Choosing a Prior Distribution

STA 732. Surya Tokdar

Thinking about a prior

A prior pdf on a parameter $\theta \in \Theta$ that indexes a statistical model $X \sim p(x|\theta)$, reflects the analyst's pre-data understanding/knowledge/belief about θ. When θ is an interpretable quantity, "belief about θ" is tangible. Example: θ = proportion of students in support of a university policy. If θ is merely an artifact of our formalization of X through the pdfs $p(x|\theta)$, then belief about θ really means belief about X consistent with the statistical model. Example: X_i = increase in sleep hours, $X_i \overset{\text{IID}}{\sim} N(\mu, \sigma^2)$.

Ideally, we would like to construct the prior pdf $\pi(\theta)$ to match an expert's belief about θ and/or X. However, belief is a mental condition, so one first need to *quantify* the expert's belief. Such a prior is usually called a subjective prior, as it is based upon an individual's *subjective* belief. A commonly used alternative is to go for a *default/non-informative/low-informative/uniform* prior that essentially reflects a lack of strong and precisely quantified prior information. Often such a prior is called an *objective* prior. We will look at both kinds and try to understand the rational behind them. (BTW, the use of the qualifiers subjective/objective is highly controversial and sometimes quite misleading.)

We shall make use of the following two examples:

1. X = number of female births out of n *placenta-previa* births. Model $X \sim Bin(n, p)$, p = female birth-rate under placenta previa (a certain pregnancy complication).

2. $Y_1, \cdots, Y_n$ = (self reported) weekly food expenditures of n Duke undergraduates. Model $Y_i \overset{\text{IID}}{\sim} N(\mu, \sigma^2)$.

Subjective prior: basic considerations

I based the following discussion on the excellent paper by **Garthwaite, Kadane and O'Hagan** (*JASA* 2005). To quantify an expert's belief about X and/or θ we must ask her questions whose answers will relate to some sorts of numerical summaries of these variables. Our task would then be to find a prior distribution $\pi(\theta)$ so that the same summaries now calculated based on the complete model $X \sim f(x|\theta), \theta \sim \pi(\theta)$ match the answers given by the expert. How do we make an expert quantify her beliefs?

A general strategy: the bisection approach

Suppose we want to quantify our beliefs about a scalar variable Z and then choose a pdf/pmf $f(z)$ that matches these quantified beliefs. There are various things we can quantify about Z, e.g., its center, spread, a range that is likely to contain most possible values, whether it is likely to be asymmetrically distributed around its center and so on. It is known, through experimentation, that we are fairly good at quantifying beliefs about "central values", but not so good at quantifying beliefs about spread or range. In particular, the question that we can most reliably answer is:

> *What is the number q_1 that we think Z is equally likely to be larger or smaller than?*

By "reliably answer" I mean that in answering this question, what we really believe and what we say we believe are usually close [psychologists have ways of figuring out what we *really believe*, or so they claim.]

Once we identify q_1, we must restrict our choice of $f(z)$ to pdfs that have q_1 as their median (i.e., 0.5-th quantile, i.e., $P(Z < q_1) = P(Z > q_1) = 0.5$ under these pdfs). Clearly, there are infinitely many pdfs that satisfy this. To make further progress, we need to answer more questions about our beliefs on Z. Now that the median has already been quantified, how can we talk about "centers" any more?

There is a fairly clever approach. We next ask this,

> *Imagine we are told $Z > q_1$ (recall q_1 is already identified). Given this information, what is the number q_2 that Z is equally likely to be larger or smaller than?*

Once we identify q_1, q_2, our choice of $f(z)$ must satisfy conditions: $P(Z < q_1) = 1/2$ and $P(Z < q_2) = 3/4$ under this pdf. So q_2 gives the 0.75-th quantile of $f(z)$.

We can repeat this on the left side, given the information $Z < q_1$ identify q_3 that Z is equally likely to be larger or smaller than. Then q_3 is the 0.25-th quantile $f(z)$. Continuing like these, we can identify the 0.875-th, the 0.125-th, the 0.9375-th, the 0.0625-th,... quantiles of $f(z)$.

