Stat 200: Introduction to Statistical Inference **Autumn 2018/19**

Lecture 6: Bayesian estimation

Lecturer: Art B. Owen *October 11*

Abstract

These notes are mnemonics about what was covered in class. They don't replace being present or reading the book. Reading ahead in the book is very effective.

6.1 Rao-Blackwell

Before starting on Bayes we briefly covered the Rao-Blackwell theorem. If $\hat\theta$ is an estimator of θ with $\mathbb{E}(\hat\theta^2) < \infty$ for all θ and T is a sufficient statistic for θ, then $\tilde\theta = \mathbb{E}(\hat\theta \mid T)$ has $\mathbb{E}((\tilde\theta-\theta)^2) \leqslant \mathbb{E}((\hat\theta-\theta)^2)$. The proof in Rice uses formulas for iterated expectation and the familiar $\mathrm{Var}(Y) = \mathrm{Var}(\mathbb{E}(Y \mid X)) + \mathbb{E}(\mathrm{Var}(Y \mid X))$.

Nowhere does that proof use the fact that T is sufficient. A more mathematical book, by Casella and Berger (Statistical Inference, 1990), states the reason. Conditioning on T is sure to give a function of the data and hence it is a statistic. You might otherwise get $\tilde\theta$ that could only be computed by using the known θ. [And if you had that, you would not have needed any data, much less a statistic.]

6.2 Bayes estimates

Bayes estimates are based on Bayes theorem that we use to reverse the direction of a conditional probability. Suppose that $A_1, \ldots, A_n$ are mutually exclusive and exhaustive events, we know $\Pr(A_i)$, we know $\Pr(B \mid A_i)$ for some event B but we want $\Pr(A_i \mid B)$. Think of n distinct diseases $i = 1, \ldots, n$ and one symptom B. Or B could be a whole collection of symptoms. Then using basic rules of probabilty

$$\Pr(A_i \mid B) = \cdots = \frac{\Pr(B \mid A_i)\Pr(A_i)}{\sum_{j=1}^n \Pr(B \mid A_j)\Pr(A_j)}.$$

Similarly for two continuous random variables

$$f_{Y|X}(y \mid x) = \cdots = \frac{f_{X|Y}(x \mid y) f_Y(y)}{\int f_{X|Y}(x \mid y) f_Y(y)\,\mathrm{d}y}.$$

Get used to seeing a potentially awkward sum or integral in the denominator.

In the Bayesian framework we suppose that the parameter θ of interest is actually the observed value of a random variable Θ. [Other places in the notes use Θ for the set of legal θ values but here it is a random variable.] The idea is that some mechanism, we call it "nature" has somehow sampled $\Theta = \theta$ from a

distribution $f_\Theta(\theta)$. Then we observe $X = x$ from $X \sim f_{X|\Theta}(x \mid \theta)$. We know X. What we want is the unobserved Θ. That now plays the role of Y above and we get

$$f_{\Theta|X}(\theta \mid x) = \frac{f_{X|\Theta}(x \mid \theta) f_\Theta(\theta)}{\int f_{X|\Theta}(x \mid \theta) f_\Theta(\theta)\,\mathrm{d}\theta}.$$

In Bayesian estimation we have two main tasks. The first is picking a good model f_Θ. That can raise thorny philosophical issues about which people can reasonably disagree. [Or unreasonably.]

The second is computing some potentially nasty integrals. Sometimes the integrals can be done in closed form. Sometimes they can be done using symbolic math. Sometimes numerical quadrature, sometimes Monte Carlo, sometimes Markov chain Monte Carlo (MCMC), sometimes approximate MCMC, and sometimes nobody can do it at all. In this course we will work from the simple end of the problem, either leaving the result as an integral, or looking at cases with easy integrals, or using some basic numerical method in the homework. This course is not about numerics.

We call f_Θ the ***prior distribution*** of Θ and $f_{\Theta|X}$ the ***posterior distribution*** of Θ. The prior is what we know (or believe) before seeing X and the posterior is our updated knowledge or belief after seeing X. We can write

$$f_{\Theta|X}(\theta \mid x) \propto f_{X|\Theta}(x \mid \theta) \times f_\Theta(\theta), \quad \text{i.e.,}$$
$$\text{posterior} \propto \text{prior} \times \text{likelihood}.$$

Many books write expressions like

$$f(\theta \mid x) \propto f(x \mid \theta) \times f(\theta), \quad \text{or}$$
$$p(\theta \mid x) \propto p(x \mid \theta) \times p(\theta),$$

dropping the function identifying subscripts. In that notation $p(x)$ and $p(\theta)$ are not the same function $p(\cdot)$. Think of them as living things that look inside their own parentheses before deciding what function to be. The class seemed about equally split on whether to use subscripts or not, and to break the tie, the textbook uses them and so will we.

