Chapter 4

Lecture 13

The score function, Fisher information and bounds

Let Θ be an open interval in $\mathbb{R}^1$ and suppose that $dP_\theta(s) = \ell_\theta(s)d\mu(s)$, where μ is a fixed measure on S. Suppose that $\theta \mapsto \ell_\theta(s)$ is differentiable for each fixed s; then $\delta \mapsto \Omega_{\delta,\theta}(s) = \frac{\ell_\delta(s)}{\ell_\theta(s)}$ is also differentiable for each fixed (s,θ). If we use dashes for derivatives with respect to the parameters as described, then

$$\Omega'_{\theta,\theta}(s) = \frac{\ell'_\theta(s)}{\ell_\theta(s)} =: \gamma_\theta^{(1)}(s)$$

is the SCORE FUNCTION at θ (given s). We also define $I(\theta) := E_\theta\big(\gamma_\theta^{(1)}(s)\big)^2$, the FISHER INFORMATION (for estimating θ) in s.

Note.

$$\begin{aligned}
&\Big(\int_S \ell_\delta(s)d\mu(s) = 1 \ \forall \delta \in \Theta\Big)\\
&\quad\Rightarrow \Big(\int_S \Omega'_{\delta,\theta}(s)dP_\theta(s) = \int_S \frac{\ell'_\delta(s)}{\ell_\theta(s)}\ell_\theta(s)d\mu(s) = \int_S \ell'_\delta(s)d\mu(s) = 0 \ \forall\delta\in\Theta\Big)\\
&\qquad\qquad \Rightarrow E_\theta\big(\gamma_\theta^{(1)}(s)\big) = E_\theta\big(\Omega'_{\theta,\theta}(s)\big) = 0 \Rightarrow I(\theta) = \mathrm{Var}_\theta(\gamma_\theta^{(1)})
\end{aligned}$$

Similarly, we have $\int_S \ell''_\delta(s)d\mu(s) = 0$, $\int_S \ell'''_\delta(s)d\mu(s) = 0$, etc. for all $\delta \in \Theta$, so that $E_\theta\big(\gamma_\theta^{(j)}(s)\big) = 0$ for $j = 1,2,3,\ldots$, where $\gamma_\theta^{(j)}(s) = \big(\frac{\partial^j \ell_\theta(s)}{\partial\theta^j}\big)/\ell_\theta(s)$. Conditions under which the interchanging of differentiation and integration (as above) is valid will be given later.

Suppose that we are interested in W_θ and want some concrete method of constructing it. We have that

$$\Omega_{\delta,\theta}(s) = \Omega_{\theta,\theta} + (\delta-\theta)\gamma_\theta^{(1)}(s) + \frac{1}{2}(\delta-\theta)^2\gamma_\theta^{(2)}(s) + \cdots,$$

which suggests that $W_\theta = \mathrm{Span}\{1, \gamma_\theta^{(1)}, \gamma_\theta^{(2)}, \ldots\}$. We will see that this equality holds exactly in a one-parameter exponential family and approximately in general in large samples. To see that $\gamma_\theta^{(j)} \in W_\theta$, we reason as follows: First, of course, we note that $1 \in W_\theta$. Then, since $\Omega_{\delta,\theta}, \Omega_{\theta,\theta} \in W_\theta$, we have that $\frac{1}{\delta-\theta}(\Omega_{\delta,\theta} - \Omega_{\theta,\theta}) \in W_\theta$ for $\delta \neq \theta$, from which it follows that $\gamma_\theta^{(1)} \in W_\theta$. Similar inductive reasoning allows us to conclude that each $\gamma_\theta^{(j)}$ is in W_θ.

