Lecture 18:
Bayesian estimation

Relevant textbook passages:
Larsen–Marx [10]: Sections 5.3, 5.8, 5.9, 6.2

18.1 ★ Exponential families

In the last lecture I mentioned that the MLE $T$ of $\theta$ has a variance that achieves the Cramér–Rao lower bound if it is unbiased has a likelihood of the form

$$f_T(t; \theta) = e^{a(\theta)t + b(\theta)}.$$  

(When $\theta$ is an $m$-vector, then $t$ is too, and $a$ is a $a(\theta)t$ is a dot product.) A family of densities $f(x; \theta)$ of the form

$$f(x; \theta) = a(\theta)b(x) \exp \left[ \sum_{i=1}^{k} g_i(\theta)h_i(x) \right]$$

is called an exponential family [2, p. 161]. Let me elaborate on this condition, as it is actually key to the concept of a sufficient statistic.

The sufficient statistics that we have seen so far are typically the sample mean, or the sample variance. The pair $(\bar{x}, S^2)$ for the normal case is a good example. The key point here is that even as the sample size gets arbitrarily large, the sufficient statistic remains 2-dimensional. This is one reason that it is a useful sufficient statistic. It can be shown, see, e.g., Pitman [13], that having an exponential family is necessary to have a sufficient statistic of fixed dimension, subject to some smoothness conditions on the likelihood function. If I get the time, I will write up an argument for this.

The other point I forgot to mention is that the quantity

$$I = E_{\theta} \left[ \left( \frac{\partial}{\partial \theta} \log f(X; \theta) \right)^2 \right]$$

that appears in the denominator of the Cramér–Rao lower bound has an interesting interpretation. The log-likelihood function $\mathcal{L}(\theta; x) = \ln L(\theta; x)$ regarded as a

---

1 This Pitman is not your textbook author, but rather his father.
function of the random variable $X$ is a random variable, and so is its derivative (with respect to $\theta$) $L'(\theta; X)$. We saw that $E_\theta \left[L'(\theta; X)\right] = 0$ for each $\theta$. Thus the term $I = E_\theta \left[L'(\theta; X)^2\right]$ is the variance of the random variable $L'(\theta; X)$. R. A. Fisher [4, 5, 6] interpreted this as the “intrinsic accuracy” of the distribution. For the case of the normal distribution, you can verify that $L'(\mu; x) = \frac{1}{\sigma^2}(x - \mu)$, so $I = \frac{1}{\sigma^2}$. The term $I$ has since become known as the Fisher information. For multinomials, it is related to entropy and Shannon’s [14, 15] information measure.

Another way to interpret the lower bound in the normal case is that the variance of an unbiased estimator of the mean is bounded below by the variance of the underlying average of the random variables on which it is based. This is hardly mysterious. And it is not surprising that the sample mean achieves that minimum variance.

18.2 Bayesian estimation

The Bayesian approach essentially treats the parameters $\theta$ as subject to the laws of probability. Whether one thinks the parameters are chosen at random by Nature or whether the randomness is purely subjective belief is essentially a religious issue, but the mathematics, or formal approach, is the same in either case.

We start with a probability density $\varphi$ on the set $\Theta$ of parameters. It is either the statistician’s degree of belief, or a probability with which Nature selects parameters. It is called the prior probability density on $\Theta$.

The joint density or likelihood $f(x; \theta)$ of the data vector $x$ under parameter $\theta$ is treated as a conditional probability density. After observing $x$, we use Bayes’ Law to compute the posterior probability on $\Theta$, which your textbook refers to as $g$, but I shall refer to as $\varphi(\cdot | x)$.

[You have already done similar calculations with some of the urn problems earlier.]

