Computing the Bayesian evidence

From Bayes' theorem

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

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

$$Z = \int_{\mathcal{X}} \mathrm{d}x \; \mathcal{L}(x|d)\pi(x). \tag{2}$$

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

$$P^*(x|d) = \mathcal{L}(x|d)\pi(x)\,, \tag{3}$$

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

$$Z = \int_{\mathcal{X}} \mathrm{d}x \; P^*(x|d)\,. \tag{4}$$

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

$$\log P^*(x|d) = \log P^*(x|d) - \frac{c}{2}(x-\hat{x})^2 + \ldots\,, \tag{5}$$

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

$$\log P^*(x|d) \approx P^*(\hat{x}|d) \exp\left(-\frac{c}{2}(x-\hat{x})^2\right)\,, \tag{6}$$

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

$$Z \approx P^*(\hat{x}|d) \int_{-\infty}^{\infty} \mathrm{d}x \;\; \exp\left(-\frac{c}{2}(x-\hat{x})^2\right) = P^*(\hat{x}|d)\sqrt{\frac{2\pi}{c}}\,. \tag{7}$$

This is the Laplace approximation for the evidence in 1-dimension.

Notice that the Laplace approximation for the evidence in equation 7 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

$$c' = \frac{c}{\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2}\,. \tag{8}$$

If we instead use c' in equation 7 then we obtain a different approximation for Z. This is undesirable because the true evidence (defined in 4) 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 8.

We can repeat this calculation for a general multivariate D-dimensional 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,

$$\log P^*(\vec{x}|d) \approx \log P^*(\hat{\vec{x}}|d) - \frac{1}{2} C_{ij}(x_i - \hat{x}_i)(x_i - \hat{x}_i)\,. \tag{9}$$

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;

$$C_{ij} = -\frac{\partial^2}{\partial x_i \partial x_j}\Big|_{\vec{x}=\hat{\vec{x}}} \log P^*(\vec{x}|d)\,. \tag{10}$$

The derivatives may be evaluated either analytically or numerically. Again, the integral of this multivariate Gaussian is a well known result and involves the determinant of the matrix $\mathbf{C} = C_{ij}$,

$$Z \approx P^*(\hat{\vec{x}}|d)\sqrt{\frac{(2\pi)^D}{\det \mathbf{C}}}\,. \tag{11}$$

This result 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 11 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. 11. 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. 2.

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 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

$$P(x|d,\beta) = \frac{\mathcal{L}(d|x)^\beta \pi(x)}{Z(\beta)}, \tag{12}$$

$$\text{where} \quad Z(\beta) = \int_\mathcal{X} \mathrm{d}x \; \mathcal{L}(d|x)^\beta \pi(x). \tag{13}$$

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. 2) and $Z(\beta = 0) = 1$ (this follows from the fact that prior is a normalised probability

[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.

distribution).

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\beta}\log Z(\beta) &= \frac{1}{Z(\beta)}\frac{\mathrm{d}Z}{\mathrm{d}\beta} &(14)\\
&= \frac{1}{Z(\beta)}\int_{\mathcal{X}}\mathrm{d}x\;\pi(x)\mathcal{L}(d|x)^{\beta}\log\mathcal{L}(d|x) &(15)\\
&= \int_{\mathcal{X}}\mathrm{d}x\;P(x|d,\beta)\log\mathcal{L}(d|x) &(16)\\
&= \mathrm{E}_x[\log\mathcal{L}(d|x)|\beta] &(17)
\end{aligned}
$$

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. 14 to find

$$\log Z = \int_0^1 \mathrm{d}\beta\;\mathrm{E}_x[\log\mathcal{L}(d|x)|\beta]. \tag{18}$$

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 $\mu = 1, 2, 3, \ldots, M$ with $\beta_1 = 0$ and $\beta_M = 1$. 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 $x_i^{(\beta_\mu)} \overset{\text{iid}}{\sim} P(x|d, \beta_\mu)$, for $i = 0, 1, \ldots, N-1$. The expectation is then approximated by the Monte-Carlo sum,

$$\mathrm{E}_x[\log\mathcal{L}(d|x)|\beta_\mu] \approx \frac{1}{N}\sum_{i=0}^{N-1}\log\mathcal{L}(d|x_i^{(\beta_\mu)})^{\beta}. \tag{19}$$

Finally, the evidence is obtained by evaluating the 1-dimensional integral in Eq. 18 using the trapezium rule with $\Delta\beta_\mu = \beta_{\mu+1} - \beta_\mu$ to obtain

$$\log Z \approx \frac{1}{2}\sum_{\mu=1}^{M-1}\big(\mathrm{E}_x[\log\mathcal{L}(d|x)|\beta_{\mu+1}] - \mathrm{E}_x[\log\mathcal{L}(d|x)|\beta_\mu]\big)\Delta\beta_\mu. \tag{20}$$

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.

It's not hard to see how, in practice, this procedure can quickly become very expensive; for an accurate trapezium rule integral 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 $\mathrm{E}_x[\log \mathcal{L}(d,x)|\beta]$.

In practice, some performance gains 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.

