

Probability and Statistics

Fall 2000

LECT. 10: BAYESIAN INFERENCE II, CB 7.2.3,9.2.4

We looked at an example where the model was a normal distribution for X given the unknown mean μ , and a normal prior distribution. The posterior distribution turned out to be normal as well. Suppose we use another distribution as the prior distribution. Say we use a Beta distribution with parameters 4 and 5 as the prior distribution. In that case the posterior is proportional to

$$f_{\mu|X}(\mu|x) \propto \exp\left(-\frac{1}{2}(x - \mu)^2\right) \cdot 1\{0 < \mu < 1\} \cdot \mu^3 \cdot (1 - \mu)^4.$$

What does this lead us to? In principle this is possible, but in practice this does not work very well. Although your prior beliefs can be anything you want, unless you choose your beliefs carefully, you are not going to get anywhere. Often it is possible to choose a conjugate prior distribution, such that the posterior distribution is in the same family as the prior distribution. We have already seen two examples of that. If your model for X is multinomial with known N and unknown probability p , and your prior distribution for p is Beta with parameters α and β , your posterior distribution for p is Beta with parameters $\alpha + X$ and $\beta + N - X$.

The second example was the normal distribution. Let us look at this case in more detail. Suppose that given μ the random variable X has a normal distribution with mean μ and known variance σ^2 . The prior distribution for μ is normal with mean μ_0 and variance τ^2 . The posterior distribution is proportional to:

$$\begin{aligned} f_{\mu|X}(\mu|x) &\propto \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right) \cdot \exp\left(-\frac{1}{2 \cdot \tau^2}(\mu - \mu_0)^2\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\frac{x^2}{\sigma^2} - \frac{2x\mu}{\sigma^2} + \frac{\mu^2}{\sigma^2} - \frac{2\mu\mu_0}{\tau^2} + \frac{\mu_0^2}{\tau^2}\right)\right) \\ &\propto \exp\left(-\frac{1}{2}\left(\mu^2 \frac{\sigma^2 + \tau^2}{\tau^2\sigma^2} - \mu \frac{2x\tau^2 + 2\mu_0\sigma^2}{\tau^2 + \sigma^2}\right)\right) \end{aligned}$$

$$\propto \exp - \frac{1}{2(1/(1/\tau^2 + 1/\sigma^2))} \left((\mu - (x/\sigma^2 + \mu_0/\tau^2)/(1/\sigma^2 + 1/\tau^2)) \right).$$

Hence the posterior distribution is normal with mean $(x/\sigma^2 + \mu_0/\tau^2)/(1/\sigma^2 + 1/\tau^2)$ and variance $(1/(1/\tau^2 + 1/\sigma^2))$. The calculations are messy but the result is quite intuitive: the posterior mean is a weighted average of the prior mean μ_0 and the observation x with weights adding up to one and proportional to the precision (defined as one over the variance), $1/\sigma^2$ for x and $1/\tau^2$ for μ_0 . The posterior precision is obtained by adding up the precision for each component. So, what you expect ex post, $E[\mu|X]$, that is, after seeing the data, is a weighted average of what you expected before, $E[\mu]$, and the data, X , with the weights determined by their respective variances.

There are a number of insights obtained by studying this example more carefully. Suppose we are really sure about the value of μ before we conduct the experiment. In that case we would set τ^2 small and the weight given to the observation would be small, and the posterior distribution would be close to the prior distribution. Suppose on the other hand we are very unsure about the value of μ . What value for τ should we choose? Obviously a large value, but what is the limit? We can in fact let τ go to infinity. Even though the prior distribution is not a proper distribution anymore if $\tau^2 = \infty$, the posterior distribution is perfectly well defined. In that case we have an improper prior distribution. We give equal prior weight to any value of μ . That would seem to capture pretty well the idea that a priori we are ignorant about μ . This is the idea of looking for an uninformative prior distribution. This is not always easy, and the subject of a large literature. For example, a flat prior distribution is not always uninformative about particular functions of parameters.

Now suppose we have two observations. Let us assume that conditional on μ these two observations are independent and identically distributed with a normal distribution with mean μ and variance σ^2 . We start with a prior distribution for μ that is normal with mean μ_0 and variance τ^2 . We can factor the conditional distribution of μ given X_1 and X_2 as

$$\begin{aligned} f(\mu|X_1, X_2) &= \frac{f(\mu, X_1|X_2)}{f(X_1|X_2)} = \frac{f(\mu, X_1, X_2)/f(X_2)}{f(X_1|X_2)} \\ &= \frac{f(\mu, X_1, X_2)}{f(X_1|X_2)f(X_2)} = \frac{f(X_1|\mu, X_2)f(X_2|\mu)f(\mu)}{f(X_1|X_2)f(X_2)}. \end{aligned}$$

By independence of X_1 and X_2 this is equal to

$$\frac{f(X_1|\mu)f(X_2|\mu)f(\mu)}{f(X_1, X_2)f(X_2)} \propto f(X_1|\mu)f(\mu)f(X_2|\mu).$$

The first two factors combined are proportional to $f(\mu|X_1)$. So, the total expression is the posterior of μ given X_2 with prior distribution for μ equal to $f(\mu|X_1)$. In other words, we can do this sequentially. First we calculate the prior distribution of μ given X_1 , taking $f(\mu)$ as the prior distribution. Then we take the resulting posterior distribution as the prior distribution and add the information in X_2 .

