Statistical Inference

Robert L. Wolpert
Institute of Statistics and Decision Sciences
Duke University, Durham, NC, USA

1. Information in Multidimensional Problems

1.1. Example: Normal

1.2. Example: Beta

2. Default Prior Distributions

2.1. Uniform

2.2. Invariant

2.3. Jeffreys

Let's restrict attention to problems $k = 1$ with dimensional parameter space $\Theta \subset \mathbb{R}$ for a moment, and imagine what happens with a reparametrization to $\eta = g(\theta)$ for some 1:1 monotonic function $g : \mathbb{R} \to \mathbb{R}$. If we impose uniformity in one parametrization, then

$$\begin{aligned} \pi_\eta(\eta) &= \frac{\pi_\theta(\theta)}{|g'(\theta)|} \\ &= 1/|g'\big(g^{-1}(\eta)\big)|, \end{aligned}$$

which will typically *not* be uniform; thus the idea of using as a default the uniform distribution entails the significant choice of parametrization. The Binomial Distribution, for example, can be parametrized equally well by the success probability p or by its logistic $\eta = \log \frac{p}{1-p}$; scaling in the Normal

Distribution can be parametrized using standard deviation σ, variance σ^2, precision $\tau = \sigma^{-2}$, or the logarithm $\omega = \log \sigma^2$. The posterior distributions, and hence the inference, can depend on these choices:

Bernoulli	p	$p^s(1-p)^f$	$\mathsf{Be}(s+1, f+1)$
	$\log \frac{p}{1-p}$	$p^{s-1}(1-p)^{f-1}$	Be(s,f)
Normal	σ	$\tau^{(n-3)/2} e^{-\tau \frac{n}{2} S^2}$	$\mathsf{Ga}(\frac{n-1}{2}, \frac{n}{2} S^2)$
	σ^2	$\tau^{(n-4)/2} e^{-\tau \frac{n}{2} S^2}$	$\mathsf{Ga}(\frac{n-2}{2}, \frac{n}{2} S^2)$
	$\log \sigma^2$	$\tau^{(n-2)/2} e^{-\tau \frac{n}{2} S^2}$	$\mathsf{Ga}(\frac{n}{2}, \frac{n}{2} S^2)$
	σ^{-2}	$\tau^{n/2} e^{-\tau \frac{n}{2} S^2}$	$\mathsf{Ga}(\frac{n+2}{2}, \frac{n}{2} S^2)$

Certainly it's disappointing that the apparently arbitrary choice of parametrization should affect the posterior and, through it, the inference.

Laplace's suggestion was to choose a parametrization for which uniformity was most plausible; Sir Harrold Jeffreys had another idea, a new recipe for a prior distribution $\pi_J(\theta)$ that would be invariant under changes in parametrization. He began by looking at how the Information matrix changes under changes in parametrization from θ to $\eta = g(\theta)$. First consider the one-dimensional version:

$$\begin{aligned} I^\theta &= -\frac{\partial^2}{\partial \theta^2} \log f(x|\theta) \\ &= -\Big(\frac{\partial \eta}{\partial \theta}\Big)^2 \frac{\partial^2}{\partial \eta^2} \log f(x|\eta) \\ &= \Big(\frac{\partial \eta}{\partial \theta}\Big)^2 I^\eta \end{aligned}$$

Thus the Jacobian $\frac{\partial \eta}{\partial \theta}$ can be evaluated by $\sqrt{I^\theta / I^\eta}$— and, in particular,

$$\pi_J(\theta) \equiv \sqrt{I(\theta)}$$

determines a prior density that transforms exactly the right way under smooth changes of variables. In k dimensions the same idea again leads to an invariant distribution, $\pi_J(\theta) \equiv \sqrt{|I(\theta)|}$, where $|I(\theta)|$ denotes the determinant of the Fisher information matrix.

2.3.1. Examples

For the Bernoulli and Binomial distributions the Information is $I(p) = \frac{1}{p(1-p)}$, so $\pi_J(p) \propto 1/\sqrt{p(1-p)}$ is the $\mathsf{Be}(1/2, 1/2)$ (or "arcsin") law, for which the posterior upon observing s successes and f failures is $\mathsf{Be}(s + 1/2, f + 1/2)$; this is halfway inbetween the earlier results treating p and $\log \frac{p}{1-p}$ as uniform.

For the normal distribution the Information matrix is $I(\mu, \tau) = \begin{pmatrix} \tau & 0 \\ 0 & \tau^{-2}/2 \end{pmatrix}$ with determinant $|I| \propto \tau^{-1}$, leading to $\pi_J(\mu, \tau) \propto \tau^{-1/2}$. Changing variables to standard deviation $\sigma = \tau^{-1/2}$ or variance $v = \sigma^2 = \tau^{-1}$ leads to $\pi_J(\mu, \sigma) \propto \sigma^{-2}$ and $\pi_J(\mu, v) \propto v^{-3/2}$, respectively. Under this distribution the posterior distribution for μ is a noncentral t centered at $\bar{X}$ with $\nu = n$ degrees of freedom.

Most authors prefer the posterior distribution under the prior $\pi(\mu, \tau) \propto \tau^{-1}$, leading to noncentral t with $n-1$ degrees of freedom; this $\pi(\mu, \tau)$ arises naturally either as right Haar measure on $\mathbb{R}^2$, treated as the group of translations and rescaling, or as the product $\pi_J(\mu)\pi_J(\tau)$ of Jeffreys prior distributions for μ and τ, each treated as the sole parameter in a one-dimensional inference problem.