STA 114: Statistics

Notes 12. The Jeffreys Prior

Uniform priors and invariance

Recall that in his female birth rate analysis, Laplace used a uniform prior on the birth rate $p \in [0,1]$. His justification was one of "ignorance" or "lack of information". He pretended that he had no (prior) reason to consider one value of $p = p_1$ more likely than another value $p = p_2$ (both values coming from the range $[0,1]$). A uniform pdf is consistent with such a consideration. But there is a logical flaw.

Consider the log-odds ratio of a female birth $\eta = \log \frac{p}{1-p}$. By the same logic, Laplace should not prefer any value of $\eta = \eta_1$ over any other $\eta = \eta_2$. So a prior plausibility score $\xi_\eta(\eta)$ on η should satisfy $\xi_\eta(\eta_1)/\xi_\eta(\eta_2) = 1$ for all η_1, η_2 (there is a technical difficulty to turn $\xi_\eta(\eta)$ into a pdf, but we ignore that for the moment). But if p is assigned the uniform prior pdf $\xi_p(p) = 1$, $p \in [0,1]$, then it induces the following prior pdf on η (by the change of variable $p = 1/(1+\exp(-\eta))$):

$$\tilde{\xi}_\eta(\eta) = \xi_p\left(\frac{1}{1+\exp(-\eta)}\right)\frac{\exp(\eta)}{(1+\exp(\eta))^2} = \frac{\exp(\eta)}{(1+\exp(\eta))^2},$$

and hence $\tilde{\xi}_\eta(\eta_1)/\tilde{\xi}_\eta(\eta_2) \neq 1$ unless $\eta_1 = \eta_2$.

That is, the logic of no-preference on p leads to a (induced) prior pdf on η that does not conform with the logic of no-preference applied directly to η, even through η is a monotone transform of p. This was held as a major criticism against Bayesian inference in the early 20th century by, among others, the most eminent statisticians of all, R. A. Fisher. This all but killed the development of Bayesian statistics until H Jeffreys revived this topic in mid 20th century.

Invariance under monotone transformation

Note that the premise of this discussion and debate is the case when there is not much prior information about the parameter. The question is, is there a prior pdf (for a given model) that would be universally accepted as a non-informative prior? Laplace's proposal was to use the uniform distribution. When the parameter space is discrete and finite, this choice is indeed non-informative and even survives the scrutiny of monotone transformations mentioned above. But when the parameter space is a continuum and one is seeking a prior pdf, uniform distributions are not universally accepted. The lack of invariance under monotone transformation being one big criticism.

Jeffreys proposed that an acceptable "non-informative prior finding principle" should be invariant under monotone transformations of the parameter. Let the statistical model be $X \sim f(x|\theta)$, $\theta \in \Theta$. And suppose the principle under consideration produces the prior

$\xi_\theta(\theta)$ for θ. Now suppose we look at a re-parametrization $\eta = h(\theta)$, given by a smooth monotone transformation h. The reparametrized model is $X \sim g(x|\eta)$, $\eta \in \mathcal{E}$, where $g(x|\eta) = f(x|h^{-1}(\eta))$ and $\mathcal{E} = h(\Theta) = \{h(\theta) : \theta \in \Theta\}$. Suppose the principle, when applied to the re-parametrized model, produces a prior pdf $\xi_\eta(\eta)$ on η.

But one could also derive a prior pdf $\tilde{\xi}_\eta(\eta)$ by starting from the prior pdf $\xi(\theta)$ on θ and using the transformation $\eta = h(\theta)$. This pdf is given by $\tilde{\xi}_\eta(\eta) = \xi_\theta(h^{-1}(\eta))/|h'(h^{-1}(\eta))|$. Jeffreys demand of invariance is same as saying that the two pdfs $\xi_\eta(\eta)$ [found by applying the principle directly on η] and $\tilde{\xi}_\eta(\eta)$ [found by applying the principle to θ and then deriving the corresponding pdf on η] should be the same. A little algebra shows that

$$\begin{aligned} & \xi_\eta(\eta) = \tilde{\xi}_\eta(\eta), \text{ for all } \eta \in \mathcal{E} \\ \iff\; & \xi_\eta(h(\theta)) = \tilde{\xi}_\eta(h(\theta)), \text{ for all } \theta \in \Theta \\ \iff\; & \xi_\eta(h(\theta)) = \frac{\xi_\theta(\theta)}{|h'(\theta)|}, \text{ for all } \theta \in \Theta \end{aligned}$$

