# Alternatives informative priors

October 24, 2017

## 1 Fischer Information

We should first define Fischer Information, because we will use it at this document. The Fisher information (for one parameter) is denoted as

$$I(\theta) = E[(\frac{d}{d\theta} log(f(X|\theta)))^2]$$

where the expectation is taken with respect to X which has a PDF $f(x|\theta)$. This quality is useful in obtaining estimators for $\theta$ with good properties, such as low variance. It is also the basis for Jeffrey's prior.

**Example**: Let $X|\theta \sim N(\theta, 1)$. Then we have:

$$f(x|\theta) = \frac{1}{\sqrt{2\pi}} exp[-\frac{1}{2}(x-\theta)^2]$$

$$log(f(x|\theta)) = -\frac{1}{2} log(2\pi) - \frac{1}{2}(x-\theta)^2$$

$$\frac{d}{d\theta} log(f(x|\theta)) = -\frac{2}{2}(x-\theta)(-1) = x - \theta$$

$$(\frac{d}{d\theta} log(f(x|\theta)))^2 = (x-\theta)^2$$

and so $I(\theta) = E[(X-\theta)^2] = Var(X) = 1$.

## 2 Non-informative priors

So far, we've seen examples of choosing priors that contain a significant amount of information. You've also seen some examples of choosing priors where we're attempting to not put too much information in to keep them vague.

Another approach is referred to as objective Bayesian statistics or inference where we explicitly try to minimize the amount of information that goes into the prior.

### 2.1 First case

This is an attempt to have the data have maximum influence on the posterior which mention further this as **non-informative priors**.

For example, let's go back to coin flipping or data comes from Bernoulli distribution with unknown parameter theta. How do we minimize our par information in $\theta$? One obvious intuitive approach is to say that all values of theta are equally likely. So we could have a prior for $\theta \sim U[0, 1]$. Saying all values of theta are equally likely seems like it would have no information in it.

Saying all values of theta are equally likely seems like it would have no information in it.

The **effective sample size** of a beta prior is the sum of it's two parameters. So in this case it has an effective sample size of two. This is equivalent to data, with one head and one tail already in it.

So this is not a completely non informative prior.

We could think about a prior that has less information.

For example, a beta $1/2$, $1/2$. This would have only half as much information as an effective sample size of just one. We can take this even further. Think about something like a beta 0.001, 0.001. This would have much less information, have a sample fairly close to 0. In this case, the data would determine the posterior and there would be very little influence from the prior.

Can we go even further? In fact we can, we can think of the limiting case. Something that we can think of as a $Beta(0, 0)$. What would that look like?

$$f(\theta) \propto \theta^{-1}(1-\theta)^{-1}$$

**This is not a proper density. If you integrate this over 0 to 1, you'll get an infinite integral, so it's not a true density in the sense of integrating to 1.** There's no way to normalize it, it has an infinite integral. This is what we refer to as an **improper prior**.

It's improper in the sense that it doesn't have a proper density. But it's not necessarily improper, in the sense that we can't use it. If we collect data, we use this prior and as long as we observe at least one head and at least one tail. Or one's success and one's failure then we can get a posterior:

$$f(\theta|y) \propto \theta^{y-1}(1-\theta)^{n-y-1}\ ,\ \sim Beta(y, n-y)$$

.

Its posterior mean will be: $\hat{\theta} = \frac{y}{n}$, which should be recognised as the **MLE**.

So by using this improper prior, we get a posterior which gives us point estimates exactly the same as the frequentest approach.

But in this case, we can also think of having a full posterior. And so if we want to make interval statements, probability statements, we can actually find an interval and say that there's 95% probability that $\theta$ is in this interval. Which is not something you can do under the frequentest approach even though we may get the exact same interval.

Key concepts here that I want to state in terms of using improper priors.

```
1) improper priors are okay as long as the posterior itself is proper. There may be

2)  For many problems there does exist a prior, typically an improper prior. That w
```

### 2.2 Second case

Let: $Y_i \sim N(\mu, \sigma^2)$

** Known $\sigma$**

We assume that $\sigma$ is known. We take a vague prior : $\mu \sim N(0, 10000^2)$. That would just spread things out across the real line. You can take a wide variety of possible values. That would be fairly non informative across a lot of possibilities.

We can then think about taking the limit. What happens if we let the variance go to infinity? In that case, we're basically spreading out this distribution across the entire real line. And so we could say, we have a density which is proportional to what? It's just constant across the whole real line.

