Chapter 6

Lecture 19

The vector-valued score function and information in the multi-parameter case

Now we have an experiment $(S, \mathcal{A}, P_\theta)$, $\theta = (\theta_1, \ldots, \theta_p) \in \Theta$ with Θ an open set in $\mathbb{R}^p$ and a smooth function $g : \Theta \to \mathbb{R}^1$. We assume that $dP_\theta(s) = \ell_\theta(s) d\mu(s)$ as before, and define $\ell(\theta \mid s) := \ell_\theta(s)$. Assume that ℓ is smooth in θ and let $g_i(\theta) = \frac{\partial}{\partial \theta_i} g(\theta)$, $\ell_i(\theta \mid s) = \frac{\partial}{\partial \theta_i} \ell(\theta \mid s)$ and $\ell_{ij}(\theta \mid s) = \frac{\partial^2}{\partial \theta_i \partial \theta_j} \ell(\theta \mid s)$ for $1 \leq i, j \leq p$. There are two approaches to the present topic in this situation:

Approach 1. Generalize the previous one-dimensional discussion: Suppose that t is unbiased for g – that is to say,

$$\int_S t(s) \ell(\delta \mid s) d\mu(s) = E_\delta(t) = g(\delta)$$

for all $\delta \in \Theta$. Then

$$E_\theta\big(t(s)\ell_i(\theta \mid s)/\ell(\theta \mid s)\big) = \int_S t(s)\ell_i(\theta \mid s) d\mu(s) = g_i(\theta)$$

for $i = 1, \ldots, p$ and hence every $t \in U_g$ has the same projection on $\text{Span}\{1, L_1, \ldots, L_p\}$, where $L(\theta \mid s) = L_\theta(s)$ and

$$L_i(\theta \mid s) = \frac{\partial}{\partial \theta_i} L(\theta \mid s) = \frac{\ell_i(\theta \mid s)}{\ell(\theta \mid s)}.$$

This approach is useful for studies of conditions which ensure that $L_1, L_2, \ldots, L_p$ are in $W_\theta = \text{Span}\{\Omega_{\delta,\theta} : \delta \in \Theta\}$.

Approach 2. Use the result for the θ-real case: Fix $\theta \in \Theta$ and a vector $c = (c_1, \ldots, c_p) \neq 0$, and suppose that δ is restricted to the line passing through θ and $\theta + c$ – in other words, that we consider only $\delta = \theta + \xi c$ for some scalar ξ. (Note that, since Θ is open, if ξ is sufficiently small then $\theta + \xi c \in \Theta$.) Then g becomes a function of ξ for which t remains unbiased. By (12),

$$\mathrm{Var}_\theta(t) \geq [\text{Fisher information in } s \text{ for } g \text{ at } \theta \text{ in the restricted problem}]^{-1}$$
$$= \left(\frac{dg}{d\xi}\bigg|_{\xi=0} \right)^2 \Big/ [\text{Fisher information for } \xi \text{ in } s \text{ for estimating } g]$$

Now, since $\delta = \theta + \xi c$,

$$\frac{dg}{d\xi}\bigg|_{\xi=0} = \sum_{i=1}^{p} \frac{\partial g}{\partial \delta_i}\bigg|_{\delta=\theta} c_i = \sum_{i=1}^{p} c_i g_i(\theta).$$

The information in the denominator is $E_\theta(dL/d\xi)^2$, and

$$\frac{dL}{d\xi}\bigg|_{\xi=0} = \sum_{i=1}^{p} c_i L_i(\theta \mid s),$$

so that the information may be expressed explicitly as

$$E_\theta \left(\frac{dL}{d\xi} \right)^2 = \sum_{i=1}^{p} \sum_{j=1}^{p} c_i c_j E_\theta\big(L_i(\theta \mid s) L_j(\theta \mid s)\big) = \sum_{i,j} c_i c_j I_{ij},$$

where I_{ij} is the (i, j)th entry of the Fisher information matrix

$$I(\theta) = \big\{ \mathrm{Cov}_\theta(L_i(\theta \mid s), L_j(\theta \mid s)) \big\}_{p \times p}$$

(where the sample space is S). Let

$$L_{ij} = \frac{\partial L_i}{\partial \theta_j} = \frac{\partial}{\partial \theta_j} \left[\frac{\ell_i}{\ell} \right] = \frac{\ell_{ij}}{\ell} - \frac{\ell_i \ell_j}{\ell^2};$$

then

$$E_\theta(L_{ij}) = \int \ell_{ij}(\theta \mid s) d\mu(s) - E_\theta(L_i L_j) = -E_\theta(L_i L_j)$$

and hence we have the p-dimensional analogue of (13):

13^p. $I(\theta) = \big\{ -E_\theta(L_{ij}(\theta \mid s) \big\}$.