It is clear that 1 and $\gamma_\theta^{(1)}$ are the most important generators if s is very informative, for then only δ near the true θ are important. In any case,

$$W_\theta^{(k)} := \text{Span}\{1, \gamma_\theta^{(1)}, \gamma_\theta^{(2)}, \ldots, \gamma_\theta^{(k)}\} \subseteq W_\theta.$$

We know that, in $V_\theta = L^2(P_\theta)$, every $t \in U_g$ projects to the same $\tilde{t} \in W_\theta$; thus every $t \in U_g$ has the same projection to $W_\theta^{(k)}$ – say $t_{\theta,k}^*$. Then we have:

11. BHATTACHARYA BOUNDS: For each $t \in U_g$,

$$\text{Var}_\theta(t) \geq E_\theta(t_{\theta,k}^*)^2 - \left[g(\theta)\right]^2$$

for $k = 1, 2, \ldots$.

Proof. This follows since

$$\text{Var}_\theta(t) + \left[g(\theta)\right]^2 = E_\theta(t^2) \geq E_\theta(t_{\theta,k}^*)^2.$$

□

Let us consider the case $k = 1$ – i.e., projection to $\text{Span}\{1, \gamma_\theta^{(1)}\}$. We have seen that $1 \perp \gamma_\theta^{(1)}$ – i.e., that $E_\theta(\gamma_\theta^{(1)}) = 0$ – and that $||\gamma_\theta^{(1)}||^2 = I(\theta)$. Hence $\{1, \gamma_\theta^{(1)}/||\gamma_\theta^{(1)}||\}$ is an orthonormal basis in $W_\theta^{(1)}$ and, for any $t \in V_\theta$, the projection $t_{\theta,1}^*$ of t to $W_\theta^{(1)}$ is

$$t_{\theta,1}^* = (1,t)1 + \left(\frac{\gamma_\theta^{(1)}}{||\gamma_\theta^{(1)}||}, t\right) \frac{\gamma_\theta^{(1)}}{||\gamma_\theta^{(1)}||}.$$

Now $(1,t) = E_\theta(t) = g(\theta)$ since t is unbiased, and

$$\begin{aligned}(\gamma_\theta^{(1)}, t) = E_\theta(t \cdot \gamma_\theta^{(1)}) = \int_S t(s) \frac{\ell_\theta'(s)}{\ell_\theta(s)} dP_\theta(s) &= \int_S t(s)\ell_\theta'(s) d\mu(s) \\ &\overset{(?)}{=} \frac{d}{d\theta} \int_S t(s)\ell_\theta(s) d\mu(s) = \frac{d}{d\theta} g(\theta) = g'(\theta).\end{aligned}$$

The above calculations give us that

$$t_{\theta,1}^* = g(\theta) + \frac{g'(\theta)}{||\gamma_\theta^{(1)}(s)||} \frac{\gamma_\theta^{(1)}(s)}{||\gamma_\theta^{(1)}(s)||};$$

since the summands are orthogonal,

$$||t_{\theta,1}^*||^2 = g(\theta)^2 + \frac{(g'(\theta))^2}{||\gamma_\theta^{(1)}(s)||^2} = g(\theta)^2 + \frac{(g'(\theta))^2}{I(\theta)}.$$

From this we see:

12 (Fisher-Darmois-Cramér-Rao). INFORMATION INEQUALITY: For $t \in U_g$,

$$\mathrm{Var}_\theta(t) \geq \frac{(g'(\theta))^2}{I(\theta)}.$$

The Fisher information can be related to the second derivative of the log-likelihood: Let $L_\theta(s) = \log_e \ell_\theta(s)$. Then $L'_\theta(s) = \frac{\ell'_\theta(s)}{\ell_\theta(s)} = \gamma_\theta^{(1)}(s)$ and

$$L''_\theta(s) = \frac{\ell''_\theta(s)}{\ell_\theta(s)} - \left(\frac{\ell'_\theta(s)}{\ell_\theta(s)}\right)^2 = \frac{\ell''_\theta(s)}{\ell_\theta(s)} - \left[\gamma_\theta^{(1)}\right]^2;$$

but $E_\theta\big(\ell''_\theta(s)/\ell_\theta(s)\big) = \int_S \ell''_\theta(s)d\mu(s) = 0$, and so

13. $E_\theta\big(L''_\theta(s)\big) = -I(\theta)$.

Exact conditions under which statements (11)–(13) hold are deferred until Lecture 5.1.

