Eventually Almost Everywhere
A blog about probability and olympiads by Dominic Yeo

Bayesian Inference and the Jeffreys Prior

Posted on **May 10, 2013**

Last term I was tutoring for the second year statistics course in Oxford. This post is about the final quarter of the course, on the subject of Bayesian inference, and in particular on the Jeffreys prior.

There are loads and loads of articles sitting around on the web contributing the debate about the relative merits of Bayesian and frequentist methods. I do not want to continue that debate here, partly because I don't have a strong opinion, but mainly because I don't really understand that much about the underlying issues.

What I will say is that after a few months of working fairly intensively with various complicated stochastic processes, I am starting to feel fairly happy throwing about conditional probability rather freely. When discussing some of the more combinatorial models for example, quite often we have no desire to compute or approximate complication normalising constants, and so instead talk about 'weights'. And a similar idea underlies Bayesian inference. As in frequentist methods we have an unknown parameter, and we observe some data. Furthermore, we know the probability that such data might have arisen under any value of the parameter. We want to make inference about the value of the parameter given the data, so it makes sense to multiply the probability that the data emerged as a result of some parameter value by some weighting on the set of parameter values.

In summary, we assign a prior distribution representing our initial beliefs about the parameter before we have seen any data, then we update this by weighting by the likelihood that the observed data might have arisen from a particular parameter. We often write this as:

$$\pi(\theta|x) \propto f(x|\theta)\pi(\theta),$$

or say that posterior = likelihood x prior. Note that in many applications it won't be necessary to work out what the normalising constant on the distribution ought to be.

That's the setup for Bayesian methods. I think the general feeling about the relative usefulness of such an approach is that it all depends on the prior. Once we have the prior, everything is concrete and unambiguously determined. But how should we choose the prior?

There are two cases worth thinking about. The first is where we have a lot of information about the problem already. This might well be the case in some forms of scientific research, where future analysis aims to build on work already completed. It might also be the case that we have already performed some Bayesian calculations, so our current prior is in fact the posterior from a previous set of experiments. In any case, if we have such an 'informative prior', it makes sense to use it in some circumstances.

In general though, it is entirely possible that neither of these situations will hold but we still want to try Bayesian analysis. The ideal situation would be if the choice of prior had no effect on the analysis, but if that were true, then we couldn't really be doing any Bayesian analysis. The Jeffreys prior is one natural candidate because it removes a specific problem with choosing a prior to express ignorance.

It sounds reasonable to say that if we have total ignorance about the parameter, then we should take the prior to be uniform on the set of possible values taken by the parameter. There are two potential objections to this. The first is that if the parameter could take any real value, then the prior will not be a distribution as the uniform distribution on the reals is not normalisable. Such a prior is called *improper*. This isn't a huge problem really though. For making inference we are only interested in the posterior distribution, and so if the posterior turns out to be normalisable we are probably fine.

The second problem is more serious. Even though we want to express ignorance of the parameter, is there a canonical choice for THE parameter? An example will make this objection more clear. Suppose we know nothing about the parameter T except that it lies in [0,1]. Then the uniform distribution on [0,1] seems like the natural candidate for the prior. But what if we considered T^100 to be the parameter instead? Again if we have total ignorance we should assign T^100 the uniform distribution on its support, which is again [0,1]. But if T^100 is uniform on [0,1], then T is massively concentrated near 1, and in particular cannot also be uniformly distributed on [0,1]. So as a minimum requirement for expressing ignorance, we want a way of generating a prior that doesn't depend on the choice of parameterisation.

The Jeffreys prior has this property. Note that there may be separate problems with making such an assumption, but this prior solves this particular objection. We define it to be $\pi(\theta) \propto [I(\theta)]^{1/2}$ where I is the *Fisher information*, defined as

$$I(\theta) = -\mathbb{E}_\theta\left[\frac{\partial^2 l(X_1,\theta)}{\partial\theta^2}\right],$$

where the expectation is over the data X_1 for fixed θ, and l is the log-likelihood. Proving that this has the property that it is invariant under reparameterisation requires demonstrating that the Jeffreys prior corresponding to $g(\theta)$ is the same as applying a change of measure to the Jeffreys prior for θ. The proof is a nice exercise in the chain rule, and I don't want to reproduce it here.

For a Binomial likelihood function, we find that the Jeffreys prior is Beta(1/2,1/2), which has density that looks roughly like a bucket suspended above [0,1]. It is certainly worth asking why the 'natural' choice for prior might put lots of mass at the edge of the domain for the parameter.

