Noninformative ("Default") Bayes

Rebecca C. Steorts
Predictive Modeling and Data Mining: STA 521

October 26 2015

Today's menu

- Subjective prior
- Default prior
- Are they really noninformative?
- Invariance property
- Jeffreys' prior

- Ideally, we would like a *subjective prior*: a prior reflecting our beliefs about the unknown parameter of interest.
- What are some examples in practice when we have subjective information?
- When may we not have subjective information?

In dealing with real-life problems you may run into problems such as

- not having past historical data,
- not having an expert opinion to base your prior knowledge on (perhaps your research is cutting-edge and new), or
- as your model becomes more complicated, it becomes hard to know what priors to put on each unknown parameter.

- Suppose we could find a distribution $p(\theta)$ that contained no or little information about θ in the sense that it didn't favor one value of θ over another (provided this is possible).
- Then it would be natural to refer to such a distribution as a *noninformative prior*. We could also argue that all or most of the information contained in the posterior distribution, $p(\theta|x)$, came from the data.
- Thus, all resulting inferences were *objective, noninformative, default* and not subjective.

Informative/subjective priors represent our prior beliefs about parameter values before collecting any data. For example, in reality, if statisticians are unsure about specifying the prior, they will turn to the experts in the field or experimenters to look at past data to help fix the prior.

Example 2.7, Carlin, Rubin, etc

Suppose that X is the number of pregnant mothers arriving at a hospital to deliver their babies during a given month. The discrete count nature of the data as well as its natural interpretation leads to adopting a Poisson likelihood,

$$p(x|\theta) = \frac{e^{-\theta}\theta^x}{x!},\ x \in \{0, 1, 2, \ldots\},\ \theta > 0.$$

A convenient choice for the prior distribution here is a Gamma(a, b) since it is conjugate for the Poisson likelihood. To illustrate the example further, suppose that 42 moms deliver babies during the month of December. Suppose from past data at this hospital, we assume a prior of Gamma$(5, 6)$. From this, we can easily calculate the posterior distribution, posterior mean and variance, and do various calculations of interest in `R`.

Noninformative/objective priors contain little or no information about θ in the sense that they do not favor one value of θ over another. Therefore, when we calculate the posterior distribution, most if not all of the inference will arise from the likelihood. Inferences in this case are *objective and not subjective*. Let's look at the following example to see why we might consider such priors.

- As we noted earlier, it would be natural to take the prior on θ as Gamma(a, b) since it is the conjugate prior for the Poisson likelihood.
- However suppose that for this data set we do not have any information on the number of pregnant mothers arriving at the hospital so there is no basis for using a Gamma prior or any other *informative* prior.
- In this situation, we could take some noninformative prior.

Comment: Since many of the objective priors are improper, so we must check that the posterior is proper.

Propriety of the Posterior

- If the prior is proper, then the posterior will *always* be proper.
- If the prior is improper, you must check that the posterior is proper.

What does a "flat prior" really mean? People really abuse the word flat and interchange it for noninformative. Let's talk about what people really mean when they use the term "flat," since it can have different meanings.
Often statisticians will refer to a prior as being flat, when a plot of its density actually looks flat, i.e., uniform. An example of this would be taking such a prior to be

$$\theta \sim \text{Unif}(0, 1).$$

We can plot the density of this prior to see that the density is flat.