Stopping

Of course we can't continue forever. Pretty soon we start answering "I don't know", "I really don't know", "leave me alone"... Wherever we stop, we'd still have a large collections of pdfs that will match the quantities we have identified as the desired quantiles. At this point, we usually choose the one (among the matching ones) that is convenient to work with[1].

More intelligently, we can start with a collection of convenient pdfs (like a conjugate family of prior pdfs for a Bayesian analysis) and keep quantifying $q_1, q_2, \cdots$ until a member of this family is uniquely determined as the only one that provides a match. If the collection of pdfs is indexed by k many unknown quantities, then we are likely to get a unique, exact match by the time we have quantified k quantiles.

[1] If we are more careful, we choose a few of such pdfs and perform our analysis under each, and then present all. If we are lucky the reports are close. Otherwise, we say there is too much prior uncertainty to come up with a singular analysis.

Belief quantification for placenta-previa births

Quantifying p

If we decide to use a $Be(a, b)$ as the prior for p, then we should be able to identify a and b by asking the expert about the median and either the first or the third quartile of θ (but in the language described above). Which quartile we go for may depend on whether the expert believes that placenta-previa might result in unusually low or high female birth rate.

Once we find these answers, say q_1 and q_2 from the expert, we look for a, b that best match these. This might require running a numerical routine as follows:

```
## get median q1 and 3-rd quartile q2 of p
fn <- function(par){
  a <- par[1]
  b <- par[2]
  return(sum((pbeta(c(q1, q2), a, b) - c(1/2, 3/4))^2))
}
optim(c(1, 1), fn)
```

For example, with $q_1 = 0.485$ and $q_2 = 0.6$, this code gives $a = 4.34$ and $b = 4.59$.

Quantifying X

The female birth-rate p can be thought of an intangible quantity that's merely an artifact of our binomial model. In this case we could try to make the expert answer questions about X directly. We can ask the expert to imagine $n = 1000$ and then quantify what number X is equally likely to be larger or smaller than, etc. Say we get the median q_1 and the 3-rd quartile q_2 through this. Now, when $X|p \sim Bin(n, p)$ and $p \sim Be(a, b)$, the marginal pmf of X is the beta-binomial pmf $g(x|a, b)$ given by

$$g(x|a, b) = \binom{n}{x} \frac{B(a + x, b + n - x)}{B(a, b)}, \quad x = 0, 1, \cdots, n.$$

To find a, b, we can modify the earlier code to:

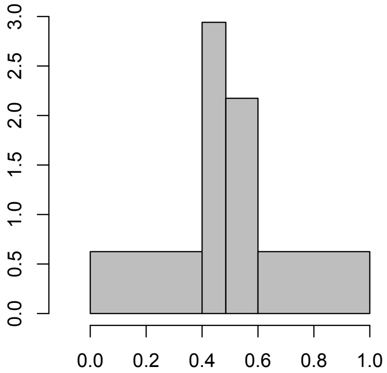
```
## get median q1 and 3-rd quartile q2 of X
fn2 <- function(par){
  a <- par[1]
  b <- par[2]
  return(sum((pbetabin.ab(c(q1, q2), n, a, b) - c(1/2, 3/4))^2))
}
optim(c(1, 1), fn2)
```

For $q_1 = 485$ and $q_2 = 600$, this code gives $a = 4.38$ and $b = 4.62$.

Other prior distributions

Let's go back to the case of quantifying p where the expert specifies the median to be 0.485 and the third quartile to be 0.6. Suppose she also specifies the first quartile to be 0.4. The first quartile of $Be(4.34, 4.59)$ is 0.371, which is not a bad match to the expert's answer, but might not be entirely satisfactory.

In this case matches could be found from several other useful collection of prior distributions. One example is the collection of prior pdfs that are piecewise uniform. The one below matches the expert's answers (uniquely if we restrict to only 4 pieces).



Belief quantification for food expenditure

We focus on the collection of prior pdfs $\{N\chi^{-2}(m, k, r, s^2) : -\infty < m < \infty, k > 0, r > 0, s^2 > 0\}$. Since this collection is indexed by 4 quantities, we need four pieces of quantification on $Y_1, Y_2, \cdots$. We will quantify three quantities for Y_1 and one quantity for $Y_1 - Y_2$. Under our model, the pdfs of these variables indeed depend on the choice of m, k, r, s. In particular:

$$\frac{Y_1 - m}{s\sqrt{1 + 1/k}} \sim t(r), \quad \text{and} \quad \frac{Y_1 - Y_2}{s\sqrt{2}} \sim t(r).$$

This is because of the following result.