I don't have a definitive answer, but I do have an intuitive idea which comes from the meaning of the Fisher information. As the second derivative of the log-likelihood, a large Fisher information means that with high probability we will see data for which the likelihood changes substantially if we vary the parameter. In particular, this means that the posterior probability of a parameter close to 0 will be eliminated more quickly by the data if the true parameter is different.

If the variance is small, as it is for parameter near 0, then the data generated by this parameter will have the greatest effect on the posterior, since the likelihood will be small almost everywhere except near the parameter. We see the opposite effect if the variance is large. So it makes sense to compensate for this by placing extra prior mass at parameter values where the data has the strongest effect. Note that in the previous example, the Jeffreys prior is in fact exactly inversely proportional to the standard deviation. For the above argument to make sense, we need it to be monotonic with respect to SD, and it just happens that in this case, being 1/SD is precisely the form required to be invariant under reparameterisation.

Anyway, I thought that was reasonably interesting, as indeed was the whole course. I feel reassured that I can justify having my work address as the Department of Statistics since I now know at least epsilon about statistics!

Related articles

- Bayesian and Frequentist Approaches: Ask the Right Question (win-vector.com)

Reblog Like

One blogger likes this.

RELATED

The Envelope 'Paradox'
In "General Interest"

Exchangeability and De Finetti's Theorem
In "Exchangeability"

The Fisher Information and Cramer-Rao Bound
In "Statistics"

This entry was posted in **Statistics**, **Teaching** and tagged **Bayesian**, **conditional probability**, **conjugate prior**, **Fisher information**, **frequentist**, **improper**, **Jeffreys**, **Jeffreys prior**, **Likelihood function**, **posterior distribution**, **prior distribution**, **Statistics** by **dominicyeo**. Bookmark the **permalink [https://eventuallyalmosteverywhere.wordpress.com/2013/05/10/bayesian-inference-and-the-jeffreys-prior/]** .

4 THOUGHTS ON "BAYESIAN INFERENCE AND THE JEFFREYS PRIOR"

Andrew
on **May 2, 2014 at 3:53 pm** said:

"The ideal situation would be if the choice of prior had no effect on the analysis, but if that were true, then we couldn't really be doing any Bayesian analysis." This is exactly what diffuse priors are supposed to do! Doing BDA is not contingent upon there being prior weights to parameter values; it is about using Bayes' rule combined with a subjective probability framework to justify using probability distributions to model parameters. Unless I misunderstood you. Also, Bayesians have no problem admitting a "true" parameter value–else, what would we be trying to estimate?!

dominicyeo
on **June 1, 2014 at 10:10 pm** said:

Hi Andrew,

Thanks for your comments – on reflection this article really isn't very good at all. Regarding your second comment, I have no idea what point I was trying to make, but it certainly doesn't make any sense to me now. [I've now removed that sentence.]

I'd be interested to hear a bit more about your first comment "doing BDA is not contingent upon there being prior weights to parameter values". What does this mean in practice? Is the claim that a diffuse prior doesn't count as a prior weighting in some sense?

Dominic

Thanks for your interest, Dominic. By definition, a prior distribution probabilistically weighs each possible value of the parameter that it is modeling. I may have been a tad reckless in my wording; I merely meant that–practically speaking–a noninformative prior really doesn't weigh the values in any meaningful way because they are all (more or less) equally probable–exactly so if a uniform distribution is used. However, mathematically speaking, there is and must be a prior weighting to the potential parameter values that is given by the prior distribution. So your last point is correct. Sometimes, though, it's merely easier to think that it doesn't really weigh the values because the weighing bears no consequence on the posterior.
Cheers!

Onur
on **August 10, 2017 at 12:37 pm** said:

Hello,

Thank you for the blog, it helped me a lot. Still, there is something that I have problem with understanding.

'If the variance is small, as it is for parameter near 0, then the data generated by this parameter will have the greatest effect on the posterior, since the likelihood will be small almost everywhere except near the parameter.'

As my understanding, that sentence states that if the fisher information is small, data has the greatest effect on the posterior. And the likelihood will be small anywhere far from the parameter. And we will use these kind of priors for our Jeffreys prior. So we prefer small amount of fisher information for Jeffreys prior which has some changes on likelihood based on our choice of parameter. In this case parameter has effect on our prior. But in jeffreys there shouldn't be. What am I missing here? Thank you ☺