Stat260: Bayesian Modeling and Inference **Lecture Date: February 17, 2010**

Lecture 7: Jeffreys Priors and Reference Priors

Lecturer: Michael I. Jordan *Scribe: Kevin McLoughlin*

1 Jeffreys Priors

Recall from last time that the Jeffreys prior is defined in terms of the Fisher information:

$$\pi_J(\theta) \propto I(\theta)^{\frac{1}{2}} \tag{1}$$

where the Fisher information $I(\theta)$ is given by

$$I(\theta) = -\mathbb{E}_\theta \left[\frac{d^2 \log p(X|\theta)}{d\theta^2} \right] \tag{2}$$

Example 1. Suppose X is binomially distributed:

$$\begin{aligned} X &\sim \text{Bin}(n, \theta), 0 \le \theta \le 1 \\ p(x|\theta) &= \binom{n}{x} \theta^x (1-\theta)^{n-x} \end{aligned}$$

We want to choose a prior $\pi(\theta)$ that is invariant under reparameterizations. We saw previously that a flat prior $\pi(\theta) \propto 1$ does not have this property. Let's derive a Jeffreys prior for θ. Ignoring terms that don't depend on θ, we have

$$\begin{aligned} \log p(x|\theta) &= x \log \theta + (n-x) \log(1-\theta) \\ \frac{d}{d\theta} \log p(x|\theta) &= \frac{x}{\theta} - \frac{n-x}{1-\theta} \\ \frac{d^2}{d\theta^2} \log p(x|\theta) &= -\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2} \end{aligned}$$

Since $\mathbb{E}_\theta X = n\theta$ under $\text{Bin}(n, \theta)$, we have

$$\begin{aligned} I(\theta) &= -\mathbb{E}_\theta \left[\frac{d^2 \log p(x|\theta)}{d\theta^2} \right] \\ &= \frac{n\theta}{\theta^2} + \frac{n - n\theta}{(1-\theta)^2} \\ &= \frac{n}{\theta} + \frac{n}{1-\theta} \\ &= \frac{n}{\theta(1-\theta)} \end{aligned}$$

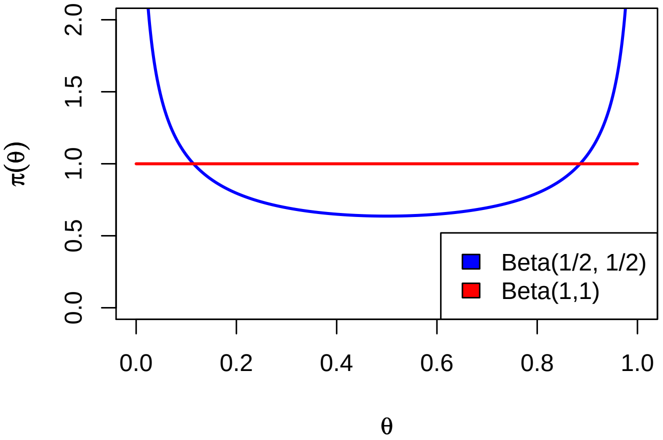

Figure 1: Jeffreys prior and flat prior densities

Therefore $\pi_J(\theta) = I(\theta)^{\frac{1}{2}} \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}}$, which is the form of a $\text{Beta}(\frac{1}{2}, \frac{1}{2})$ density.

Figure 1 compares the prior density $\pi_J(\theta)$ with that for a flat prior (which is equivalent to a $\text{Beta}(1, 1)$ distribution).

Note that in this case the prior is inversely proportional to the standard deviation. Why does this make sense?

We see that the data has the least effect on the posterior when the true $\theta = \frac{1}{2}$, and has the greatest effect near the extremes, $\theta = 0$ or 1. The Jeffreys prior compensates for this by placing more mass near the extremes of the range, where the data has the strongest effect. We could get the same effect by (for example) setting $\pi(\theta) \propto \frac{1}{\text{Var}(\theta)}$ instead of $\pi(\theta) \propto \frac{1}{\text{Var}(\theta)^{\frac{1}{2}}}$. However, the former prior is not invariant under reparameterization, as we would prefer.

1.1 Jeffreys priors and conjugacy

Jeffreys priors are widely used in Bayesian analysis. In general, they are not conjugate priors; the fact that we ended up with a conjugate Beta prior for the binomial example above is just a lucky coincidence. For example, with a Gaussian model $X \sim \mathcal{N}(\mu, \sigma^2)$ we showed in the last lecture that

$$\pi_J(\mu) \propto 1$$
$$\pi_J(\sigma) \propto \frac{1}{\sigma}$$

which do not look anything like a Gaussian or an inverse gamma, respectively.

However, it can be shown that Jeffreys priors are *limits* of conjugate prior densities. For example, a Gaussian density $\mathcal{N}(\mu_0, \sigma_0^2)$ approaches a flat prior as $\sigma_0 \to \infty$, while the inverse gamma $\sigma \propto \sigma^{-(a+1)}e^{-b/\sigma} \to 1/\sigma$ as $a, b \to 0$.

