Estimating the model evidence by thermodynamic integration.

This write up follows what is presented in Goggans and Chi (2004). When necessary, I will point out the differences between their work and how I implemented this algorithm to our toy model selection problem.

Notation: D is the data, the models which we are considering are $M_1, \ldots, M_{N_m}$, and model M_k has parameters θ_k. The goal is to select the best model from the set $\{M_k\}$ based on our observations D. In probabilistic terms, this means we want to find which model has the maximum posterior probability:

$$\max_k p(M_k \mid D). \tag{1}$$

Using Bayes' rule, we can rewrite the posterior probability as

$$p(M_k \mid D) = \frac{p(D \mid M_k)p(M_k)}{p(D)}. \tag{2}$$

Taking the log of both sides gives us that

$$\log p(M_k \mid D) = \log p(D \mid M_k) + \log p(M_k) - \log p(D). \tag{3}$$

Let us assume that we know the prior probability of each model, i.e., $p(M_k)$ is known. *For our model selection problem, I assumed that the prior model probability is uniform over all models which corresponds to* $p(M_k) = 1/N_m$. This means that the log posterior probability can be split into two parts:

$$\underbrace{\log p(M_k \mid D)}_{\text{Want to know}} = \underbrace{\log p(D \mid M_k)}_{\text{Unknown}} + \underbrace{\log p(M_k) - \log p(D)}_{\text{Constant for all } k} \tag{4}$$

This means that estimating the log posterior for model M_k is tantamount to estimating the log likelihood for model M_k. This model likelihood $p(D \mid M_k)$ is also called the evidence.

This here is the one key difference between my implementation and the work presented in Goggans and Chi (2004). They assume that we can only evaluate the model likelihood $p(D \mid M_k)$ *and model prior* $p(M_k)$ *up to some proportional constant. They claim this is sufficient but I have been unable to convince myself of their argument. In practice, this assumption is absolutely needed since we cannot typically evaluate the model likelihood exactly. Fortunately however, in our toy model selection problem, we are able to evaluate exactly and can side step this issue. If and when we choose to implement this in more complex model selection problems, we will need to bridge this gap. For now, we continue with deriving the thermodynamic integration algorithm and save this issue for another day.*

The goal now is to estimate the model likelihood $p(D \mid M_k)$. We can do this by considering M_k's model parameters θ_k. Through Bayes' rule, we have that

$$p(\theta_k \mid D, M_k) = \frac{p(D \mid \theta_k, M_k)p(\theta_k \mid M_k)}{p(D \mid M_k)}. \tag{5}$$

Thus, the model likelihood $p(D \mid M_k)$ is the normalizing factor of $p(D \mid \theta_k, M_k)p(\theta_k \mid M_k)$. This is helpful to use because we are well practiced in working with these quantities.

It is here where the assumptions in Goggans and Chi (2004) are important. In practice, we are unable to evaluate $p(D \mid \theta_k, M_k)$ *exactly and we must rely on proportional evaluation. Again, in our toy model selection problem, we are able to side step this important issue because the problem is not too complex.*

Thermodynamic integration is a technique to estimate the model likelihood $p(D \mid M_k)$ through MCMC estimates of $p(\theta_k \mid D, M_k)$. In our derivation of thermodynamic integration below, we drop the subscript k for ease of notation. Thus, we are concerned with estimating the model likelihood $p(D \mid M)$ by means of estimating the parameter posterior $p(\theta \mid D, M)$. We will start with what we want and derive what we need.

$$\begin{aligned}
\log p(D \mid M) &= \log p(D \mid M) - 0 && (6)\\
&= \log p(D \mid M) - \log(1) && (7)\\
&= \log p(D \mid M) - \log\left[\int p(\theta \mid M) d\theta\right] && (8)\\
&= \log\left[\int p(D \mid \theta, M) p(\theta \mid M) d\theta\right] - \log\left[\int p(\theta \mid M) d\theta\right] && (9)\\
&= \log\left[\int p(D \mid \theta, M)^1 p(\theta \mid M) d\theta\right] - \log\left[\int p(D \mid \theta, M)^0 p(\theta \mid M) d\theta\right] && (10)\\
&= \log\left[\int p(D \mid \theta, M)^\beta p(\theta \mid M) d\theta\right]\Big|_{\beta=0}^{\beta=1} && (11)\\
&= \int_{\beta=0}^{1} f(\beta) d\beta && (12)
\end{aligned}$$

where

$$f(\beta) = \frac{d}{d\beta}\left(\log\left[\int p(D \mid \theta, M)^\beta p(\theta \mid M) d\theta\right]\right) \quad (13)$$

Note that the step from (7) to (8) implies that our parameter prior is normalized. ***This assumption is required for thermodynamic integration.*** All of this means that if we want to find the log evidence, we need to be able to evaluate this function $f(\beta)$. Since this function is scalar with respect to β, estimating the integral from $\beta = 0$ to $\beta = 1$ can be done by Riemann sums (assuming we can evaluate f).