RESULT 1. If $(W, V) \sim N\chi^{-2}(m, k, r, s^2)$ and $U|(W = w, V = v) \sim N(aw, bv)$ then $\frac{U-am}{s\sqrt{(b+a^2/k)}} \sim t(r)$.

Proof. Must have $U = aW + \sqrt{bV}Z$ for a $Z \sim N(0, 1)$ that is independent of W and V. So, given $V = v$, $U \sim N(am, a^2v/k + bv) = N(am, v/k')$ where $1/k' = b + a^2/k$. Hence $(U, V) \sim N\chi^{-2}(am, k', r, s^2)$, from which the result follows. □

We will start by quantifying the median, the 0.75-th and the 0.875-th quantiles q_1, q_2, q_3 for Y_1. This follows the bisection approach discussed above, but only on one side (we do not get 0.25-th quantile, etc.). This is because we are restricted only to pdfs of Y_1 that are symmetric around the median. We also apply the bisection method on $Y_1 - Y_2$ to quantify its 0.75-th quantile (the median must be quantified 0, by symmetry of Y_1 and Y_2).

Solving for m, k, r, s

Because $\frac{Y_1 - m}{s\sqrt{1+1/k}} \sim t(r)$, for any fraction $u \in (0,1)$ the u-th quantile of Y_1 must equal $m + s\sqrt{1+1/k}\ \Phi_r^{-1}(u)$ where $\Phi_r^{-1}(u)$ is the u-th quantile of the $t(r)$ pdf. First note that $\Phi_r^{-1}(0.5) = 0$ for any r. So

$$q_1 = m + s\sqrt{1+1/k}\ \Phi_r^{-1}(0.5) = m$$

and so $\boxed{m = q_1}$.

Next, in our old notations, $\Phi_r^{-1}(0.75) = \Phi_r^{-1}(1 - 0.5/2) = z_r(0.5)$ and similarly, $\Phi_r^{-1}(0.875) = z_r(0.25)$ and so

$$\begin{aligned} q_2 &= m + s\sqrt{1+1/k}\ z_r(0.5) \\ q_3 &= m + s\sqrt{1+1/k}\ z_r(0.25) \end{aligned}$$

and so $\boxed{\frac{z_r(0.5)}{z_r(0.25)} = \frac{q_2 - m}{q_3 - m} = \frac{q_2 - q_1}{q_3 - q_1}}$. The ratio $z_r(0.5)/z_r(0.25)$ is a continuous, increasing function in r and ranges between 0 (for $r \to 0$) and $z(0.5)/z(0.25) = 0.5863347$ [for $r \to \infty$, as $z_r(\alpha)$ becomes $z(\alpha)$]. See Figure 1. Therefore it is important that we have $\frac{q_2 - q_1}{q_3 - q_1}$ within this range. Otherwise, there is no $N\chi^{-2}(m, k, r, s^2)$ that matches our prior belief. In case of a mismatch we may revisit some of our answers about q_1, q_2 and q_3. The most suspect would be q_3 and a revised answer maybe considered for which a match occurs. If $\frac{q_2 - q_1}{q_3 - q_1}$ is inside the range $[0, 0.5863347]$ then we can identify r as follows.

```
ratio <- (q2 - q1) / (q3 - q1)
fn <- function(r) return(qt(0.75, r) / qt(0.875, r) - ratio)
r.sol <- uniroot(fn, interval = c(1e-3, 1e3))
r <- r.sol$root
```

Now that we have m and r, we can also identify $s' = s\sqrt{1+1/k}$ from the identity $q_2 = m + s\sqrt{1+1/k}\ z_r(0.5)$. Namely, $\boxed{s' = (q_2 - q_1)/z_r(0.5)}$. But we cannot disentangle s and k from this. In fact no amount of further quantification on Y_1 can identify s and k separately from s'.

