Jeffreys prior

In Bayesian probability, the **Jeffreys prior**, named after Sir Harold Jeffreys, is a non-informative (objective) prior distribution for a parameter space; it is proportional to the square root of the determinant of the Fisher information matrix:

$$p\left(\vec{\theta}\right) \propto \sqrt{\det \mathcal{I}\left(\vec{\theta}\right)}.$$

It has the key feature that its *functional dependence on the likelihood* L is invariant under reparameterization of the parameter vector $\vec{\theta}$ (the functional form of the prior density function itself is *not* invariant under reparameterization, of course: only the measure that is identically zero has that property; see below). This makes it of special interest for use with *scale parameters*.[1]

Contents

Reparameterization

One-parameter case

For an alternative parameterization φ we can derive

$$p(\varphi) \propto \sqrt{I(\varphi)}$$

from

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

using the change of variables theorem for transformations and the definition of Fisher information:

$$\begin{aligned}
p(\varphi) &= p(\theta) \left| \frac{d\theta}{d\varphi} \right| \\
&\propto \sqrt{I(\theta) \left(\frac{d\theta}{d\varphi} \right)^2} = \sqrt{\mathrm{E}\left[\left(\frac{d \ln L}{d\theta} \right)^2 \right] \left(\frac{d\theta}{d\varphi} \right)^2} \\
&= \sqrt{\mathrm{E}\left[\left(\frac{d \ln L}{d\theta} \frac{d\theta}{d\varphi} \right)^2 \right]} = \sqrt{\mathrm{E}\left[\left(\frac{d \ln L}{d\varphi} \right)^2 \right]} \\
&= \sqrt{I(\varphi)}.
\end{aligned}$$

That is, the functional form of the prior $p(\cdot)$ can be derived from that of the likelihood $L(\cdot)$ using the same procedure for both parametrizations.

Note, however, that the form of the prior is different for the two parametrizations. For example, if $p(\theta) = 1/\theta$ (as in the case of the normal distribution, see below), and $\varphi = \ln(\theta)$, then $p(\varphi) = 1$, which is obviously different from $1/\varphi$.

Multiple-parameter case

For an alternative parameterization $\vec{\varphi}$ we can derive

$$p(\vec{\varphi}) \propto \sqrt{\det I(\vec{\varphi})}$$

from

$$p(\vec{\theta}) \propto \sqrt{\det I(\vec{\theta})}$$

using the change of variables theorem for transformations, the definition of Fisher information, and that the product of determinants is the determinant of the matrix product:

$$\begin{aligned}
p(\vec{\varphi}) &= p(\vec{\theta}) \left| \det \frac{\partial \theta_i}{\partial \varphi_j} \right| \\
&\propto \sqrt{\det I(\vec{\theta}) \det^2 \frac{\partial \theta_i}{\partial \varphi_j}} \\
&= \sqrt{\det \frac{\partial \theta_k}{\partial \varphi_i} \det \mathrm{E}\left[\frac{\partial \ln L}{\partial \theta_k} \frac{\partial \ln L}{\partial \theta_l} \right] \det \frac{\partial \theta_l}{\partial \varphi_j}} \\
&= \sqrt{\det \mathrm{E}\left[\sum_{k,l} \frac{\partial \theta_k}{\partial \varphi_i} \frac{\partial \ln L}{\partial \theta_k} \frac{\partial \ln L}{\partial \theta_l} \frac{\partial \theta_l}{\partial \varphi_j} \right]} \\
&= \sqrt{\det \mathrm{E}\left[\frac{\partial \ln L}{\partial \varphi_i} \frac{\partial \ln L}{\partial \varphi_j} \right]} = \sqrt{\det I(\vec{\varphi})}.
\end{aligned}$$

Attributes

From a practical and mathematical standpoint, a valid reason to use this non-informative prior instead of others, like the ones obtained through a limit in conjugate families of distributions, is that its construction from the likelihood does not depend on the set of parameter variables that is chosen to describe parameter space. It is not the only prior with this property, however. As is clear from the derivation above, instead of $\ln(L)$ we could use any other smooth function $f(L)$, and the resulting prior would still have the same kind of invariance property.