Clearly, this is an improper prior because if you integrate the real line you get an infinite answer. However, if we go ahead and plug this into finding a posterior

$$f(\mu|y) \propto f(\mu|y)f(y)$$

$$\propto \exp\{-\frac{1}{2\sigma^2}\sum(y_i - \mu)^2\}$$

$$\propto exp\{-\frac{1}{2\frac{\sigma^2}{n}}(\mu - \bar{y})^2\}$$

.

$$\mu|y \sim N(\bar{y}, \frac{\sigma^2}{n})$$

** Unknown $\sigma$**

The standard non informative prior is :

$$f(\sigma^2) \propto \frac{1}{\sigma^2}\ , \Gamma^{-1}(0,0)$$

This is an improper prior and it's uniform on the log scale of sigma squared.

Posterior for $\sigma^2$:

$$\sigma^2|y \sim \Gamma^{-1}(\frac{n-1}{2}, \frac{1}{2}\sum(y_i - \bar{y})^2)$$

# 3 Jeffrey's Prior

Choosing a uniform prior depends upon the particular parameterization. For example, $y_i \sim N(\mu, \sigma^2)$

Suppose, I used a prior which is uniform on the log scale for $\sigma^2$, so $f(\sigma^2) \propto \frac{1}{\sigma^2}$.

Suppose somebody else decides, they just want to put a uniform prior on $\sigma^2$ itself, $f(\sigma^2) \propto 1$. These are both uniform on certain scales, or certain parameterizations, but they are different priors.

So when we compute the posteriors, we will get different posteriors. The key thing is that **uniform priors are not invariant with respect to transformation**. Depending upon how you parameterize the problem, you can get different answers by using a uniform prior.

**Jeffreys prior is one attempt to round to this**

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

The Jeffreys prior is exactly the prior we have seen before. It's uniform for $\mu$, and then for $\sigma^2$ it's uniformed on the log scale. This prior will then be invariant transformation will be putting the same information into the prior. Even if we use a different parameterization for the normal.

In the example of $Y_i \sim B(\theta)$ $f(\theta) \propto \theta^{-\frac{1}{2}}(1-\theta)^{-\frac{1}{2}} \sim Beta(\frac{1}{2}, \frac{1}{2})$.

This is a rare example where the Jeffreys prior turns out to be a proper prior. You'll note that this prior actually does have some information in it. It's equivalent to an **effective sample size of one data point**. However this information will then be the same, not depending on the prioritization we use. This case we have $\theta$ as a probability. But another alternative, used in probabilities calculations, sometimes we model things on a logistics scale. And in that case, we can transfer everything and using the Jeffreys prior, we'll maintain the exact same information.

Other possible approaches to objective basing inference include priors such as reference priors and maximum entropy priors.

Finally, I'd like to mention a related concept which is empirical basing analysis. The idea in empirical base is that you use the data to help inform your prior, such as using the mean of the data to set the mean of the prior distribution. This approach often leads to reasonable point estimates in your posterior. However, it's sort of cheating, because you're using the data twice. And as a result, it may lead to improper uncertainty estimates.

**Question**

Jeffreys priors are "transformation invariant" in the sense that if we calculate the Jeffreys proor for $\theta$ and then reparameterize to use $\phi = g(\theta)$, we get the same result as if we had first reparameterized adn then found the Jeffrey's prior for $\phi$. Why might this property be desirable?

**Answer**

Different investigators might parameterize a problem in different ways. Using the Jeffreys prior that ensures they both obtain the same answer.

# 4 Review

** Question 1 **

Suppose we flip a coin five times to estimate $\theta$, the probability of obtaining heads. We use a Bernoulli likelihood for the data and a non-informative (and improper) Beta(0,0) prior for $\theta$. We observe the following sequence: (H, H, H, T, H).

Because we observed at least one H and at least one T, the posterior is proper. What is the posterior distribution for $\theta$?

**Answer 1**

```
In [2]: a <- 0
        b <- 0
        y <- c(1, 1, 1, 0, 1)
        n <- length(y)
        post_a <- a + sum(y); post_a
        post_b <- b + n - sum(y) ; post_b
        #Beta(4,1)
```

4

1

** Question 2**

Continuing the previous question, what is the posterior mean for $\theta$?

**Answer 2**

```
In [7]: post_a/(post_a + post_b)
```

0.8

Consider again the thermometer calibration problem.

Assume a normal likelihood with unknown mean $\theta$ and known variance $\sigma$ˆ{2}=0.25. Now use the non-informative (and improper) flat prior for $\theta$ across all real numbers. This is equivalent to a conjugate normal prior with variance equal to $\infty$.