The posterior density given observation $x$ is given by Bayes’ Law as

$$
\varphi(\theta_0 | x) = \frac{f(x; \theta_0)\varphi(\theta_0)}{\int_\Theta f(x; \theta)\varphi(\theta)\,d\theta}
\propto f(x; \theta_0)\varphi(\theta_0).
$$

A uniform prior or uninformative prior has $\varphi(\theta)$ independent of $\theta$, in which case the posterior density is proportional to the likelihood function.

Actually, the description of Bayesian methodology here is perhaps simplistic. There is still an ongoing debate about Bayesian methods, what they are, and whether they should be used. The exchange between Brad Efron [3] Herman Chernoff [1], and Dennis Lindley [11], among others, though nearly twenty years old still is relevant.
18.3 Conjugate priors

An important concept for Bayesian estimation is that of a **conjugate prior**. A parametric family of distributions is conjugate to a likelihood function if the posterior belongs to the family whenever the prior does. In other words, the prior looks like the result of having seen a prior history of the data generating process.

Morris DeGroot [2] devotes Chapter 9 to conjugate priors. Here are a few examples.

**18.3.1 Example (Binomial \((n, p)\) Likelihood)** The likelihood function for \(p\) given a sample of \(n\) independent Bernoulli\((p)\) trials \(X_1, \ldots, X_n\) can be based on the sufficient statistic \(k = \sum_{i=1}^{n} X_i:\)

\[
L(p; k, n) = \binom{n}{k} p^k (1-p)^{n-k}.
\]

If the prior density \(\varphi\) on \(p\) is a \(\text{Beta}(s, f)\) density

\[
\varphi(p) \propto p^{s-1} (1-p)^{f-1}
\]

on \([0, 1]\), then the posterior density satisfies

\[
\varphi(p \mid k) \propto p^k (1-p)^{n-k}: p^{s+k-1-1} (1-p)^{f+(n-k)-1},
\]

which is a \(\text{Beta}(s+k, f+n-k)\) density. The interpretation of the parameters in the Beta distribution is this. If the expected number of successes is \(s+1\) and the expected number of failures is \(f+1\), then the density of \(p\) is \(\text{Beta}(s, f)\). The mean is \(s/(s+f)\). So starting with a \(\text{Beta}(s, f)\) is like starting with a prior history of \(s-1\) successes and \(f-1\) failures. The \(\text{Beta}(1, 1)\) distribution is the uniform \(U[0, 1]\) distribution.

Suppose we start with a prior heavily biased toward 0, say \(\text{Beta}(1, 19)\) which has mean \(1/20\) and density:

![Beta(1,19) Probability Density](image)

After the coin tossing data (29,140 successes and 29,228 failures) the posterior density is \(\text{Beta}(29140, 29228)\):
That is, the data essentially drown out the prior.
The next figure shows the effect on the posterior of more data.

Here are a couple of other examples of conjugate priors, taken from DeGroot [2]. He describes many more examples, including the conjugate for a Normal with known standard deviation.

- **Exponential**($\lambda$) (cf. DeGroot [2, p. 166])

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with an Exponential density. The likelihood function in terms of the sufficient statistic $T = \sum_{i=1}^{n} X_i$ is

$$L(\lambda \mid T, n) \propto \lambda^n e^{-\lambda T}.$$  

The conjugate prior $\varphi$ for $\lambda$ on $(0, \infty)$ is a Gamma($n_0, T_0$) density with $n_0 > 0$, $T_0 > 0$,

$$\varphi(\lambda) \propto \lambda^{n_0-1} e^{-\lambda T_0},$$

and the posterior density for $\lambda$ is a Gamma($n_0 + n, T_0 + T$) density,

$$\varphi(\lambda \mid k, n) \propto \lambda^{n_0+n-1} e^{-\lambda(T_0+T)}.$$
• Poisson($\mu$) (cf. DeGroot [2, p. 164])

