L2: Bayesian Theory—Priors, Part 1

September 29, 2020

Notation for prior, likelihood, and posterior

We will (usually) use the following notation (following Reich and Ghosh 2019):

- *Prior distribution* for θ: $\pi(\theta)$
- *Likelihood function of* $\mathbf{Y}$ *given* θ: $f(\mathbf{Y}|\theta)$
- *Marginal distribution* for data, $\mathbf{Y}$: $m(\mathbf{Y})$
- *Posterior distribution* for θ: $p(\theta|\mathbf{Y})$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)$ denotes a data vector.

Then Bayes Theorem can be written as follows:

$$p(\theta|\mathbf{y}) = \frac{\pi(\theta)f(\mathbf{Y}|\theta)}{m(\mathbf{Y})} \propto \pi(\theta)f(\mathbf{Y}|\theta) \tag{1}$$

Comments.

- Alternative terminology for the likelihood function of Y given θ, $f(y|\theta)$, is the *Sampling distribution*, or the *Data distribution*. These terms will be used interchangeably.
- The likelihood function sometimes written as $L(\theta|\mathbf{Y})$.
- Alternative terminology for the Marginal distribution is the *Normalising Constant*, as it is this factor that ensures that the Posterior distribution integrates to 1.

1 Overview

Prior distributions or simply priors, $\pi(\theta)$, are an important component of Bayesian statistics and have been the focus of considerable discussion and debate.

Priors are chosen by the data analyst—so how does one choose them?

As Gelman et al. (2017) say: "A key sticking point of Bayesian analysis is the choice of prior distribution ...".

1.1 What does the prior mean?

In brief the prior is a probability distribution for an unobserved quantity, in this case a parameter θ. It reflects our knowledge about θ prior to observing data.

- Remember that in the setting we are considering, the parameter itself is not random, but it is unknown, i.e., we are uncertain as to its value.

- The prior distribution is meant to be a measure of one's uncertainty about the parameter value.
- In a pure Bayesian sense, the prior distribution for a parameter reflects one's personal beliefs, they are *subjective* probabilities about the possible value of the parameter (see Lecture Note 1). Thus different people typically have different priors[1].
- Often, however, Bayesian inference is carried out because it is potentially easier than frequentist inference, or lends itself to more easily understood explanations. In this case one simply wants a prior that allows Bayesian inference without "unduly" influencing the posterior, namely, one such that the data dominate the posterior[2].

  Reich & Ghosh (p 59) go so far as to say "In the absence of prior information, selecting the prior can be viewed as a nuisance to be avoided if possible".
- So at one extreme a prior represents very specific beliefs and at the other extreme it is simply a necessary component of Bayesian inference.

1.2 Subjective vs Objective priors

One categorisation of priors is that of subjective priors versus objective priors. The line between the two is not necessarily sharply defined.

Subjective priors are the personal priors that have been selected by an individual. Two different heart surgeons, for example, may have individual and different prior probability distributions for the average number of hours required for a particular type of heart by-pass surgery. For example, one surgeon thinks the average is around 4 hours (with a range of 3.5 to 4.5 hours), while another surgeon thinks the average is around 4.5 hours (with a range of 4.3 to 4.7 hours, thus a narrower range: more certainty).

Objective priors are priors chosen according to some "rule" or procedure whereby two individuals carrying out the same rule will end up with the same prior distribution. Some examples of objective priors are Jeffreys prior, reference priors, maximum entropy priors, empirical Bayes priors, and penalised complexity priors (see Section 2.3 of Reich and Ghosh). We will later discuss Jeffreys priors.

1.3 "Non-informative" priors

Another categorisation of priors is so-called "non-informative" priors. Quotation marks are used here because the label is somewhat imprecise and nearly all priors are informative in the sense that the prior has *some* influence on the posterior.

With these caveats in mind: in the absence of expert knowledge, or in situations where one does not want to have the results depend too heavily upon the prior, one would like to use what is sometimes labelled a *non-informative* prior. What is loosely meant is that the posterior will not be *too* influenced by the prior and is *dominated* by the likelihood.

Here we consider two examples.

- The set of possible parameter values is finite: $\theta_1, \ldots, \theta_K$. A non-informative prior gives equal prior probability to each possible value, namely $\pi(\theta_i)=1/K$, $i=1, \ldots, K$.
- The parameter θ is continuous but with bounded support, e.g., (0,20), then a Uniform(0,20) prior would be consider non-informative.