** Question 3**

You collect the following n=5 measurements: (94.6, 95.4, 96.2, 94.9, 95.9). What is the posterior distribution for $\theta$?

**Answer 3**

N(95.4, 0.25)

Recall from the lesson that with a flat prior on $\theta$, the posterior distribution is $N(\bar{y}, \sigma^2)$.

```
In [11]: measurements <- c(94.6, 95.4, 96.2, 94.9, 95.9)
         n <- length(measurements)
         mean <- mean(measurements); mean
```

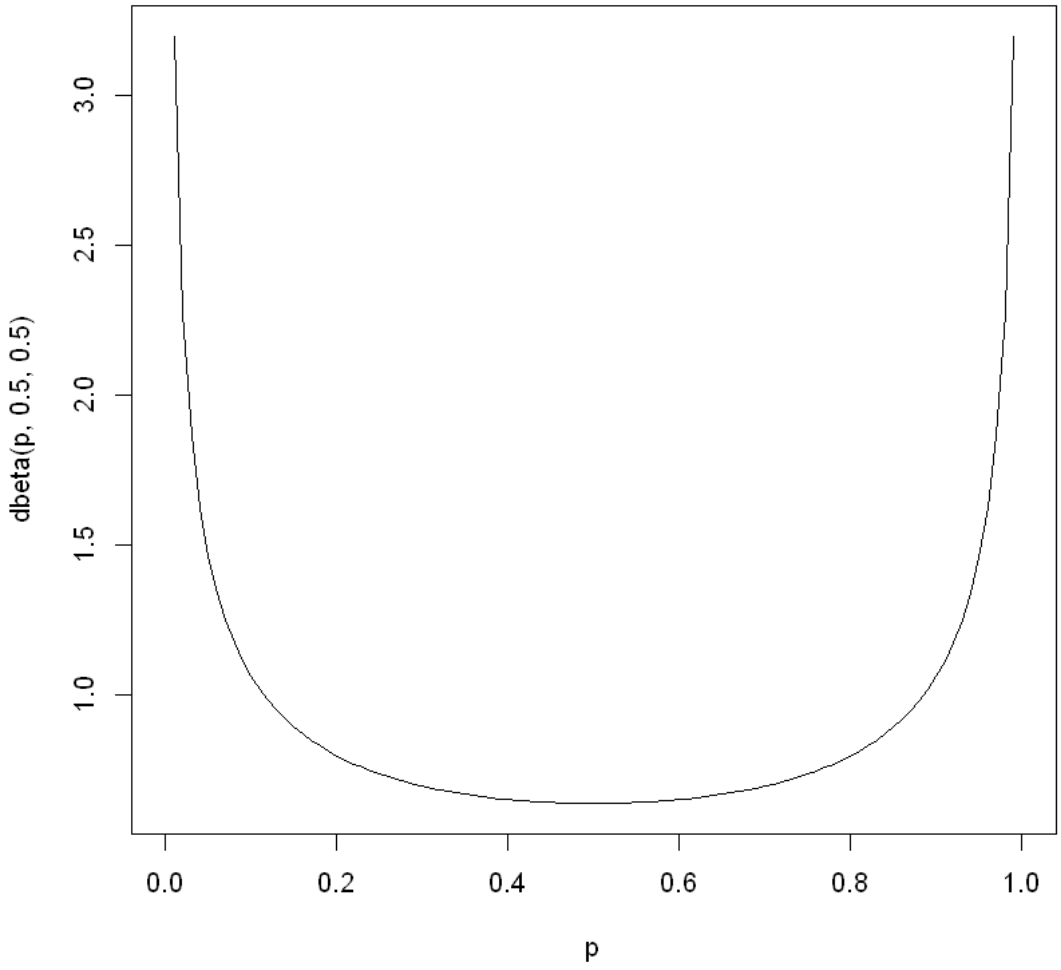
95.4

** Question 4**

Plot the Jeffreys prior for a Bernoulli/binomial success probability p.

**Answer 4**

```
In [9]: p=seq(from=0,to=1,by=.01)
        plot(p,dbeta(p,0.5,0.5),type="l")
```




** Question 5**

Scientist A studies the probability of a certain outcome of an experiment and calls it $\theta$. To be non-informative, he assumes a Uniform(0,1) prior for $\theta$.

Scientist B studies the same outcome of the same experiment using the same data, but wishes to model the odds $\phi = \dfrac{\theta}{1-\theta}$. Scientiest B places a uniform distribution on $\phi$. If she reports her inferences in terms of the probability $\theta$, will they be equivalent to the inferences made by Scientist A?

**Answer 5**

No, they did not use the Jeffreys prior.