Lecture 14

Heuristics for maximum likelihood estimate:

i. $W_\theta = \mathrm{Span}\{1, \gamma_\theta^{(1)}, \gamma_\theta^{(2)}, \ldots\}$.

ii. $W_\theta \approx \mathrm{Span}\{1, \gamma_\theta^{(1)}\}$ if s is highly informative.

iii. The MLE $\hat\theta(s) \dot\in W_\theta$ (whatever θ may be!).

The last item gives us that:

iv. $\hat\theta$ is approximately the UMVUE of its own expected value function (the same is true of estimates related to $\hat\theta$ in certain ways).

Let $\hat\theta(s)$ be the MLE of θ and assume that $\hat\theta$ is close to θ. Since $\hat\theta(s)$ maximizes L_δ, we have

$$0 = L'_{\hat\theta} = L'_\theta + (\hat\theta - \theta)L''_\theta + \cdots \approx L'_\theta + (\hat\theta - \theta)L''_\theta.$$

Assume also that the experiment (that is, $(S, \mathcal{A}, P_\theta)$, $\theta \in \Theta$) is highly informative in the sense that $I(\theta)$ is large (for a given θ). We know that $E_\theta(L'_\theta) = 0$ and $\mathrm{Var}_\theta(L'_\theta) = I(\theta)$; hence, informally, L'_θ is "about" 0, "give or take" about $\sqrt{I(\theta)}$. From (13), $E_\theta(-L''_\theta) = I(\theta)$ – i.e., $E_\theta\big(-\frac{L''_\theta}{I(\theta)}\big) = 1$. Assume that the random variable $-\frac{L''_\theta}{I(\theta)} \approx 1$. Then

$$\hat\theta \approx \theta - \frac{L'_\theta}{L''_\theta} = \theta + \frac{L'_\theta}{\sqrt{I(\theta)}}\frac{1}{\sqrt{I(\theta)}}\frac{1}{-L''_\theta/I(\theta)} \approx \theta + \frac{1}{\sqrt{I_\theta}}\frac{\gamma_\theta^{(1)}}{||\gamma_\theta^{(1)}||}, \qquad (*)$$

and hence $\hat{\theta}$ is nearly in $W_\theta^{(1)} \subseteq W_\theta$; so $\hat{\theta}$ is nearly LMVU, and hence $\hat{\theta}$ is nearly the UMVUE (of θ). From (*),

$$E_\theta(\hat{\theta}) \approx \theta \quad \text{and} \quad \mathrm{Var}_\theta(\hat{\theta}) \approx \frac{1}{I(\theta)}.$$

The MLE of $g(\theta)$ is $g(\hat{\theta})$. Assuming that g is continuously differentiable, we have

$$g(\hat{\theta}) \approx g(\theta) + g'(\theta)(\hat{\theta} - \theta).$$

So $g(\hat{\theta})$ is nearly in W_θ (since $1 \in W_\theta$ and $\hat{\theta}$ is nearly in W_θ). Hence

$$E_\theta g(\hat{\theta}) \approx g(\theta) \quad \text{and} \quad \mathrm{Var}_\theta\big(g(\hat{\theta})\big) \approx \frac{(g'(\theta))^2}{I(\theta)},$$

where $\frac{[g'(\theta)]^2}{I(\theta)}$ is the lower bound in (12).

Note. $\frac{I(\theta)}{[g'(\theta)]^2}$ is the information in s for estimating $g(\theta)$.

Suppose that $(S_1, \mathcal{A}_1, P_\theta^{(1)})$ and $(S_2, \mathcal{A}_2, P_\theta^{(2)})$, $\theta \in \Theta$, are independent experiments concerning θ, with sample points s_1 and s_2. Let $s = (s_1, s_2)$, $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2$ and $P_\theta = P_\theta^{(1)} \times P_\theta^{(2)}$, and let $I_i(\theta)$ be the information in s_i for estimating θ ($i = 1, 2$). Then the information in s for estimating θ is $I(\theta) = I_1(\theta) + I_2(\theta)$. (This result extends inductively to any finite number of independent experiments.)

