Journal of Statistics Education

The Counter-intuitive Non-informative Prior for the Bernoulli Family

Mu Zhu
University of Waterloo

Arthur Y. Lu
Renaissance Technologies Corp.

Journal of Statistics Education Volume 12, Number 2 (2004),
http://www.amstat.org/publications/jse/v12n2/zhu.pdf

Key Words: Beta Distribution; Conjugate priors; Maximum likelihood estimation; Posterior mean.

Abstract

In Bayesian statistics, the choice of the prior distribution is often controversial. Different rules for selecting priors have been suggested in the literature, which, sometimes, produce priors that are difficult for the students to understand intuitively. In this article, we use a simple heuristic to illustrate to the students the rather counter-intuitive fact that flat priors are not necessarily non-informative; and non-informative priors are not necessarily flat.

1 Introduction

In Bayesian analysis, selecting priors using Jeffreys' rule (e.g., Tanner 1996; Gelman, et al. 1995) can yield some rather counter-intuitive results that are hard for students to grasp. Consider the following simple scenario: Let $X_1, X_2, ..., X_n$ be i.i.d. observations from the Bernoulli(p) distribution. To estimate p, a Bayesian analyst would put a prior distribution on p and use the posterior distribution of p to draw various conclusions, e.g., estimating p with the posterior mean. When there is no strong prior opinion on what p is, it is desirable to pick a prior that is *non-informative*.

In this simple case, it is most intuitive to use the uniform distribution on [0,1] as a non-informative prior; it is non-informative because it says that all possible values of p are equally likely *a priori*. However, a non-informative prior constructed using Jeffrey's rule is of the form (see e.g., Gelman 1995)

$$\pi(p) \propto \frac{1}{\sqrt{p(1-p)}}. \tag{1}$$

Jeffrey's rule is motivated by an invariance argument: Suppose one picks $\pi_p(p)$ as the prior for p according to a certain rule. In order for $\pi_p(p)$ to be non-informative, it is argued that the parameterization must not influence the choice of $\pi_p(p)$, i.e., if one re-parameterizes the problem in terms of $\theta = h(p)$, then the rule must pick

$$\pi_\theta(\theta) = \pi_p(h^{-1}(\theta))\left|\frac{dp}{d\theta}\right|$$

as the prior for θ.

Given p, let $f(x|p)$ be the likelihood function. Jeffrey's rule is to pick

$$\pi_p(p) \propto (I(p))^{\frac{1}{2}}$$

as a prior, where

$$I(p) = -\mathrm{E}\left(\frac{d^2 \log f(x|p)}{dp^2}\right)$$

is the Fisher Information. To see that this is invariant with respect to parameterization, suppose we re-parameterize in terms of $\theta = h(p)$, then

$$\begin{aligned} I(\theta) &= -\mathrm{E}\left(\frac{d^2 \log f(x|\theta)}{d\theta^2}\right) \\ &= -\mathrm{E}\left(\frac{d^2 \log f(x|p)}{dp^2}\left(\frac{dp}{d\theta}\right)^2\right) \\ &= I(p)\left(\frac{dp}{d\theta}\right)^2. \end{aligned}$$

Applying Jeffrey's rule, one would pick a prior on θ as

$$\begin{aligned} \pi_\theta(\theta) &\propto (I(\theta))^{\frac{1}{2}} \\ &= (I(p))^{\frac{1}{2}}\left|\frac{dp}{d\theta}\right| \\ &= \pi_p(p)\left|\frac{dp}{d\theta}\right|, \end{aligned}$$

which satisifies the invariance argument.

Jeffrey's prior for this simple problem can be quite counter-intuitive. Under the prior (1), it appears that some values of p are more likely than others (see e.g., Figure 1). Therefore

intuitively, it appears that this prior is actually quite informative. This is a very difficult point to explain to the students.

In this article, we construct a simple (albeit naive) argument and illustrate to the students why the uniform prior is not necessarily the most non-informative. We will not rely on the Fisher information or Jeffreys' invariance argument. Instead, we rely on a very simple and naive heuristic to judge the non-informativeness of a prior: Since the maximum likelihood estimator (MLE) is not affected by any prior opinion, we simply ask: is there a prior which would produce a Bayesian estimate (e.g., posterior mean) that coincides with the MLE? If so, that prior could be regarded as non-informative since the prior opinion exerts no influence on the final estimate whatsoever. Using this naive heuristic, we can see that the uniform prior is actually more informative than Jeffreys' prior (1); whereas the least informative prior is, surprisingly enough, an extremely "opinionated" distribution approaching two point masses at 0 and 1!

