integral by running multiple Markov chains at different temperatures and aggregating the results. Because thermodynamic integration requires an MCMC chain to be run at inverse temperature $\beta = 1$, a nice feature of the algorithm produces samples from the posterior at the same time as calculating the Bayesian evidence.

It's not hard to see how, in practice, this procedure can quickly become very expensive. In order for the trapezium rule integral to be accurate, a large number of temperatures need to be used. For each temperature, a Markov chain needs to be run to produce a large number of (independent) samples from the modified posterior to calculate (and reliably estimate the associated error) on the expectation quantity $\mathrm{E}_x[\log\mathcal{L}(d,x)|\beta]$.

Some performance gain can be achieved by running the Markov chains sequentially in order of decreasing temperature (i.e. increasing β) and using information from the previous temperature to initialise the next Markov chain thereby reducing the burn in period. Further improvements are possible by running all the different temperature chains in parallel and allowing them to exchange information during their evolution (in a way that ensures the detailed balance condition is preserved); this approach is known as *parallel-Tempered MCMC*.

0.4 The Savage-Dickey density ratio

The Savage-Dickey density ratio is not actually a method for calculating the evidence, $Z = P(d|M)$. Instead, it allows to calculate the *evidence ratio* (also known as the *Bayes factor*) between two models, $\mathcal{B}_{1,2} = P(d|M_1)/P(d|M_2)$, in the special situation where one of the models, say M_1, is *nested* inside the other model, M_2.

The term *nested models* applies to the situation where a simple model M_1 can be recovered from a more complicated model M_2 by fixing the value(s) of one of more of its parameters. For example, suppose that model M_2 has parameters (ϵ, ϕ) (where $\epsilon \in \mathbb{R}$ and $\phi \in \mathbb{R}^d$) and that model M_1 has parameters ϕ (i.e. the dimensionality of M_1 is one lower than that of M_2) and that when the extra parameter of M_2 takes a specific values, say $\epsilon = 0$, the two models makes the same prediction for the data; i.e.

$$\mathcal{L}(d|\phi, M_1) = \mathcal{L}(d|\epsilon = 0, \phi, M_2). \tag{22}$$

The Savage-Dickey method also assumes that we use consistent priors on the shared parameters of the two models. The priors must satisfy

$$\pi(\phi|M_1) = \pi(\phi|\epsilon = 0, M_2). \tag{23}$$

In practice, this is usually achieved by using a separable prior for the more complicated model, $\pi(\epsilon, \phi|M_2) = f(\epsilon)g(\phi)$ and then using $\pi(\phi|M_1) = g(\phi)$ as the prior for the simpler model.

The evidence for the simpler model M_1 can be written as

$$\begin{aligned} Z_{M_1} &= P(d|M_1) & (24)\\ &= \int \mathrm{d}\phi \; \mathcal{L}(d|\phi, M_1)\pi(\phi|M_1) & (25)\\ &= \int \mathrm{d}\phi \; \mathcal{L}(d|\phi, \epsilon = 0, M_2)\pi(\phi|\epsilon = 0, M_2) & (26)\\ &= P(d|\epsilon = 0, M_2) & (27)\\ &= \frac{P(\epsilon = 0|d, M_2)}{P(\epsilon = 0|M_2)} P(d|M_2), & (28) \end{aligned}$$

where on the second line we have used Eqs. 22 and 23 and on the final line we have used Bayes' theorem. The evidence for the more complicated model M_2 is $Z_{M_2} = P(d|M_2)$. Therefore, the Bayes factor is given by

$$\begin{aligned} \mathcal{B}_{1,2} &= \frac{Z_{M_1}}{Z_{M_2}} & (29)\\ &= \frac{P(\epsilon = 0|d, M_2)}{P(\epsilon = 0|M_2)}. & (30) \end{aligned}$$

This implies that Bayes factor is given by the ratio of the 1-dimensional posterior PDF evaluated at $\epsilon = 0$ to the 1-dimensional prior PDF evaluated at $\epsilon = 0$.

The prior is generally chosen to be a simple analytical function in which case the denominator of Eq. 30 is easy to evaluate. The posterior in the numerator is harder. An MCMC method can be used to sample the posterior for model M_2 to obtain independent samples $(\epsilon_i, \phi_I) \sim P(\epsilon, \phi | d, M_2)$ for $i = 1, 2, \ldots, N$. Discarding the ϕ_i values, the remaining samples ϵ_i can be used to estimate the posterior PDF on $P(\epsilon | d, M_2$ which can then be evaluated.