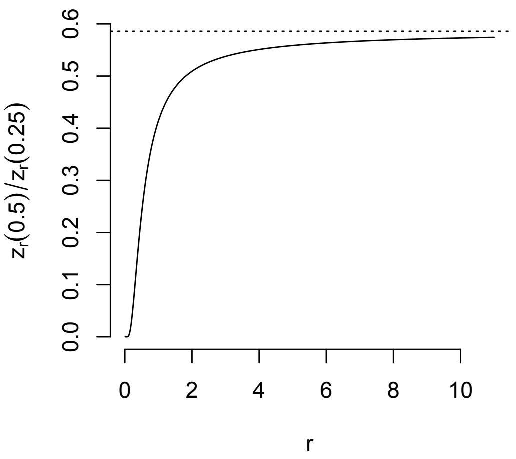


Figure 1: The ratio $\frac{z_r(0.5)}{z_r(0.25)}$ as a function of r.

So we now turn to $Y_1 - Y_2$ whose 0.75-th quantile must equal $0 + \sqrt{2} s z_r(0.5)$. Equating this to q_4, and using the value of r that we obtained before, we can now identify s by $\boxed{s = q_4/\{z_r(0.5)\sqrt{2}\}}$. Combine this with the identified value of $s' = s\sqrt{1 + 1/k}$ to identify k as: $\boxed{k = s^2/(s'^2 - s^2)}$. This is a legitimate value for k, provided $s' > s$. If we do not get this then again we need to see if we can revise our quantified beliefs.

Default, low-information priors

There's a general notion that with sufficient data, the likelihood function dominate the prior function, and so similar posterior distributions are obtained even when one starts with different choices of $\pi(\theta)$. This is not true in general, but certainly holds in the IID and related settings. In particular, in an IID setting, i.e., with data of the form $X = (X_1, \cdots, X_n)$ with $X_i \overset{\text{IID}}{\sim} g(x_i|\theta)$, $\theta \in \Theta$, if we have

$$\sqrt{n}(\hat{\theta}_{\text{MLE}}(X) - \theta) \xrightarrow{d} N(0, I_1^F(\theta)^{-1})$$

whenever $X \sim f(x|\theta) = \prod_{i=1}^n g(x_i|\theta)$, then for any prior $\pi(\theta)$, the posterior pdf $\pi(\theta|x)$ is well approximated by $N(\hat{\theta}_{\text{MLE}}(x), I_x^{-1})$ for almost all data $X = x$. Note that the approximating normal density does not involve the prior.

This property is often used to argue that a careful specification of the prior pdf through detailed elicitation (which is usually time consuming, especially if you have to convince the expert that there's some merit to this soul searching exercise) is perhaps unnecessary, as any two low-information prior distributions are likely to produce very similar posterior pdfs. Instead one can use an off-the shelf default, low information prior and report the corresponding posterior. Note that the argument here seems to be that the posterior will stand the sensitivity test if one was performed.

Uniform priors

Technically, a low information prior is one whose pdf $\pi(\theta)$ is flat compared to the likelihood function. The flattest of flats obtains by choosing $\pi(\theta) = \text{const}$, for all $\theta \in \Theta$. If Θ is bounded, then this is same as choosing π as the uniform distribution on Θ.

Interestingly, two earliest instances of Bayesian analyses, due to Bayes and Laplace, dealt with $X \sim Bin(n, p)$ and used $\pi(p) = Unif(0, 1)$. But they arrived at this prior from two different angles. Bayes reasoned that he believed $P(X = 0) = P(X = 1) = \cdots = P(X = n)$ (and hence each probability is $1/(n + 1)$). This is indeed the case when $\pi(p) = Unif(0, 1)$ [but not a unique choice, there are other pdfs for which this happens].

On the other hand, Laplace (unaware of Bayes' work) said he had no reasons to prefer any $p = p_1$ over any other $p = p_2$ as long as both $p_1, p_2 \in (0, 1)$. The only pdf $\pi(p)$ over $(0, 1)$ with $\pi(p_1)/\pi(p_2) = 1$ for all p_1, p_2 is the uniform pdf.