The above lower bound for $\mathrm{Var}_\theta(t)$ can now be written as

$$\Big[\sum_i c_i g_i(\theta) \Big]^2 \Big/ \Big(\sum_{i,j} c_i c_j I_{ij} \Big).$$

Let us assume that I is positive definite. It will be shown below that

$$\sup_c \{\text{the bound above}\} = \sum_{i,j} g_i(\theta) I^{ij}(\theta) g_j(\theta), \qquad (*)$$

where $\{I^{ij}(\theta)\} = I^{-1}(\theta)$; and the supremum is achieved when c is a multiple of $h(\theta) I^{-1}(\theta)$, where $h(\theta) = (g_1(\theta), \ldots, g_p(\theta)) = \nabla g(\theta)$.

Thus we have the p-dimensional analogue of (12):

12^p. If $t \in U_g$, then $\mathrm{Var}_\theta(t) \geq h(\theta)I^{-1}(\theta)h(\theta)'$.

Assume that this bound is attained, at least approximately; then, for the estimation of g, there exists a one-dimensional problem (namely, the one obtained by restricting δ to $\{\theta+\xi c^* : \xi \in \mathbb{R}\}$, where $c^* = h(\theta)I^{-1}(\theta)$) which is as difficult as the p-dimensional problem.

Proof of (*). For $u = (u_1, \dots, u_p)$ and $v = (v_1, \dots, v_p)$ in $\mathbb{R}^p$, let $(u|v) := \sum_{i=1}^p u_i v_i = uv'$ and $||u|| := (u|u)^{1/2}$. Let I be a (fixed) positive definite symmetric $p \times p$ matrix and set $(u|v)_* := \sum_{i,j} u_i I_{ij} v_j = uIv'$ and $||u||_* := (u|u)_*^{1/2}$. Let $g = (g_1, \dots, g_p)$ be a fixed point in $\mathbb{R}^p$. Consider the maximization over $\underline{a} = (a_1, \dots, a_p) \in \mathbb{R}^p$ of

$$\frac{(\sum_{i=1}^p a_i g_i)^2}{\sum_{i,j} a_i I_{ij} a_j} = \frac{(\underline{a}g')^2}{||\underline{a}||_*^2} = \frac{(\underline{a}I|gI^{-1})^2}{||\underline{a}||_*^2} = \frac{(\underline{a}|gI^{-1})_*^2}{||\underline{a}||_*^2} = \left(\frac{a}{||a||_*}\middle|gI^{-1}\right)_*^2.$$

The unique (up to scalar multiples) maximizing value is given by $\underline{a} = gI^{-1}$ and the maximum value is

$$\left(\frac{gI^{-1}}{||gI^{-1}||_*}\middle|gI^{-1}\right)_*^2 = \left[\frac{(gI^{-1})I(gI^{-1})'}{||gI^{-1}||_*}\right]^2 = ||gI^{-1}||_*^2 = gI^{-1}g'.$$

□

Lecture 20

We have seen that, with $\theta = (\theta_1, \dots, \theta_p)$ and fixed g, the "most difficult" one-dimensional problem is with $\delta \in \Theta$ unknown but restricted to

$$\{\theta + \xi c^* : |\xi| \text{ is sufficiently small}\},$$

where $c^* = c^*(\theta) = h(\theta)I^{-1}(\theta)$ and $h(\theta) = \operatorname{grad} g(\theta) = (g_1(\theta), \dots, g_p(\theta))$, $g_i = \frac{\partial g}{\partial \theta_i}$; i.e.,

$$t \in U_g \Rightarrow \mathrm{Var}_\theta(t) \geq \mathrm{Var}_\theta(\tilde{t}) \geq \mathrm{Var}_\theta(t_{\theta,1}^*) = h(\theta)I^{-1}(\theta)h'(\theta),$$

where $\tilde{t}$ is the projection (of *any* $t \in U_g$) to W_θ and $t_{\theta,1}^*$ is the projection (again, of *any* $t \in U_g$) to $\mathrm{Span}\{1, dL/d\xi|_{\xi=0}\}$. Now (remembering that $\delta = \theta + \xi c^*$)

