Jeffreys' Prior Asymptotically and Approximately Maximizes Expected Information*

Yaohan Chen

School of Economics, Singapore Management University

February 15, 2020

1. Two Alternative Representation of Expected Information

Recall that a specific model $\mathcal{M}$ with $\boldsymbol{x}$ as observed data and θ as the associated parameters of interest is defined as

$$\mathcal{M} \equiv \{p(\boldsymbol{x} \mid \theta) \mid \boldsymbol{x} \in \mathcal{X}, \ \theta \in \Theta\} \tag{1}$$

and the corresponding expected information for a given model $\mathcal{M}$ for prior $q(\theta)$ is

$$\begin{aligned}
I\{q \mid \mathcal{M}\} &= \iint_{\mathcal{X}\times\Theta} p(\boldsymbol{x} \mid \theta) q(\theta) \log\left(\frac{p(\theta \mid \boldsymbol{x})}{q(\theta)}\right) d\boldsymbol{x} d\theta \\
&= \iint_{\mathcal{X}\times\Theta} p(\theta \mid \boldsymbol{x}) p(\boldsymbol{x}) \log\left(\frac{p(\theta \mid \boldsymbol{x})}{q(\theta)}\right) d\boldsymbol{x} d\theta \\
&= H\{q(\theta)\} - \int p(\boldsymbol{x}) H\{p(\theta \mid \boldsymbol{x})\} d\boldsymbol{x} \\
&= H\{q(\theta)\} - \int q(\theta) \int p(\boldsymbol{x} \mid \theta) H\{p(\theta \mid \boldsymbol{x})\} d\boldsymbol{x} d\theta \qquad (2) \\
&= H\{q(\theta)\} + \int q(\theta) \int p(\boldsymbol{x} \mid \theta) \log p(\theta \mid \boldsymbol{x}) d\boldsymbol{x} d\theta \qquad (3)
\end{aligned}$$

where

$$H\{q(\theta)\} = -\int q(\theta) \log q(\theta) d\theta$$

is called the entropy of $q(\theta)$. As long as we can have

$$p(\theta \mid \boldsymbol{x}) = \frac{q(\theta) p(\boldsymbol{x} \mid \theta)}{\int q(\theta) p(\boldsymbol{x} \mid \theta) d\theta} \quad p(\boldsymbol{x}) = \int q(\theta) p(\boldsymbol{x} \mid \theta) d\theta \quad \Longrightarrow \quad p(\boldsymbol{x}) p(\theta \mid \boldsymbol{x}) = q(\theta) p(\boldsymbol{x} \mid \theta)$$

*In this summary note I basically discussed some important concepts and associated techniques used in Bayesian Analysis. All errors are my own. You may contact me at yaohan.chen.2017@phdecons.smu.edu.sg

which is guaranteed by the assumption that

$$\int q(\theta)p(\boldsymbol{x} \mid \theta)d\theta < \infty$$

Hence (2) and (3) are two representation for expected information respectively. Rewrite (2) and (3) as following respectively,

$$\begin{aligned} H\{q(\theta)\} - \int q(\theta) \int p(\boldsymbol{x} \mid \theta) H\{p(\theta \mid \boldsymbol{x})\} d\boldsymbol{x} d\theta \\ = \int q(\theta) \log \left\{ \exp\left(- \int p(\boldsymbol{x}|\theta) H\{p(\theta \mid \boldsymbol{x})\} d\boldsymbol{x} \right) \Big/ q(\theta) \right\} d\theta \end{aligned} \tag{4}$$

$$\begin{aligned} H\{q(\theta)\} + \int q(\theta) \int p(\boldsymbol{x} \mid \theta) \log p(\theta \mid \boldsymbol{x}) d\boldsymbol{x} d\theta \\ = \int q(\theta) \log \left\{ \exp\left(\int p(\boldsymbol{x} \mid \theta) \log p(\theta \mid \boldsymbol{x}) d\boldsymbol{x} \right) \Big/ q(\theta) \right\} d\theta \end{aligned} \tag{5}$$

It is easy to observe that (4) and (5) can be formally written as

$$I\{q \mid \mathcal{M}\} = \int q(\theta) \log\left(\frac{f(\theta)}{q(\theta)}\right) d\theta$$

with

$$f(\theta) = \exp\left\{ - \int p(\boldsymbol{x} \mid \theta) H\{p(\theta \mid \boldsymbol{x})\} d\boldsymbol{x} \right\} \text{ in (4)} \quad \text{and} \quad f(\theta) = \exp\left\{ \int p(\boldsymbol{x} \mid \theta) \log p(\theta \mid \boldsymbol{x}) d\boldsymbol{x} \right\} \text{ in (5)}$$

Since we want to select q to maximize $I\{q \mid \mathcal{M}\}$, and note that due to Jensen Inequality,

$$\int q(\theta) \log\left(\frac{f(\theta)}{q(\theta)}\right) \leqslant \log\left(\int q(\theta) \cdot \frac{f(\theta)}{q(\theta)} d\theta\right) = \log\left(\int f(\theta) d\theta\right)$$

and this implies that $f(\theta) \propto q(\theta)$ and $q(\theta)$ should be **necessarily** of the following form

$$q(\theta) \propto \exp\left\{ - \int p(\boldsymbol{x} \mid \theta) H\{p(\theta \mid \boldsymbol{x})\} d\boldsymbol{x} \right\} \tag{6}$$

or

$$q(\theta) \propto \exp\left\{ \int p(\boldsymbol{x} \mid \theta) \log p(\theta \mid \boldsymbol{x}) d\boldsymbol{x} \right\} \tag{7}$$

and this may serve as a heuristic justification for the construction of $f_k(\theta)$ in the main theorem in Berger et al. (2009), thus

$$f_k(\theta) = \exp\left\{ \int_{\mathcal{T}_k} p(\boldsymbol{t}_k \mid \theta) \log\left[\pi^*(\theta \mid \boldsymbol{t}_k)\right] \, d\boldsymbol{t}_k \right\}$$