Sometimes the Jeffreys prior cannot be normalized, and is thus an improper prior. For example, the Jeffreys prior for the distribution mean is uniform over the entire real line in the case of a Gaussian distribution of known variance.

Use of the Jeffreys prior violates the strong version of the likelihood principle, which is accepted by many, but by no means all, statisticians. When using the Jeffreys prior, inferences about $\vec{\theta}$ depend not just on the probability of the observed data as a function of $\vec{\theta}$, but also on the universe of all possible experimental outcomes, as determined by the experimental design, because the Fisher information is computed from an expectation over the chosen universe. Accordingly, the Jeffreys prior, and hence the inferences made using it, may be different for two experiments involving the same $\vec{\theta}$ parameter even when the likelihood functions for the two experiments are the same—a violation of the strong likelihood principle.

Minimum description length

In the minimum description length approach to statistics the goal is to describe data as compactly as possible where the length of a description is measured in bits of the code used. For a parametric family of distributions one compares a code with the best code based on one of the distributions in the parameterized family. The main result is that in exponential families, asymptotically for large sample size, the code based on the distribution that is a mixture of the elements in the exponential family with the Jeffreys prior is optimal. This result holds if one restricts the parameter set to a compact subset in the interior of the full parameter space. If the full parameter is used a modified version of the result should be used.

Examples

The Jeffreys prior for a parameter (or a set of parameters) depends upon the statistical model.

Gaussian distribution with mean parameter

For the Gaussian distribution of the real value x

$$f(x \mid \mu) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$$

with σ fixed, the Jeffreys prior for the mean μ is

$$\begin{aligned}
p(\mu) \propto \sqrt{I(\mu)} &= \sqrt{\mathrm{E}\left[\left(\frac{d}{d\mu} \log f(x \mid \mu) \right)^2 \right]} = \sqrt{\mathrm{E}\left[\left(\frac{x-\mu}{\sigma^2} \right)^2 \right]} \\
&= \sqrt{\int_{-\infty}^{+\infty} f(x \mid \mu) \left(\frac{x-\mu}{\sigma^2} \right)^2 dx} = \sqrt{1/\sigma^2} \propto 1.
\end{aligned}$$

That is, the Jeffreys prior for μ does not depend upon μ; it is the unnormalized uniform distribution on the real line — the distribution that is 1 (or some other fixed constant) for all points. This is an improper prior, and is, up to the choice of constant, the unique *translation*-invariant distribution on the reals (the Haar measure with respect to addition of reals), corresponding to the mean being a measure of *location* and translation-invariance corresponding to no information about location.

Gaussian distribution with standard deviation parameter

For the Gaussian distribution of the real value x

$$f(x \mid \sigma) = \frac{e^{-(x-\mu)^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}},$$

with μ fixed, the Jeffreys prior for the standard deviation $\sigma > 0$ is

$$\begin{aligned} p(\sigma) \propto \sqrt{I(\sigma)} &= \sqrt{\mathrm{E}\left[\left(\frac{d}{d\sigma}\log f(x \mid \sigma)\right)^2\right]} = \sqrt{\mathrm{E}\left[\left(\frac{(x-\mu)^2-\sigma^2}{\sigma^3}\right)^2\right]} \\ &= \sqrt{\int_{-\infty}^{+\infty} f(x \mid \sigma)\left(\frac{(x-\mu)^2-\sigma^2}{\sigma^3}\right)^2 dx} = \sqrt{\frac{2}{\sigma^2}} \propto \frac{1}{\sigma}. \end{aligned}$$

Equivalently, the Jeffreys prior for $\log\sigma = \int d\sigma/\sigma$ is the unnormalized uniform distribution on the real line, and thus this distribution is also known as the **logarithmic prior**. Similarly, the Jeffreys prior for $\log\sigma^2 = 2\log\sigma$ is also uniform. It is the unique (up to a multiple) prior (on the positive reals) that is *scale*-invariant (the Haar measure with respect to multiplication of positive reals), corresponding to the standard deviation being a measure of *scale* and scale-invariance corresponding to no information about scale. As with the uniform distribution on the reals, it is an improper prior.