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with a Poisson probability mass function, with unknown parameter $\mu$, and let $k = \sum_{i=1}^{n} X_i$. The likelihood function satisfies

$$L(\mu; k, n) \propto \mu^k e^{-n\mu}.$$ 

The conjugate prior $\varphi$ for $\mu$ on $(0, \infty)$ is a Gamma($k_0, n_0$) density with $k_0 > 0$, $n_0 > 0$

$$\varphi(\mu) \propto \mu^{k_0-1} e^{-n_0\mu},$$

and the posterior density for $\mu$ is a Gamma($k_0 + k, n_0 + n$) density,

$$\varphi(\mu \mid k, n) \propto \mu^{k_0+k-1} e^{-(n_0+n)\mu}$$

### 18.4 Loss functions

Bayesian posteriors are often used to create point estimates by minimizing the posterior expected values of a **loss function**. A loss function $L$ is a function of both the parameter and the estimate, satisfying $L(\hat{\theta}, \theta) \geq 0$ and $L(\theta, \theta) = 0$. The associated **risk function** is defined by

$$\int_{\Theta} L(\hat{\theta}, \theta) \varphi(\theta \mid x) d\theta.$$ 

When $\hat{\theta}$ is chosen to minimize the risk, it is called a **Bayesian estimate**.

**Aside**: It is unfortunate that $L$ is used to denote the loss function in this context, and it is used to denote the likelihood function in other contexts. What can I say? Decision theorists and economists frequently use **utility functions**, which are measures of gains rather than losses, and maximize expected utility rather than minimize expected loss.

#### 18.4.1 Proposition (Larsen and Marx [10, Theorem 5.8.1, pp. 342–344])

a. When $L(\hat{\theta}, \theta) = |\hat{\theta} - \theta|$, the risk minimizing $\hat{\theta}$ is the median of $\varphi(\theta \mid x)$.

b. When the loss is the square error, $L(\hat{\theta}, \theta) = (\hat{\theta} - \theta)^2$, the risk minimizing $\hat{\theta}$ is the mean of $g$.

**Proof**: a. The risk function is given by

$$\int |\hat{\theta} - \theta| \varphi(\theta \mid x) d\theta = \int \left( (\hat{\theta} - \theta) \mathbf{1}_{(-\infty, \hat{\theta})}(\theta) + (\theta - \hat{\theta}) \mathbf{1}_{[\hat{\theta}, \infty)}(\theta) \right) \varphi(\theta \mid x) d\theta$$

Differentiating with respect to $\hat{\theta}$ gives

$$\int \left( \mathbf{1}_{(-\infty, \hat{\theta})}(\theta) - \mathbf{1}_{[\hat{\theta}, \infty)}(\theta) \right) \varphi(\theta \mid x) d\theta$$

which is negative for $\hat{\theta} < \text{median}$, positive for $\hat{\theta} > \text{median}$, equal to zero at the median of the posterior.
b. The risk function is given by

\[ \int (\hat{\theta} - \theta)^2 \varphi(\theta \mid x) d\theta. \]

The first order condition for a minimum is obtained by differentiating under the integral sign to get:

\[ \int 2(\hat{\theta} - \theta) \varphi(\theta \mid x) d\theta = 0. \]

The solution is \( \hat{\theta} = \int \theta \varphi(\theta \mid x) d\theta \).

That is, \( \hat{\theta} \) is the posterior mean.

---

### 18.5 ★ Appendix: The Bayesian Bookie

#### 18.5.1 Statistical inference: the game

Freedman and Purves [7] caricature statistical inference in terms of the following game.

1. The Master of Ceremonies chooses an urn \( \theta_0 \) from a set \( \Theta \) of urns, draws a sample \( x \) from the urn according to the probability measure \( p_{\theta}(x) \), and exhibits the sample to the Bettor and the Bookie.

2. A Bookie posts prices \( q \) for lottery tickets that pay off 1 unit in case \( \theta \) is the urn, for each \( \theta \in \Theta \). The Bookie must buy and/or sell tickets at these prices.

3. The Bettor buys a portfolio of lottery tickets. The Bettor may also sell tickets to the Bookie at the same price the Bookie sells them.

