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

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

Computing the evidence, $Z = P(\text{data}|\text{model})$, is the key step in Bayesian model selection. This involves evaluating the integral in Eq. 2 that is, unfortunately, usually impossible to perform analytically and difficult to evaluate using standard numerical integration routines for all but the simplest models.

0.1 Analytic computation

Analytically computing the Bayesian evidence is generally only possible when using a conjugate prior.

0.2 Laplace's approximation

In general, finding the Bayesian evidence involves numerically evaluating a complicated, high-dimensional integral over the parameter space of the model. 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.

However, in this section we will describe a useful method for approximating the evidence called the *Laplace approximation.* The essential idea is that we will attempt to approximate the shape of the posterior (i.e. the product of the likelihood and the prior) using a multivariate normal distribution and then integrate this function instead.

To recap, 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;

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

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

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

where $c = -\mathrm{d}^2/\mathrm{d}x^2|_{x=\hat{x}} \log P^*(x|d)$. 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

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

A Gaussian is simple to integrate analytically; we have

$$Z \approx P^*(\hat{x}|d)\sqrt{\frac{2\pi}{c}}\,. \tag{7}$$

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;

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

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 well when the posterior has a single, narrow region of significant support that is well approximated by a multivariate Gaussian. As a consequence of the central limit theorem, this will typically be the case in the limit of large sample size, or high signal-to-noise ratio. If the posterior contains multiple, widely-separated modes then Laplace's approximation can be applied separately to each peak.

0.3 Thermodynamic integration

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

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 joint posterior on β and the model parameters x; 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 is obtained by setting $\beta = 0$ ($\mathcal{L}(d|x,\beta=0) = \mathcal{L}(d|x)^0 = 1$; 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 distribution).

[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

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

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

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

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

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

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

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

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

0.5 Avoiding the evidence altogether

0.6 Nested sampling