$$\left.\frac{dL}{d\xi}\right|_{\xi=0} = \sum_{i=1}^p c_i^* L_i(\theta \mid s) =: L'$$

and, under P_θ (i.e., for $\xi = 0$) $1 \perp L'$, so $\{1, L'/||L'||\}$ is an orthonormal basis for $\mathrm{Span}\{1, L'\}$ and

$$t_{\theta,1}^* = g(\theta) \cdot 1 + \left(t, \frac{L'}{||L'||}\right) \cdot \frac{L'}{||L'||} = g(\theta) + \frac{1}{||L'||}\left.\frac{dg}{d\xi}\right|_{\xi=0}\frac{L'}{||L'||}$$

$$= g(\theta) + \left(\sum_{i=1}^p c_i^* L_i(\theta \mid s)\right)\frac{\sum_i c_i^* g_i(\theta)}{\sum_{i,j} c_i^* I_{ij}(\theta) c_j^*} = g(\theta) + \left(\sum_{i=1}^p c_i^* L_i(\theta \mid s)\right)\frac{c^* h'}{c^* I c^{*\prime}}.$$

Note that $c^{*\prime} = I^{-1}h$, so $c^* I c^{*\prime} = hI^{-1}h' = c^* h'$ and so the above formula becomes

$$t^*_{\theta,1} = g(\theta) + \sum_{i=1}^{p} c_i^* L_i.$$

We have

$$\mathrm{Var}_\theta(t^*_{\theta,1}) = \frac{(\sum c_i^* g_i(\theta))^2}{(\sum_{i,j} c_i^* I_{ij}(\theta) c_j^*)} = \frac{(hI^{-1}h')^2}{(hI^{-1})I(hI^{-1})'} = hI^{-1}h'.$$

More heuristic (as in the one-dimensional parameter case)

"ML estimates are nearly unbiased and nearly attain the bound in 12^p."

We assume that the ML estimate $\hat{\theta}$ of θ exists. Since Θ is open and $L(\cdot \mid s)$ is continuously differentiable, we have that

$$L_i(\hat{\theta}) = \left.\frac{\partial L(\theta \mid s)}{\partial \theta_i}\right|_{\theta=\hat{\theta}} = 0.$$

Choose and fix $\theta \in \Theta$, and regard it as the actual parameter value. If we assume that $\hat{\theta}$ is close to θ, then

$$L_i(\hat{\theta}) \approx L_i(\theta) + \sum_{j=1}^{p} (\hat{\theta}_j - \theta_j) L_{ji}(\theta), \quad i = 1, \ldots, p.$$

Assume that the sample is highly informative, i.e., that

$$L_{ji}(\theta \mid s) \approx -I_{ij}(\theta).$$

(We know that $E_\theta\big(L_{ji}(\theta \mid s)\big) = -I_{ji}(\theta)$. We are thus assuming that

$$\{L_{ji}\} = \{-I_{ji}(1 + \varepsilon_{ji})\},$$

where $\varepsilon_{ji}(\theta, s) \to 0$ in probability. This happens typically when the data is highly informative.) From this it follows that

$$L_i(\theta) \approx \sum_{j=1}^{p} (\hat{\theta}_j - \theta_j) I_{ji}(\theta), \quad i = 1, \ldots, p$$

– i.e., $(\hat{\theta} - \theta)I = (L_1, \ldots, L_p)$.

Definition. $L^{(1)}(\theta \mid s) := \big(L_1(\theta \mid s), \ldots, L_p(\theta \mid s)\big)$ is the SCORE VECTOR.

Thus the ML estimate of a given g is

$$\hat{t}(s) = g(\hat{\theta}(s)) \approx g(\theta) + \sum_{j=1}^{p} (\hat{\theta}_j(s) - \theta_j) g_j(\theta) = g(\theta) + (\hat{\theta}(s) - \theta) h'(\theta)$$

$$\approx g(\theta) + L^{(1)}(\theta \mid s) I^{-1}(\theta) h'(\theta) = t^*_{\theta,1}$$

under P_θ. Since $E_\theta\big(L^{(1)}(\theta \mid s)\big) = 0$, we have $E_\theta(\hat{t}) \approx g(\theta)$. Since θ is arbitrary, $\hat{t}$ is approximately unbiased for g, i.e., $\hat{t} \mathrel{\dot\in} U_g$. Since