4. The MC reveals the urn \( \theta_0 \), and the tickets are payed off.

The reason this is a caricature is that in the real world of statistical inference, there is never an MC to reveal \( \theta_0 \).

#### 18.5.2 Strategies

The Bettor and the Bookie choose their strategies for the game in advance of playing, so they must decide what to do for each possible sample that could be observed.

The Bookie chooses \( q \geq 0 \in \mathbb{R}^{\Theta \times X} \). For each \( x \in X \) and \( \theta \in \Theta \), \( q(\theta, x) \) is the price he sets, after having seen the sample \( x \), for a lottery ticket that pays $1 if the chosen urn is \( \theta \).

Then Bettor then chooses wagers \( w \in \mathbb{R}^{\Theta \times X} \), and buys \( w(\theta, x) \) \( \theta \)-tickets, after having the seen the sample \( x \) and the prices \( q \).

Under these strategies, the expected payoff to the Bettor when \( \theta \) is the selected urn is just

\[ \sum_{x \in X} \left( \sum_{t \in \Theta} \left( 1_t(\theta) - q(t, x) \right) w(t, x) \right) p_{\theta}(x). \]
18.5.1 Bayesian updating theorem  

Either

(i) The Bookie chooses some prior \( P \) on \( \Theta \) and sets prices according to the posterior \( P(\theta|x) \),

\[
P(\theta|x) = \frac{p_\theta(x)P(\theta)}{\sum_{t \in \Theta} p_t(x)P(t)}.
\]

Or else

(ii) There is a betting strategy that gives the Bettor a positive expected payoff regardless of which urn \( \theta \) is selected by the MC.

Note that this result does not say that the MC actually selected the urn at random according to \( P \)—it is merely a device to calculate the prices to avoid (ii).

**Proof:** Condition (ii) is equivalent to the matrix inequality

\[
\begin{bmatrix}
(\theta, x) \\
\vdots \\
\vdots
\end{bmatrix}
\begin{bmatrix}
\vdots \\
1_\theta(t) - q(\theta, x) p_t(x) \\
\vdots \\
w(\theta, x)
\end{bmatrix}
\gg 0,
\]

where rows are indexed by \( t \in \Theta \) and columns are indexed by \( (\theta, x) \in \Theta \times X \).

Gordan’s Alternative 18.5.2 below, asserts that the alternative to (ii) is the existence of a probability vector \( P \in \mathbb{R}^\Theta \) such that for each column \( (\theta, x) \in \Theta \times X \),

\[
\sum_{t \in \Theta} (1_\theta(t) - q(\theta, x)) p_t(x) P(t) = 0.
\]

In other words,

\[
p_\theta(x) P(\theta) = \sum_{t \in \Theta} q(\theta, x) p_t(x) P(t),
\]

or

\[
q(\theta, x) = \frac{p_\theta(x) P(\theta)}{\sum_{t \in \Theta} p_t(x) P(t)} = P(\theta|x),
\]

which is (i).  

The proof relied on this result due to P. Gordan [9], which is a form of a Theorem of the Alternative. See David Gale [8, Chapter 2] or my on-line notes for a proof.

18.5.2 Gordan’s Alternative  

Let \( A \) be an \( m \times n \) matrix. Exactly one of the following alternatives holds. Either there exists \( x \in \mathbb{R}^n \) satisfying

\[
Ax \gg 0.
\]

or else there exists \( p \in \mathbb{R}^m \) satisfying

\[
\begin{align*}
pA &= 0 \\
p &= 0.
\end{align*}
\]
Figure 18.1. Geometry of the Gordan Alternative

Bibliography

http://www.jstor.org/stable/2683106


http://www.jstor.org/stable/2683105


DOI: 10.1017/S0305004100009580

DOI: 0.1098/rspa.1934.0050

http://www.jstor.org/stable/2239586