We emphasize here that it is not our intention to imply that our naive heuristic is the best or even an appropriate point of view for judging the non-informativeness of different priors in Bayesian analysis. We only provide this argument as an interesting demonstration that can be used in the classroom.

2 The Maximum Likelihood Estimator (MLE)

Without considering any prior opinion, a typical approach for estimating p is the method of maximum likelihood. Let X_i be a Bernoulli random variable with $P(X_i = 1) = p, P(X_i = 0) = 1 - p$; the log-likelihood function for the Bernoulli distribution is

$$l(p) = \sum_{i=1}^{n} \log p^{x_i}(1-p)^{1-x_i} = \sum_{i=1}^{n} x_i \log p + (1 - x_i) \log(1-p).$$

We shall write $x = \sum_{i=1}^{n} x_i$ throughout the article. To maximize this log-likelihood, the first order condition is

$$\frac{x}{p} - \frac{n-x}{1-p} = 0,$$

which gives

$$\hat{p}_{\text{mle}} = \frac{x}{n}.$$

3 The Bayesian Estimator

The Bayesian approach to estimation starts with a prior distribution on the parameter of interest. Often, we have no prior knowledge on p. To reflect such lack of knowledge, the most intuitive choice is to put a uniform prior on p, i.e., $\pi_0(p) = 1$ for $p \in [0, 1]$. This says that, *a priori*, p could be anything between 0 and 1 with equal chance. Then, the posterior

distribution of p is given by (Bayes' Theorem):

$$\pi_1(p|x_1, ..., x_n) = \frac{f(x_1, ..., x_n|p)\pi_0(p)}{\int_0^1 f(x_1, ..., x_n|p)\pi_0(p)dp}.$$

We shall find this posterior distribution more generally below using the idea of the conjugate prior.

3.1 The Beta Conjugate Prior

Consider the Beta(α, β) distribution as the prior for p, i.e.,

$$\pi_0(p) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}.$$

The uniform distribution is a special case of the Beta distribution, with $\alpha = \beta = 1$. The reason why one would consider using the Beta distribution as the prior is because the Beta distribution and the Bernoulli distribution form a *conjugate* pair, so that the posterior distribution is still a Beta (e.g., DeGroot 1970). This gives us some analytic convenience. To see this, note that

$$\begin{aligned} \pi_1(p) &\propto f(x_1, x_2, ..., x_n|p)\pi_0(p) && (2) \\ &= p^x(1-p)^{n-x} p^{\alpha-1}(1-p)^{\beta-1} && (3) \\ &= p^{\alpha+x-1}(1-p)^{\beta+n-x-1} && (4) \end{aligned}$$

is Beta$(\alpha + x, \beta + n - x)$. The following properties of the Beta distribution are useful: If $p \sim$ Beta(α, β), then

$$\mathrm{E}(p) = \frac{\alpha}{\alpha + \beta}, \quad (5)$$

and

$$\mathrm{Var}(p) = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \quad (6)$$

We can now write down a general formula for obtaining the Bayesian posterior mean, $\hat{p}_{\text{bayes}}$: if $\pi_0(p)$ is Beta(α, β), then

$$\hat{p}_{\text{bayes}} = \frac{\alpha + x}{\alpha + \beta + n}. \quad (7)$$

Therefore, under the uniform prior (i.e., $\alpha = \beta = 1$), the posterior mean is

$$\hat{p}_{\text{bayes}} = \frac{1 + x}{2 + n}.$$

Remark 1. For the simple purpose of demonstration, we use the posterior mean as the Bayesian estimate. However, we emphasize that generally one could use the posterior median or mode as a Bayesian point estimate as well.