$$\hat{t}(s) \approx g(\theta) + L^{(1)}(\theta \mid s) I^{-1}(\theta) h'(\theta) = g(\theta) + c^* \big(L^{(1)}(\theta \mid s)\big)'$$

under P_θ, we know that $\hat{t} \mathrel{\dot\in} \mathrm{Span}\{1, L_1, \ldots, L_p\}$, so that $\hat{t} \approx t^*_{\theta,1}$ under P_θ and

$$\mathrm{Var}_\theta(\hat{t}) \approx \mathrm{Var}_\theta(t^*_{\theta,1}) = h(\theta) I^{-1}(\theta) h'(\theta).$$

This is, if true, remarkable, for it happens for *every* g and *every* $\theta \in \Theta$.

Example 3. Suppose that the X_i are iid $N(\mu, \sigma^2)$ and $\theta = (\theta_1, \theta_2) = (\mu, \sigma^2)$. Some functions g which may be of interest are $g(\theta) = \mu$, $g(\theta) = \sigma^2$ (or $g(\theta) = \sigma$), $g(\theta) = \mu/\sigma$ (or $g(\theta) = \sigma/\mu$, if $\mu \neq 0$) and $g(\theta) =$ the real number c such that $P_\theta(X_i < c) = \alpha$ (for some fixed $0 < \alpha < 1$) – i.e., $g(\theta) = \mu + z_\alpha \sigma$, where z_α is the normal α fractile.

Let us compute I. Since s consists of n iid parts, $I(\theta)$ for s is simply $nI_1(\theta)$, where $I_1(\theta)$ is I for X_1. If X_1 is the entire data, then

$$L = C - \frac{1}{2}\log\tau - \frac{1}{2\tau}(X_1 - \mu)^2,$$

where C is a constant and $\tau := \sigma^2 = \theta_2$; thus

$$L_1 = \frac{X_1 - \mu}{\tau} \quad \text{and} \quad L_2 = -\frac{1}{2\tau} + \frac{1}{2\tau^2}(X_1 - \mu)^2.$$

Homework 4

3. Check that

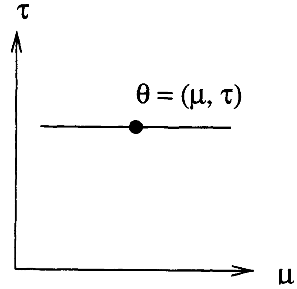
$$I_1(\theta) = \begin{pmatrix} 1/\tau & 0 \\ 0 & 1/2\tau^2 \end{pmatrix}.$$

Lecture 21

Example 3 (continued). We return to the situation $s = (X_1, \ldots, X_n)$; then

$$I(s) = n\begin{pmatrix} 1/\tau & 0 \\ 0 & 1/2\tau^2 \end{pmatrix} \quad \text{and} \quad I^{-1}(s) = \begin{pmatrix} \tau/n & 0 \\ 0 & 2\tau^2/n \end{pmatrix}.$$

Consider $g(\theta) = \mu = \theta_1$; then the most difficult one-dimensional problem is



This one-dimensional problem is in a one-parameter exponential family with sufficient statistic $\overline{X}$, and $\overline{X}$ is a UMVUE in this one-dimensional problem which attains the C-R bound – i.e., $\overline{X}$ is unbiased and $\mathrm{Var}_\theta(\overline{X}) = h(\theta)I^{-1}(\theta)h'(\theta)$, where $h = (1,0)$; thus

$$\mathrm{Var}_\theta(\overline{X}) = \tau/n \ \forall\theta \in \Theta.$$

The following are some gs (and their corresponding C-R bounds) for which the C-R bound is *not* attained:

i. $g(\theta) = \sigma^2$; the C-R bound is $\frac{2\tau^2}{n}$.

ii. $g(\theta) = \sigma$; the C-R bound is $\frac{\tau}{2n}$.

iii. $g(\theta) = \mu + z_\alpha\sigma$, $h = (1, z_\alpha/2\sqrt{\tau})$; the C-R bound is $\frac{\tau}{n} + \tau\frac{z_\alpha^2}{2n}$.

