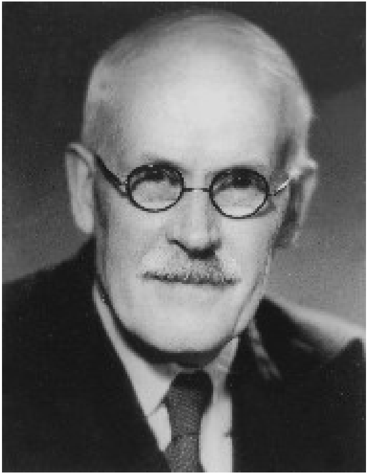
The Jeffreys Prior

Yingbo Li

Clemson University

MATH 9810

Sir Harold Jeffreys



Harold Jeffreys (1891 - 1989)

- English mathematician, statistician, geophysicist, and astronomer
- His book *Theory of Probability*, which first appeared in 1939, played an important role in the revival of the Bayesian view of probability.

Information Matrix

Suppose data X has density $f(x \mid \boldsymbol{\theta})$ which is twice differentiable in the coordinates of the unknown parameter $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$.

Expected Fisher Information: $p \times p$ matrix $\mathbf{I}$ with j, k entry

$$
\begin{aligned}
I_{j,k}(\boldsymbol{\theta}) &= -E_{\mathbf{X}|\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log f(x \mid \boldsymbol{\theta})\right] \\
\text{if indpt.} &= -\sum_{i=1}^{n} E_{X_i|\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log f_i(x_i \mid \boldsymbol{\theta})\right] \\
\text{if i.i.d.} &= -nE_{X_1|\boldsymbol{\theta}}\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log f(x_1 \mid \boldsymbol{\theta})\right]
\end{aligned}
$$

Jeffreys-rule prior

Derived to have invariance under 1-1 transformations

- Jeffreys proposed

$$p(\boldsymbol{\theta}) \propto \sqrt{\det\left[\mathbf{I}(\boldsymbol{\theta})\right]},$$

  where $\mathbf{I}(\boldsymbol{\theta})$ is Fisher information. This is called the *Jeffreys-rule prior* (often incorrectly shortened to the Jeffreys prior).

- If $\boldsymbol{\psi}$ is 1-1 transformation of $\boldsymbol{\theta}$:

$$p(\boldsymbol{\psi}) = p(\boldsymbol{\theta})[\det \mathbf{J}],$$

  where $J_{ij} = \partial\theta_i / \partial\psi_j$.

- Since

$$\mathbf{I}^*(\boldsymbol{\psi}) = \mathbf{J}\mathbf{I}(\boldsymbol{\theta})\mathbf{J}^T \Longrightarrow \det \mathbf{I}^* = [\det \mathbf{I}](\det \mathbf{J})^2$$

  Jeffreys prior $p(\boldsymbol{\psi}) \propto \sqrt{\det\left[\mathbf{I}^*(\boldsymbol{\psi})\right]}$ is invariant.

Example: Jeffreys Prior for Bernoulli Data

The Jeffreys prior for iid Bernoulli data $X_i \mid \theta \stackrel{\text{iid}}{\sim} \text{Ber}(\theta)$ is

$$p(\theta) \propto \frac{1}{\sqrt{\theta(1-\theta)}} = \text{Beta}(1/2, 1/2)$$

Note: Jeffreys would often deviate from using the Jeffreys-rule prior (so Jeffreys prior is the name best used for the priors he recommended).

- For a Poisson mean λ, he curiously recommended

$$p(\lambda) \propto 1/\lambda,$$

  instead of the Jeffreys-rule prior

$$p(\lambda) \propto 1/\sqrt{\lambda};$$

  and even though $p(\lambda) \propto 1/\lambda$ doesnt work when $x = 0$ is observed.
- For normal problems with unknown variance, he recommended the "independent Jeffreys prior."