3.2 The Effects of Different Priors

How do the parameters α and β (i.e., using different priors in the Beta family) affect the outcome? For this discussion, we focus on a particular sub-family of Beta distributions with $\alpha = \beta = c$, i.e., $\pi_0(p)$ is Beta(c,c). Again, the uniform distribution is a member of this sub-family, with $c = 1$. Furthermore, if the prior on $p \sim \text{Beta}(c, c)$, then

$$\text{E}(p) = \frac{c}{c+c} = \frac{1}{2} \quad \text{and} \quad \text{Var}(p) = \frac{c^2}{4c^2(2c+1)} = \frac{1}{4(2c+1)}. \tag{8}$$

It is clear from (7) that the prior parameter c influences the posterior mean as if an extra $2c$ observations, equally split between zeros and ones, were added to the sample. Therefore, the larger c is, the more influence the prior will have on the posterior mean. The uniform prior ($c = 1$) adds two observations; Jeffrey's prior, which according to equation (1) corresponds to $c = \frac{1}{2}$, adds one extra observation. It is in this sense that Jeffrey's prior is actually less influential than the uniform prior.

Since the prior variance is clearly a decreasing function in c (8), this also says that the larger the prior variance, the less influential the prior is, which makes intuitive sense: a large prior variance would normally indicate a relatively weak prior opinion. In view of this, two extreme cases become quite interesting: $c \to \infty$ and $c \to 0$.

Case 1: $c \to \infty$**.** It is easy to see from (7) that as $c \to \infty$, we have $\hat{p}_{\text{bayes}} = \frac{1}{2}$, the same as the prior mean regardless of what the observed outcomes are. In other words, our prior opinion of p is so strong that it can not be changed by the observed outcomes. From (8), we see that the prior variance approaches 0 as $c \to \infty$. This is, again, consistent with our intuition: the small prior variance means that one's prior belief is heavily concentrated on the point $p = \frac{1}{2}$, so heavy that the observed outcomes could not alter this belief in any way.

Case 2: $c \to 0$**.** Following the same logic, it is clear from (7) that the least influential prior in our sub-family would have been the one with $c = 0$. Using such a prior, the posterior mean would have been the same as the MLE; i.e., it would have been entirely determined by the observed outcomes. But the Beta(0,0) distribution is not defined. Therefore, we consider the distribution Beta(ϵ, ϵ) for arbitrarily small $\epsilon > 0$. To understand the behavior of this distribution, we can examine the limiting distribution as $c \to 0$:

$$B_{0,0} = \lim_{c \to 0} \text{Beta}(c, c).$$

Theorem 1 *The limiting distribution $B_{0,0}$ consists of two equal point masses at 0 and 1.*

From (8), it can be seen that the variance of $B_{0,0}$ is $\frac{1}{4}$; the above theorem (see Appendix for a proof) is due to the following fact: for a symmetric distribution with a compact support on the unit interval to have variance $\frac{1}{4}$, it must consist of just two equal point masses at 0 and 1.

Theorem 1 says that the prior distribution Beta(ϵ, ϵ) with arbitrarily small $\epsilon > 0$ approaches two point masses at 0 and 1. Such a prior belief, of course, seems extremely strong, since it says p is essentially either 0 or 1. Intuitively, one would consider such a strong prior belief to be extremely unreasonable, but this is the prior that would yield a posterior mean as close as possible to the MLE. In this sense, the prior Beta(ϵ, ϵ) for arbitrarily small $\epsilon > 0$, which would otherwise appear strong, could actually be regarded as the least influential prior in this family.

Remark 2. Theorem 1 states that the limiting distribution $B_{0,0}$ is the Bernoulli$(\frac{1}{2})$ distribution, which, strictly speaking, is not a member of the Beta family. Moreover, note that if $B_{0,0}$ is actually used as a prior, then the posterior distribution is not defined unless all the observations $X_1, X_2, ..., X_n$ are identical. Therefore $B_{0,0}$ is in itself quite an influential prior, but Beta(ϵ, ϵ) is not, although for arbitrarily small ϵ, it encodes essentially the same prior opinion as $B_{0,0}$, whose predictive distribution puts half probability on all ones and half on all zeros.

4 The Non-informative Prior

The lesson from this discussion is extremely interesting; it tells us that *flat* priors (such as the uniform prior) are not always the same thing as *non-informative* priors. A seemingly informative prior can actually be quite weak in the sense that it does not influence the posterior opinion very much. It is clear in our example that the MLE is the result of using a weak prior, whereas the most intuitive non-informative prior (the uniform prior) is not as weak or non-informative as one would have thought.

