Lecture Notes 17
Bayesian Inference

Relevant material is in Chapter 11.

1 Introduction

So far we have been using *frequentist (or classical) methods.* In the frequentist approach, probability is interpreted as long run frequencies. The goal of frequentist inference is to create procedures with long run guarantees. Indeed, a better name for frequentist inference might be *procedural inference.* Moreover, the guarantees should be uniform over θ if possible. For example, a confidence interval traps the true value of θ with probability $1-\alpha$, no matter what the true value of θ is. **In frequentist inference, procedures are random while parameters are fixed, unknown quantities.**

In the *Bayesian approach*, probability is regarded as a measure of **subjective degree of belief**. In this framework, everything, including parameters, is regarded as random. There are no long run frequency guarantees. Bayesian inference is quite controversial.

Note that when we used Bayes estimators in minimax theory, we were not doing Bayesian inference. We were simply using Bayesian estimators as a method to derive minimax estimators.

One very important point, which causes a lot of confusion, is this:

Using Bayes' Theorem $\neq$ Bayesian inference

The difference between Bayesian inference and frequentist inference is the **goal.**

Bayesian Goal: Quantify and analyze subjective degrees of belief.

Frequentist Goal: Create procedures that have frequency guarantees.

Neither method of inference is right or wrong. Which one you use depends on your goal. If your goal is to quantify and analyze your subjective degrees of belief, you should use Bayesian inference. If our goal create procedures that have frequency guarantees then you should use frequentist procedures.

Sometimes you can do both. That is, sometimes a Bayesian method will also have good frequentist properties. Sometimes it won't.

A summary of the main ideas is in Table 1.

	Bayesian	Frequentist
Probability	subjective degree of belief	limiting frequency
Goal	analyze beliefs	create procedures with frequency guarantees
θ	random variable	fixed
X	random variable	random variable
Use Bayes' theorem?	Yes. To update beliefs.	Yes, if it leads to procedure with good frequentist behavior. Otherwise no.

Table 1: Bayesian versus Frequentist Inference

To add to the confusion:

Bayes' nets: are directed graphs endowed with distributions. This has nothing to do with Bayesian inference.

Bayes' rule: is the optimal classification rule in a binary classification problem. This has nothing to do with Bayesian inference.

2 The Mechanics of Bayes

Let $X_1, \ldots, X_n \sim p(x|\theta)$. In Bayes we also include a prior $p(\theta)$. It follows from Bayes' theorem that the posterior distribution of θ given the data is

$$p(\theta|X_1, \ldots, X_n) = \frac{p(X_1, \ldots, X_n|\theta)p(\theta)}{p(X_1, \ldots, X_n)}$$

where

$$p(X_1, \ldots, X_n) = \int p(X_1, \ldots, X_n|\theta)p(\theta)d\theta.$$

Hence,

$$p(\theta|X_1, \ldots, X_n) \propto L(\theta)p(\theta)$$

where $L(\theta) = p(X_1, \ldots, X_n|\theta)$ is the likelihood function. The interpretation is that $p(\theta|X_1, \ldots, X_n)$ represents your subjective beliefs about θ after observing $X_1, \ldots, X_n$.

A commonly used point estimator is the posterior mean

$$\overline{\theta} = \mathbb{E}(\theta|X_1, \ldots, X_n) = \int \theta p(\theta|X_1, \ldots, X_n)d\theta = \frac{\int \theta L(\theta)p(\theta)}{\int L(\theta)p(\theta)}.$$

For interval estimation we use $C = (a, b)$ where a and b are chosen so that

$$\int_a^b p(\theta|X_1, \ldots, X_n) = 1 - \alpha.$$

This interpretation is that

$$P(\theta \in C|X_1,\ldots,X_n) = 1-\alpha.$$

This does **not** mean that C traps θ with probability $1-\alpha$. We will discuss the distinction in detail later.