The Jeffreys priors

In addition to making the demand of invariance, Jeffreys also described how to construct such a prior. The construction is based on the Fisher information function of a model. Consider a model $X \sim f(x|\theta)$, where $\theta \in \Theta$ is scalar and $\theta \mapsto \log f(x|\theta)$ is twice differentiable in θ for every x. The Fisher information of the model at any θ is defined to be:

$$I^F(\theta) = E_{[X|\theta]}\left\{\frac{\partial}{\partial\theta}\log f(X|\theta)\right\}^2 = \mathrm{E}_{[X|\theta]}\{\dot{\ell}_X(\theta)\}^2.$$

Under some regularity conditions (look back at our approximate normality of MLE result), this equals

$$I^F(\theta) = -E_{[X|\theta]}\frac{\partial^2}{\partial\theta^2}\log f(X|\theta) = -E_{[X|\theta]}\ddot{\ell}_X(\theta).$$

If data X has n components $(X_1, \cdots, X_n)$ and the model is $X_i \overset{\text{IID}}{\sim} g(x_i|\theta)$, then $f(x|\theta) = \prod_{i=1}^n g(x_i|\theta)$ and so

$$I^F(\theta) = \sum_{i=1}^n \Big[- E_{[X_i|\theta]}\frac{\partial^2}{\partial\theta^2}\log g(X_i|\theta)\Big] = nI_1^F(\theta)$$

where $I_1^F(\theta)$ is the single observation Fisher information of $X_i \sim g(x_i|\theta)$ at θ.

The Jeffreys proposal of a non-informative prior pdf for the model $X \sim f(x|\theta)$ is

$$\xi^J(\theta) = \text{const.} \times \sqrt{I^F(\theta)}.$$

If $\int_\Theta \sqrt{I^F(\theta)}d\theta$ is finite number, then the constant is taken to be one over this number, so that $\xi^J(\theta)$ defines a pdf over Θ. If this integral is infinite, the constant is left unspecified, and the corresponding function $\xi^J(\theta)$ is called an "improper" prior pdf of $\theta \in \Theta$. An improper prior pdf is accepted so long as it produces a proper posterior pdf for every possibly observation $X = x$. That is

$$\xi^J(\theta|x) = \frac{f(x|\theta)\xi^J(\theta)}{\int_\Theta f(x|\theta')\xi^J(\theta')d\theta'}$$

must be a pdf on Θ, which means the integral $\int f(x|\theta)\xi^J(\theta)d\theta$ must be finite. Even though an improper pdf is not really a pdf, it still expresses relative plausibility scores through the well defined ratios $\xi^J(\theta_1)/\xi^J\theta_2$ (the arbitrary constant cancels from numerator and denominator, so its exact value does not matter).

Below we show that the principle behind the construction Jeffreys prior is invariant to smooth, monotone transformation of the parameter. Here we briefly comment why it is "non-informative". It turns out that the Jeffreys prior is indeed the uniform prior over the parameter space Θ, but not under the Euclidean geometry (pdfs depend on the geometry, as they give limits of probability of a set over the volume of the set, and volume calculation depends on geometry). The geometry that one needs to consider stems from defining a distance between $\theta_1, \theta_2 \in \Theta$ in terms of the distance between the two pdfs $f(x|\theta_1)$ and $f(x|\theta_2)$. An advantage of this definition of distance is that it remains invariant to reparametrization under monotone transformation.

The Jeffreys prior is invariant under monotone transformation

Consider a model $X \sim f(x|\theta)$, $\theta \in \Theta$ and its reparametrized version $X \sim g(x|\eta)$, $\eta \in \mathcal{E}$, where $\eta = h(\theta)$ with h a differentiable, monotone transformation (θ is assumed scalar). To distinguish between the two models, we let $I^F_\theta(\theta)$ and $I^F_\eta(\eta)$ denote the two Fisher information functions. Then,

$$\begin{aligned}
I^F_\theta(\theta) &= \int \left\{\frac{\partial}{\partial\theta}\log f(x|\theta)\right\}^2 f(x|\theta)dx \\
&= \int \left\{\frac{\partial}{\partial\theta}\log g(x|h(\theta))\right\}^2 g(x|h(\theta))dx \\
&= \int \left\{\frac{\partial}{\partial\eta}\log g(x|\eta)\Big|_{\eta=h(\theta)} h'(\theta)\right\}^2 g(x|h(\theta))dx \quad \text{[chain rule of differentiation]} \\
&= \{h'(\theta)\}^2 I^F_\eta(h(\theta)).
\end{aligned}$$

And therefore,

$$\xi^J_\eta(h(\theta)) = \text{const.} \times \sqrt{I^F_\eta(h(\theta))} = \text{const.} \times \frac{\sqrt{I^F_\theta(\theta)}}{|h'(\theta)|} = \frac{\xi^J_\theta(\theta)}{|h'(\theta)|}$$

as demanded by Jeffreys.