Proof. $dP_\theta^{(i)}(s) = \ell_\theta^{(i)}(s_i) d\mu^{(i)}(s_i)$ for $i = 1, 2$, so $dP_\theta(s) = \ell_\theta^{(1)}(s_1)\ell_\theta^{(2)}(s_2) d\nu(s)$ and hence

$$L_\theta(s) = \log \ell_\theta^{(1)}(s_1) + \log \ell_\theta^{(2)}(s_2) = L_\theta^{(1)}(s) + L_\theta^{(2)}(s).$$

The result now follows from (13). □

Example 1(a). $s = (X_1, \ldots, X_n)$, $X_i \overset{\text{iid}}{\sim} N(\theta, 1)$. The information in s for estimating θ is the sum of the information in $X_1, \ldots, X_n$, respectively, for estimating θ, which sum is (since the X_i are iid) n times the information in X_1, which product is (since X_1 is distributed as $N(\theta, 1)$) just n. $L'_\theta(X_1) = X_1 - \theta = \gamma_\theta^{(1)}(X_1)$ and $\mathrm{Var}_\theta(\gamma_\theta^{(1)}) = 1 = I_1(\theta)$. (We check that $\frac{L'_\theta(s)}{\sqrt{I(\theta)}}$ is about 0, give or take about 1; and $\frac{L''_\theta(s)}{I(\theta)} \approx 1$ (indeed, here it is identically 1).)

Example 2. $X_1, \ldots, X_n, \ldots$ are iid as

$$\begin{cases} 0 & \text{with probability } 1 - \theta \\ 1 & \text{with probability } \theta, \end{cases}$$

and $\Theta = (0, 1)$. $s = (X_1, \ldots, X_N)$, N the stopping time. The three cases we discussed are:

a. $N \equiv n$ (n a fixed positive integer).

b. N is the first time k successes (i.e., 1s) are recorded (k a fixed positive integer).

c. Two-stage scheme.

In all cases (even other than (a)–(c) above), $\ell_\theta(s) = \theta^{T(s)}(1-\theta)^{N(s)-T(s)}$, where $T(s) = \sum_{i=1}^{N(s)} X_i$, and hence

$$L'_\theta(s) = \frac{T(s)}{\theta} - \frac{V(s)}{1-\theta}$$

and

$$L''_\theta(s) = -\frac{T(s)}{\theta^2} - \frac{V(s)}{(1-\theta)^2},$$

where $V(s) = N(s) - T(s)$. Since $E_\theta(L'_\theta) = 0$, we have

$$\frac{E_\theta(T)}{\theta} = \frac{E_\theta(V)}{1-\theta}, \tag{4.1}$$

$$I(\theta) = \mathrm{Var}_\theta(L'_\theta) = \frac{\mathrm{Var}_\theta(T)}{\theta^2} + \frac{\mathrm{Var}_\theta(V)}{(1-\theta)^2} - \frac{2}{\theta(1-\theta)}\mathrm{Cov}_\theta(T,V) \tag{4.2}$$

and

$$I(\theta) = E_\theta(-L''_\theta) = \frac{1}{\theta^2}E_\theta(T) + \frac{1}{(1-\theta)^2}E_\theta(V). \tag{4.3}$$

Exercise: What happens in case (a) (i.e., $N \equiv n$)? This is like Example 1(a) except that $I(\theta)^{-1} = \frac{\theta(1-\theta)}{n}$ depends on θ.

Suppose now that we are in case (b). Then $T \equiv k$ and $V(s) = N(s) - k$. Hence, from (4.1), $\frac{k}{\theta} = \frac{E_\theta(N-k)}{1-\theta}$ and therefore $E_\theta(N) = \frac{k}{\theta}$ (which we could also compute directly) and

$$I(\theta) = \frac{k}{\theta^2} + \frac{1}{(1-\theta)^2}E_\theta(N-k) = \frac{k}{\theta^2(1-\theta)}$$

(from equation (4.3) above). Hence the heuristics apply when k is large. In all cases $\hat\theta(s)$ is $\frac{T(s)}{N(s)}$.

