Reference priors

Consider an inference problem in which we have data X parameterized by Θ, with sufficent statistic $T(X)$.

We want to find a prior distribution $\pi(\theta)$ that maximizes the its KL divergence from the posterior distribution $\pi(\theta \mid t)$, averaged over the distribution of T.

$$D_{KL}(p(\theta \mid t) \| p(\theta)) = \int p(\theta \mid t) \log \frac{p(\theta \mid t)}{p(\theta)} d\theta.$$

$$\begin{aligned} I(\Theta, T) &= \int D_{KL}(p(\theta \mid t) \| p(\theta)) p(t) dt \\ &= \int \int p(\theta, t) \log \frac{p(\theta, t)}{p(\theta) p(t)} d\theta dt. \end{aligned}$$

The *reference prior* is

$$p^*(\theta) = \underset{p(\theta)}{\arg\max}\, I(\Theta, T).$$

In fact, to make the reference prior more analytically tractable, we consider the asymptotic limit of $I(\Theta, T)$ as the number of observations goes to infinity.

Can be argued that it is also philosophically "right" thing to do, since we want to not only consider the information we obtain from a particular experiment, but also the information we might obtain from future experiments.

Handwavy proof

Let's rewrite $I(\Theta, T)$ as

$$\begin{aligned} I(\Theta, T) &= \int p(t) \int p(\theta \mid t) \log \frac{p(\theta \mid t)}{p(\theta)} d\theta dt \\ &= \int p(\theta) \log \frac{f_k(\theta)}{p(\theta)} d\theta, \end{aligned}$$

where

$$f_k(\theta) = \exp\left(\int p(t \mid \theta) \log p(\theta \mid t) dt\right).$$

Using functional form of a Lagrange multiplier, with the constraint $\int p(\theta) d\theta = 1$:

$$\mathcal{L}(p(\theta), \lambda) = \int p(\theta) \log \frac{f_k(\theta)}{p(\theta)} d\theta + \lambda \left(\int p(\theta) d\theta - 1 \right).$$

Via methods of calculus of variations, it is possible to show that

$$p^*(\theta) \propto f_k(\theta).$$

Hence we need to compute

$$f_k(\theta) = \exp\left(\int p(t \mid \theta) \log p(\theta \mid t) dt\right).$$

Under some conditions, the posterior distribution $p(\theta \mid t)$ is asymptotically normal with mean θ_0. This is a consequence of the Bernstein-von Mises theorem i.e. the "Bayesian Central Limit Theorem". (more on this in the future)

Skipping the details, we can show that

$$f_k(\theta) \approx \exp\left(\int p(t \mid \theta) \log(I^{1/2}(\theta)) dt\right) = I^{1/2}(\theta),$$

where $I(\theta)$ is the Fisher information.

Thus the Jeffreys prior is a reference prior when there are no nuisance parameters. They are not the same in general.

It is generally believed that the reference prior has better properties than the Jeffreys prior.

Like in the maximum entropy case (a.k.a the Jaynes prior), it is possible add constraints to the optimization problem.

Suppose we have constraints

$$E_{p(\theta)}[\theta^i] = c_i, \quad i =\ldots.$$

Then the reference prior is given by

$$p^{**}(\theta) \propto p^*(\theta) \exp\left(\sum_i \lambda_i \theta^i\right),$$

for some λ_i that remain to be determined.

There are number of other methods to construct non-informative priors.

For some arguments for (and against) the use of "non-informative" priors, see the article by James Berger: The case for objective Bayesian analysis.