CHIB'S EVIDENCE ESTIMATOR

FRANK WOOD

Let's say that we are trying to pick between models $\mathcal{M}_1$ and $\mathcal{M}_2$. We don't know which model describes the observed data $\mathcal{X}$ better. Each of these models has some latent variables and parameters which we'll group together (as good Bayesians do) and call θ_1 and θ_2 (the models can have different parameter spaces).

If $\frac{P(\mathcal{X}|\mathcal{M}_1)}{P(\mathcal{X}|\mathcal{M}_2)} > 1$ then it is generally reasonable and agreed that model $\mathcal{M}_1$ should be preferred. There are elegant Occam's razor arguments [1] for doing model selection in this way. For the purposes of this tutorial, we'll just assume that it's reasonable to pick amongst models in this way. It should be noted that $P(\mathcal{X}|\mathcal{M}_1)$ is generally not available but instead must be computed. Namely, we usually marginalize out the latent variables via some integration computation, i.e. $P(\mathcal{X}|\mathcal{M}_1) = \int P(\mathcal{X}, \theta_1|\mathcal{M}_1)d\theta_1$. This is usually analytically and computationally intractable. Rejoice when it is tractable.

0.1. **A useful and (in retrospect) obvious identity.**

$$P(\mathcal{X}|\mathcal{M}_1) = \frac{P(\mathcal{X}, \theta_1|\mathcal{M}_1)}{P(\theta_1|\mathcal{X}, \mathcal{M}_1)} \tag{1}$$

Bayes' rule can be used to estimate evidence if it's re-arranged in this way. Since we'll be working with a single model at a time we can drop the model designation entirely and work with

$$P(\mathcal{X}) = \frac{P(\mathcal{X}, \theta)}{P(\theta|\mathcal{X})} \tag{2}$$

which is cleaner to read. When computing ratios the different models and parameter vectors are obvious from setting.

Note that this ratio is the same regardless of the value of θ so, theoretically, we can stick any-old value of θ in there and simply compute the evidence. Let's pick a θ^* where both $P(\mathcal{X}, \theta^*)$ and $P(\theta^*|\mathcal{X})$ are nonzero and preferably large. Somewhere near a posterior mode should probably do.

Also note that $P(\mathcal{X}, \theta^*)$ is a number which is easy to compute. Simply take the observations $\mathcal{X}$ and the value of θ^* you picked and plug them into the joint distribution specified that constitutes the generative model. The denominator is tricky. That can't be computed for the same reason that the evidence is hard to compute – it's really a chicken and egg problem: if you have the evidence then you can compute the posterior (the denominator) and, if you have the posterior (particularly in analytic form) you can compute the evidence. So, which first *amigo*?

0.2. **Approximating the posterior via sampling.** Assuming that you know how to sample from $P(\theta|\mathcal{X})$ (it's proportional to the joint – Gibbs sampling and Markov Chain Monte Carlo (MCMC) are usually the ticket here, though sequential Monte Carlo (SMC) isn't ruled out). All of them give you a shot at estimating the value of $P(\theta^*|\mathcal{X})$ for a particular value of θ^*. The simplest way is to sample from $P(\theta|\mathcal{X})$ then count the number of times $\theta = \theta^*$ in the resulting set $\mathcal{S} = \{\theta^{(1)}, \ldots, \theta^{(S)}\}$.[1] This, for reasons given in the footnote, is either impossible (continuous case) or highly inefficient in practical situations (if you can easily enumerate the values of θ then this whole procedure is moot).

If you remember the definition of stationarity and how it applies in the MCMC sampling then knowing

$$P(\theta^*|\mathcal{X}) = \sum_{\theta} T(\theta^* \leftarrow \theta) P(\theta|\mathcal{X}) \tag{3}$$

is really helpful in coming up with a more reasonable way to estimate the number $P(\theta^*|\mathcal{X})$ that we're after.

Why? Well, if $T(\theta^* \leftarrow \theta)$ is a procedure that corresponds to a conditional probability distribution that has $P(\theta|\mathcal{X})$ as its stationary distribution then $P(\theta^*|\mathcal{X})$ is the same distribution as $P(\theta|\mathcal{X})$. More relevant is the fact that samples drawn according to $P(\theta|\mathcal{X})$ can be transformed into samples from $P(\theta^*|\mathcal{X})$ via this rule. This is how samplers work.[2]

Remember: here we're after a *number* $P(\theta^*|\mathcal{X})$ not *samples* from this distribution. We can use (3) to compute such a number from samples (better if they're iid and exact) from $P(\theta|\mathcal{X})$ because we can compute the number (the conditional probability) $T(\theta^* \leftarrow \theta^{(s)})$ for each sample $\theta^{(s)} \sim P(\theta|\mathcal{X})$.

Computing these number is straightforward for Gibbs and Metropolis-Hastings samplers. The transition kernel in both cases explicitly gives you the equations for computing these numbers. In the case of a Gibbs sampler, for instance, where component-wise updates on the current state of the sampler are performed, $T(\theta^* \leftarrow \theta^{(s)})$ is computed like

$$T(\theta^* \leftarrow \theta^{(s)}) = \prod_j P(\theta^*_j|\tilde{\theta}_j)$$

where $\tilde{\theta}_j = (\theta^*_{1:j-1}, \theta^{(s)}_{j:J})$ and j indexes the dimension of a J dimensional parameter space. Each of these terms can be computed directly.

There is, unfortunately, sometimes a relatively large problem with Chib estimates and this has to do with label switching in models where there are symmetric posterior modes. Think about (2), in particular estimating the denominator. Theoretically any posterior

[1] This is one of those cases where measure crops up again. If $P(\theta...)$ is continuous then θ^* will never appear in the set. This can be fixed up by either discretizing space really finely or by thinking about θ as being some small subset of the parameter space. This is easier to rationalize by taking the former perspective.

[2] The average amount of time a single particle (sample) spends in a particular state is proportional to its probability under the stationary distribution of the process. The whole point is that we often can't directly generate samples from the stationary distribution of a process but want to generate samples from it anyway.

sampler should mix over all symmetric modes of the distribution. For argument's sake though, let's say that the sampler immediately mixes to a different mode than the one in which θ^* is near the peak and then stays there for the finite run of collected samples. Then, because all of the transition terms in(3) are essentially zero, the estimated posterior probability of θ^* will be substantially incorrectly calculated. Chib apparently circumvented this problem by relabeling samples from the posterior to ensure that they were in the same mode as θ^*. This, though, artificially inflates the posterior probability of θ^* by (as suggested by [**?**]) a factor equal to the number of symmetric modes in the posterior. This means that in this case the overall Chib evidence estimate is lower than the true evidence by this amount ($K!$ for a mixture model with K components for instance). In models without these symmetries this problem doesn't exist. It may still be difficult to approximate the posterior probability of a particular parameter setting well however.

0.3. **Bibliography.**

References

[1] C. E. Rasmussen and Z. Ghahramani. Occam's razor. In *Advances in Neural Information Processing Systems 13*. MIT Press, Cambridge, MA, 2001.