Bayesian Estimation

Robert Jacobs
Department of Brain & Cognitive Sciences
University of Rochester
Rochester, NY 14627, USA

August 8, 2008

Bayesian estimation and maximum likelihood estimation make very different assumptions. Suppose that we are trying to estimate the value of some parameter, such as the population mean μ_X of some random variable labeled X. Maximum likelihood estimation assumes that this mean has a fixed value, albeit an unknown value. Because the value of the population mean μ_X is unknown, we collect a sample and then say that our estimate of μ_X is equal to the sample mean.

In contrast, Bayesian estimation does not assume that the population mean has a fixed value. Instead it assumes that this mean is itself a random variable with some probability distribution. For example, μ_X may be a random variable with a normal distribution. The mean and variance of this distribution may be unknown (though we may have some prior beliefs about their values) and, thus, we may collect a sample so as to estimate μ_X's mean and variance.

That is, suppose that the random variable X has a normal distribution:

$$X \sim N(\mu_X, \sigma_X^2). \tag{1}$$

We want to estimate the distribution of the population mean μ_X (for simplicity, assume that σ_X^2 is some known constant). Assume that this distribution is normal with known parameters:

$$\mu_X \sim N(\mu_0, \sigma_0^2) \tag{2}$$

where μ_0 and σ_0^2 are fixed constants that represent our prior beliefs. We collect a sample of the random variable X, labeled $\mathcal{X} = \{x^{(t)}\}_{t=1}^{T}$. According to Bayes' rule,

$$\begin{aligned}
\text{Posterior} &\propto \text{Prior} \times \text{Likelihood} && (3)\\
p(\mu_X|\mathcal{X}) &\propto p(\mu_X)\, l(\mu_X|\mathcal{X}) && (4)\\
p(\mu_X|\mathcal{X}) &\propto p(\mu_X)\, p(\mathcal{X}|\mu_X) && (5)
\end{aligned}$$

where $p(\mathcal{X}|\mu_X)$ is computed based on Equation (1), and $p(\mu_X)$ is computed based on Equation (2). Note that Bayesian estimation does not give us a point estimate of the parameter that we are studying (e.g., $\mu_X = 1/T \sum_{t=1}^{T} x^{(t)}$). Instead, it gives us the entire distribution of the parameter. In many cases, this is an important advantage of Bayesian estimation over maximum likelihood estimation.

An attractive feature of Bayesian estimation is that it can be applied sequentially (under the independence assumption given below). Above we noted that

$$p(\mu_X|\mathcal{X}) \propto p(\mu_X)\, p(\mathcal{X}|\mu_X). \tag{6}$$

Now suppose that a second set of data is collected, labeled $\mathcal{Y} = \{x^{(t)}\}_{t=1}^{T}$ (we assume that the second set of data $\mathcal{Y}$ is independent of the first set of data $\mathcal{X}$). Then

$$p(\mu_X|\mathcal{X}, \mathcal{Y}) \propto p(\mu_X)\, p(\mathcal{X}, \mathcal{Y}|\mu_X). \tag{7}$$

But independence implies that

$$p(\mathcal{X}, \mathcal{Y}|\mu_X) = p(\mathcal{X}|\mu_X)\, p(\mathcal{Y}|\mu_X) \tag{8}$$

and so we can re-write $p(\mu_X|\mathcal{X}, \mathcal{Y})$ as

$$\begin{aligned} p(\mu_X|\mathcal{X}, \mathcal{Y}) &\propto p(\mu_X)\, p(\mathcal{X}|\mu_X)\, p(\mathcal{Y}|\mu_X) & (9) \\ &\propto p(\mu_X|\mathcal{X})\, p(\mathcal{Y}|\mu_X). & (10) \end{aligned}$$

That is, the posterior distribution that we computed after seeing the first dataset $\mathcal{X}$ $[p(\mu_X|\mathcal{X})]$ is now used as the prior distribution for the new computations based on the new dataset $\mathcal{Y}$.

It may be useful to go through an example in order to illustrate some points regarding Bayesian estimation. Let's suppose that we have a data set $\mathcal{Y}$ consisting of n items $(y_1, \ldots, y_n)$. We assume that the data are independent and identically distributed according to a Normal distribution whose mean is denoted θ and whose variance is denoted σ^2:

$$y_i \sim N(\theta, \sigma^2). \tag{11}$$

Furthermore, before we've seen the data items, we believe that the mean parameter θ is also a random variable with a Normal distribution whose mean is denoted μ_0 and whose variance is denoted τ_0^2:

$$\theta \sim N(\mu_0, \tau_0^2). \tag{12}$$

Based on the data set, let's compute the posterior density of the mean parameter θ:

$$\begin{aligned} p(\theta|\mathcal{Y}) &\propto p(\theta)\, p(\mathcal{Y}|\theta) & (13) \\ &= p(\theta) \prod_{i=1}^{n} p(y_i|\theta) & (14) \\ &\propto \exp(-\frac{1}{2\tau_0^2}(\theta - \mu_0)^2) \prod_{i=1}^{n} \exp(-\frac{1}{2\sigma^2}(y_i - \theta)^2) & (15) \\ &\propto \exp(-\frac{1}{2}[\frac{1}{\tau_0^2}(\theta - \mu_0)^2 + \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \theta)^2]). & (16) \end{aligned}$$

Now we do some algebraic simplification. The above equations state that the posterior distribution for θ is a Normal distribution. We want an easy way to calculate the mean (denoted μ_n) of this Normal distribution. Let $\bar{y}$ denote the average of the n data items in $\mathcal{Y}$. Then it can be shown that the posterior mean, μ_n, is a linear weighted sum of the prior mean, μ_0, and the average of the data, $\bar{y}$:

$$\mu_n = w_0\mu_0 + w_n\bar{y} \tag{17}$$

where the linear coefficients are given by

$$w_0 = \frac{\frac{1}{\tau_0^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}} \tag{18}$$

$$w_n = \frac{\frac{n}{\sigma^2}}{\frac{1}{\tau_0^2} + \frac{n}{\sigma^2}}. \tag{19}$$

Note that when the prior variance τ_0^2 is large, or when the number of data items n is large, then w_n will tend to be large (i.e. we give a lot of weight to the data) and w_0 will be small (i.e. we discount the prior information). (Keep in mind here that w_0 and w_n are non-negative and sum to one.) On the other hand, when the prior variance is small, or when the number of data items is small, then w_0 will tend to be large (i.e. we give a lot of weight to the prior distribution) and w_n will be small (i.e. we discount the data). Intuitively, these calculations are sensible, right?

