Bayesian inference

Fredrik Ronquist

October 5, 2005

1 Introduction

The last few decades has seen a growing interest in Bayesian inference, an alternative approach to statistical inference. There are many similarities between maximum likelihood and Bayesian inference but also important differences. Both methods use the same stochastic models of molecular evolution but the underlying inference principles are different, as we will explore in this lecture.

2 A probability exercise

It may be helpful to start with a simple exercise in probability theory. Assume we have 100 balls. The balls are either white or black and either small or large. Thus, there are four kinds of balls: (1) white and small; (2) white and large; (3) black and small; and (4) black and large. There are 10 balls of the first type, 15 of the second, 50 of the third and 25 of the fourth (Table 1).

Table 1: A collection of balls

Color	Small	Large	Total
White	10	15	25
Black	50	25	75
Total	60	40	100

Now, let us calculate the joint probability of a ball picked randomly from this set being small and white, $P(\text{small}, \text{white})$. If the probabilities of being small or large and white or black had been independent, we could have calculated the joint probability simply as the product of the probabilities of each of the two events, that is

$$P(\text{small}, \text{white}) = P(\text{small})P(\text{white})$$

Clearly, this is not the case for our collection of balls. An alternative approach we can use then, is to obtain the desired probability as the product of the probability of the ball being small times the probability of it being white given that it is small. In equation form we could write

$$P(\text{white}, \text{small}) = P(\text{small})P(\text{white}|\text{small})$$

If we plugged numbers into the equation, we would get $P(\text{white}, \text{small}) = P(\text{small})P(\text{white}|\text{small}) = (60/100)(10/60) = 10/100$.

The joint probability of the ball being small and white can also be calculated by multiplying the probability of the ball being white with the probability of it being small given that it is white. This equation would be

$$P(\text{white}, \text{small}) = P(\text{white})P(\text{small}|\text{white})$$

Plugging numbers into the equation, we would get $P(\text{white}, \text{small}) = P(\text{white})P(\text{small}|\text{white}) = (25/100)(10/25) = 10/100$, the same result as obtained previously.

Since both methods give the same joint probability, we can equate them, in which case we get

$$P(\text{small})P(\text{white}|\text{small}) = P(\text{white})P(\text{small}|\text{white})$$

After moving one factor over to the right side of the equation, we get

$$P(\text{white}|\text{small}) = \frac{P(\text{white})P(\text{small}|\text{white})}{P(\text{small})}$$

Replacing white and small with two general events labeled A and B, we get the general probability statement

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}$$

3 Bayes' theorem

The simple probability statement derived above is known as Bayes' rule or Bayes' theorem. When it is used in statistical inference, it is applied to data, D, and a set of parameter values θ of the model that generated the data. We can then write Bayes' theorem as

$$P(\theta|D) = \frac{P(\theta)P(D|\theta)}{P(D)}$$

Each of the probabilities in this equation has a name. The probability on the left is called the *posterior probability* of the parameter values given the data. The first probability in the numerator is the *prior probability* of the parameter values, while the second is easily recognized as the *likelihood*, or more formally, the probability of the data given the observations. Finally, the probability in the denominator is known as the *model likelihood*; it is the total probability of the data given the model (the conditioning on a model is implicit in all of the probabilities). We can view the latter probability simply as a normalizing constant, with the task of ensuring that the posterior probability is a true probability distribution, that is, that it integrates (or sums) to 1 over parameter space. Because of the normalizing function of the denominator, it is common to see Bayes' theorem expressed as

$$P(\theta|D) \propto P(\theta)P(D|\theta)$$

that is, the posterior probability is proportional to the prior probability times the likelihood.

Looking closer at the denominator of Bayes' theorem, $P(D)$, it is clear that it is the sum or integral of the posterior probabilities over all possible values of θ, that is

$$P(D) = \int_\theta P(\theta)P(D|\theta)\,\mathrm{d}\theta$$

4 Bayesian inference

How is Bayes' theorem used in Bayesian inference? As usual, we are considering a model with some parameters that are free to vary. The Bayesian approach forces us to first formulate some prior beliefs about the value of these parameters before seeing the data. These beliefs should take the form of a probability distribution on the parameters, $P(\theta)$. Once we have formulated our prior beliefs, we use the likelihood of the observed data to update the prior probability distribution to a posterior probability distribution specifying what we should believe about the parameters after seeing the data.

The posterior probability distribution essentially represents the end result of a Bayesian analysis; it is possible to summarize it in terms such as the mode (the parameter value with the *maximum posterior probability*) the mean or the 95 % *credibility interval* (note that Bayesian statisticians use the term credibility interval for an interval that is superficially similar to the confidence interval used in maximum likelihood inference) (Fig. 1). It is important to recognize these commonly used indicators as imperfect summaries of the posterior probability distribution. Unlike maximum likelihood analysis, the aim of a Bayesian analysis is not to provide so-called *point estimates* of the model parameters; the result of the analysis is the posterior probability distribution itself.