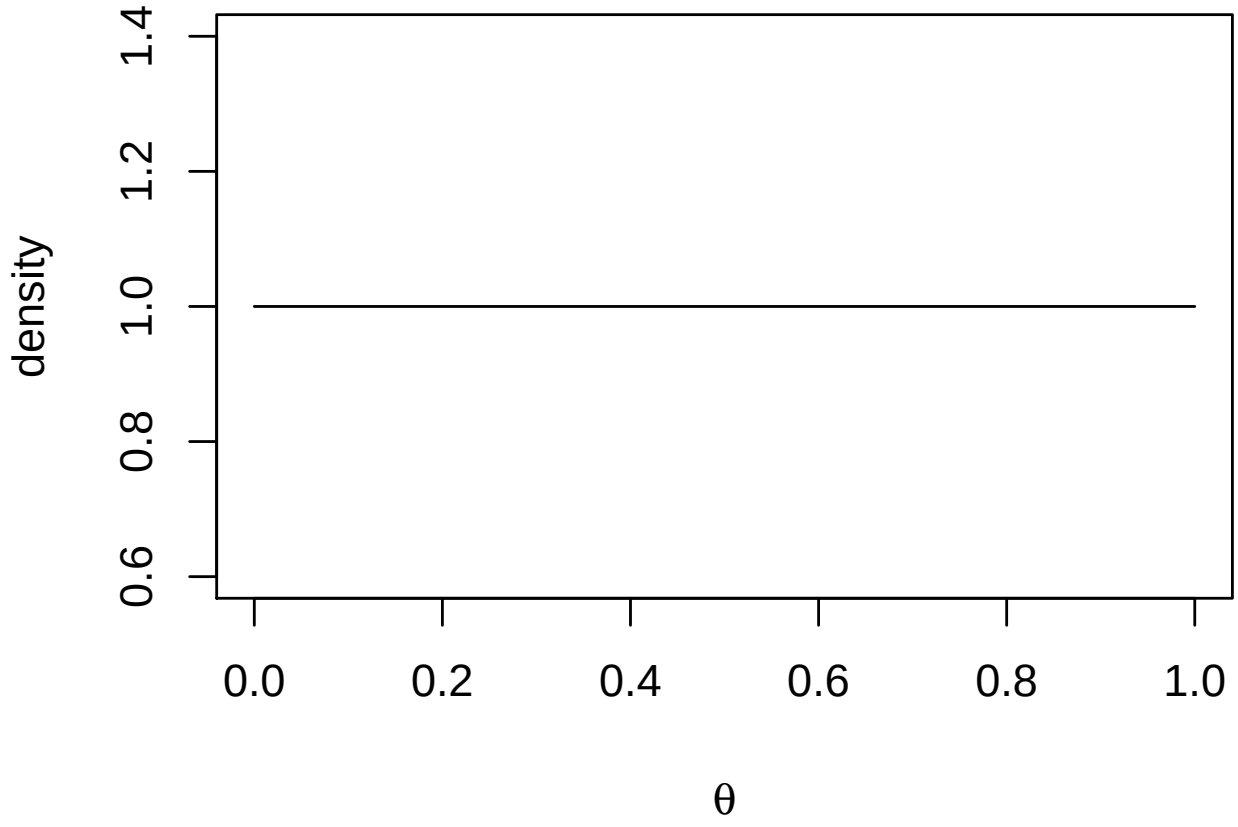

Figure 1: Unif(0,1) prior

What happens if we consider though the transformation to $1/\theta$. Is our prior still flat?

Suppose we consider Jeffreys' prior, $p_J(\theta)$, where $X \sim \text{Bin}(n, \theta)$.

We calculate Jeffreys' prior by finding the Fisher information. The Fisher information tells us how much information the data gives us for certain parameter values.

- Here, $p_J(\theta) \propto \text{Beta}(1/2, 1/2)$.
- Let's consider the plot of this prior. Flat here is a purely abstract idea.
- In order to achieve objective inference, we need to compensate more for values on the boundary than values in the middle.