6.3 Examples

Let θ be the speed of light (meters per second in a vacuum). Suppose that $X \sim N(\theta, \sigma^2)$ where σ^2 is a known value describing the accuracy of our experiment. Then take a prior distribution $\Theta \sim U[A, B]$ for $A = 299 \times 10^8$, and $B = 301 \times 10^8$. Now we get a posterior distribution

$$f_{\Theta|X}(\theta \mid x) = \frac{\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(x-\theta)^2/\sigma^2} \times 1_{A \leqslant \theta \leqslant B}}{\int_A^B \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(x-\theta)^2/\sigma^2}\,\mathrm{d}\theta}.$$

The prior here might represent somebody's rough idea of the truth before getting any data. Notice that here the prior belief has $A \leqslant \theta \leqslant B$ with absolute certainty. No amount of data can overturn that certainty. This is a good thing if what we are certain about is correct. A more cautious approach is to use a heavier tailed prior just in case.

Now let Θ be the probability that a ride-hailing driver gets 5 points from the customer. Let $X \sim \mathrm{Bin}(n, \theta)$ be the observe number of 5s in n rides (IID Bernoulli trials). Perhaps our prior belief is well described by

$U[1/2, 1]$. Alternatively, if we have by now a good estimate $\hat{\theta}_i$ for $N = 10{,}000$ other drivers similar to the one we are interested in, then we can use a discrete prior distribution

$$p_\Theta(\theta) = \frac{1}{N}\sum_{i=1}^{N} 1_{\theta=\hat{\theta}_i}.$$

Suppose for instance that out of those N historical drivers exactly 30 of them hat $\hat{\theta} = 0.93$. Then

$$p_\Theta(0.93) = \frac{1}{N}\sum_{i=1}^{N} 1_{0.93=\hat{\theta}_i} = \frac{30}{10{,}000} = 0.003.$$

6.4 Conjugate distributions

Suppose that $X \sim \mathrm{Bin}(n, \theta)$ and the prior distribution for Θ is $\mathrm{Beta}(\alpha, \beta)$. In class we did examples like this, multiplying prior by likelihood. We can ignore any constant factors. Constant means they don't depend on θ so for our purposes a function of x is a constant. We get a posterior

$$f_{\Theta|X}(\theta \mid x) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}\binom{n}{x}\theta^x(1-\theta)^{n-x}$$

and so we have

$$f_{\Theta|X}(\theta \mid x) \propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1}.$$

The $\mathrm{Beta}(\alpha + x, \beta + n - x)$ PDF is proportional to this. There can be only one PDF proportional to that. Therefore the posterior distribution actually is $\mathrm{Beta}(\alpha + x, \beta + n - x)$.

This posterior distribution is in the same family as the prior distribution. That is an enormous convenience. We say that the Beta distribution is the ***conjugate prior*** to the Binomial distribution. Conjugacy is an important special case where the numerical problems in Bayes are easy to handle.

In class we had the example of boy and girl births in Laplace's time in 18th century France. Also Laplace's example of whether we get another sunrise in the next 24 hours, where the posterior is $\mathrm{Beta}(N + 1, 1)$ if you start with a $U[0, 1] = \mathrm{Beta}(1, 1)$ prior and get N events in a row. That is called ***Laplace's rule of succession***. It's useful for generating nonzero probability estimates of p when you get $X = 0$ for $X \sim \mathrm{Bin}(n, p)$. Somebody writing a language processing system might have needed a model for whether the next word they saw would be "XKCD".

Now that the posterior is $\mathrm{Beta}(\alpha + x, \beta + n - x)$ we can interpret α and β. Every time we get new data with x successes in n trials, we add x to the first parameter and $n - x$. So we can think of α and β as roughly, the number of successes and failures that had been observed before any data were gathered at all. Note that the Beta distribution allows fractional α and β so they might not be actual counts.

If the prior and likelihood are both of the form $\exp(-A\theta^2 - B\theta - C)$ for numbers A, B and C that might depend on x, then so is the posterior. This means that a Normal prior and normal likelihood give a normal posterior. See several cases worked out in Rice Ch 8.

6.5 Flat priors

You could take a flat prior like $U[0, 1]$ or $U[-10^6, 10^6]$ to model not knowing anything about Θ. That is mildly problematic becuase Θ^3 for example gets a non-uniform distribution so suddenly you know something

about Θ^3 without knowing about Θ. People may choose a flat prior on whichever function of Θ is most directly connected to their interests.

There is no uniform distribution over $[0, \infty)$ or $\mathbb{R}$ or $\{0, 1, 2, \dots\}$. A flat prior over any of those sets does not exist as a distribution. Sometimes we use an ***improper prior*** proportional to some positive constant, such as 1, over those sets. This is also called a ***noninformative prior***. The posterior distribution is then

$$f_{\Theta|X}(\theta \mid x) \propto 1 \times f_\Theta(\theta), \quad \text{i.e.,}$$
$$\text{posterior}(\theta) \propto \text{likelihood}(\theta).$$

The posterior distribution is then proper if and only if the likelihood has a finite integral over θ.

6.6 Bayesian estimates

If we have the whole posterior distribution then we have several ways to pick an estimate $\hat{\theta}$. It could be the mean, median or mode of the posterior distribution. The posterior mean minimizes the posterior mean squared error

$$\mathbb{E}_{\Theta|X}((\hat{\theta} - \Theta)^2 \mid x).$$

The posterior median minimizes the posterior mean absolute error

$$\mathbb{E}_{\Theta|X}(|\hat{\theta} - \Theta| \mid x).$$

The posterior mode maximizes the posterior probability of a correctly guessed Θ (or a guess within some small error $\mathrm{d}x$).