[1] See page 2 of Gelman, et al 2017 where they refer to one extreme of choosing priors as a "maximalist" position where priors are chosen completely independently of the sample data, one knows nothing about the likelihood.

[2] Gelman, et al 2017 refer to this as a "minimalist" perspective on priors.

Synonyms for such priors include "vague", "diffuse", or "uninformative" priors[3].

1.4 Terminology: hyperparameters

The parameters of the prior distribution that are specified by the individual are called *hyperparameters*.

For example, if the sampling distribution is $y \sim \text{Binomial}(n, \phi)$ and the prior distribution is $\phi \sim \text{Beta}(2, 7)$, then 2 and 7 are the hyperparameters.

Another example, if the sampling distribution is $y \sim \text{Poisson}(\lambda)$ and the prior distribution is $\lambda \sim \text{Gamma}(20,5)$, then 20 and 5 are hyperparameters.

Hierarchical models. For more complex models, e.g., random effects models, the hyperparameters can appear at different locations in the overall model structure. For example, each spring, young salmon are released from a hatchery and the number of fish released in year t that survive one year is Binomial(n_t, ϕ_t) where n_t is the number released in year t and ϕ_t is the probability of surviving in year t. The survival probability ϕ_t varies between years is itself a random variable. A Beta distribution is to model the between year ("environmental") variation. Thus:

$$\begin{aligned} y_t|\phi_t &\sim \text{Binomial}(n_t, \phi_t) \\ \phi_t|\alpha, \beta &\sim \text{Beta}(\alpha, \beta) \end{aligned}$$

The unknown parameters are now α and β and prior distributions must be specified for them. Suppose that Gamma distributions are used as the priors for both:

$$\begin{aligned} \alpha &\sim \text{Gamma}(a_1, b_1) \\ \beta &\sim \text{Gamma}(a_2, b_2) \end{aligned}$$

In this case the hyperparameters are a_1, b_1, a_2, and b_2.

Note that this is a case with truly random parameters, namely, the ϕ_t's.

1.5 Terminology: improper priors

When the support of a parameter is finite, and one would like to use a non-informative prior, then the notion of using a flat or uniform prior over the support is attractive. This is in effect saying that $\pi(\theta) \propto c$, where c is some constant. For example, if θ can take on values from -10 to 10, then $\pi(\theta) = 1/20$.

In the case of infinite support, one cannot define a constant pdf (or pmf) as $\int_{-\infty}^{\infty} c \, d\theta$ is not integrable (the integral is not finite). Of course, what one can do in practice is choose lower and upper bounds that are "extreme" enough that one is fairly certain that θ will not be outside these bounds and then use a Uniform prior; e.g., bounds $= [-c, c]$, then $\theta \sim Uniform(-c, c)$ and $\pi(\theta) = 1/2c$.

However, sometimes the specified priors are not proper probability distributions in that the prior does not integrate to 1. For example, suppose the sampling model for data y is Normal(μ, 4^2). A "non-informative" prior for μ is to say that all values are equally probable, in other words the prior for μ is uniform or "flat" over the real number line, say $\pi(\mu) \propto 1$. However, such a "prior" is not a probability distribution, as it does not integrate to one. It is what is called an *improper prior*. Improper priors can be viewed the result of taking a Uniform$(-c, c)$ distribution and letting c go to ∞.

[3]Some statisticians dislike any of these labels due to the lack of precise meaning.

Improper priors are acceptable so long the resulting posterior is *not* improper. Referring to the above normal distribution example, n=1 and y_1=3, then the posterior for μ:

$$p(\mu|y=3) = \frac{\frac{1}{\sqrt{2\pi 4^2}} e^{-\frac{(3-\mu)^2}{2*4^2}} \times \mathbf{1}}{\int \frac{1}{\sqrt{2\pi 4^2}} e^{-\frac{(3-\mu)^2}{2*4^2}} \times \mathbf{1}\ d\mu} = \frac{1}{\sqrt{2\pi 4^2}} e^{-\frac{(\mu-3)^2}{2*4^2}}$$

or $\mu|y \sim$ Normal(3, 4^2) which integrates to 1 and is thus a *proper posterior*. There is nothing inherently wrong with using improper priors but one must check that the posteriors are proper.