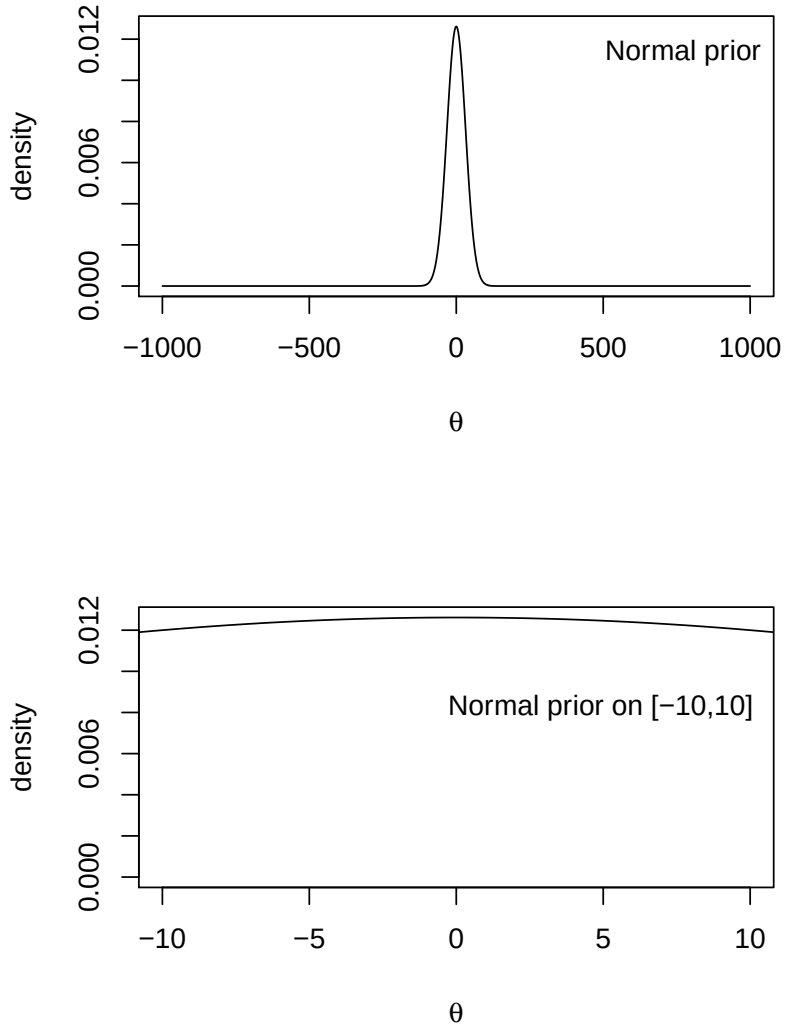

Figure 2: Normal priors

The Uniform Prior

(Thomas Bayes)

- In 1763, Thomas Bayes considered the question of what prior to use when estimating a binomial success probability p.
- He described the problem quite differently back then by considering throwing balls onto a billiard table. He separated the billiard table into many different intervals and considered different events.
- By doing so (and not going into the details of this), he argued that a Uniform(0,1) prior was appropriate for p.

(Laplace) In 1814, Pierre-Simon Laplace wanted to know the probability that the sun will rise tomorrow. He answered this question using the following Bayesian analysis:

- Let X represent the number of days the sun rises. Let p be the probability the sun will rise tomorrow.
- Let $X|p \sim \text{Bin}(n, p)$.
- Suppose $p \sim \text{Uniform}(0, 1)$.
- Based on reading the Bible, Laplace computed the total number of days n in recorded history, and the number of days x on which the sun rose. Clearly, $x = n$.

Then

$$\begin{aligned}\pi(p|x) &\propto \binom{n}{x} p^x (1-p)^{n-x} \cdot 1 \\ &\propto p^{x+1-1}(1-p)^{n-x+1-1}\end{aligned}$$

This implies

$$p|x \sim \mathsf{Beta}(x+1, n-x+1)$$

Then

$$\hat{p} = E[p|x] = \frac{x+1}{x+1+n-x+1} = \frac{x+1}{n+2} = \frac{n+1}{n+2}.$$

- Thus, Laplace's estimate for the probability that the sun rises tomorrow is $(n+1)/(n+2)$, where n is the total number of days recorded in history.
- For instance, if so far we have encountered 100 days in the history of our universe, this would say that the probability the sun will rise tomorrow is $101/102 \approx 0.9902$.
- However, we know that this calculation is ridiculous.
- Here, we have extremely strong subjective information (the laws of physics) that says it is extremely likely that the sun will rise tomorrow.
- Thus, objective Bayesian methods shouldn't be recklessly applied to every problem we study—especially when subjective information this strong is available.