In the first step the prior is

$$\mu \sim \mathcal{N}(\mu_0, \tau^2).$$

Our previous calculations showed that the posterior distribution is

$$\mu|X_1 \sim \mathcal{N}(\tilde{\mu}_0, \tilde{\tau}^2) = \mathcal{N}\left(\frac{\mu_0/\tau^2 + x_1/\sigma^2}{1/\tau^2 + 1/\sigma^2}, \frac{1}{1/\tau^2 + 1/\sigma^2}\right).$$

Taking this as the prior distribution and X_2 as the observation we get for the posterior distribution of μ given x_1 and x_2 :

$$\mu|X_1, X_2 \sim \mathcal{N}\left(\frac{\tilde{\mu}_0/\tilde{\tau}^2 + x_2/\sigma^2}{1/\tilde{\tau}^2 + 1/\sigma^2}, \frac{1}{1/\tilde{\tau}^2 + 1/\sigma^2}\right) = \mathcal{N}\left(\frac{\mu_0/\tau^2 + (x_1 + x_2)/\sigma^2}{1/\tau^2 + 2/\sigma^2}, \frac{1}{1/\tau^2 + 2/\sigma^2}\right).$$

With N observations X_1, \dots, X_N we would get

$$\mu|X_1, \dots, X_N \sim \mathcal{N}\left(\frac{\mu_0/\tau^2 + \sum x_i/\sigma^2}{1/\tau^2 + N/\sigma^2}, \frac{1}{1/\tau^2 + N/\sigma^2}\right).$$

Note that when N is very large, the distribution of $\sqrt{N}(\mu - \bar{x})$ is approximately

$$\mathcal{N}(0, \sigma^2).$$

In other words, in large samples the influence of the prior distribution disappears, unless the prior distribution is chosen particularly badly, e.g., equal to zero on an important part

of the parameter space. This is true in general, i.e., for different models and different prior distributions.

Let us return to the Bernoulli example. Suppose that conditional on $P = p$, X_1, X_2, \dots, X_N are independent with Bernoulli distributions with probability p . Let the prior distribution of P be Beta with parameters α and β . Now consider the conditional distribution of P given X_1, \dots, X_N :

$$f_{P|X_1, \dots, X_N}(p|x) \propto p^{\alpha-1+\sum_{i=1}^N X_i} \cdot (1-p)^{\beta-1+N-\sum_{i=1}^N X_i},$$

which is a Beta distribution with parameters $\alpha - 1 + \sum_{i=1}^N X_i$ and $\beta - 1 + N - \sum_{i=1}^N X_i$. Let $Y_N = \sum_{i=1}^N X_i$. Then the mean and variance are

$$E[P|X_1, \dots, X_N] = \frac{\alpha + Y_N}{\alpha + \beta + N},$$

and

$$V(P) = \frac{(\alpha + Y_N)(\beta + N - Y_N)}{(\alpha + \beta + N)^2(\alpha + \beta + 1 + N)}.$$

What happens if n gets large? Let $\hat{p} = Y_N/N$ be the relative frequency of success. Then the mean and variance converge to

$$E[P|X_1, \dots, X_N] \approx \hat{p},$$

and

$$V(P) \approx 0.$$

As the sample size gets larger, the posterior distribution becomes concentrated at a value that does not depend on the prior distribution. This in fact can be taken a step further. In this example, the limiting distribution of $\sqrt{N} \cdot (P - \hat{p})$ conditional on the data, can be shown to be normal with mean zero and variance $\hat{p}(1 - \hat{p})$, again irrespective of the choice of α and β . The interpretation of this result is very important: in large sample the choice of prior

distribution is not very important in the sense that the information in the prior distribution gets dominated by the sample information. That is, unless your prior beliefs are so strong that they cannot be overturned by evidence (i.e., the prior distribution is zero over some important range of the parameter space), at some point the evidence in the data outweighs any prior beliefs you might have started out with.

Are we really restricted to choosing the prior distributions in these conjugate families? No. First of all the posterior distributions are well defined irrespective of conjugacy. It is merely a computational device. If you are outside the conjugate families, you typically have to resort to numerical methods for calculating posterior moments. Recently many methods have been developed that make this process much easier. This is especially important in cases where conjugate families do not exist.