We've also seen that the least influential prior, Beta(ϵ, ϵ) for arbitrarily small $\epsilon > 0$, is also the one with the largest variance in the sub-family, whereas the most "stubborn" prior (when $c \to \infty$) is also the one with the smallest variance (8). Generally, a larger variance would also imply a flatter distribution. But since the family of Beta distributions has compact support on the unit interval, the variance is maximized by a rather extreme prior instead of the usual flat prior, Beta(1,1).

Remark 3. Another common prior that is used in the literature for this problem (see e.g., Zellner 1996, p. 40) is an improper prior of the form

$$\pi_0(p) \propto \frac{1}{p(1-p)}, \tag{9}$$

also called the Haldane prior. It is improper because $\int_0^1 \pi_0(p)dp = \infty$ and hence it is not a proper distribution function. Zellner (1996) notes that this improper prior corresponds to

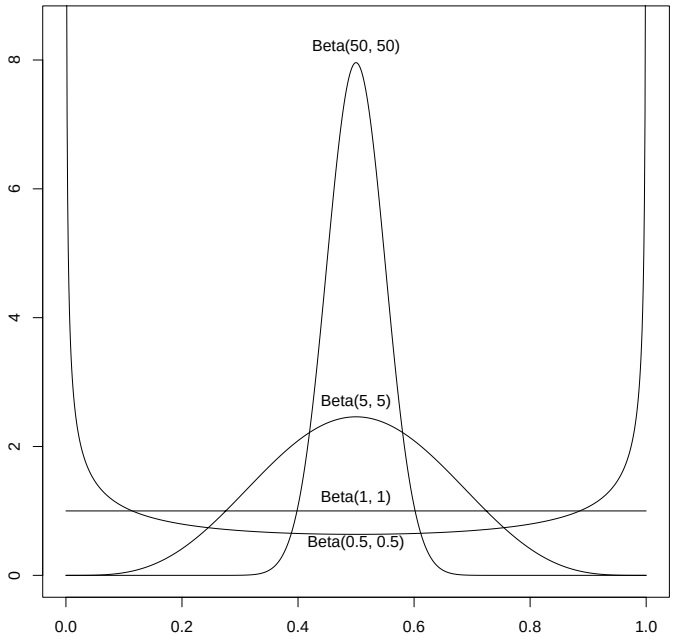

Figure 1: The Beta family of distributions on $[0, 1]$. *When* $\alpha = \beta = c$, *the Beta distribution is symmetric. When* $c = 1$, *it is the uniform distribution. When* $c > 1$, *it has a maximum at* $\frac{1}{2}$. *When* $c < 1$, *it has a minimum at* $\frac{1}{2}$ *and two maxima at* 0 *and* 1.

putting a flat prior on

$$\eta = \log \frac{p}{1-p},$$

the log-odds. In other words, on the log-odds scale, this improper prior *is* the usual flat, non-informative prior. Other than the improperness, however, we can see that (9) is closely related to the limiting distribution $B_{0,0}$: they are two different ways of expressing the otherwise undefined Beta(0,0) distribution.

Remark 4. One can use this simple heuristic in other situations as well to evaluate the non-informativeness of different priors. For example, consider $X_1, X_2, ..., X_n \sim \mathrm{N}(\mu, \sigma^2)$, where σ^2 is known. Let $\pi(\mu) \sim \mathrm{N}(\theta_0, \sigma_0^2)$ be the prior on μ. Then it can be shown (e.g., Tanner 1996, p. 17) that the posterior mean is

$$\theta_0 \left(\frac{1/\sigma_0^2}{1/\sigma_0^2 + n/\sigma^2} \right) + \bar{x} \left(\frac{n/\sigma^2}{1/\sigma_0^2 + n/\sigma^2} \right).$$

It is then clear that the posterior mean agrees with the MLE $\bar{x}$ if and only if $\sigma_0 \to \infty$, i.e., if we put a prior on μ that is essentially flat. In this case, using our simple heuristic, we arrive at the intuitive conclusion that the most non-informative prior is indeed flat.