1.6 Competing priors and their effect

Different individuals with different degrees of knowledge and experience of the underlying random process (the data distribution) are likely to have different priors. For example, an experienced heart surgeon who is assessing the probability of survival for a month following coronary by-pass surgery for 60 year old females, θ, may have a prior that is more concentrated over a range of values, e.g., most of the probability is between 60% and 80%, than a first year medical student, e.g., a prior with most of the probability spread relatively evenly between 10% and 90% (see left panel of Figure 1). The fundamental, and sometime controversial, issue is how much effect the choice of the prior has on the posterior. For example, after observing the survival or not of n=3 60 year old females who had by-pass surgery, where 1 survived, the posterior distribution for the experienced heart surgeon and the first year medical student might be much different. For example, suppose the surgeon's prior for θ was Beta(29.3, 12.6) and the medical student's was Beta(2.63, 2.63). The differences between their priors are shown in the left panel of Figure 1. After observing n=3 with y=1 surviving, thus 33% survival, the posterior distributions are Beta(30.3, 14.6) for the surgeon and Beta(3.63, 4.63) for the medical student. The right panel of the figure shows that the medical student's posterior was more influenced by the data than the surgeon. Whose posterior distribution are you more likely to believe? And why?

1.7 Mathematical perspectives

Mathematically it can be seen that the effect of the prior on the posterior depends upon the prior itself and the data distribution (likelihood):

$$p(\theta|y) \propto \pi(\theta) f(y|\theta) = \pi(\theta) f(y|\theta)$$

The influence of the data distribution is, in many cases, a function of sample size—typically, as n increases, the influence of the likelihood on the posterior increases.

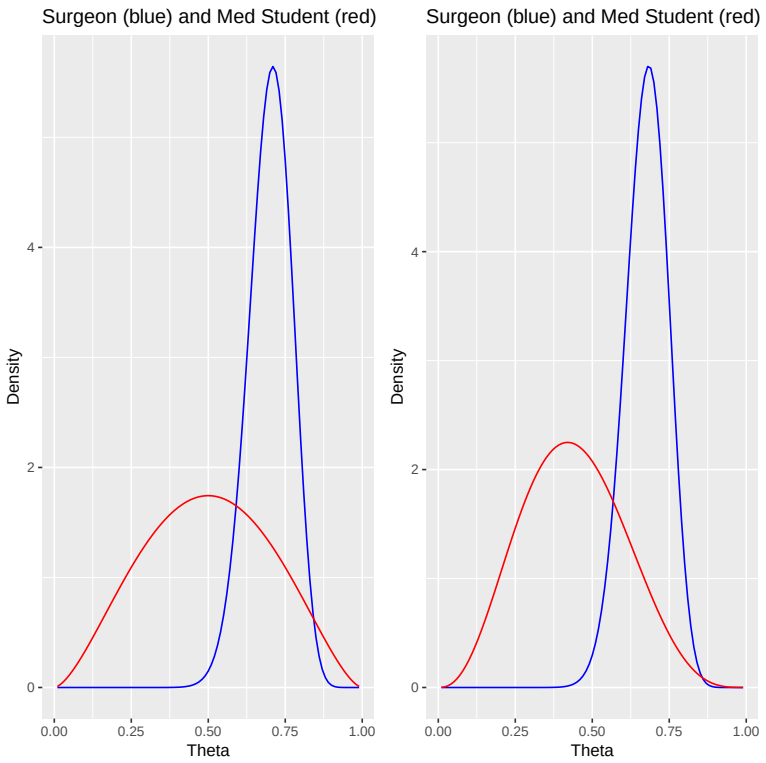
- *Graphical display.* In situations where there is just one parameter or relatively few parameters, visual examination of the prior and posterior distributions is a practical, and feasible, procedure for seeing the relative influence. Figure 2 shows the effects of different priors. As sample size n increases, the likelihood begins to *dominate* the prior in the sense that the posterior is very similar to the likelihood.
- *Implied sample size of a prior.* In some settings one can quantify the relative contribution of the prior and the likelihood to specific features of the posterior like the posterior mean.

  For a binomial parameter θ with Beta(α, β) prior and a sample size n with y successes, the posterior mean for θ can be written:

$$E[\theta|y] = \frac{\alpha + y}{\alpha + \beta + n} = \frac{\alpha + \beta}{\alpha + \beta + n} E[\theta] + \frac{n}{\alpha + \beta + n} \overline{y}$$

  Thus the prior mean is contributing a relative weight of $\alpha+\beta$ to the posterior mean. The sum $\alpha+\beta$ can be viewed as a sample size. Thus if $\alpha + \beta$ is large relative to n, the prior can dominate the posterior.