To see this, it is enough to check case (i), since the reasoning for the other cases is similar. Here

$$\ell(\theta \mid s) = C\tau^{-n/2}e^{-\frac{1}{2\tau}[n(\overline{X}-\mu)^2+nv]},$$

where C is a constant and $v = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$;

$$L(\theta \mid s) = C' - \frac{n}{2}\log\tau - \frac{1}{2\tau}\big[n(\overline{X}-\mu)^2 + nv\big],$$

where $C' = \log C$; $L_1(\theta \mid s) = \frac{n}{\tau}(\overline{X}-\mu)$ and

$$L_2(\theta \mid s) = -\frac{n}{2\tau} + \frac{1}{2\tau^2}\big[n(\overline{X}-\mu)^2 + nv\big].$$

Let $\delta = (\mu_*, \tau_*)$; then

$$E_\delta\big(L_1(\theta \mid s)\big) = \frac{n}{\tau}(\mu_* - \mu)$$

and

$$\begin{aligned} E_\delta\big(L_2(\theta \mid s)\big) &= -\frac{n}{2\tau} + \frac{1}{2\tau^2}\left[\tau_*(n-1) + n\frac{\tau_*}{n} + n(\mu_*-\mu)^2\right] \\ &= -\frac{n}{2\tau} + \frac{1}{2\tau^2}\big[n\tau_* + n(\mu_*-\mu)^2\big]. \end{aligned}$$

From these equations it is easily seen that there do *not* exist constants $a(\theta)$, $b(\theta)$ and $c(\theta)$ such that

$$E_\delta\big[a(\theta) + b(\theta)L_1(\theta \mid s) + c(\theta)L_2(\theta \mid s)\big] = \tau_*$$

for all $\delta = (\mu_*, \tau_*)$ – i.e., there is no unbiased estimate of τ_* in $\mathrm{Span}\{1, L_1(\theta \mid \cdot), L_2(\theta \mid \cdot)\}$, so that the C-R bound is not attainable for $g(\theta) = \tau$.

On the other hand, $\overline{X} = \mu + \frac{\tau}{n}L_1(\theta \mid s)$ is in $\mathrm{Span}\{1, L_1, L_2\}$ and is unbiased for μ, and so attains the C-R bound for μ. It is easy to check that the ML estimate is $\hat{\theta} = (\overline{X}, v)$, so the MLE for μ is $\overline{X}$; it is exactly unbiased, and its variation is the C-R bound. The MLE for $\tau = \sigma^2$ is $v = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$; we have that $E_\theta(v) = \frac{n-1}{n}\tau = \tau - \frac{\tau}{n}$ (note that $\frac{\tau}{n}$ is small when I is "large"),

$$\mathrm{Var}_\theta(v) = \frac{\tau^2}{n^2}\,\mathrm{Var}_\theta(X_{n-1}^2) = \frac{2(n-1)}{n^2}\tau^2,$$

which is *less* than the C-R bound $\frac{2\tau^2}{n}$ for τ (so v is *not* unbiased), and

$$\mathrm{MSE}_\theta(v) = \frac{2(n-1)}{n^2}\tau^2 + \frac{\tau^2}{n^2} = \frac{2\tau^2}{n} - \frac{\tau^2}{n^2} < \frac{2\tau^2}{n}.$$

Homework 4

4. The ML estimate for $\sigma = \sqrt{\tau}$ is $\sqrt{v}$. Show that $E_\theta(\sqrt{v}) = \sigma + o(1)$ and $\mathrm{Var}_\theta(\sqrt{v}) = \frac{\tau}{2n} + o(1)$ as $n \to \infty$. (HINT: z is an $X_k^2 \Leftrightarrow \frac{1}{2}z$ is a $\Gamma(k/2)$ variable. A $\Gamma(m)$ variable has density $\frac{e^{-x}x^{m-1}}{\Gamma(m)}$ in $(0,\infty)$. $\Gamma(m+1) = \sqrt{2\pi m}\cdot m^m e^{-m} + o(1/m)$ as $m \to \infty$, so
$$\frac{\Gamma(m+h)}{\Gamma(m)} = m^h\big(1 + o(1)\big)$$
as $m \to \infty$ for a fixed h.)