Both men used a "no preference" argument to come up with the prior, with Bayes applying this to X while Laplace doing the same to p. Laplace's reasoning, however, leaves a loophole [contributed and publicized by Fisher who declared Laplace's methods and in general Bayes methods to be a fundamental mistake in mathematics.] This is described below; the key idea is of *invariance*.

Invariance

Consider $X \sim p(x|\theta)$, $\theta \in [a, b]$ where we have chosen the low information, uniform prior $\pi(\theta) = 1/(b - a)$ on $[a, b]$. Now consider a re-parametrization, say $\eta = e^{\theta}$. We can rewrite our model as, $X \sim q(x|\eta)$, $\eta \in [c, d]$, where $q(x|\eta) = p(x|\log(\eta))$, $c = e^{a}$, $d = e^{b}$. The low-information uniform prior for this model is $\pi_q(\eta) = 1/(d - c)$.

Quite remarkably, by simply reparametrizing the model, we have arrived at a completely different prior! Choosing $\pi(\theta) = \frac{1}{b-a}$ on θ is equivalent to choosing $\pi_q(\eta) = \frac{1}{\eta(b-a)}$ on η by change of variable. This is not the uniform prior on $[c, d]$.

The Jeffreys prior

The first systematic solution to the invariance problem was proposed by Jeffreys and it remains popular to date. Harold Jeffreys (1946) proposed to choose the prior pdf as:

$$\pi_J(\theta) = \text{const} \times \{\det I^F(\theta)\}^{1/2}$$

where $I^F(\theta) = $ Fisher information: $= -\mathbb{E}_{[X|\theta]} \frac{\partial^2}{\partial \theta^2} \log p(X|\theta)$. If $X = (X_1, \cdots, X_n)$ with X_i being IID, then $I^F(\theta) = nI_1^F(\theta)$ and the Jeffreys prior can be expressed as: $\pi_J(\theta) = \text{const} \times \{I_1^F(\theta)\}^{1/2}$.

The Jeffreys prior is invariant under re-parametrization. The Jeffreys priors $\pi_{J,\Theta}(\theta)$ and $\pi_{J,\mathcal{E}}(\eta)$ for the models $X \sim p(x|\theta), \theta \in \Theta$ and $X \sim q(x|\eta), \eta \in \mathcal{E}$ where $q(x|\eta) = p(x|g(\eta))$ for some differentiable, bijection $g : \mathcal{E} \to \Theta$ are indeed related to each other as

$$\pi_{J,\mathcal{E}}(\eta) = \pi_{J,\Theta}(g(\eta))|\det g'(\eta)|$$

as they should be under simple change of variable. That is if you took the Jeffreys prior for a model and then applied change of variable according to a reparametrization, you'd get the Jeffreys prior for the reparametrized model.

Although Jefrreys' prior construction rule appears as a nifty trick to avoid the reparametrization problem, it has a deeper connection to uniform distributions over a parameter space. A statistical model $\{f(x|\theta) : \theta \in \Theta\}$ is best viewed as a manifold indexed by θ and a correct measure of distance between a θ and $\theta + d\theta$ is given by the Riemannian metric $\rho^2(\theta, \theta + d\theta) = d\theta^T I^F(\theta) d\theta$. This metric has a close relation with the Kullback-Leibler divergence, which is the most fundamental measure of difference between two pdfs for (IID) statistical inference. Recall that, the KL-divergence is defined as: $K(f, g) = \int f(x) \log\{f(x)/g(x)\}dx = \int f(x)\{\log f(x) - \log g(x)\}dx$. Now for a regular model,

$$\log f(x|\theta + d\theta) \approx \log f(x|\theta) + \left\{\frac{\partial}{\partial\theta} \log f(x|\theta)\right\}^T d\theta + \frac{1}{2} d\theta^T \left\{\frac{\partial^2}{\partial\theta^2} \log f(x|\theta)\right\} d\theta$$

and hence,

$$\begin{aligned} K(f(\cdot|\theta), f(\cdot|\theta + d\theta)) &\approx \int f(x|\theta)\{\log f(x|\theta) - \log f(x|\theta + d\theta)\}dx \\ &= -\left[\int f(x|\theta)\left\{\frac{\partial}{\partial\theta} \log f(x|\theta)\right\} dx\right]^T d\theta \\ &\qquad - d\theta^T \left[\int f(x|\theta)\left\{\frac{\partial^2}{\partial\theta^2} \log f(x|\theta)\right\} dx\right] d\theta \\ &= d\theta^T I^F(\theta) d\theta \end{aligned}$$