Any method can be used to estimate the density $P(\epsilon | d, M_2$ from the samples ϵ_i. For example, kernel density estimation (KDE) is a common choice. Care must be taken when the value $\epsilon = 0$ lies on a boundary of the prior $f(\epsilon)$ to avoid biasing the density estimate.

0.5 Avoiding the evidence altogether

The evidence may be so difficult to calculate that the best approach might be to find a way to avoid having to calculate it altogether.

0.5.1 Defining an Augmented model

Suppose we have two models: M_1 with parameters x_1 and M_2 with parameters x_2. Model 1 has the likelihood $\mathcal{L}(d | x_1, M_1)$ and prior $\pi(x_1 | M_1)$. Similarly, model 2 has the likelihood $\mathcal{L}(d | x_2, M_2)$ and prior $\pi(x_2 | M_2)$.

An augmented model, M_*, with parameters (ϵ, x_1, x_2) is defined with a prior $\pi(\epsilon, x_1, x_2 | M_*) = \mathbb{1}_{(0,1)}(\epsilon) \pi(x_1 | M_1) \pi(x_2 | M_2)$ and a likelihood

$$\mathcal{L}(d | \epsilon, x_1, x_2, M_*) = \begin{cases} \mathcal{L}(d | x_1, M_1) & \text{if } \epsilon < 1/2 \\ \mathcal{L}(d | x_2, M_2) & \text{if } \epsilon > 1/2 \end{cases} . \tag{31}$$

The idea now is that we sample in the parameter space of the augmented model, $(\epsilon, x_1, x_2)_i \sim P(\epsilon, x_1, x_2 | d)$, and use the ratio of the number of samples with $\epsilon_i < 1/2$ to the number of samples with $\epsilon_i > 1/2$ as an estimate of the Bayes factor.

Formally, this result can be derived by considering the relative probability that epsilon in greater/less than 1/2 in the augmented model. Consider the probability

$$\begin{align}
\mathrm{Prob}(\epsilon < 1/2|M_*) &= \int_0^{1/2} \mathrm{d}\epsilon \int \mathrm{d}x_1 \int \mathrm{d}x_1 \; P(\epsilon, x_1, x_2|M_*) \tag{32}\\
&= \frac{1}{Z_{M_*}} \int_0^{1/2} \mathrm{d}\epsilon \int \mathrm{d}x_1 \int \mathrm{d}x_2 \; \mathcal{L}(d|\epsilon, x_1, x_2, M_*)\pi(\epsilon, x_1, x_2|M_*) \tag{33}\\
&= \frac{1}{Z_{M_*}} \int_0^{1/2} \mathrm{d}\epsilon \int \mathrm{d}x_1 \int \mathrm{d}x_2 \; \mathcal{L}(d|\epsilon, x_1, x_2, M_*)\pi(x_1|M_1)\pi(x_2|M_2) \tag{34}\\
&= \frac{1}{Z_{M_*}} \int_0^{1/2} \mathrm{d}\epsilon \int \mathrm{d}x_1 \int \mathrm{d}x_2 \; \mathcal{L}(d|x_1, M_1)\pi(x_1|M_1)\pi(x_2|M_2) \tag{35}\\
&= \frac{1}{2Z_{M_*}} \int \mathrm{d}x_1 \; \mathcal{L}(d|x_1, M_1)\pi(x_1|M_1) \tag{36}\\
&= \frac{Z_{M_1}}{2Z_{M_*}}. \tag{37}
\end{align}$$

Similarly,

$$\mathrm{Prob}(\epsilon > 1/2|M_*) = \frac{Z_{M_2}}{2Z_{M_*}}. \tag{38}$$

Therefore, taking the ratio of Eqs. 37 and 38 gives

$$\mathcal{B}_{1,2} = \frac{Z_{M_1}}{Z_{M_2}} = \frac{\mathrm{Prob}(\epsilon < 1/2|M_*)}{\mathrm{Prob}(\epsilon > 1/2|M_*)}. \tag{39}$$

This method can be trivially extended to include any number of competing models.

0.5.2 The Bayes factor from importance sampling

Suppose we have two models with identical parameters x. Model 1 has the likelihood $\mathcal{L}(d|x, M_1)$ and model two has the likelihood $\mathcal{L}(d|x, M_2)$. We will use the same prior for both models; $\pi(x|M_1) = \pi(x|M_2)$.