Figure 1: Some common summaries used for posterior probability distribution. They work well for simple posteriors (a) but can be misleading for more complex posteriors such as bimodal distributions (b).

Further differences between maximum likelihood and Bayesian inference are evident when analyzing multi-parameter models. Recall that the maximum likelihood approach distinguished between structural and nuisance parameters in multi-parameter models. There is no such difference in Bayesian analysis. The result of a Bayesian analysis is a joint probability distribution on all parameters in the model. If we are interested in one particular parameter, we would calculate the marginal probability distribution for that parameter by integrating (or summing) out all other parameters in the model. Assume for instance that we were interested in the parameter θ_i and that θ_{-i} is used to denote all other parameters in the model. Then we would calculate the marginal probability distribution

$$P(\theta_i|D) = \int_{\theta - i} P(\theta|D)\,\mathrm{d}\theta_{-i}$$

There is nothing that stops us from calculating several of these marginal distributions, one at a time, if we were interested in more than one parameter in the model. A common way of describing the Bayesian approach is to say that we integrate out the uncertainty concerning all other parameters when we are trying to draw inferences about a parameter of interest.

Note that the marginal distribution for a parameter in the model, the distribution of interest in Bayesian inference, is identical up to a normalizing constant to the integrated likelihood that would result if we were to integrate out all parameters in the likelihood function except the one we focused on. It follows that the maximum posterior probability estimate is the same as the maximum likelihood estimate when a pure integrated likelihood approach is used. Note, however, the reluctance of Bayesians to focus on a point estimate such as the maximum posterior probability value, which represents the mode in the marginal posterior probability distribution.

Bayesian inference shares many of the properties of maximum likelihood inference. For instance, Bayesian inference provides consistent and efficient parameter estimates under general conditions similar to those for likelihood inference given suitably chosen priors. In fact, the consistent use of integration instead of maximization over parameters that are not of interest can result in Bayesian inference being more robust than maximum likelihood.

5 An example of Bayesian inference

Coin tossing demonstrates several aspects of Bayesian inference well. Assume that we want to estimate the probability p of obtaining heads with a particular coin. We throw it 11 times and get five heads and six tails.

Using the binomial probability model, we can now either use maximum likelihood or Bayesian inference to estimate p. As we saw previously, the maximum likelihood estimate of p is $\hat{p} = 5/11 = 0.45$. This estimate is not necessarily the best estimate of p given all of the knowledge we have. For instance, if we were to bet over the true value of p, and given that there is nothing special about the coin, most people would bet their money on a value closer to 0.5 than to 0.45. This is because we use background knowledge; most coins are likely to have p values very close to 0.5 and the outcome of 11 throws simply is not enough to change that prior expectation considerably. When it is important to estimate the value of some unknown, and it is always important when money is involved, we tend to abandon maximum likelihood estimates and reason like Bayesians.

We can think of the Bayesian approach as combining a prior probability distribution with the likelihood function (Fig. 2). In the simplest case, our prior would put equal probability on all possible values of p, that is, our prior would be a uniform distribution on the interval $(0, 1)$. This would result in a posterior distribution that is simply a scaled variant of the likelihood function. However, a more realistic prior would put more probability on values of p close to 0.5. A suitable probability distribution to use in this case for formulating the prior would be the beta distribution. For instance, a Beta$(10, 10)$ distribution would be equivalent to incorporating information from about 20 (actually 18) previous coin tosses, half of which were heads. If we combine this prior with the likelihood function resulting from the binomial probability model on 5 heads and 6 tails, we would get a posterior that would be a Beta$(15, 16)$ distribution. Note how the parameters of the beta simply reflect the total number of heads and tails that we are basing the posterior on (actually the number of heads plus one and the number of tails plus one). If we used an even more informative prior, such as a Beta$(100, 100)$ distribution, the posterior would become more influenced, dominated, by the prior.

Figure 2: Coin tossing in a Bayesian context. By combining a given data set with different prior probability distributions, we get different posteriors (a-c). However, as we collect more and more data, the posteriors are influenced more and more by the likelihood and the influence of the prior gradually decreases (d-i)

As we increase the number of observations (throws of a particular coin) for a fixed prior, the more influenced the posterior will be by the likelihood. Eventually, the posterior will be determined entirely by the likelihood. For instance, if 10,000 throws still indicated that the coin had a heads probability of 0.45, then the posterior would spike on that value even for a $\text{Beta}(100, 100)$ prior distribution that put a lot of probability on fair coins ($p = 0.5$).