Example 1 *Let $X_1,\ldots,X_n \sim \text{Bernoulli}(p)$. Let the prior be $p \sim \text{Beta}(\alpha,\beta)$. Hence*

$$p(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$$

and

$$\Gamma(\alpha) = \int_0^\infty t^{\alpha-1}e^{-t}dt.$$

Set $Y = \sum_i X_i$. Then

$$p(p|X) \propto \underbrace{p^Y 1-p^{n-Y}}_{\text{likelihood}} \times \underbrace{p^{\alpha-1} 1-p^{\beta-1}}_{\text{prior}} \propto p^{Y+\alpha-1} 1-p^{n-Y+\beta-1}.$$

Therefore, $p|X \sim \text{Beta}(Y+\alpha, n-Y+\beta)$. (See page 325 for more details.) The Bayes estimator is

$$\widetilde{p} = \frac{Y+\alpha}{(Y+\alpha)+(n-Y+\beta)} = \frac{Y+\alpha}{\alpha+\beta+n} = (1-\lambda)\widehat{p}_{mle} + \lambda\,\overline{p}$$

where

$$\overline{p} = \frac{\alpha}{\alpha+\beta}, \quad \lambda = \frac{\alpha+\beta}{\alpha+\beta+n}.$$

This is an example of a conjugate prior.

Example 2 *Let $X_1,\ldots,X_n \sim N(\mu,\sigma^2)$ with σ^2 known. Let $\mu \sim N(m,\tau^2)$. Then*

$$\mathbb{E}(\mu|X) = \frac{\tau^2}{\tau^2+\frac{\sigma^2}{n}}\overline{X} + \frac{\frac{\sigma^2}{n}}{\tau^2+\frac{\sigma^2}{n}}m$$

and

$$\text{Var}(\mu|X) = \frac{\sigma^2\tau^2/n}{\tau^2+\frac{\sigma^2}{n}}.$$

Example 3 *Suppose that $X_1,\ldots,X_n \sim \text{Bernoulli}(p_1)$ and that $Y_1,\ldots,Y_m \sim \text{Bernoulli}(p_2)$. We are interested in $\delta = p_2 - p_1$. Let us use then prior $p(p_1,p_2) = 1$. The posterior for p_1, p_2 is*

$$p(p_1,p_2|\text{Data}) \propto p_1^X(1-p_1)^{n-X}p_2^Y(1-p_2)^{m-Y} \propto g(p_1)h(p_2)$$

where $X = \sum_i X_i$, $Y = \sum_i Y_i$, g is a $\text{Beta}(X+1,n-X+1)$ density and h is a $\text{Beta}(Y+1,m-Y+1)$ density. To get the posterior for δ we need to do a change variables: $(p_1,p_2,) \to (\delta,p_2)$ to get $p(\delta,p_2|\text{Data})$. Then we integrate:

$$p(\delta|\text{Data}) = \int p(\delta,p_2|\text{Data})d\,p_2.$$

(An easier approach is to use simulation.)

3 Where Does the Prior Come From?

This is the million dollar question. In principle, the Bayesian is supposed to choose a prior π that represents their prior information. This will be challenging in high dimensional cases to say the least. Also, critics will say that someone's prior opinions should not be included in a data analysis because this is not scientific.

There has been some effort to define "noninformative priors" but this has not worked out so well. An example is the *Jeffreys prior* which is defined to be

$$p(\theta) \propto \sqrt{I(\theta)}.$$

You can use a flat prior but be aware that this prior doesn't retain its flatness under transformations. In high dimensional cases, the prior ends up being highly influential. The result is that Bayesian methods tend to have poor frequentist behavior. We'll return to this point soon.

It is common to use flat priors even if they don't integrate to 1. This is possible since the posterior might still integrate to 1 even if the prior doesn't.

4 Large Sample Theory

There is a Bayesian central limit theorem. In nice models, with large n,

$$p(\theta|X_1, \ldots, X_n) \approx N\left(\widehat{\theta}, \frac{1}{I_n(\widehat{\theta})}\right) \tag{1}$$

where $\widehat{\theta}_n$ is the mle and I is the Fisher information. In these cases, the $1-\alpha$ Bayesian intervals will be approximately the same as the frequentist confidence intervals. That is, an approximate $1-\alpha$ posterior interval is

$$C = \widehat{\theta} \pm \frac{z_{\alpha/2}}{\sqrt{I_n(\widehat{\theta})}}$$

which is the Wald confidence interval. However, this is only true if n is large and the dimension of the model is fixed.