If the two models make similar predictions for the data then we can use importance sampling to evaluate the Bayes factor between the two models.

The evidence for the second model is given by

$$\begin{aligned}
Z_{M_2} &= \int \mathrm{d}x \; \mathcal{L}(d|x, M_2)\pi(x|M_2), && (40)\\
&= \int \mathrm{d}x \; \frac{\mathcal{L}(d|x, M_2)}{\mathcal{L}(d|x, M_1)}\mathcal{L}(d|x, M_1)\pi(x|M_2), && (41)\\
&= \int \mathrm{d}x \; \frac{\mathcal{L}(d|x, M_2)}{\mathcal{L}(d|x, M_1)}\mathcal{L}(d|x, M_1)\pi(x|M_1), && (42)\\
&= Z_{M_1} \int \mathrm{d}x \; \frac{\mathcal{L}(d|x, M_2)}{\mathcal{L}(d|x, M_1)}P(x|d, M_1). && (43)
\end{aligned}$$

On the second line we have simply inserted a factors of $\mathcal{L}(d|x, M_1)$ into the numerator and denominator, on the third line we have used $\pi(x|M_1) = \pi(x|M_2)$ and on the forth line we have used Bayes' theorem to write the posterior for model M_2 as $P(x|d, M_2) = \mathcal{L}(d|x, M_1)\pi(x|M_1)/Z_{M_1}$. Therefore the Bayes factor is given by

$$\begin{aligned}
\mathcal{B}_{1,2} &= \frac{Z_{M_1}}{Z_{M_2}} && (44)\\
&= \int \mathrm{d}x \; \frac{\mathcal{L}(d|x, M_2)}{\mathcal{L}(d|x, M_1)}P(x|d, M_1) && (45)\\
&= \mathrm{E}_{x\sim P(x|d,M_1)}\left[\frac{\mathcal{L}(d|x, M_2)}{\mathcal{L}(d|x, M_1)}\right] && (46)
\end{aligned}$$

The right-hand side of this expression is the expectation of the likelihood ratio over values of the model parameters distributed according to the posterior of model M_2. We can therefore pick one of the models, say M_1, and run an MCMC to obtain posterior samples $x_i \sim P(x|d, M_1)$ which can then be used to calculate the Bayes factor using

$$\mathcal{B}_{1,2} \approx \frac{1}{N}\sum_{i=0}^{N-1} \frac{\mathcal{L}(d|x_i, M_2)}{\mathcal{L}(d|x_i, M_1)} \qquad (47)$$

This method works best when the two models are similar; i.e. when $\mathcal{L}(x|d, M_1) \approx \mathcal{L}(x|d, M_2)$.

0.6 Nested sampling

At this point, it might seem like we're starting to loose our way. We began this section with the claim that the Bayesian evidence is the most important quantity for model comparison and we need a reliable way to calculate it. However, the last few sections have been spent describing hacks and swindles specifically designed to avoid ever having to actually calculate an evidence!

The only thing we've seen so far that resembles a general algorithm for calculating the evidence was thermodynamic integration (Sec. 0.3) and although that method has been successfully used in some problems, the computational costs involved in obtained accurate evidences coupled with the need for extensive fine tuning in the algorithm (e.g. the choice of the temperature ladder, β_μ) has prevented this from being becoming widely used.

We will now remedy this problem by describing the *nested sampling algorithm* which over the last few years has become the most common method used for calculating the Bayesian evidence.

Nested sampling is numerical integration algorithm for evaluating the evidence integral in Eq. 3 in high-dimensional Bayesian inference problems where analytic approximations fail[2]. As a by product, nested sampling also produces samples from the posterior distribution; therefore, nested sampling can also be considered a stochastic sampling algorithm.

As always, the inputs to the Bayesian inference problem are the prior, $\pi(x)$, and the likelihood $\mathcal{L}(d|x)$, regarded as a function of the model parameters, $x \in \mathcal{X}$. The posterior is given by Bayes' theorem $P(x|d) = \mathcal{L}(d|x)\pi(x)/Z$, where the normalisation constant, known as the *evidence*, is given by the integral

$$Z = \int_{\mathcal{X}} \mathrm{d}x \; \mathcal{L}(d|x)\pi(x). \tag{48}$$

We will be interested in problems where $\mathcal{X}$ is high dimensional and this integral cannot be evaluated (or suitably approximated) analytically.