Using the chain rule, we have that

$$\begin{aligned}
f(\beta) &= \frac{d}{d\beta}\left(\log\left[\int p(D \mid \theta, M)^\beta p(\theta \mid M) d\theta\right]\right) && (14)\\
&= \frac{1}{\int p(D \mid \theta, M)^\beta p(\theta \mid M) d\theta} \frac{d}{d\beta}\left[\int p(D \mid \theta, M)^\beta p(\theta \mid M) d\theta\right] && (15)\\
&= \frac{1}{\int p(D \mid \theta, M)^\beta p(\theta \mid M) d\theta} \int \frac{d}{d\beta}\left[p(D \mid \theta, M)^\beta p(\theta \mid M)\right] d\theta && (16)\\
&= \frac{1}{\int p(D \mid \theta, M)^\beta p(\theta \mid M) d\theta} \int \frac{d}{d\beta}\left[p(D \mid \theta, M)^\beta\right] p(\theta \mid M) d\theta && (17)\\
&= \frac{1}{\int p(D \mid \theta, M)^\beta p(\theta \mid M) d\theta} \int \log\Big(p(D \mid \theta, M)\Big) p(D \mid \theta, M)^\beta p(\theta \mid M) d\theta && (18)\\
&= \int \left[\log\Big(p(D \mid \theta, M)\Big) \frac{p(D \mid \theta, M)^\beta p(\theta \mid M)}{\int p(D \mid \theta, M)^\beta p(\theta \mid M) d\theta}\right] d\theta && (19)\\
&= \int \log\Big(p(D \mid \theta, M)\Big) p(\theta \mid \beta, D, M) d\theta && (20)
\end{aligned}$$

In the last step, we introduce the power posterior $p(\theta \mid \beta, D, M)$ where β is an annealing parameter in the interval $[0, 1]$. The power posterior is defined as

$$p(\theta \mid \beta, D, M) = \frac{\big[p(D \mid \theta, M)\big]^{\beta} p(\theta \mid M)}{p(D \mid \beta, M)} = \frac{p(D \mid \theta, M)^{\beta} p(\theta \mid M)}{\int p(D \mid \theta, M)^{\beta} p(\theta \mid M) d\theta}. \tag{21}$$

Note that sampling the power posterior is as easy (or difficult) as sampling the parameter posterior $p(\theta \mid D, M)$ since the only difference is an exponent on the parameter likelihood probability.

The last key step in understanding thermodynamic integration is making sense of the function

$$f(\beta) = \int \log \Big(p(D \mid \theta, M)\Big) p(\theta \mid \beta, D, M) d\theta. \tag{22}$$

This is just the expected value of the log parameter likelihood distribution with respect to the power posterior. We can estimate this quantity using a Monte Carlo approximation

$$f(\beta) \simeq \frac{1}{N_e} \sum_{j=1}^{N_e} \log p(D \mid \theta[j], M); \quad \theta[j] \sim p(\theta \mid \beta, D, M). \tag{23}$$

It is here that we see why I believe we need exact evaluation of the parameter likelihood distribution. If we are only able to evaluate proportionally then $f(\beta)$ would be shifted which would ultimately alter the model evidence.

We now put all the pieces together in the order by which we would numerical estimate the model evidence.

1. Choose a partition of $[0, 1]$: $\{\beta_1, \ldots, \beta_n\}$.
2. For each β_j, sample the power posterior $p(\theta \mid \beta_j, D, M) \propto \big[p(D \mid \theta, M)\big]^{\beta_j} p(\theta \mid M)$.
3. Estimate $f(\beta_j)$ by equation (23).
4. Estimate the log model evidence, $\log p(D \mid M)$, by approximating the integral in equation (12)

Emcee in Python has a MCMC sampler, the Parallel Transport Sampler or PTsampler, which efficiently samples the power posteriors for a given partition of β's. You can then estimate the model evidence using the steps above by calling the thermodynamic integration attribute.

By doing this for each model M_k, we can choose which model is best based on the data D based on equation (4).

References:

- Goggans, P. M., & Chi, Y. (2004, April). Using thermodynamic integration to calculate the posterior probability in Bayesian model selection problems. In AIP Conference Proceedings (Vol. 707, No. 1, pp. 59-66).
- `http://eriqande.github.io/sisg_mcmc_course/thermodynamic-integration.nb.html`
  This helped my understanding and has some further references.