Let's summarize this point: **In low dimensional models, with lots of data and assuming the usual regularity conditions, Bayesian posterior intervals will also be frequentist confidence intervals. In this case, there is little difference between the two.**

Here is a rough derivation of (1). Note that

$$\log p(\theta|X_1, \ldots, X_n) = \sum_{i=1}^{n} \log p(X_i|\theta) + \log p(\theta) - \log C$$

where C is the normalizing constant. Now the sum has n terms which grows with sample size. The last two terms are O(1). So the sum dominates, that is,

$$\log p(\theta|X_1,\ldots,X_n) \approx \sum_{i=1}^n \log p(X_i|\theta) = \ell(\theta).$$

Next, we note that

$$\ell(\theta) \approx \ell(\widehat{\theta}) + (\theta-\widehat{\theta})\ell'(\widehat{\theta}) + \frac{(\theta-\widehat{\theta})^2\ell''(\widehat{\theta})}{2}.$$

Now $\ell'(\widehat{\theta}) = 0$ so

$$\ell(\theta) \approx \ell(\widehat{\theta}) + \frac{(\theta-\widehat{\theta})^2\ell''(\widehat{\theta})}{2}.$$

Thus, approximately,

$$p(\theta|X_1,\ldots,X_n) \propto \exp\left(-\frac{(\theta-\widehat{\theta})^2}{2\sigma^2}\right)$$

where

$$\sigma^2 = -\frac{1}{\ell''(\widehat{\theta})}.$$

Let $\ell_i = \log p(X_i|\widehat{\theta}_0)$ where θ_0 is the true value. Since $\widehat{\theta} \approx \theta_0$,

$$\ell''(\widehat{\theta}) \approx \ell''(\theta_0) = \sum_i \ell_i'' = n\frac{1}{n}\sum_i \ell_i'' \approx -nI_1(\theta_0) \approx -nI_1(\widehat{\theta}) = -I_n(\widehat{\theta})$$

and therefore, $\sigma^2 \approx 1/I_n(\widehat{\theta})$.

5 Bayes Versus Frequentist

In general, Bayesian and frequentist inferences can be quite different. If C is a $1-\alpha$ Bayesian interval then

$$P(\theta \in C|X_1,\ldots,X_n) = 1-\alpha.$$

This does **not imply** that

$$\text{frequentist coverage} = \inf_\theta P_\theta(\theta \in C) = 1-\alpha..$$

Typically, a $1-\alpha$ Bayesian interval has coverage lower than $1-\alpha$. Suppose you wake up everyday and produce a Bayesian 95 percent interval for some parameter. (A different parameter everyday.) The fraction of times your interval contains the true parameter will not be 95 percent. Here are some examples to make this clear.

Example 4 Normal means. *Let $X_i \sim N(\mu_i, 1)$, $i = 1, \ldots, n$. Suppose we use the flat prior $p(\mu_1, \ldots, \mu_n) \propto 1$. Then, with $\mu = (\mu_1, \ldots, \mu_n)$, the posterior for μ is multivariate Normal with mean $X = (X_1, \ldots, X_n)$ and covariance matrix equal to the identity matrix. Let $\theta = \sum_{i=1}^n \mu_i^2$. Let $C_n = [c_n, \infty)$ where c_n is chosen so that $\mathbb{P}(\theta \in C_n | X_1, \ldots, X_n) = .95$. How often, in the frequentist sense, does C_n trap θ? Stein (1959) showed that*

$$\mathbb{P}_\mu(\theta \in C_n) \to 0, \quad \text{as } n \to \infty.$$

Thus, $\mathbb{P}_\mu(\theta \in C_n) \approx 0$ even though $\mathbb{P}(\theta \in C_n | X_1, \ldots, X_n) = .95$.