Note that in some people's minds Bayesian estimation is controversial. Suppose, for example, that we want to estimate the probability that when a coin is flipped, it lands heads side up. Our prior assumption is that the coin is fair, meaning that the probability of a heads is 0.5. We flip the coin ten times, and find that it lands heads side up on eight flips and tails side up on two flips. According to maximum likelihood estimation, we would estimate the probability of a head as 8/10 = 0.8. Bayesian estimation, however, would average the data (8/10 = 0.8) with the mean of the prior distribution (0.5) so it might estimate the probability of a head as, for example, equal to 0.65.

Is this an okay thing to do? That is, is it okay to discount the data in light of your prior theory? Some statisticians have said no. They claim that if the theory is not consistent with the data, then we must listen to the data and throw the theory away. Bayesians are accused of discounting the data and, thus, of being bad scientists who are wed to preconceived ideologies that they will not give up even if the data contradicts them. Bayesians defend themselves by pointing out that statisticians who advocate maximum likelihood estimation are "slaves" to their data.

One way of possibly resolving this controversy is to note that when there is a relatively small amount of data, the posterior distribution obtained from Bayes' rule is close to the prior distribution, whereas when there is a large amount of data, the posterior distribution is close to the likelihood distribution. So when there is a small amount of data, Bayesians stay close to their preconceived theories; when there is a large amount of data, Bayesians drop their theories in favor of the estimate based upon the actual data. That is, in the large data case, the Bayesian estimate and the maximum likelihood estimate tend to be very close.

For the sake of clarity, let's go through an example (taken from Gelman, Carlin, Stern, and Rubin, 1995). The following example is not typical of statistical applications of the Bayesian method, because it deals with a very small amount of data and concerns a single individual's state (gene carrier or not) rather than the estimation of a parameter that describes an entire population. Nonetheless, it is useful for our purposes.

Human males have one X-chromosome and one Y-chromosome, whereas females have two X-chromosomes, each chromosome being inherited from one parent. Hemophilia is a disease that exhibits X-chromosome-linked recessive inheritance, meaning that a male who inherits the gene that causes the disease on the X-chromosome is affected, whereas a female carrying the gene on only one of her two X-chromosomes is not affected.

The prior distribution: Consider a woman who has an affected brother, which implies that her mother must be a carrier of the hemophilia gene with one ‘good’ and one ‘bad’ hemophilia gene. We also know that her father is not affected; thus the woman herself has a fifty-fifty chance of having the gene. The unknown quantity of interest, the state of the woman, has just two values: the woman is either a carrier of the gene ($\theta = 1$) or not ($\theta = 0$). Based on the information provided thus far, the prior distribution for the unknown θ is $p(\theta = 1) = p(\theta = 0) = 0.5$.

The model and likelihood: Suppose that the woman has two sons, neither of whom is affected. Let $y_i = 1$ or 0 denote an affected or unaffected son, respectively. The outcomes of the two sons, conditional on the unknown θ, are independent (we are assuming that the sons are not identical twins). Then the likelihood function is

$$p(y_1 = 0, y_2 = 0|\theta = 1) \quad = \quad (0.5)\,(0.5) = 0.25 \tag{20}$$

$$p(y_1 = 0, y_2 = 0|\theta = 0) \quad = \quad (1.0)\,(1.0) = 1.00. \tag{21}$$

These expressions follow from the fact that if the woman is a carrier, then each of her sons will have a 50% chance of inheriting the gene, and so being affected, whereas if she is not a carrier then there is a probability of one that a son of hers will be unaffected (we are ignoring the possibility of mutations).

The posterior distribution: We now use Bayes’ rule to combine the information in the data with the prior probability. Using y to denote the joint data (y_1, y_2), we get

$$p(\theta = 1|y) \quad = \quad \frac{p(y|\theta = 1)\,p(\theta = 1)}{p(y|\theta = 1)\,p(\theta = 1)\;+\;p(y|\theta = 0)\,p(\theta = 0)} \tag{22}$$

$$= \quad \frac{(0.25)(0.5)}{0.25)(0.5)\;+\;(1.0)(0.5)} \tag{23}$$

$$= \quad \frac{0.125}{0.625} \tag{24}$$

$$= \quad 0.20. \tag{25}$$

Intuitively it is clear that if a woman has unaffected children, it is less probable that she is a carrier, and Bayes’ rule provides a formal mechanism for determining the extent of the correction.

Adding more data: Suppose that the woman has a third son, who is also unaffected. The entire calculation does not need to be re-done; rather we use the previous posterior distribution as the new prior distribution, to obtain:

$$p(\theta = 1|y_1, y_2, y_3) = \frac{(0.5)(0.20)}{(0.5)(0.20)\;+\;(1.0)(0.8)} = 0.111. \tag{26}$$

Alternatively, if we suppose that the third son is affected, it is easy to check that the posterior probability of the woman being a carrier becomes 1.

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