Figure 1: (Hypothetical) Prior (left panel) and posterior (right panel) distributions for the probability of by-pass surgery survival for experienced heart surgeon and first year medical student. The posterior is based on a sample of n=3 with y=1 survivors, thus the observed survival fraction is $1/3 = 0.33$.



For example, suppose n=10. If $\alpha+\beta$=2, the prior is like a sample of size 2, while if $\alpha+\beta$=200, then the prior has a much larger contribution to the posterior mean. Or another view, if $\alpha+\beta$=2, the relative contribution of the prior to the posterior mean is $2/12 = 0.17$, while if $\alpha+\beta$=200, it's 200/210=0.95.

1.8 Priors for multiple parameters

Given how difficult it can be to specify a prior for one parameter, one can imagine the difficulty with multiple parameters, as one needs to specify a joint prior distribution. The simplest setting is to specify independent marginal priors for each parameter, and then the joint prior is simply the product. For example with two parameters:

$$\pi(\theta_1, \theta_2) = \pi(\theta_1)\pi(\theta_2)$$

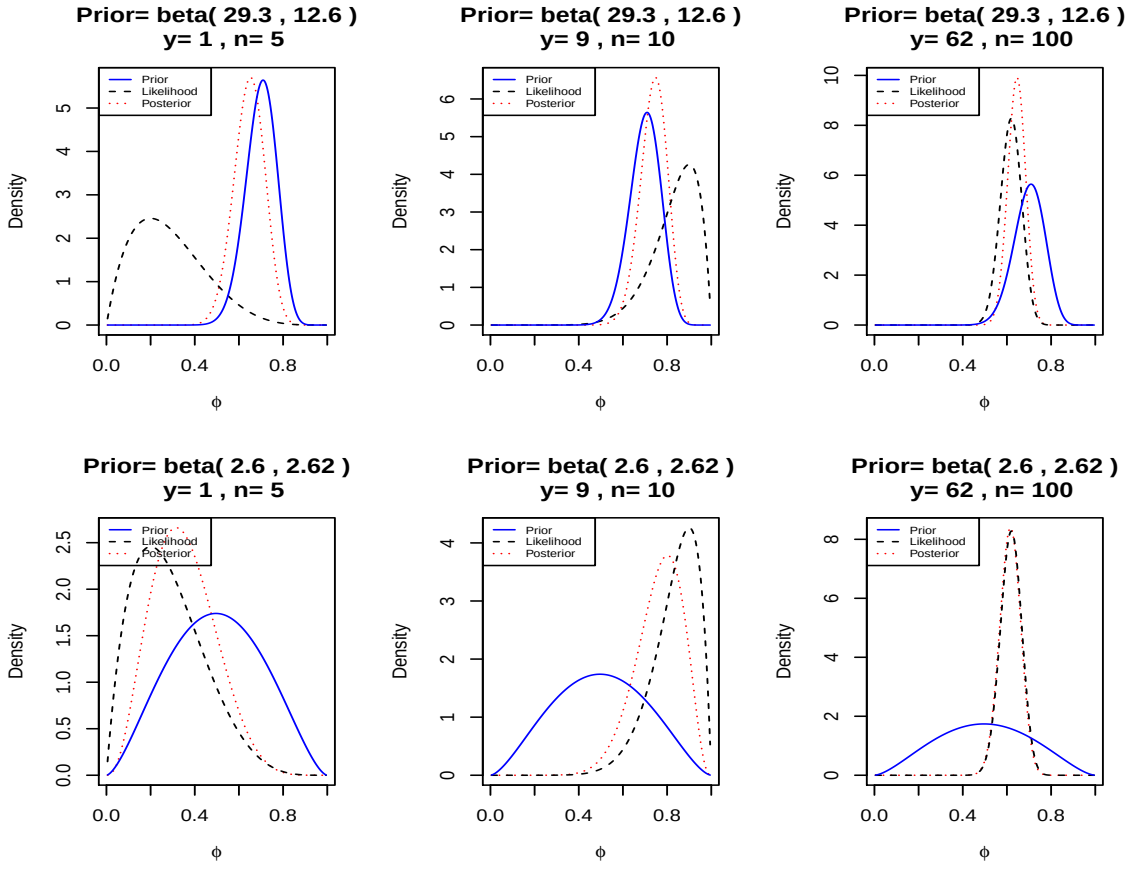
Another approach is to construct the joint prior is to begin with a marginal prior for one parameter, then define a conditional prior for a second parameter given the first, then another conditional prior for the third parameter given the first and second priors, and so on. For example,

$$\pi(\theta_1, \theta_2, \theta_3) = \pi(\theta_1)\pi(\theta_2|\theta_1)\pi(\theta_3|\theta_1, \theta_2)$$

With many parameters, this can be difficult and even with a few parameters, say 3 or 4, determining which parameter to start with, and the subsequent conditional priors can be challenging. Thus, the use of independent priors is common.

