Handout 20: Jeffreys Prior and Empirical Priors

Perhaps the biggest challenge when working with Bayesian statistics is selecting an appropriate prior distribution. We have seen that in simple cases, the use of conjugate priors allows us to compute Bayesian estimators analytically. But, the question of selecting the hyperparameters remains. Also, in many cases, we need to work with more complex models that do not have conjugate priors. How to select the prior in those cases? There is no single solution, but let's see two common approaches.

Empirical Priors In many cases, we can use a larger dataset to estimate the prior distribution before looking at our specific example. For example, if we used a Bayesian estimator to predict whether a startup will be successful, we could use previous data about all startups in the previous 10 years to generate a prior distribution. Similarly, we could do this to predict the success of a new drug or the sale price of a home. While only possible in certain cases, when feasible, it offers an easily implemented and defensible approach to constructing a prior.[1]

[1] There is also a technique called Empirical Bayes, which is much more complex. It was once very popular, but has been replaced by other techniques in recent years. Empirical priors, on the other hand, are commonly used and very useful.

Jeffreys Prior Another approach is to try to select a neutral (non-informative) prior that indicates that we do not know anything specific regarding our initial knowledge about the parameters of interest. One way to do this is to put equal weight (in other words, a uniform distribution) on all the values. This often works relatively well, but in some cases we can improve this approach.

A minor problem with uniform priors is that they depend on the parameterization choosen for the probability distribution. For example, we usually define the normal distribution through the variance σ^2. However, we could also do this by defining the standard deviation σ or the precision σ^{-1}. Putting an equal weight on all values of the parameter means something different in each of these cases. A (perhaps surprising) solution is to set the prior proportional to the Fisher information:

$$p(\theta) \propto \sqrt{\mathcal{I}(\theta)}.$$

This is called the Jeffreys Prior. What does this offer us? Well, it turns out that this prior will result in the same results regardless of the parameterization choosen. In other words, if we have a different parameterization in terms of $\phi = g(\theta)$, then the transformation of the prior $p(\theta)$ will be the same as the prior on ϕ based on $\mathcal{I}(\phi)$.

In some cases, the Jeffreys Prior is an **improper prior**.[2] In other words, it is not something that can be normalized to be a proper probability. This is generally okay, though, and we can compute the posterior in the normal way.

[2] Some sources warn against improper priors based on certain edge-cases, but I personally think they are okay when motivated by the non-informative properties of the Jeffreys Prior.

Proof of Invariance To see why the Jeffreys Prior relates the Fisher information, consider a re-parameterization $\varphi = g(\theta)$. If we have a prior $p_\theta(\theta)$, an equivalent prior in terms of φ is given by the normal change of variables formula:

$$p_\varphi(\varphi) = p_\theta(\theta) \cdot \left| \frac{\partial \varphi}{\partial \theta} \right|$$

Now, look at how the value of $I(\varphi)$ related to $I(\theta)$:

$$\begin{aligned} I(\varphi) &= \mathbb{E}\left[\frac{\partial}{\partial \varphi} \log(f) \right]^2 \\ &= \mathbb{E}\left[\frac{\partial \theta}{\partial \varphi} \frac{\partial}{\partial \theta} \log(f) \right]^2 \\ &= \mathbb{E}\left[\frac{\partial}{\partial \theta} \log(f) \right]^2 \times \left[\frac{\partial \theta}{\partial \varphi} \right]^2 \\ &= I(\theta) \times \left[\frac{\partial \theta}{\partial \varphi} \right]^2 \end{aligned}$$

Taking the square-root of both sides yields:

$$\sqrt{I(\varphi)} = \sqrt{I(\theta)} \times \left| \frac{\partial \theta}{\partial \varphi} \right|$$

And we see that making the prior proportional to the square-root of the Fisher information produces an equivalent prior regardless of the parameterization.