Criticism of the Uniform Prior

- The Uniform prior of Bayes and Laplace and has been criticized for many different reasons.
- We will discuss one important reason for criticism and not go into the other reasons since they go beyond the scope of this course.
- In statistics, it is often a good property when a rule for choosing a prior is *invariant* under what are called one-to-one transformations. Invariant basically means unchanging in some sense.
- The invariance principle means that a rule for choosing a prior should provide equivalent beliefs even if we consider a transformed version of our parameter, like p^2 or $\log p$ instead of p.

Jeffreys' Prior

One prior that is invariant under one-to-one transformations is Jeffreys' prior.

What does the invariance principle mean?

Suppose our prior parameter is θ, however we would like to transform to ϕ.

Define $\phi = f(\theta)$, where f is a one-to-one function.

Jeffreys' prior says that if θ has the distribution specified by Jeffreys' prior for θ, then $f(\theta)$ will have the distribution specified by Jeffreys' prior for ϕ. We will clarify by going over two examples to illustrate this idea.

Example: Uniform

Note, for example, that if θ has a Uniform prior, Then one can show $\phi = f(\theta)$ will not have a Uniform prior (unless f is the identity function).

Example: Jeffreys'

Define

$$I(\theta) = -E\left[\frac{\partial^2 \log p(y|\theta)}{\partial \theta^2}\right],$$

where $I(\theta)$ is called the Fisher information. Then *Jeffreys' prior* is defined to be

$$p_J(\theta) = \sqrt{I(\theta)}.$$

For homework you will prove that the uniform prior in not invariant to transformation but that Jeffrey's is.

Example: Jeffreys'

Suppose

$$X|\theta \sim \text{Binomial}(n, \theta).$$

Let's calculate the posterior using Jeffreys' prior. To do so we need to calculate $I(\theta)$. Ignoring terms that don't depend on θ, we find

$$\log p(x|\theta) = x \log(\theta) + (n - x) \log(1 - \theta) \implies$$
$$\frac{\partial \log p(x|\theta)}{\partial \theta} = \frac{x}{\theta} - \frac{n - x}{1 - \theta}$$
$$\frac{\partial^2 \log p(x|\theta)}{\partial \theta^2} = -\frac{x}{\theta^2} - \frac{n - x}{(1 - \theta)^2}$$

Example: Jeffreys'

Since, $E(X) = n\theta$, then

$$I(\theta) = -E\left[-\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2}\right] = \frac{n\theta}{\theta^2} + \frac{n-n\theta}{(1-\theta)^2} = \frac{n}{\theta}\frac{n}{(1-\theta)} = \frac{n}{\theta(1-\theta)}.$$

This implies that

$$\begin{aligned} p_J(\theta) &= \sqrt{\frac{n}{\theta(1-\theta)}} \\ &\propto \text{Beta}(1/2, 1/2). \end{aligned}$$