Note: independent priors do *Not* imply independent posteriors.

Figure 2: (Hypothetical) Changes in Prior (blue solid line) and Posterior (red dotted line) distributions for the probability of by-pass surgery survival for an experienced heart surgeon and first year medical student as sample size increases.



2 Conjugate Priors

As we saw in Lecture 1 (and the above heart by-pass surgery example), there are situations where the posterior distribution is the same as the prior distribution (other than updated hyperparameters in the posterior). Such priors are called *Conjugate Priors*.

Before computer intensive algorithms for computing posterior distributions, like Markov chain Monte Carlo, became available, the use of Conjugate Priors was more a technical convenience than a desired reflection of parameter uncertainty. While much more complicated posteriors can now be calculated, conjugate priors are worth knowing about and can be useful.

Note: conjugate distributions typically are subjective distributions in the sense that an individual chooses the hyperparameters for the prior distribution.

2.1 Examining the likelihood

One way of determining whether or not there is a probabiity distribution that could serve as a conjugate prior is to examine the kernel of the likelihood, to see if one can identify a known probability distribution for θ. Some examples:

$$\begin{aligned}
\text{Binomial pmf for } y \ &\propto \theta^y(1-\theta)^{n-y} \ \approx \ \theta^{\alpha-1}(1-\theta)^{\beta-1} : \text{ Beta pdf for } \theta \\
\text{Poisson pmf for } y &\propto \exp(-\theta)\theta^y \ \approx \ \exp(-\beta\theta)\theta^{\alpha-1} : \text{ Gamma pdf for } \theta \\
\text{Exponential pdf for } y &\propto \theta\exp(-\theta y) \ \approx \ \exp(-\beta\theta)\theta^{\alpha-1} : \text{ Gamma pdf for } \theta
\end{aligned}$$

Reich and Ghosh (2019, Section 2.1.5) and discuss the following general procedure for finding what they call "natural" conjugate priors.

- For a given data distribution $f(Y|\theta)$, imagine m distinct and fixed values denoted y_1^0,y_2^0, ..., y_m^0. They call these *pseudo-observations*.
- Postulate a prior distribution for θ with these m pseudo-observations as hyperparameters and that the prior will be proportional to the data distribution (likelihood):

$$\pi(\theta|y_1^0, y_2^0, \ldots, y_m^0, m) \propto \prod_{i=1}^{m} f(y_i^0|\theta) \tag{2}$$

- Given a sample of n independent and real (observed) values from $f(\mathbf{Y}|\theta)$, the posterior is then:

$$\begin{aligned}
p(\theta|Y) &\propto \pi(\theta|y_1^0, y_2^0, \ldots, y_m^0, m) \prod_{i=1}^{n} f(y_i|\theta) \\
&= \prod_{j=1}^{m+n} f(y_j^*|\theta)
\end{aligned}$$

  where for j=1,...,m, y_j^*=y_i^0, $i = 1, \ldots, m$ and for j=$m+1$,...,$m+n$, y_j^*=y_i, $i = 1, \ldots, n$.

- The posterior is the same form as the prior and thus the prior is conjugate

As an example consider the Poisson(θ) data distribution. The prior is constructed as follows.

$$\pi(\theta|y_1^0, \ldots, y_m^0, m) \propto \prod_{i=1}^{m} e^{-\theta}\theta^{y_i^0} = e^{-m\theta}\theta^{\sum_{i=1}^m y_i^0}$$

which is seen to be the kernel of a Gamma(s_0,m), where $s_0 = \sum_{i=1}^m y_i^0$.

Using the "natural conjugate" construction, the following conjugate priors for data distributions with a single unknown parameter result.

Sampling Dist'n $f(y\|\theta)$	Parameter	Prior Dist'n $\pi(\theta)$
Bernoulli(θ)	θ	Beta(α, β)
Poisson(θ)	θ	Gamma(α, β)
Exponental(θ)	θ	Gamma(α, β)
Normal(μ,σ_o^2)	μ (σ_o^2 known)	Normal(μ_h, σ_h^2)
Normal(μ_o,σ^2)	σ^2 (μ_o known)	Inverse Gamma(α, β)
Normal(μ_o,$\frac{1}{\tau}$)	τ (μ_o known)	Gamma(α, β)

In the last example, $\tau = 1/\sigma^2$ is the *precision* of a Normal distribution.