1.2 Limitations of Jeffreys priors

Jeffreys priors work well for single parameter models, but not for models with multidimensional parameters. By analogy with the one-dimensional case, one might construct a naive Jeffreys prior as the joint density:

$$\pi_J(\theta) = |I(\theta)|^{1/2}$$

where the Fisher information *matrix* is given by:

$$I(\theta)_{ij} = -\mathbb{E}_\theta \left[\frac{\partial^2 \log p(X|\theta)}{\partial\theta_i \partial\theta_j} \right]$$

Let's see what happens when we apply a Jeffreys prior for θ to a multivariate Gaussian location model. Suppose $X \sim N(\theta, I)$ for some p-dimensional random vector X, and we are interested in performing inference on $||\theta||^2$. In this case the Jeffreys prior for θ is flat. It turns out that the posterior has the form of a noncentral χ^2 distribution with p degrees of freedom. The posterior mean given one observation of X is $\mathbb{E}(||\theta||^2|X) = ||X||^2 + p$. This is not a good estimate because it adds p to the square of the norm of X whereas we might normally want to shrink our estimate towards zero. By contrast, the minimum variance frequentist estimate of $||\theta||^2$ is $||X||^2 - p$.

Intuitively, a multidimensional flat prior carries a lot of information about the expected value of a parameter. Since most of the mass of a flat prior distribution is in a shell at infinite distance, it says that we expect the value of θ to lie at some extreme distance from the origin, which causes our estimate of the norm to be pushed further away from zero.

Example 2. Consider a naive Jeffreys prior for a two-parameter Gaussian: $X \sim N(\mu, \sigma^2)$, and let $\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}$. We take derivatives to compute the Fisher information matrix:

$$\begin{aligned} I(\theta) &= -\mathbb{E}_\theta \begin{pmatrix} \frac{1}{\sigma^2} & \frac{2(X-\mu)}{\sigma^2} \\ \frac{2(X-\mu)}{\sigma^2} & \frac{3}{\sigma^4}(X-\mu)^2 - \frac{1}{\sigma^2} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{\sigma^2} \end{pmatrix} \end{aligned}$$

since $\mathbb{E}_\theta(X-\mu) = 0$ and $\mathbb{E}_\theta(X-\mu)^2 = \sigma^2$. Therefore

$$\pi_J(\theta) = |I(\theta)|^{1/2} \propto \frac{1}{\sigma^2}.$$

Unfortunately, this prior turns out to have poor convergence properties.

Jeffreys himself proposed using the prior $\pi_J(\theta) \propto \frac{1}{\sigma}$, which is a product of the separate priors for μ and σ. This prior is better motivated and gives better results as well. It also turns out to be the same as the reference prior, which we will discuss next.

2 Reference Priors

Reference priors were proposed by Jose Bernardo in a 1979 paper, and further developed by Jim Berger and others from the 1980's through the present. They are credited with bringing about an "objective Bayesian renaissance"; an annual conference is now devoted to the objective Bayesian approach.

The idea behind reference priors is to formalize what exactly we mean by an "uninformative prior": it is a function that maximizes some measure of distance or divergence between the posterior and prior, as data observations are made. Any of several possible divergence measures can be chosen, for example the Kullback-Leibler divergence or the Hellinger distance. By maximizing the divergence, we allow the data to have the maximum effect on the posterior estimates.

For one dimensional parameters, it will turn out that reference priors and Jeffreys priors are equivalent. For multidimensional parameters, they differ.

One might ask, how can we choose a prior to maximize the divergence between the posterior and prior, without having seen the data first? Reference priors handle this by taking the *expectation* of the divergence, given a model distribution for the data. This sounds superficially like a frequentist approach - basing inference on "imagined" data. But once the prior is chosen based on some model, inference proceeds in a standard Bayesian fashion. (This contrasts with the frequentist approach, which continues to deal with imagined data even after seeing the real data!)

2.1 Reference priors and mutual information

Consider an inference problem in which we have data X parameterized by Θ, with sufficient statistic $T = T(X)$. We want to find a reference prior $p(\theta)$ that maximizes its K-L divergence from the posterior $p(\theta|t)$, averaged over the distribution of T. This K-L divergence is

$$\int p(\theta|t)\log\frac{p(\theta|t)}{p(\theta)}d\theta$$

Its expectation over the distribution of T can be written:

$$\begin{aligned} I(\Theta,T) &= \int p(t)\int p(\theta|t)\log\frac{p(\theta|t)}{p(\theta)}d\theta dt \\ &= \int\int p(\theta,t)\log\frac{p(\theta,t)}{p(\theta)p(t)}d\theta dt \end{aligned}$$

This may be recognized as the mutual information between θ and t. Therefore, choosing a reference prior involves finding $p^*(\theta)$ that maximizes the mutual information:

$$p^*(\theta) = \arg\max_{p(\theta)} I(\Theta,T) \qquad (3)$$

We note that defining reference priors in terms of mutual information implies that they are invariant under reparameterization, since the mutual information itself is invariant.

Solving equation (3) is a problem in the calculus of variations. In the next lecture we'll derive reference priors for a variety of common situations.