2. A Heuristic Normal Example

2.1. Entropy for normal distribution

Normal density has the following form with mean μ and variance σ^2

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

Entropy is given as

$$\begin{aligned}\int -f(x)\log f(x)dx &= -\int \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}\left[-\log\left(\sqrt{2\pi}\sigma\right) - \frac{(x-\mu)^2}{2\sigma^2}\right] dx \\ &= \log\left(\sqrt{2\pi}\sigma\right) + \frac{1}{2} \end{aligned} \quad (8)$$

which is obviously only correlated with σ.This result will be used in the demonstration of the next subsection.

2.2. Jeffreys' prior should be adopted

Suppose that $\boldsymbol{x} = \{x_1, \cdots, x_n\}$ is sample from iid normal distribution with mean μ and variance σ^2. Model is

$$\mathcal{M} \equiv \{p(\boldsymbol{x} \mid \mu, \sigma), \boldsymbol{x} \in \mathcal{X}\} \quad (9)$$

and

$$p(\boldsymbol{x} \mid \mu, \sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left\{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}\right\} \quad (10)$$

Denote θ as the parameter of interest (could be either μ or σ) and

$$\mathcal{I}(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta'} \log f(x \mid \mu, \sigma)\right]$$

as the Fisher Information Matrix with

$$f(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}.$$

Straightforward calculation gives

$$\mathcal{I}(\theta) = \begin{cases} \frac{1}{\sigma^2} & \text{if } \theta = \mu \\ \frac{2}{\sigma^2} & \text{if } \theta = \sigma \end{cases} \quad (11)$$

Further note that for a given prior $q(\theta)$, in general the associated posterior is

$$q^*(\theta) \propto e^{\log p(\boldsymbol{x}|\theta)} q(\theta)$$

Denoting $l(\theta) = \log p(\boldsymbol{x} \mid \theta)$ and expanding it around $\hat{\theta}$ to the second order

$$l(\theta) \approx l(\hat{\theta}) + l'(\hat{\theta})(\theta - \hat{\theta}) + l''(\hat{\theta})(\theta - \hat{\theta})^2 \Big/ 2$$

where $\hat{\theta}$ is the associated MLE which necessarily implies that $l'(\hat{\theta}) = 0$ and

$$\hat{\theta} = \begin{cases} \frac{\sum_1^n x_i}{n} & \text{if } \theta = \mu \\ \sqrt{\frac{\sum_1^n (x_i - \mu)^2}{n}} & \text{if } \theta = \sigma \end{cases}$$

Hence approximately for large n,

$$q^*(\theta) \propto e^{l(\hat{\theta})} q(\hat{\theta}) \exp\left\{ l''(\hat{\theta})(\theta - \hat{\theta})^2 \Big/ 2 \right\}.$$

Note that

$$l''(\hat{\theta}) = \sum_1^n \frac{\partial^2}{\partial\theta\partial\theta'} \log f(x_i \mid \mu, \sigma) \Bigg|_{\theta = \hat{\theta}}$$

and

$$\frac{1}{n} \sum_1^n \frac{\partial^2}{\partial\theta\partial\theta'} \log f(x_i \mid \mu, \sigma) \Bigg|_{\theta = \hat{\theta}} \xrightarrow{\text{P}} \mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta'} \log f(x \mid \mu, \sigma) \right] \Bigg|_{\theta = \hat{\theta}} = -\mathcal{I}(\hat{\theta})$$

Hence

$$q^*(\theta) \propto \exp\left\{ -n\mathcal{I}(\hat{\theta})(\theta - \hat{\theta})^2 \Big/ 2 \right\} \tag{12}$$

Thus the posterior density will be approximately proportional to $\mathcal{N}\left(\hat{\theta}, \frac{\mathcal{I}(\hat{\theta})^{-1}}{n}\right)$. For a given $\tilde{\theta}$,

$$\hat{\theta} \xrightarrow{p(\boldsymbol{x}|\tilde{\theta})} \tilde{\theta}$$

Thus for n large enough, approximately it is possible to have entropy for posterior density from (8) and (11) as

$$H\{p(\theta \mid \boldsymbol{x})\} \approx \log\left(\sqrt{2\pi} \frac{1}{n\mathcal{I}(\hat{\theta})} \right) + \frac{1}{2} \approx C_1 \log \tilde{\sigma} - \log n + C_2$$

where C_1 and C_2 is just constant. And from (6) it could be heuristically claimed that prior maximizing expected information should be

$$q(\theta) \propto \exp\{-\mathbb{E}[H\{\theta \mid \boldsymbol{x}\}]\} \propto \exp\{-C_1 \log \tilde{\sigma} + \log n - C_2\} \propto \frac{1}{\tilde{\sigma}} \tag{13}$$

Suppose that prior is continuous, then asymptotically and approximately we should have

$$q(\theta) \propto \frac{1}{\sigma} \tag{14}$$

which is the Jeffreys' Prior. More general results is available from Clarke (1994)

References

Berger, J. O., Bernardo, J. M., and Sun, D. (2009). The formal definition of reference priors. *The Annals of Statistics*, 37(2):905–938. 2

Clarke, B. S. (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41:37–60. 5