Exercise: Derive $\mathrm{Var}_\theta(N)$ from (4.2) and check the behavior of $L'_\theta/\sqrt{I(\theta)}$ and $L''_\theta/\sqrt{I(\theta)}$.

Example 1(e). $s = (X_1, \cdots, X_n)$, with the X_i iid with density $ae^{-b(x-\theta)^4}$ $(a, b > 0)$.

Homework 3

1. a. Find a and b such that $\mathrm{Var}_\theta(X_1) = 1$ (to make it comparable to Example 1(a)).
   b. Find $I(\theta)$.
   c. $E_\theta(\overline{X}) \equiv \theta$, so $\overline{X}$ is unbiased for θ. Is $\overline{X}$ the UMVUE? (Note the answer is no.)
   d. What is the UMVUE?
   e. Give an explicit method for finding $\hat\theta(s)$.

Lecture 15

13(a). Suppose $t \in U_g$ is such that

$$\mathrm{Var}_\theta(t) = \frac{[g'(\theta)]^2}{I(\theta)} \ \forall \theta \in \Theta;$$

then $\{P_\theta : \theta \in \Theta\}$ is a one-parameter exponential family with statistic t – i.e.,

$$\frac{dP_\theta}{d\mu}(s) = \ell_\theta(s) = \varphi(s)e^{A(\theta)+B(\theta)t(s)},$$

where A and B are smooth functions; moreover, $g(\theta) = -A'(\theta)/B'(\theta)$.

Proof. By the same argument as used in the proof of (12), we have that $t \in W_\theta^{(1)}$ for all θ – i.e., that

$$t(s) = a(\theta) + b(\theta)L'_\theta(s) \qquad \text{a.e.}(P_\theta)$$

for all θ. From this it follows that $L'_\theta(s) = \alpha(\theta) + \beta(\theta)t(s)$ (if $b \equiv 0$ then $\mathrm{Var}_\theta(t) = 0$ for all θ. We rule out this case) and hence that

$$L_\theta(s) = A(\theta) + B(\theta)t(s) + C(s),$$

where $A(\theta) = \int \alpha(\theta)d\theta$. This gives the required form for $\ell_\theta(s)$. Also,

$$0 = E_\theta(L'_\theta) = \alpha(\theta) + \beta(\theta)E_\theta(t) = \alpha(\theta) + \beta(\theta)g(\theta)$$

and so $g(\theta) = -\alpha(\theta)/\beta(\theta) = -A'(\theta)/B'(\theta)$. □

Note. For a near-rigorous proof, see R. A. Wijsman 1973 AS, pp. 538–542, and V. M. Joshi 1976 AS, pp. 998–1002.

Note. The necessary conditions on $\{P_\theta : \theta \in \Theta\}$ and g are also sufficient for the attainment of the C-R bound. We will see this later.

Example 1(a). Since

$$\ell_\theta(s) = \varphi_1(s)e^{-\frac{n}{2}(\overline{X}-\theta)^2} = \varphi_2(s)e^{-\frac{n\theta^2}{2}+(n\theta)\overline{X}},$$

the C-R bound is attained by $\overline{X}$ for estimating $g(\theta) = \theta$. This implies that $\overline{X}$ is LMVU at θ, which in turn implies that it is UMVU. Also, the C-R bound is not attained by any unbiased estimate of any g which is not an affine function of θ. In particular, since $\overline{X}^2 - \frac{1}{n}$ is an unbiased estimate of $g(\theta) = \theta^2$, it does not attain the C-R bound since $\theta^2 \neq -A'(\theta)/B'(\theta)$. We have seen before, however, that $\overline{X}^2 - \frac{1}{n}$ is the UMVUE.