The Jeffreys prior is uniform over Θ with respect to the geometry of Θ induced by the metric ρ [which is usually different from the Euclidean geometry of Θ.]. To see why $\pi(\theta) = \text{const} \times \{\det I^F(\theta)\}^{1/2}$ is uniform in ρ notice that for a small $\Delta > 0$, the ρ-ball of radius Δ around θ corresponds to the ellipsoid $\{\tilde{\theta} : (\tilde{\theta} - \theta)^T I^F(\theta)(\tilde{\theta} - \theta) \le \Delta^2\}$ in the Euclidean metric, with volume $= V_p\{\det I^F(\theta)\}^{-1/2}\Delta^p$, if $\dim(\theta) = p$ where V_p is the volume of the unit sphere in p dimensions. And so the mass Jeffreys' prior assigns to this ρ-ball of radius Δ is const × $V_p\Delta^p$, no matter what θ is.

Improper priors

For unbounded Θ, a flat choice $\pi(\theta) = \text{const}$ or the Jeffreys' construct $\pi(\theta) = \text{const} \times \{I^F(\theta)\}^{1/2}$ need not give a function that is a pdf on Θ. These functions are non-negative, but may integrate to infinity. Some Bayesians accept such a $\pi(\theta)$, usually called an improper prior, as long as the posterior $\pi(\theta|x) = \text{const} \times f(x|\theta)\pi(\theta)$ is proper, i.e., a pdf on Θ.

Reference priors

The concept of low information was given a formal treatment in Bernardo (1979) and subsequently in a series of papers by Berger and Bernardo. A good reference is **Bergern, Bernardo and Sun** (*Ann. Statis.* 2009). First, you measure information gain from data $X = x$ by a divergence measure between the prior and the posterior, e.g., the Kullback-Leibler divergence $K(\pi, \pi(\cdot|x))$ of $\pi(\theta|x)$ from $\pi(\theta)$. The bigger the divergence the larger is the information gain. For a given prior $\pi(\theta)$, the expected information gain under the model is

$$\text{IG}(\pi) = \int K(\pi, \pi(\cdot|x)) f(x|\theta)\pi(\theta)d\theta$$

The reference prior $\pi_R(\theta)$ is the prior π that maximizes $\text{IG}(\pi)$ (usually in a certain asymptotic sense).

For one-dimensional θ, the reference prior is usually same as the Jeffreys prior. Important differences may arise when $\dim(\theta) > 1$. A catalog of low-information priors are available at **Berger and Yang** (*ISDS Tech. Rep.* 1998).

Default priors for binomial model

Clearly the flat prior is $\pi(p) = \mathit{Unif}(0, 1)$. The Jeffreys prior is $\pi(p) = \text{const} \times p^{-1/2}(1 - p)^{-1/2} = \mathit{Be}(1/2, 1/2)$, which is also the reference prior. If instead one uses an improper uniform prior on $\eta = \log \frac{p}{1-p}$ then the corresponding prior on p is $\pi(p) = \text{const} \times p^{-1}(1 - p)^{-1}$ which can be identified as "$\mathit{Be}(0, 0)$", with a slight abuse of notation, as the limiting case of $\mathit{Be}(a, b)$ with $a, b \to 0$.

The table below shows posterior summaries of p based on various choices of priors discussed so far. The actual data was $n = 980$ with $x = 437$.