Example (Normal model)**.** Consider data $X = (X_1, \cdots, X_n)$, modeled as $X_i \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$ with σ^2 assumed known, and $\mu \in (-\infty, \infty)$. The Fisher information function in μ of a single observation is in μ is given by

$$I^F_1(\mu) = -\mathrm{E}_{[X_1|\mu]}\frac{\partial^2}{\partial\mu^2}\frac{(X_1-\mu)^2}{2\sigma^2} = \frac{1}{\sigma^2}$$

and hence Fisher information at μ of the model for X is $I^F(\mu) = nI^F_1(\mu) = n/\sigma^2$. Therefore the Jeffreys prior for μ is

$$\xi^J(\mu) = \text{const.} \times \sqrt{n/\sigma^2} = \text{const}, \quad -\infty < \mu < \infty.$$

This is a "flat" prior over the parameter space $(-\infty, \infty)$. Unfortunately, this does not lead to a pdf for any value of the constant as $\int_{-\infty}^{\infty} d\mu = \infty$. So this is an improper prior.

The posterior associated with the Jeffreys prior is

$$\xi^J(\mu|x) = \frac{\exp\{-\frac{(\bar{x}-\mu)^2}{2\sigma^2/n}\}}{\int_{-\infty}^{\infty} \exp\{-\frac{(\bar{x}-\mu')^2}{2\sigma^2/n}\} d\mu'} = \mathsf{Normal}(\bar{x}, \sigma^2/n)$$

which is a proper pdf. Thus the Jeffreys prior is an "acceptable one" in this case.

It is an interesting fact that summaries of $\xi^J(\mu|x)$ numerically match summaries from classical inference. For example, the posterior mean and median is $\bar{x}$ which happens to be $\hat{\mu}_{\text{MLE}}(x)$. Also, a $100(1-\alpha)\%$ central posterior credible interval is $\bar{x} \mp \sigma z(\alpha)/\sqrt{n}$ which matches the $100(1-\alpha)\%$ confidence interval for μ.

Multiparameter model and the Jeffreys prior

When the model is indexed by multiple parameters, we need some extension of our definitions of the Fisher information and the Jeffreys prior. For simplicity we only consider a two-parameter model $X \sim f(x|\theta_1, \theta_2)$. Then Fisher information is defined as

$$I^F(\theta_1, \theta_2) = \begin{pmatrix} \mathrm{E}_{[X|\theta_1,\theta_2]}\{-\frac{\partial^2}{\partial\theta_1^2} \log f(x|\theta_1, \theta_2)\} & \mathrm{E}_{[X|\theta_1,\theta_2]}\{-\frac{\partial^2}{\partial\theta_1\partial\theta_2} \log f(x|\theta_1, \theta_2)\} \\ \mathrm{E}_{[X|\theta_1,\theta_2]}\{-\frac{\partial^2}{\partial\theta_2\partial\theta_1} \log f(x|\theta_1, \theta_2)\} & \mathrm{E}_{[X|\theta_1,\theta_2]}\{-\frac{\partial^2}{\partial\theta_2^2} \log f(x|\theta_1, \theta_2)\} \end{pmatrix}.$$

Next, Jeffreys' prior is defined as

$$\xi^J(\theta_1, \theta_2) = \text{const} \times \sqrt{\det[I^F(\theta_1, \theta_2)]}$$

where $\det[A]$ denotes the determinant of a matrix A. For a 2×2 matrix $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ the determinant is: $\det[A] = ad - bc$.