To study the Bhattacharya bounds, note that $\ell'_\theta = \ell_\theta \cdot [-n\theta + n\overline{X}]$ and $\ell''_\theta = \ell_\theta \cdot [-n\theta + n\overline{X}]^2 + \ell_\theta \cdot [-n]$, so that ℓ'_θ/ℓ_θ is affine in $\overline{X}$ and $\ell''_\theta/\ell_\theta$ is quadratic. This implies that

$$W_\theta^{(1)} = \text{Span}\{1, \ell'_\theta/\ell_\theta\} = \text{Span}\{1, \overline{X}\}$$

and

$$W_\theta^{(2)} = \text{Span}\{1, \ell'_\theta/\ell_\theta, \ell''_\theta/\ell_\theta\} = \text{Span}\{1, \overline{X}, \overline{X}^2\},$$

whence $\overline{X}^2 - \frac{1}{n} \in W_\theta^{(2)}$ attains the Bhattacharya bound and is the UMVUE. In fact, W_θ is the space of *all* functions of $\overline{X}$, and hence any function of $\overline{X}$ (but not θ) is the UMVUE of its expectation.

Example 1(b). $s = (X_1, X_2, \ldots)$ are iid from $\frac{1}{2}e^{-|x-\theta|}$ on $\mathbb{R}^1$. Here W_θ is well-defined (i.e., (8)–(10) hold), but (11)–(13) are not applicable since ℓ_θ is not sufficiently smooth. In such a situation, the following is useful.

14 (Chapman-Robbins). Given $(S, \mathcal{A}, P_\theta)$, $\theta \in \Theta$, with Θ an open interval in $\mathbb{R}^1$, if $t \in U_g$ then

$$\text{Var}_\theta(t) \geq \overline{\lim_{\delta\to\theta}} \left(\frac{g(\delta) - g(\theta)}{\delta - \theta}\right)^2 \Big/ E_\theta\left(\frac{\Omega_{\delta,\theta} - 1}{\delta - \theta}\right)^2$$

for all θ such that $\Omega_{\delta,\theta} = \frac{dP_\delta}{dP_\theta}$ exists for all δ in a neighborhood of θ.

Proof.

$$E_\delta(t) = g(\delta) \Rightarrow \int_S t \cdot \Omega_{\delta,\theta} dP_\theta = g(\delta) \Rightarrow \int_S t(\Omega_{\delta,\theta} - 1) dP_\theta = g(\delta) - g(\theta).$$

Dividing by $\delta - \theta$, we find that

$$\begin{aligned}
\int t\left(\frac{\Omega_{\delta,\theta} - 1}{\delta - \theta}\right) dP_\theta &= \frac{g(\delta) - g(\theta)}{\delta - \theta} \\
&\Rightarrow \int_S (t - g(\theta))\left(\frac{\Omega_{\delta,\theta} - 1}{\delta - \theta}\right) dP_\theta = \frac{g(\delta) - g(\theta)}{\delta - \theta} \\
&\Rightarrow \left(\frac{g(\delta) - g(\theta)}{\delta - \theta}\right)^2 \leq \text{Var}_\theta(t) \cdot E_\theta\left(\frac{\Omega_{\delta,\theta} - 1}{\delta - \theta}\right)^2.
\end{aligned}$$

□

Note. If g is differentiable at θ, then

$$\text{Var}_\theta(t) \geq (g'(\theta))^2 / \underline{\lim_{\delta\to\theta}} E_\theta\left(\frac{\Omega_{\delta,\theta} - 1}{\delta - \theta}\right)^2.$$

If, further, $\Omega_{\delta,\theta}$ is differentiable (see (12E) below for exact conditions), then this is the same as $\frac{[g'(\theta)]^2}{I(\theta)}$.

Homework 3

2. What is the Chapman-Robbins bound for $g(\theta) = \theta$ in Example 1(b)?

3. In Example 1(c), s consists of n iid observations from $\frac{1}{\pi(1+(x-\theta)^2)}$. For any g, the C-R bound is not attained by any t; but $\hat{\theta}$ has nearly the variance $\frac{1}{I(\theta)}$ if $I(\theta)$ is large. Here $I(\theta) = nI_1(\theta)$. Show that $I_1(\theta) = \frac{1}{2}$.