Prior	Posterior median (95% interval)	$P(p < 0.5\|x)$	$P(p < 0.485\|x)$
$\mathit{Be}(4.34, 4.59)$	0.45 (0.42 , 0.48)	1	0.991
$\mathit{Unif}(0, 1)$	0.45 (0.42 , 0.48)	1	0.991
$\mathit{Be}(1/2, 1/2)$	0.45 (0.42 , 0.48)	1	0.991
$\mathit{Be}(0, 0)$	0.45 (0.42 , 0.48)	1	0.992

Default priors for linear Gaussian model

For the linear model $Y_i = z_i^T\beta + \epsilon_i$, $\epsilon_i \overset{\text{IID}}{\sim} N(0, \sigma^2)$ where $(\beta, \sigma^2) \in \mathbb{R}^p \times \mathbb{R}_+$, the Fisher information matrix in (β, σ^2) is $(n/2) \times \sigma^{-2}(Z^TZ)$. Consequently, the Jeffreys prior is $\pi_J(\beta, \sigma^2) = \text{const} \times \sigma^{-(p+2)}$. In particular, for our model $Y_1, \cdots, Y_n \overset{\text{IID}}{\sim} N(\mu, \sigma^2)$, the Jeffreys prior is $\pi_J(\mu, \sigma^2) = \text{const}/\sigma^3$. The reference prior, instead, is $\pi_R(\mu, \sigma^2) = \text{const}/\sigma^2$. Both are improper priors.

Many scholars suggest (including Jeffreys) that π_R is a "better" choice that π_J. This can be formalized through decision theory (Berger's book, Ch 6.6).

The posterior distribution of (β, σ^2) under the reference prior is $\pi_R(\beta, \sigma^2 | y, z) = N_p\chi^{-2}(\hat{\beta}_{\text{LS}}, Z^T Z, n-p, s^2_{y|z})$. Consequently, a $100(1-\alpha)\%$ posterior credible interval for $\eta = a^T\beta$ is given by

$$a^T\hat{\beta}_{\text{LS}} \mp s_{y|z}\sqrt{a^T(Z^T Z)^{-1}a} \times z_{n-p}(\alpha).$$

This is also the formula of the $100(1-\alpha)\%$ ML confidence interval for β. So the Bayes 95% credible interval also has a 95% confidence coefficient. Put differently, the 95% ML confidence interval also has a Bayes interpretation.

Some comments

Subjective priors

The qualifier "subjective" stems from the generally held notion that *knowledge is objective while belief is subjective.* We shouldn't however forget that knowledge is a belief that's unanimously agreed upon. Scientific knowledge is a belief or theory that couldn't be falsified despite best efforts. The very need of choosing a prior on Θ says that we are not universally agreed upon a single value from it. So it's perfectly natural that subjective beliefs about Θ will differ from one expert to another.

The onus is on the expert to convince others that her belief is justified in light of the current (pre-data) knowledge.

Arguably, there will be some aspects of the prior that would be more convincing than others. For these less precisely understood parts, other specifications can be considered and a sensitivity analysis can be performed. If the conclusion on the important features of the model/future data remain about the same with different specifications, then there is no need to make these parts more precise. If on the other hand, the conclusion on the important features is sensitive to these different specification, then there is a need to better understand these secondary aspects. This can be done with additional studies, more expert elicitation, etc. If this fails to resolve the issue, then we have to acknowledge that precise conclusions cannot be drawn based on current knowledge.

An issue with default prior finding rules

What we discussed earlier are not really prior pdfs but *rules* to find a prior pdf for a given statistical model. That the final outcome depends on the model could be somewhat unpleasant.

Take for example a study where one wants to study the proportion p of men in a population with a rare genetic condition. One could either survey n men and record in X the number of men who exhibit the condition. Alternatively, one could keep surveying until 10 men with the rare conditions are found, in this case data is $Y =$ the number of men surveyed until the 10th hit.

The Jeffreys prior for the first scenario is $\pi(p) = Be(1/2, 1/2)$ while for the second scenario is $\pi(p) = Be(0, 1/2)$. But certainly our understanding about the quantity p doesn't change based on how we decide to learn about it!

"An" analysis and not "the" analysis

Ultimately we should all remember that we only perform one possible analysis of the data under one set of assumptions and knowledge. Whether one such set will satisfy all is debatable. Depends on how varied current knowledge and opinion is about the subject matter. Knowledge can be shared to bring everyone on the same page. Opinions can be debated to rule out ones that are obviously insane. What remain are all potential sets of assumptions under which the analysis can be re-done. If the answers change from one set to another in an important way, then it is important to acknowledge that.

It is in this context that the qualifier "objective", often assigned to default, low-information priors, can be very misleading. An analysis done with a default prior is by no means more "objective" than other possible choice of the prior distribution. It is just one of many.