Example (Normal model with unknown μ and σ^2). For the normal model $X_1, \cdots, X_n \overset{\text{IID}}{\sim} \mathsf{Normal}(\mu, \sigma^2)$, $\mu \in (-\infty, \infty)$, $\sigma^2 \in (0, \infty)$,

$$\log f(x|\mu, \sigma^2) = \text{const} - \frac{n}{2} \log \sigma^2 - \frac{(n-1)s_x^2 + n(\bar{x}-\mu)^2}{2\sigma^2}.$$

The second derivatives are

$$\begin{aligned} \frac{\partial^2}{\partial\mu^2} \log f(x|\mu, \sigma^2) &= -\frac{n}{\sigma^2} \\ \frac{\partial^2}{\partial\mu\partial(\sigma^2)} \log f(x|\mu, \sigma^2) &= -\frac{n(\bar{x}-\mu)}{\sigma^4} \\ \frac{\partial^2}{\partial(\sigma^2)\partial\mu} \log f(x|\mu, \sigma^2) &= \text{same as above} \\ \frac{\partial^2}{\partial(\sigma^2)^2} \log f(x|\mu, \sigma^2) &= \frac{n}{2\sigma^4} - \frac{(n-1)s_x^2 + n(\bar{x}-\mu)^2}{\sigma^6} \end{aligned}$$

So

$$I^F(\mu,\sigma^2) = \begin{pmatrix} \mathrm{E}_{[X|\mu,\sigma^2]}\frac{n}{\sigma^2} & \mathrm{E}_{[X|\mu,\sigma^2]}\frac{n(\bar{X}-\mu)}{\sigma^4} \\ \mathrm{E}_{[X|\mu,\sigma^2]}\frac{n(\bar{X}-\mu)}{\sigma^4} & \mathrm{E}_{[X|\mu,\sigma^2]}\{-\frac{n}{2\sigma^4}+\frac{(n-1)s_X^2+n(\bar{X}-\mu)^2}{\sigma^6}\} \end{pmatrix} = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}.$$

To derive the last equality in the above we have used,

$$\mathrm{E}_{[X|\mu,\sigma^2]}\bar{X} = \mu$$

$$\mathrm{E}_{[X|\mu,\sigma^2]}(\bar{X}-\mu)^2 \;=\; \sigma^2/n$$

$$\mathrm{E}_{[X|\mu,\sigma^2]}\sum_{i=1}^{n}(X_i-\bar{X})^2 \;=\; (n-1)\sigma^2$$

which follow from the facts (i) $\bar{X} \sim \mathsf{Normal}(\mu,\sigma^2/n)$ which has mean μ and variance σ^2 and (ii) $\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2 \sim \chi^2_{n-1}$ which has mean $n-1$.

Finally, we get the Jeffreys prior

$$\xi^J(\mu,\sigma^2) = \mathrm{const} \times \sqrt{\det[I^F(\mu,\sigma^2)]} = \mathrm{const} \times \sqrt{n^2/2\sigma^6} = \mathrm{const}\left(\frac{1}{\sigma^2}\right)^{3/2}.$$

The corresponding posterior pdf is

$$\begin{aligned} \xi^J(\mu,\sigma^2|x) &= \mathrm{const} \times (\sigma^2)^{-3/2} \times (\sigma^2)^{-n/2}\exp\left\{-\frac{(n-1)s_x^2+n(\bar{x}-\mu)^2}{2\sigma^2}\right\} \\ &= \mathrm{const} \times (\sigma^2)^{-(n+3)/2}\exp\left\{-\frac{(n-1)s_x^2+n(\bar{x}-\mu)^2}{2\sigma^2}\right\} \end{aligned}$$

which matches the $\mathsf{N}\chi^{-2}(\bar{x},n,n,\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2)$.

The reference prior for the two-parameter normal model

There are other formalizations of “low-informativeness” than the concept of uniform distribution over the parameter space. One such formalization leads to what is known as the reference prior. These priors, too, satisfy various invariance principles like the Jeffreys prior, and the two are often equal. We can’t discuss reference priors in details here, but note the reference prior for the two parameter normal model $X_1,\cdots,X_n \overset{\mathrm{IID}}{\sim} \mathsf{Normal}(\mu,\sigma^2)$, $\mu \in (-\infty,\infty)$, $\sigma^2 \in (0,\infty)$. This is given by,

$$\xi^R(\mu,\sigma^2) = \mathrm{const.} \times \frac{1}{\sigma^2}$$

and is a very popular “default” choice of a non-informative prior for this model. The posterior pdf associated with this prior is $\xi^R(\mu|x) = \mathsf{N}\chi^{-2}(\bar{x},n,n-1,s_x^2)$.

An interesting property of this posterior pdf is that it produces summaries of μ that are same as our ML summaries (the same hold for the posterior of the Jeffreys prior for the single parameter normal model with known σ^2, the Jeffresy and the reference priors are the same in this case). In particular a $100(1-\alpha)\%$ central credible interval of $\xi_1^R(\mu|x)$ is precisely the ML $100(1-\alpha)\%$-CI: $\bar{x} \mp s_x z_{n-1}(\alpha)/\sqrt{n}$.