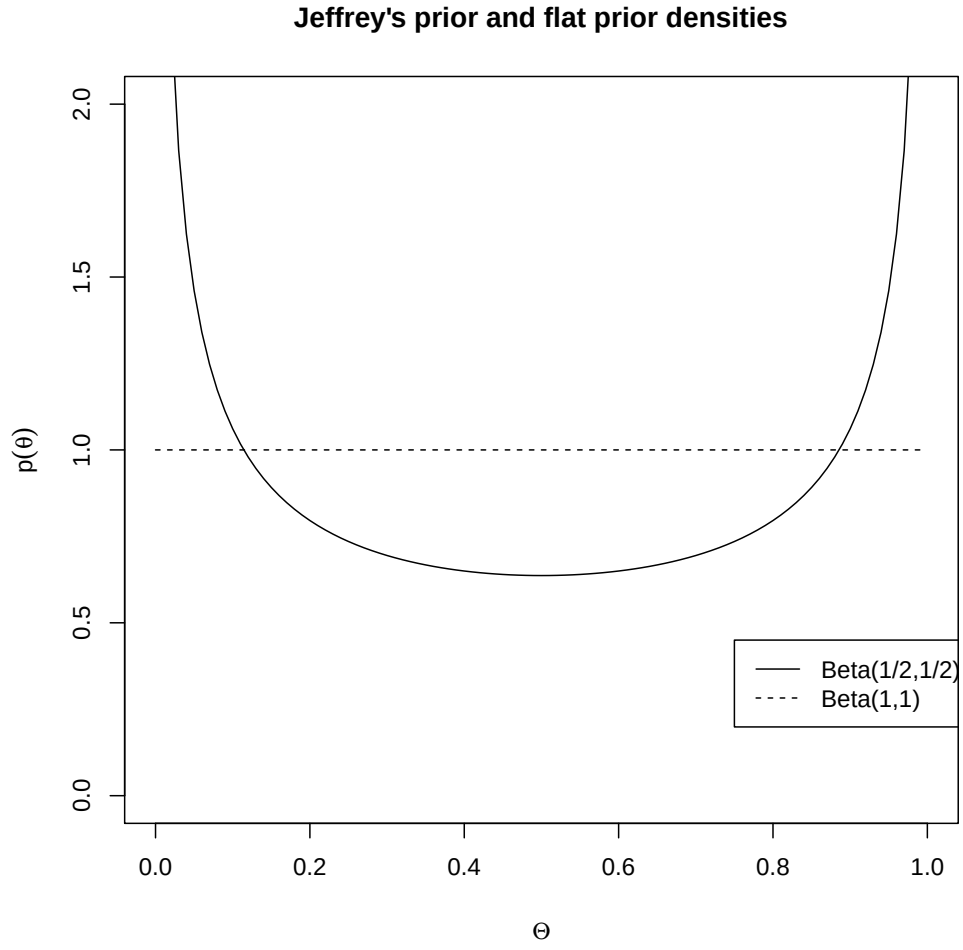

Figure 3: Jeffreys' prior and flat prior densities

Figure 3 compares the prior density $\pi_J(\theta)$ with that for a flat prior, which is equivalent to a Beta(1,1) distribution.

- We see that the data has the least effect on the posterior when the true $\theta = 1$, and has the greatest effect near the extremes, $\theta = 0$ or 1.
- 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) letting the prior be $\pi(\theta) \propto \frac{1}{\text{Var}\theta}$ instead of $\pi(\theta) \propto \frac{1}{[\text{Var}\theta]^{1/2}}$.
- However, the former prior is not invariant under reparameterization, as we would prefer.

We then find that

$$\begin{aligned} p(\theta \mid x) &\propto \theta^x (1-\theta)^{n-x} \theta^{1/2-1} (1-\theta)^{1/2-1} \\ &= \theta^{x-1/2} (1-\theta)^{n-x-1/2} \\ &= \theta^{x-1/2+1-1} (1-\theta)^{n-x-1/2+1-1}. \end{aligned}$$

Thus, $\theta|x \sim \text{Beta}(x + 1/2, n - x + 1/2)$, which is a proper posterior since the prior is proper.

Jeffreys' and Conjugacy

- 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 $X \sim N(\mu, \sigma^2)$, it can be shown that $\pi_J(\mu) = 1$ and $\pi_J(\sigma) = \frac{1}{\sigma}$, 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 $N(\mu_o, \sigma_o^2)$ approaches a flat prior as $\sigma_o^2 \to \infty$, while the inverse gamma $\sigma^{-(a+1)} e^{-b/\sigma} \to \sigma^{-1}$ as $a, b \to 0$.

Limitations of Jeffreys'

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 $|\cdot|$ denotes the determinant and the (i, j)th element of the Fisher information matrix is given by

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

Let's see what happens when we apply a Jeffreys' prior for θ to a multivariate Gaussian location model. Suppose

$$X \sim N_p(\theta, I),$$

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 non-central χ^2 distribution with p degrees of freedom.
- The posterior mean given one observation of X is $E(||\theta||^2 \mid X) = ||X||^2 + p$.
- 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 $||X||^2 - p$.