Note: key to this approach is that the kernel in Eq'n 2 is either something recognizable from a "known" distribution or, if not, it is something integrable that can easily be normalized to produce a proper probability distribution, i.e., it sums or integrates to one[4].

2.2 Technical Aside: Exponential family distributions

Definition A probability distribution which has a single parameter is in the exponential family if the pdf/pmf can be written as follows:

$$f(y \mid \theta) = h(y)g(\theta)\exp\left(\eta(\theta)\cdot T(y)\right) \tag{3}$$

This includes a large number of distributions, e.g., binomial, Poisson, exponential, Normal with known variance, Normal with known mean. For example, the binomial distribution can be written:

$$f(y|\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} = \binom{n}{y}(1-\theta)^n \exp\left(\ln\left(\frac{\theta}{1-\theta}\right)y\right)$$

where

$$h(y) = \binom{n}{y};\quad g(\theta) = (1-\theta)^n;\quad \eta(\theta) = \ln\left(\frac{\theta}{1-\theta}\right);\quad T(y) = y$$

Note that $\eta(\theta)$ is referred to as the *canonical parameterisation*.

For multiple parameters:

$$f(y \mid \boldsymbol{\theta}) = h(y)g(\boldsymbol{\theta})\exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})\cdot \mathbf{T}(y)\right) \tag{4}$$

One feature of exponential family distributions is that their support cannot depend upon the parameter θ. This rules out the Uniform(0,θ) distribution for example. One class of modelling techniques, generalized linear models (GLMs), is based on exponential family distributions. Another feature is that when expressed in the above form, the sufficient statistic, $T(y)$, is readily identified by the Factorization Theorem. There are many other important features. Here we bring these up primarily for one reason:

Conjugate priors. Exponential family distributions have conjugate priors.

$$\pi(\boldsymbol{\eta} \mid \boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu)g(\boldsymbol{\eta})^\nu \exp\left(\boldsymbol{\eta}^{\mathrm{T}}\boldsymbol{\chi}\right), \qquad \boldsymbol{\chi} \in \mathbb{R}^s \tag{5}$$

where s is the dimension of $\boldsymbol{\eta}$, and $\nu > 0$ and $\boldsymbol{\chi}$ are hyperparameters.

[4]It might be interesting to note that any non-negative integrable function $h(\theta)$ can be made a pdf. Given $h(\theta)$ where $h(\theta) \geq 0$ for all θ in "some" domain, where

$$\int h(\theta)d\theta = K < \infty$$

then

$$p(\theta) = \frac{1}{K}h(\theta)$$

is a pdf.

3 Mixture Priors

For the case of a single parameter θ, one can imagine situations where no single conjugate prior adequately describes one's prior knowledge. Mixture priors are one way of "hedging one's bets" so to speak. The simplest mixture prior is a mixture of two probability distributions (note: these need not be conjugate):

$$\pi(\theta) = q\pi_1(\theta) + (1-q)\pi_2(\theta)$$

where $0 \leq q \leq 1$.

More generally, a mixture of K distributions:

$$\pi(\theta) = \sum_{k=1}^{K} q_k \pi_k(\theta)$$

where $0 \leq q_k \leq 1$ and $\sum_{k=1}^{K} q_k = 1$.

The posterior will also be a mixture of what would be the individual posteriors, but the weightings will change, from q_k to Q_k.

$$p(\theta|\mathbf{Y}) = \sum_{k=1}^{K} Q_k p_k(\theta|\mathbf{Y})$$

The weightings, Q_k, will involve the marginal distributions from the individual priors and the marginal distribution across all priors.