6 Accumulating scientific knowledge

One of the most attractive aspects of Bayesian theory is that it provides a rational and consistent framework for accumulating scientific knowledge. According to the Bayesian view, knowledge is accumulated by successively updating probability distributions to take new data into account. In fact, we can show that a successive series of Bayesian updates based on stepwise accumulation of evidence is mathematically equivalent to a single update based on all the evidence. Assume that we start with an initial prior $P(\theta)$ and an initial set of observations D_1. The posterior probability distribution after this step would be:

$$P(\theta|D_1) = \frac{P(\theta)P(D_1|\theta)}{P(D_1)}$$

If we use this posterior probability distribution as the prior in the next step, where we take a second set of observations D_2 into account, we get:

$$\begin{aligned} P(\theta|D_2, D_1) &= \frac{P(\theta|D_1)P(D_2|\theta)}{P(D_2)} \\ &= \frac{P(\theta)P(D_1|\theta)P(D_2|\theta)}{P(D_1)P(D_2)} \\ &= \frac{P(\theta)P(D_1, D_2|\theta)}{P(D_1, D_2)} \end{aligned}$$

The last rearrangement can be made because D_1 and D_2 are independent sets of observations. Note that the last equation is the same as applying Bayes' theorem directly on the collected set of observations, $D_1 + D_2$. In other words, there is a rational procedure for accumulating scientific knowledge in the Bayesian framework and the end result is the same regardless of the order or exact way in which the data are added as long as the posterior probability distribution obtained in one analysis is used as the prior in the next analysis. In maximum likelihood analyses, background knowledge is typically ignored.

7 Specifying priors

In Bayesian texts, you often see different terms used to describe important properties of priors. Prior probability distributions are typically proper probability distributions but sometimes it is possible to use *improper priors* that do not integrate to a bounded value. For instance, a uniform distribution on the interval $(-\infty, \infty)$ would be a typical example of an improper prior. Sometimes, it is possible to get away with improper priors because the likelihood function dies off quickly enough to make the posterior proper. However, it is safer to avoid improper priors.

In Bayesian texts that discuss analytical calculation of posteriors, you often find discussions of *conjugate priors*. Conjugate priors are those that, for some specified probability model, result in the posterior being a distribution of the same form as the prior. For instance, the beta distribution is a conjugate prior for the binomial probability model because the beta times the binomial is a beta distribution. In most cases, the posteriors are estimated numerically and then it is less important to be able to identify conjugate priors.

Many scientists find it difficult to specify proper priors. This can be because there is a lack of background information or because there are difficulties in summarizing the background information in terms of a proper probability distribution. In these cases, there is often a desire to formulate a "flat" or "noninformative" prior. In many cases, there is a natural probability distribution that can be used for this purpose. In the coin-tossing example, for instance, assuming a uniform prior probability distribution on the value of p is a natural uninformative or vague prior.

There is an entire school of *objective Bayesians* arguing that Bayesian inference should always be based on noninformative priors. It turns out to be difficult to define exactly what a noninformative prior is. The most successful attempt is that of Jeffreys, known as Jeffreys' prior. Jeffreys attempted to find a prior that would make the posterior insensitive to parametrization of the problem. This can be achieved by setting $P(\theta) \propto J[(\theta)]^{1/2}$, where $J(\theta)$ is the *Fisher information* for θ:

$$J(\theta) = -E\left[\left. \frac{\mathrm{d}^2 \ln P(D|\theta)}{\mathrm{d}\theta^2} \right| \theta \right]$$

In other words, the idea is to find a prior distribution where the probability depends on the curvature (second derivative) of the expected value of the likelihood function. The Jeffreys prior is easy to apply to single-parameter models. It can be extended to multiparameter models as well but the results are controversial.

8 Calculating the posterior

Although Bayes' theorem was formulated already in the 18th century, Bayesian inference has been little used until relatively recently. The main reason for this is the difficulty in calculating the posterior analytically. It was not until computers and good numerical methods for approximating the posterior became available that the method took off. Most important among the numerical techniques for estimating the posterior is the Markov chain Monte Carlo technique, known as MCMC. The idea is to construct a Markov chain with the stationary distribution equal to the posterior distribution of interest. This chain is then started at an arbitrarily chosen point in parameter space and run until it reaches stationarity, at which time samples of the posterior probability distribution can be collected. The next two lectures will describe the MCMC technique in detail, especially as it is used in Bayesian inference.

8.1 Study Questions

1. Describe the different probabilities in Bayes' rule
2. What are the most important differences between maximum likelihood inference and Bayesian inference?
3. What is a conjugate prior?
4. What is a noninformative prior? How can it be defined?
5. How is knowledge accumulated in the Bayesian approach?