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.

$$\mathcal{L}(d|\phi, M_1) = \mathcal{L}(d|\epsilon = 0, \phi, M_2). \tag{21}$$

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

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

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

$$\begin{aligned} Z_{M_1} &= P(d|M_1) && (23) \\ &= \int \mathrm{d}\phi \; \mathcal{L}(d|\phi, M_1)\pi(\phi|M_1) && (24) \\ &= \int \mathrm{d}\phi \; \mathcal{L}(d|\phi, \epsilon = 0, M_2)\pi(\phi|\epsilon = 0, M_2) && (25) \\ &= P(d|\epsilon = 0, M_2) && (26) \\ &= \frac{P(\epsilon = 0|d, M_2)}{P(\epsilon = 0|M_2)} P(d|M_2), && (27) \end{aligned}$$

where on the second line we have used Eqs. 21 and 22 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

$$\begin{aligned} \mathcal{B}_{1,2} &= \frac{Z_{M_1}}{Z_{M_2}} && (28) \\ &= \frac{P(\epsilon = 0|d, M_2)}{P(\epsilon = 0|M_2)}. && (29) \end{aligned}$$

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. 29 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 $\pi(x_2|M_2)$.

We can define an augmented model, M_*, with parameters (ϵ, x_1, x_2). We choose the prior of this new model to be $\pi(\epsilon, x_1, x_2|M_*) = \mathbb{1}_{(0,1)}(\epsilon)\pi(x_1|M_1)\pi(x_2|M_2)$ and its likelihood to be

$$\mathcal{L}(d|\epsilon, x_1, x_2, M_*) = \begin{cases} \mathcal{L}(d|M_1) & \text{if } \epsilon < 1/2 \\ \mathcal{L}(d|M_2) & \text{if } \epsilon > 1/2 \end{cases} . \tag{30}$$

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

$$\begin{aligned}
\mathrm{Prob}(\epsilon < 1/2|M_*) &= \int_0^{1/2} \mathrm{d}\epsilon \int \mathrm{d}x_1 \int \mathrm{d}x_1 \; P(\epsilon, x_1, x_2|M_*) && (31)\\
&= \frac{1}{Z_{M_*}} \int_0^{1/2} \mathrm{d}\epsilon \int \mathrm{d}x_1 \int \mathrm{d}x_2 \; \mathcal{L}(d|\epsilon, x_1, x_2, M_*)\pi(\epsilon, x_1, x_2|M_*) && (32)\\
&= \frac{1}{Z_{M_*}} \int_0^{1/2} \mathrm{d}\epsilon \int \mathrm{d}x_1 \int \mathrm{d}x_2 \; \mathcal{L}(d|\epsilon, x_1, x_2, M_*)\pi(x_1|M_1)\pi(x_2|M_2) && (33)\\
&= \frac{1}{Z_{M_*}} \int_0^{1/2} \mathrm{d}\epsilon \int \mathrm{d}x_1 \int \mathrm{d}x_2 \; \mathcal{L}(d|x_1, M_1)\pi(x_1|M_1)\pi(x_2|M_2) && (34)\\
&= \frac{1}{2Z_{M_*}} \int \mathrm{d}x_1 \; \mathcal{L}(d|x_1, M_1)\pi(x_1|M_1) && (35)\\
&= \frac{Z_{M_1}}{2Z_{M_*}}. && (36)
\end{aligned}$$

Similarly,

$$\mathrm{Prob}(\epsilon > 1/2|M_*) = \frac{Z_{M_2}}{2Z_{M_*}}. \tag{37}$$

Therefore, taking the ratio of Eqs. 36 and 37 gives

$$\mathcal{B}_{1,2} = \frac{Z_{M_1}}{Z_{M_2}} = \frac{\mathrm{Prob}(\epsilon < 1/2|M_*)}{\mathrm{Prob}(\epsilon > 1/2|M_*)}. \tag{38}$$

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

$$\begin{aligned}
Z_{M_2} &= \int \mathrm{d}x\ \mathcal{L}(d|x, M_2)\pi(x|M_2), && (39)\\
&= \int \mathrm{d}x\ \frac{\mathcal{L}(d|x, M_2)}{\mathcal{L}(d|x, M_1)}\mathcal{L}(d|x, M_1)\pi(x|M_2), && (40)\\
&= \int \mathrm{d}x\ \frac{\mathcal{L}(d|x, M_2)}{\mathcal{L}(d|x, M_1)}\mathcal{L}(d|x, M_1)\pi(x|M_1), && (41)\\
&= Z_{M_1}\int \mathrm{d}x\ \frac{\mathcal{L}(d|x, M_2)}{\mathcal{L}(d|x, M_1)}P(x|d, M_1). && (42)
\end{aligned}$$

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

$$\begin{aligned}
\mathcal{B}_{1,2} &= \frac{Z_{M_1}}{Z_{M_2}} && (43)\\
&= \int \mathrm{d}x\ \frac{\mathcal{L}(d|x, M_2)}{\mathcal{L}(d|x, M_1)}P(x|d, M_1) && (44)\\
&= \mathrm{E}_{x\sim P(x|d,M_1)}\left[\frac{\mathcal{L}(d|x, M_2)}{\mathcal{L}(d|x, M_1)}\right] && (45)
\end{aligned}$$

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

$$\mathcal{B}_{1,2} \approx \frac{1}{N} \sum_{i=0}^{N-1} \frac{\mathcal{L}(d|x_i, M_2)}{\mathcal{L}(d|x_i, M_1)} \tag{46}$$

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