Let $\mathcal{L}_{\max}$ denote the maximum likelihood value over the parameter space,

$$\mathcal{L}_{\max} = \max_{x \in \mathcal{X}} \mathcal{L}(d|x). \tag{49}$$

For simplicity, it is assumed that there is a unique global maximum (this assumption can be relaxed later).

Let $\xi(L)$ be the prior probability (sometimes called the *probability mass*) associated with likelihoods greater than a given value L;

$$\begin{aligned} \xi(L) &= \int_{\{x:\mathcal{L}(d|x)>L\}} \mathrm{d}x \; \pi(x) && (50) \\ &= \int_{\mathcal{L}>L} \mathrm{d}x \; \pi(x), && (51) \end{aligned}$$

[2] Skilling (2004) "Nested Sampling", AIP Conf. Proc. 735, 395–405 https://doi.org/10.1063/1.1835238

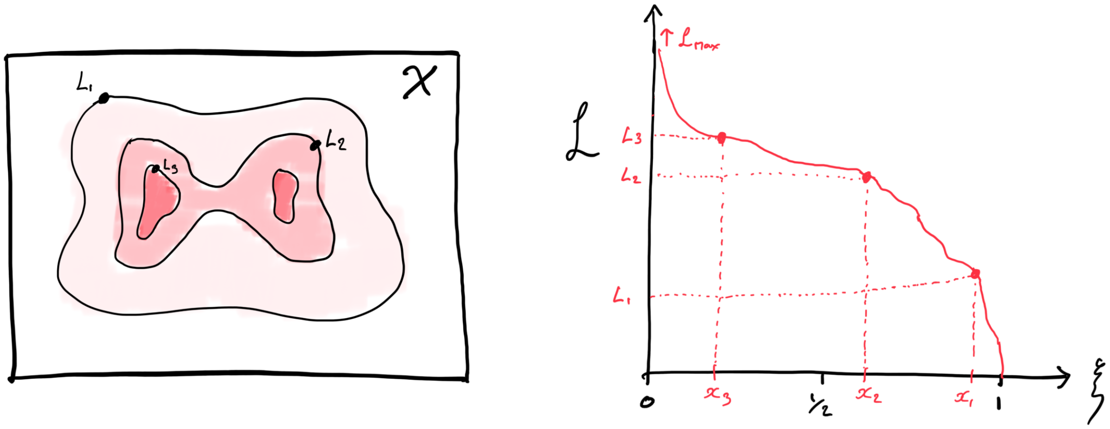


Figure 1: Sketches to illustrate the nested sampling function $\xi(L)$. *Left:* iso-probability contours of the likelihood function $\mathcal{L}(d|x)$ in the sample space $x \in \mathcal{X}$. *Right:* the likelihood as a function of ξ. For most realistic problems the function $L(\xi)$ is very strongly peaked at small values of ξ; parameters with high likelihood are rare!

where on the second line we have just introduced a shorthand notation for the constraint $\mathcal{L}(d|x) > L$ in the domain of the integral. The definition of $\xi(L)$ and its relationship to the usual likelihood as a function of the parameters $\mathcal{L}(d|x)$ is illustrated in the sketch in Fig. 1. From its definition and the fact that $\pi(x)$ is a probability density, it follows that $\xi(L)$ is a *decreasing* function satisfying

$$\xi(0) = 1, \quad \text{and} \quad \xi(\mathcal{L}_{\text{max}}) = 0. \tag{52}$$

In problems where x is continuous (and provided there are no regions where the likelihood function is exactly flat) the function $\xi(L)$ is *strictly decreasing* and can be inverted to give $L(\xi)$. (In discrete problems where a finite amount of prior probability is assigned to the same value of the likelihood, a small amount of noise, or *jitter*, can be added to the likelihood to obtain a strictly decreasing $\xi(L)$.)

The functions $\mathcal{L}(d|x)$ and $L(\xi)$ are both called the *likelihood*. The former is expressed as a function of the model parameters x while the latter is expressed as a function of the enclosed prior probability mass $0 \leq \xi \leq 1$. The function $L(\xi)$ is explored further with an example in Box 0.1.

The quantity dξ is the prior mass associated with likelihoods in the range $(L, L + \text{d}L)$. Therefore, the range dξ contributes an amount $L\text{d}\xi$ to the total evidence. Summing all these contributions gives

$$Z = \int_0^1 \mathrm{d}\xi \; L(\xi). \tag{53}$$

At first, this appears somewhat miraculous; we have changed the high-dimensional integral in Eq. 48 into the 1-dimensional integral in Eq. 53. (This is reminiscent of what was done in the method of thermodynamic integration; see Eq. 19.) Of course, we have just moved the difficulty into finding and inverting the function $\xi(L)$.