For example, consider the case of a mixture of two distributions. Let

$$m_i(\mathbf{Y}) = \int \pi_i(\theta) f(\mathbf{Y}|\theta) d\theta, \ i = 1, 2$$

and note that the marginal for the mixture is as follows:

$$m(\mathbf{Y}) = \int \pi(\theta) f(\mathbf{Y}|\theta) d\theta = \int \left[q_1\pi_1(\theta) + q_2\pi_2(\theta)\right] f(\mathbf{Y}|\theta) d\theta$$

The posterior distribution:

$$\begin{aligned} p(\theta|\mathbf{Y}) &= \frac{p(\theta, \mathbf{Y})}{m(\mathbf{Y})} = \frac{\pi(\theta) f(\mathbf{Y}|\theta)}{m(\mathbf{Y})} = \frac{\left[q_1\pi_1(\theta) + q_2\pi_2(\theta)\right] f(\mathbf{Y}|\theta)}{m(\mathbf{Y})} \\ &= \frac{q_1\pi_1(\theta) f(y|\theta)}{m(\mathbf{Y})} + \frac{q_2\pi_2(\theta) f(y|\theta)}{m(\mathbf{Y})} \\ &= \frac{q_1 m_1(\mathbf{Y})}{m(\mathbf{Y})} \frac{\pi_1(\theta) f(\mathbf{Y}|\theta)}{m_1(\mathbf{Y})} + \frac{q_2 m_2(\mathbf{Y})}{m(\mathbf{Y})} \frac{\pi_2(\theta) f(\mathbf{Y}|\theta)}{m_2(\mathbf{Y})} \\ &= Q_1 p_1(\theta|\mathbf{Y}) + Q_2 p_2(\theta|\mathbf{Y}) \end{aligned}$$

where $Q_i = \frac{q_i m_i(\mathbf{Y})}{m(\mathbf{Y})}$.

Example. Returning to the heart surgeon and the medical student example, suppose that you thought that both of them had knowledge and opinions that were credible. Perhaps the heart surgeon was overly optimistic and perhaps the medical student had had training in relevant fields that the surgeon knew nothing about. You decide to give 60% weight (q_1=0.6) to the surgeon, and 40% (q_2=0.4) to the student. Then the resulting mixture prior is

$$\pi(\theta) = 0.6 * \text{Beta}(29.3, 12.557) + 0.4 * \text{Beta}(2.625, 2.625)$$

The denominator of the posterior:

$$\begin{aligned}\int \left[q_1\pi_1(\theta) + q_2\pi_2(\theta)\right] f(y|\theta)d\theta &= \binom{n}{y}\left[q_1 \int \frac{1}{B(a_1,b_1)}\theta^{a_1+y-1}(1-\theta)^{b_1+n-y-1}d\theta + \right. \\ &\left. q_2 \int \frac{1}{B(a_2,b_2)}\theta^{a_2+y-1}(1-\theta)^{b2+n-y-1}d\theta\right] \\ &= \binom{n}{y}\left[q_1\frac{B(a_1+y,b_1+n-y)}{B(a_1,b_1)} + q_2\frac{B(a_2+y,b_2+n-y)}{B(a_2,b_2)}\right]\end{aligned}$$

where

$$B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$

Then the posterior:

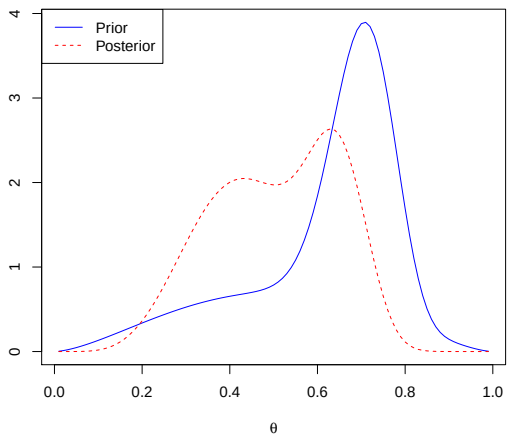
$$p(\theta|y) = Q_1 \text{ Beta}(a_1+y, b_1+n-y) + Q_2 \text{ Beta}(a_2+y, b_2+n-y)$$

where

$$Q_1 = \frac{q_1\frac{B(a_1+y,b_1+n-y)}{B(a_1,b_1)}}{\left[q_1\frac{B(a_1+y,b_1+n-y)}{B(a_1,b_1)} + q_2\frac{B(a_2+y,b_2+n-y)}{B(a_2,b_2)}\right]}$$

And after calculation, $Q_1 = 0.3442245$ and $Q_2 = 0.6557755$. Figure 3 shows the prior and the posterior when n=10 and y=4.

Figure 3: Mixture prior and posterior distributions for the probability of by-pass surgery survival for experienced heart surgeon and first year medical, where n=10,y=4.



4 References

- Hidden dangers of specifying noninformative priors. Seaman, Seaman, and Stamey. 2012. The American Statistician.
- The prior can generally only be understood in the context of the likelihood. Gelman, Simpson, and Betancourt. arXiv- 28 Aug 2017.
- Penalising model component complexity: A principled, practical approach to constructing priors. Simpson, Rue, Martins, Riebler, and Sorbye. 2015. Statistical Science. Describes a method for constructing penalised complexity priors that may become a commonly used approach.