Poisson distribution with rate parameter

For the Poisson distribution of the non-negative integer n,

$$f(n \mid \lambda) = e^{-\lambda}\frac{\lambda^n}{n!},$$

the Jeffreys prior for the rate parameter $\lambda > 0$ is

$$\begin{aligned} p(\lambda) \propto \sqrt{I(\lambda)} &= \sqrt{\mathrm{E}\left[\left(\frac{d}{d\lambda}\log f(n \mid \lambda)\right)^2\right]} = \sqrt{\mathrm{E}\left[\left(\frac{n-\lambda}{\lambda}\right)^2\right]} \\ &= \sqrt{\sum_{n=0}^{+\infty} f(n \mid \lambda)\left(\frac{n-\lambda}{\lambda}\right)^2} = \sqrt{\frac{1}{\lambda}}. \end{aligned}$$

Equivalently, the Jeffreys prior for $\sqrt{\lambda} = \int d\lambda/\sqrt{\lambda}$ is the unnormalized uniform distribution on the non-negative real line.

Bernoulli trial

For a coin that is "heads" with probability $\gamma \in [0, 1]$ and is "tails" with probability $1-\gamma$, for a given $(H, T) \in \{(0, 1), (1, 0)\}$ the probability is $\gamma^H(1-\gamma)^T$. The Jeffreys prior for the parameter γ is

$$\begin{aligned} p(\gamma) \propto \sqrt{I(\gamma)} &= \sqrt{\mathrm{E}\left[\left(\frac{d}{d\gamma}\log f(x \mid \gamma)\right)^2\right]} = \sqrt{\mathrm{E}\left[\left(\frac{H}{\gamma} - \frac{T}{1-\gamma}\right)^2\right]} \\ &= \sqrt{\gamma\left(\frac{1}{\gamma} - \frac{0}{1-\gamma}\right)^2 + (1-\gamma)\left(\frac{0}{\gamma} - \frac{1}{1-\gamma}\right)^2} = \frac{1}{\sqrt{\gamma(1-\gamma)}}. \end{aligned}$$

This is the arcsine distribution and is a beta distribution with $\alpha = \beta = 1/2$. Furthermore, if $\gamma = \sin^2(\theta)$ the Jeffreys prior for θ is uniform in the interval $[0, \pi/2]$. Equivalently, θ is uniform on the whole circle $[0, 2\pi]$.

N-sided die with biased probabilities

Similarly, for a throw of an N-sided die with outcome probabilities $\vec{\gamma} = (\gamma_1, \ldots, \gamma_N)$, each non-negative and satisfying $\sum_{i=1}^{N}\gamma_i = 1$, the Jeffreys prior for $\vec{\gamma}$ is the Dirichlet distribution with all (alpha) parameters set to one half. This amounts to using a pseudocount of one half for each possible outcome.

Alternatively, if we write $\gamma_i = {\phi_i}^2$ for each i, then the Jeffreys prior for $\vec{\phi}$ is uniform on the (*N*−1)-dimensional unit sphere (*i.e.*, it is uniform on the surface of an *N*-dimensional unit ball).

References

1. Jaynes, E. T. (1968) "Prior Probabilities", *IEEE Trans. on Systems Science and Cybernetics*, **SSC-4**, 227 pdf (http://bayes.wustl.edu/etj/articles/prior.pdf).

Further reading

- Jeffreys, H. (1946). "An Invariant Form for the Prior Probability in Estimation Problems". *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*. **186** (1007): 453–461. doi:10.1098/rspa.1946.0056 (https://doi.org/10.1098%2Frspa.1946.0056). JSTOR 97883 (https://www.jstor.org/stable/97883).
- Jeffreys, H. (1939). *Theory of Probability*. Oxford University Press.