Example 5 Sampling to a Foregone Conclusion. *Let $X_1, X_2, \ldots \sim N(\theta, 1)$. Suppose we continue sampling until $T > k$ where $T = \sqrt{n}|\overline{X}_n|$ and k is a fixed number, say, $k = 20$. The sample size N is now a random variable. It can be shown that $\mathbb{P}(N < \infty) = 1$. It can also be shown that the posterior $p(\theta | X_1, \ldots, X_N)$ is the same as if N had been fixed in advance. That is, the randomness in N does not affect the posterior. Now if the prior $p(\theta)$ is smooth then the posterior is approximately $\theta | X_1, \ldots, X_N \sim N(\overline{X}_n, 1/n)$. Hence, if $C_n = \overline{X}_n \pm 1.96/\sqrt{n}$ then $\mathbb{P}(\theta \in C_n | X_1, \ldots, X_N) \approx .95$. Notice that 0 is never in C_n since, when we stop sampling, $T > 20$, and therefore*

$$\overline{X}_n - \frac{1.96}{\sqrt{n}} > \frac{20}{\sqrt{n}} - \frac{1.96}{\sqrt{n}} > 0. \tag{2}$$

Hence, when $\theta = 0$, $\mathbb{P}_\theta(\theta \in C_n) = 0$. Thus, the coverage is

$$\text{Coverage} = \inf_\theta \mathbb{P}_\theta(\theta \in C_n) = 0.$$

This is called sampling to a foregone conclusion and is a real issue in sequential clinical trials.

Example 6 *Let $\mathcal{C} = \{c_1, \ldots, c_N\}$ be a finite set of constants. For simplicity, assume that $c_j \in \{0, 1\}$ (although this is not important). Let $\theta = N^{-1} \sum_{j=1}^N c_j$. Suppose we want to estimate θ. We proceed as follows. Let $S_1, \ldots, S_n \sim$ Bernoulli(π) where π is known. If $S_i = 1$ you get to see c_i. Otherwise, you do not. (This is an example of survey sampling.) The likelihood function is*

$$\prod_i \pi^{S_i} (1 - \pi)^{1 - S_i}.$$

The unknown parameter does not appear in the likelihood. In fact, there are no unknown parameters in the likelihood! The likelihood function contains no information at all. The posterior is the same as the prior.

But we can estimate θ. Let

$$\widehat{\theta} = \frac{1}{N\pi} \sum_{j=1}^N c_j S_j.$$

Then $\mathbb{E}(\widehat{\theta}) = \theta$. *Hoeffding's inequality implies that*

$$\mathbb{P}(|\widehat{\theta} - \theta| > \epsilon) \leq 2e^{-2n\epsilon^2\pi^2}.$$

Hence, $\widehat{\theta}$ *is close to* θ *with high probability. In particular, a* $1 - \alpha$ *confidence interval is* $\widehat{\theta} \pm \sqrt{\log(2/\alpha)/(2n\pi^2)}$.

6 Bayesian Computing

If $\theta = (\theta_1, \ldots, \theta_p)$ is a vector then the posterior $p(\theta|X_1, \ldots, X_n)$ is a multivariate distribution. If you are interested in one parameter, θ_1 for example, then you need to find the marginal posterior:

$$p(\theta_1|X_1, \ldots, X_n) = \int p(\theta_1, \ldots, \theta_p|X_1, \ldots, X_n) d\theta_2 \ \cdots \ d\theta_p.$$

Usually, this integral is intractable. In practice, we resort to Monte Carlo methods. These are discussed in 36/10-702.

7 Conclusion

Bayesian and frequentist inference are answering two different questions.

Frequentist inference answers the question: How do I construct a procedure that has frequency guarantees?

Bayesian inference answers the question: How do I update my subjective beliefs after I observe some data?

In parametric models, if n is large and the dimension of the model is fixed, Bayes and frequentist procedures will be similar. Otherwise, they can be quite different.