A R code for some of the examples

A.1 Draw prior, likelihood, and posterior for the binomial with Beta prior

```
#---- modified triplot function from the LearnBayes package
triplot.mod <- function (prior, data, where = "topright",mytitle=NULL,mycex=0.8)
 {
  a <- prior[1]; b <- prior[2]
  s <- data[1];  f <- data[2]; n <- s+f
  p <- seq(0.005, 0.995, length = 500)
  prior <- dbeta(p, a, b)
  like  <- dbeta(p, s + 1, f + 1)
  post  <- dbeta(p, a + s, b + f)
  m     <-max(c(prior, like, post))

  if(is.null(mytitle)) {
     mytitle <- paste("Prior= beta(", a, ",", b, ") , y=",
                      s, ", n=", n)
  }
  plot(p, post, type = "l", ylab = "Density", lty = 3, lwd = 1,
       main = mytitle, ylim = c(0, m), col = "red",
       xlab=expression(phi))
  lines(p, like, lty = 2, lwd = 1, col = "black")
  lines(p, prior, lty = 1, lwd = 1, col = "blue")
  legend(where, c("Prior", "Likelihood", "Posterior"),
    lty = 1:3,lwd = c(1,1,1), col = c("blue","black","red"),cex=mycex)
}
```

The code to produce Figure 2:

```
# Comparing Surgeon and Medical Student priors and posteriors
theta.seq <- seq(0.01,0.99,by=0.01)
a1 <- 29.3; b1 <- 12.557
a2 <- 2.625; b2 <- 2.625
pi1 <- 0.6
n.vec <- c(5,10,100)
y.vec <- c(1,9,62)
par(mfrow=c(2,3),oma=c(0,0,3,0))
for(i in 1:3) {
 triplot.mod(prior=c(a1,b1),data=c(y.vec[i],n.vec[i]-y.vec[i]))
 }
 for(i in 1:3) {
 triplot.mod(prior=c(a2,b2),data=c(y.vec[i],n.vec[i]-y.vec[i]))
 }
 par(mfrow=c(1,1)
```

A.2 Binomial with mixture prior

The function `binomial.beta.mix` in the *LearnBayes* package will calculate the posterior weights Q_i's it can accommodate mixtures of 2, 3, 4, and more Beta distributions. The code below was used to produce Figure 3.

```
library(LearnBayes)
# Mixture Prior for Binomial Data distribution

p.vec <- c(0.6,0.4)

a.vec <- c(29.3,2.625)
b.vec <- c(12.557,2.625)

# Data output
n <- 10
y <- 4

#calculate posterior mixture distribution
posterior.out <- binomial.beta.mix(probs=p.vec,
        betapar=cbind(a.vec,b.vec),data=c(y,n-y))

#------
print(posterior.out$probs)
#  [1] 0.3442245 0.6557755  #these are the posterior weights

print(posterior.out$betapar)
# these are the 2 beta distribution parameters
#        [,1]   [,2]
# [1,] 33.300 18.557
# [2,]  6.625  8.625

# Plot the prior and posterior distributions
# The function d.mix.beta calculates the pdf
d.mix.beta <- function(theta,a.vec,b.vec,q.vec) {
  number.dists <- length(a.vec)
  out.density <- q.vec[1]*dbeta(theta,a.vec[1],b.vec[1])
  for(i in 2:number.dists)
    out.density <- out.density+q.vec[i]*dbeta(theta,a.vec[i],b.vec[i])
  return(out.density)
 }

theta.seq <- seq(0.01,0.99,by=0.01)
out.prior <- d.mix.beta(theta=theta.seq,a.vec=a.vec,b.vec=b.vec,q.vec=p.vec)
out.posterior <- d.mix.beta(theta=theta.seq,a.vec=posterior.out$betapar[,1],
            b.vec=posterior.out$betapar[,2],q.vec=posterior.out$probs)
my.ylim <- range(c(out.prior,out.posterior))
plot(theta.seq,out.prior,ylim=my.ylim,xlab=expression(theta),ylab="",
     col="blue",lty=1,type="l")
lines(theta.seq,out.posterior,col="red",lty=2)
legend("topleft",legend=c("Prior","Posterior"),col=c("blue","red"),
       lty=c(1,2))
```

September 29, 2020