Remark 5. In Remark 1, we emphasized that the choice to focus on the posterior mean as the Bayesian estimate is based on convenience. The posterior Beta distribution is generally not symmetric, so the posterior mean, median and mode do not coincide. This means some of the phenomena illustrated above will not hold if a different posterior point estimate is used. We do not worry about this point so much here since the material presented here is only meant as a classroom demonstration to help the students appreciate why non-informative priors are not always flat. However, the posterior mean, median and mode do agree for the normal family, but of course, there the prior regarded as non-informative by the current heuristic argument is essentially flat (see Remark 4).

5 Conclusion

That the posterior mean coincides with the MLE is not necessarily the right criterion to judge whether a prior is non-informative, but it provides an extremely simple and effective demonstration of why sometimes flat priors are not necessarily non-informative, while non-informative priors are not always flat, without having to resort to deep statistical theory. This simple argument, in our experiences, has proven relatively easy for the students to appreciate. The proofs below in the appendix can also be easily understood with basic statistics knowledge at the level of, say, Rice 1995. Since it is clear from this discussion that different priors may lead to different posterior estimates, it is often desirable to report several posterior estimates using different priors, especially when the choice of the prior is fairly arbitrary, i.e., when there is no strong prior opinion.

Acknowledgment

The first author would like to thank Hugh Chipman for an interesting discussion on this subject as well as the the Natural Sciences and Engineering Research Council of Canada for providing partial research support. The authors are also grateful to the Editor, an Associate Editor and two referees for their suggestions, corrections and encouragement.

Appendix

This appendix contains a relatively simple proof of Theorem 1. Let X be a random variable with $B_{0,0}$ as its distribution; let $f(x)$ be its probability function. Clearly $f(x)$ is symmetric about $\frac{1}{2}$. From (8), we know that $\mathrm{Var}(X) = \frac{1}{4}$. Now suppose $f(x)$ is not just two point masses at 0 and 1. Then there exist $0 < \epsilon < \frac{1}{2}$ and $\delta > 0$ such that

$$g(\epsilon) \equiv \int_{\epsilon}^{1-\epsilon} f(x)dx > \delta.$$

Because $f(x)$ is symmetric about $\frac{1}{2}$, it follows that

$$\begin{aligned}
\mathrm{Var}(X) &= \int_0^1 \left(x - \frac{1}{2}\right)^2 f(x)dx \\
&= 2\int_0^{\epsilon} \left(x - \frac{1}{2}\right)^2 f(x)dx + \int_{\epsilon}^{1-\epsilon} \left(x - \frac{1}{2}\right)^2 f(x)dx \\
&\leq 2\left(0 - \frac{1}{2}\right)^2 \int_0^{\epsilon} f(x)dx + \left(\epsilon - \frac{1}{2}\right)^2 \int_{\epsilon}^{1-\epsilon} f(x)dx \\
&= \frac{1}{2}\left(\frac{1 - g(\epsilon)}{2}\right) + \left(\epsilon^2 - \epsilon + \frac{1}{4}\right) g(\epsilon) \\
&= \frac{1}{4} - \epsilon(1-\epsilon)g(\epsilon) \\
&< \frac{1}{4} - \epsilon(1-\epsilon)\delta.
\end{aligned}$$

Since $0 < \epsilon < \frac{1}{2} \implies \epsilon(1-\epsilon) > 0$ and $\delta > 0$, this is a contradiction. Therefore $f(x)$ must be just two point masses at 0 and 1. The symmetry about $\frac{1}{2}$ immediately implies that the two point masses are equal.

References

DeGroot, M. H. (1970), *Optimal Statistical Decisions*, New York: McGraw Hill.

Gelman, A. B., Carlin, J. S., Stern, H. S., and Rubin, D. B. (1995), *Bayesian Data Analysis*, London; New York: Chapman and Hall.

Rice, J. A. (1995), *Mathematical Statistics and Data Analysis 2nd ed.*, Belmont, CA: Duxbury Press.

Tanner, A. T. (1996), *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*, New York: Springer-Verlag.

Zellner, A. (1996), *An Introduction to Bayesian Inference in Econometrics*, New York; Chichester: John Wiley.

Mu Zhu
Department of Statistics and Actuarial Science
University of Waterloo
Waterloo, ON N2L 3G1
Canada
m3zhu@uwaterloo.ca

Arthur Y. Lu
Renaissance Technologies Corp.
600 Route 25A
East Setauket, NY 11733
alu@rentec.com