Box 0.1: An example of the nested sampling function $\xi(L)$

Consider a d-dimensional inference problem $x \in \mathcal{X} = \mathbb{R}^n$ with likelihood

$$\mathcal{L}(d|\mathbf{x}) = \exp\left(-|\mathbf{x}|^2/2\right) \tag{i}$$

and a prior on that is uniform inside an n-ball of radius R centred on the origin,

$$\pi(x) = \frac{1}{V_n R^n}\begin{cases} 1 & \text{if } |x| < R \\ 0 & \text{else} \end{cases} . \tag{ii}$$

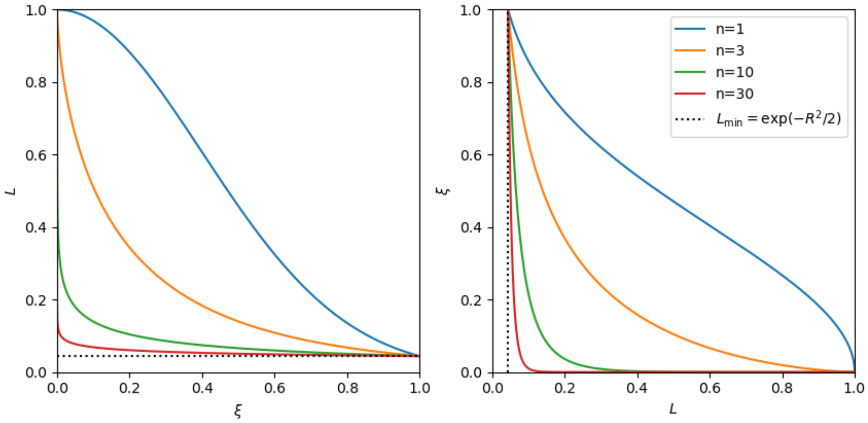
For this problem, the nested sampling function $\xi(L)$ can be calculated analytically. The iso-likelihood contours are d-dimensional spheres centred on the origin with the likelihood decreasing monotonically with radius. The sphere with radius $|x| = r$ has a likelihood value of $L = \exp(-r^2/2)$. Therefore, from the definition in Eq. 51,

$$\begin{aligned} \xi(L) &= \frac{n}{R^n}\int_0^{\sqrt{-2\log L}} \mathrm{d}r \; r^{n-1} && \text{(iii)} \\ &= \left(\frac{\sqrt{-2\log L}}{R}\right)^n . && \text{(iv)} \end{aligned}$$

And the inverse function is

$$L(\xi) = \exp\left(\frac{-R^2\xi^{2/n}}{2}\right). \tag{v}$$

These functions are plotted for $R = 2.5$ in $n = 1$, 3, 10, and 30 dimensions. Note how in high dimensions regions of parameter space with high likelihood are rare.



Given a decreasing sequence of points in the prior volume, $0 < \xi_M < \xi_{M-1} < \ldots < \xi_1 < 1$, and their associated likelihoods $L_\mu = L(\xi_\mu)$, the 1-dimensional evidence integral in Eq. 53 can be approximated using the trapezium rule;

$$Z \approx \frac{1}{2} \sum_{\mu=1}^{M} (L_{\mu-1} + L_\mu) \Delta\xi_\mu, \tag{54}$$

where $\Delta\xi_\mu = \xi_\mu - \xi_{\mu-1}$ and where we define $\xi_0 = 1$ with the associated $L_0 = 0$. By relabelling terms in the sum, we can arange this into the slightly more convenient form

$$Z \approx \sum_{\mu=1}^{M} w_\mu L_\mu, \tag{55}$$

where $w_\mu = (\xi_{\mu-1} - \xi_{\mu+1})/2$ and where we define $\xi_{M+1} = \xi_M$.

The nested sampling algorithm begins by initialising the parameter space location of a number of *live points* by drawing from the prior; $x_j \sim \pi$, independently for $j = 0, 1, \ldots, N_{\text{live}} - 1$. (It is assumed that we already have the ability to sample from the prior, $x \sim \pi$; in practice, this is not a significant restriction because the prior is usually chosen to have a simple analytic form.) The number of live points, N_{live}, should be chosen to be a large number in order to obtain an accurate estimate for the evidence; for most Bayesian inference problems, typically N_{live} is chosen to be in the range 10^3–10^4.