Example: Normal Data

Suppose observations $X_i \mid \mu, \sigma^2 \overset{\text{iid}}{\sim} \mathsf{N}(\mu, \sigma^2)$, for $i = 1, 2, \ldots, n$. Denote parameters $\boldsymbol{\theta} = (\mu, \sigma^2)$, then

- Fisher Information

$$\mathbf{I}(\boldsymbol{\theta}) = \begin{pmatrix} n/\sigma^2 & 0 \\ 0 & n/(2\sigma^4) \end{pmatrix}$$

- Jeffreys-rule prior:

$$p(\boldsymbol{\theta}) \propto (1/\sigma^6)^{1/2} = 1/\sigma^3$$

- *Independent Jeffreys prior*:

$$p(\boldsymbol{\theta}) \propto 1/\sigma^2$$

  (ultimately recommended by Jeffreys)

Strengths of the Jeffreys-rule Prior

- Almost always defined
- Invariant to transformations
- Almost always yields a proper posterior
  - Mixture models are the main known examples in which improper posteriors result.
- Great for one-dimensional parameters
- It arises from many other approaches, such as minimum description length and various entropy approaches.

Weaknesses of the Jeffreys-rule Prior

- It depends on the statistical model, and hence appears to violate the likelihood principle; but all reasonable objective Bayesian theories do this.

  Example: Suppose X is Negative Binomial, i.e.

  $$p(x \mid r, \theta) = \frac{\Gamma(x+r)}{\Gamma(x+1)\Gamma(r)} \theta^r (1-\theta)^x, \text{ for } x = 0, 1, \ldots$$

  The Jeffreys-rule prior is $p(\theta) \propto \frac{1}{\sqrt{\theta(1-\theta)}}$.
- It requires the Fisher information to exist. *Example of non-existence:* Uniform $[0, \theta]$ distribution.
- Often fails badly for higher-dimensional parameters

Example of Failure – the Neyman-Scott Problem

Suppose we observe

$$X_{ij} \sim \mathsf{N}(\mu_i, \sigma^2), \quad i = 1, \ldots, n; \; j = 1, 2$$

- Defining $\bar{X}_i = (X_{i1} + X_{i2})/2, \bar{\mathbf{X}} = (\bar{X}_1, \ldots, \bar{X}_n), S^2 = \sum_{i=1}^{n} (X_{i1} - X_{i2})^2$, and $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$, the likelihood function

$$p(\mathbf{X} \mid \boldsymbol{\mu}, \sigma^2) \propto \frac{1}{\sigma^{2n}} \cdot \exp\left[-\frac{1}{\sigma^2}\left(|\bar{\mathbf{X}} - \boldsymbol{\mu}|^2 + \frac{S^2}{4}\right)\right]$$

- The Fisher information matrix is

$$\mathbf{I}(\boldsymbol{\mu}, \sigma^2) = \mathsf{diag}\{2/\sigma^2, \dots, 2/\sigma^2, n/\sigma^4\},$$

  the last entry corresponding to the information for σ^2.
- Jeffreys rule prior:

$$p(\boldsymbol{\mu}, \sigma^2) = 1/\sigma^{(n+2)}$$

- Jeffreys rule prior: $p(\boldsymbol{\mu}, \sigma^2) = 1/\sigma^{(n+2)}$ is bad. For instance, the marginal posterior for σ^2 is

$$p(\sigma^2 \mid \mathbf{X}) = \frac{1}{(\sigma^2)^{n+1}} \cdot \exp\left(-\frac{S^2}{4\sigma^2}\right).$$

  Here S^2 depends on $\mathbf{X}$. Posterior mean

$$E(\sigma^2 \mid \mathbf{X}) = \frac{S^2}{4(n-1)}$$

- Suppose data $\mathbf{X}$ are generated from the sampling distribution under the true parameter σ_0^2. Because the frequentist density of $S^2/(2\sigma_0^2)$ is Chi-Squared with n degrees of freedom, as $n \to \infty$, the posterior mean $S^2/[4(n-1)] \to \sigma_0^2/2$.

- Thus under the Jeffreys rule prior, the posterior distribution of σ^2 is inconsistent, i.e., it does not concentrate around the true value σ_0^2.
- The independent Jeffreys prior is fine here.
- There also exist cases where the independent Jeffreys prior doesn't work.