The nested sampling algorithm then proceeds iteratively. At each iteration the live point with the lowest value of the likelihood $L_i = \min(\{\mathcal{L}(d|x_j)|j = 0, 1, \ldots, N_{\rm live}\})$ is replaced by a new live point with a higher likelihood. The position of this new live point is drawn from the prior, $x \sim \pi$, subject to the constraint that $\mathcal{L}(d|x)$ is larger than the likelihood of the point that has just been discarded. By proceeding in this way, over time the collection of live points climbs up the likelihood surface, eventually clustering near the peak $\mathcal{L}_{\rm max}$. This process results in an increasing sequence of likelihood values $L_i < L_{i+1}$ from the deceased points.

In order to use the sequence of likelihoods L_i in the expression for the evidence in Eq. 55 we need to know the associated values of ξ_i. For each deceased point, we need to know the prior probability mass associated with higher values of the likelihood. This is tackled probabilistically; at each iteration the prior mass shrinks by a factor, $\xi_i = t\xi_{i-1}$, where $0 < t < 1$ is a random variable that follows the distribution of the largest of $N_{\rm live}$ samples drawn uniformly from the interval $(0, 1)$. Therefore, we have

$$P(t) = \begin{cases} N_{\rm live} t^{N_{\rm live}-1} & \text{for } 0 < t < 1 \\ 0 & \text{otherwise} \end{cases} . \tag{56}$$

It is easy to show that $E[\log t] = -1/N_{\rm live}$ and $\mathrm{Var}[\log t] = 1/N_{\rm live}^2$. Therefore, we take $\xi_i \approx \exp(-i/N_{\rm live})$ and this will be an accurate estimate if $N_{\rm live}$ is large.

The nested sampling algorithm should be continued until the evidence integral is judged to have converged. The remaining contribution to the evidence from the surviving live points can be conservatively estimated as $\Delta Z \approx L_* \xi_i$, where L_* is the maximum likelihood value amongst the current set of live points. The algorithm is usually terminated when this drops below some user-specified tolerance.

The output of the nested sampling algorithm is the estimate for the evidence integral, Z, and the listed of weighted samples $\{(W_i, x_i)|\text{for}\, i = 0, 1, \ldots, N_{\rm iter}\}$, where $N_{\rm iter}$ is the number of iterations performed by the algorithm Alg. 0.1.

It should be clear from the above discussion that the main challenge for an practical implementation of the nested sampling algorithm is drawing samples from the prior subject to the constraint $L > L'$, where L' is current lowest likelihood value among all of the live points. This can be done in various different ways. In fact any of the MCMC algorithms described this course could be used for this purpose.

The nested sampling algorithm proceeds as follows.

Algorithm 0.1 The nested sampling algorithm

```
for $j = 0, 1, \ldots, N_{\mathrm{live}} - 1$ do
    $x_j \sim \pi(x)$                                        ▷ Initialise live points from prior
    $L_j \leftarrow \mathcal{L}(d|x_j)$
end for
$Z \leftarrow 0$                                             ▷ Initialise evidence estimate
samples ← []                                                 ▷ Empty list to store weighted samples
$i \leftarrow 0$
Stop ← False
while Stop == False do                                       ▷ Iterate $i = 0, 1, 2, \ldots$
    idx ← argmin($L_0, L_1, \ldots, L_{N_{\mathrm{live}}-1}$)     ▷ Current worst live point
    $L_i \leftarrow L_{\mathrm{idx}}$                        ▷ Likelihood of deceased point
    $\xi_i \leftarrow \exp(-i/N_{\mathrm{live}})$            ▷ Estimate of enclosed prior mass
    $w_i \leftarrow \big(\exp[-(i-1)/N_{\mathrm{live}}] - \exp[-(i+1)/N_{\mathrm{live}}]\big)/2$
    $Z \leftarrow Z + w_i L_i$                               ▷ Increment the evidence
    AppendTo(samples, $[L_i w_i, x_{\mathrm{idx}}]$)         ▷ Store weighted sample
    $x \sim \pi(x|\mathcal{L}(d|x) > L_i)$                   ▷ Sample the constrained prior
    $x_{\mathrm{idx}} \leftarrow x$                          ▷ Replace the deceased point
    $L_* \leftarrow \max(L_0, L_1, \ldots, L_{N_{\mathrm{live}}-1})$
    Stop ← Bool($L_* \xi_i <$ tol)                           ▷ Stopping condition
    $i \leftarrow i + 1$
end while
```