Lecture 22

Note. In the general case of $(S, \mathcal{A}, P_\theta)$, $\theta \in \Theta$, the above considerations are somewhat more general than are required for strict unbiased estimation. In particular, associated with each $\theta \in \Theta$ there is a set W_θ of estimates which has the following properties:

Corollary to (8). If we are estimating a scalar $g(\theta)$ corresponding to any *estimate t, then there is an estimate $\tilde{t} \in W_\theta$ such that $E_\delta(t) = E_\delta(\tilde{t})$ for all $\delta \in \Theta$ and*

$$E_\theta\big(t - g(\theta)\big)^2 =: R_t(\theta) \geq R_{\tilde{t}}(\theta) := E_\theta\big(\tilde{t} - g(\theta)\big)^2,$$

with the inequality strict unless $P_\delta(t = \tilde{t}) = 1$ for all $\delta \in \Theta$.

In general, W_θ depends on θ and we must be content with $C = \bigcap_{\theta\in\Theta} W_\theta$. In some important special cases, however – for example, in an exponential family – W_θ is independent of θ. In any case, though, the MLE and related estimates have the property that, if "$I(\theta)$" is large, any smooth function $f(\hat{\theta})$ is approximately in W_θ for any fixed θ.

Example 3 (continued). $\theta = (\mu, \tau)$, where $\tau = \sigma^2$. Choose and fix θ; then what is W_θ? There are three methods available:

Method 1. Look at $\Omega_{\delta,\theta}$. W_θ is the subspace spanned by $\{\Omega_{\delta,\theta} : \delta \in \Theta\}$.

Method 2. (Let θ be real, under regularity conditions.) $\frac{d^j}{d\delta^j}\Omega_{\delta,\theta}\big|_{\delta=\theta} \in W_\theta$. This is the method which leads to the Cramér-Rao and Bhattacharya inequalities.

Method 3. (Due to Stein.) $\int_{\delta_1}^{\delta_2} \Omega_{\delta,\theta} d\delta \in W_\theta$.

We use Method 2. Since $\ell(\theta \mid s) = e^{L(\theta|s)}$, we have $\ell_i(\theta \mid s) = e^{L(\theta|s)} L_i(\theta \mid s)$,

$$\ell_{ij}(\theta \mid s) = e^{L(\theta|s)} \big[L_{ij}(\theta \mid s) + L_i(\theta \mid s) L_j(\theta \mid s)\big],$$

etc., and hence $\ell_i/\ell = L_i$, $\ell_{ij}/\ell = L_{ij} + L_i L_j$, etc. Thus ℓ_i/ℓ, ℓ_{ij}/ℓ, etc. are in W_θ. Here we have

$$L_1 = \frac{n(\overline{X} - \mu)}{\tau} \qquad L_2 = \frac{n[v + (\overline{X} - \mu)^2]}{2\tau^2} - \frac{n}{2\tau}$$

$$L_{11} = -\frac{n}{\tau} \qquad L_{21} = -\frac{n(\overline{X} - \mu)}{\tau^2}$$

$$L_{12} = -\frac{n(\overline{X} - \mu)}{\tau^2} \qquad L_{22} = -\frac{n[v + (\overline{X} - \mu)^2]}{\tau^3} - \frac{n}{2\tau^2}.$$

Since $\ell_{11}/\ell = L_{11} + L_1^2$ is an affine function of $(\overline{X} - \mu)^2$, we have

$$\mathrm{Span}\{1, \overline{X}, v, (\overline{X} - \mu)^2\} = \mathrm{Span}\{1, L_1(\theta \mid \cdot), L_2(\theta \mid \cdot), \ell_{11}(\theta \mid \cdot)/\ell(\theta \mid \cdot)\} \subseteq W_\theta,$$

whence $\overline{X}$ is the LMVUE of $E_\delta(\overline{X}) = \mu_*$, v is the LMVUE of $E_\delta(v) = \frac{n-1}{n}\tau_*$ and $\frac{nv}{n-1}$ is the LMVUE of $E_\delta\big(nv/(n-1)\big) = \tau_*$ (remember $\delta = (\mu_*, \tau_*)$.) Since $\overline{X}$, v and $\frac{nv}{n-1}$ do not depend on θ, they are in fact in $C = \bigcap_{\theta \in \Theta} W_\theta$ and hence are the UMVUEs of their expected values. (Neither $\sqrt{v}$ nor $\frac{\overline{X}}{\sqrt{v}}$ (the latter is the MLE of μ/σ) is available by this method, but one can show by the above method that any function of $\overline{X}$ and v is in C. If Θ is the set of all pairs (μ, σ^2), then we are in the two-parameter exponential family case and a result to be stated later applies.)

Regularity conditions

Θ is open in $\mathbb{R}^p$ and $dP_\theta(s) = \ell(\theta \mid s) d\mu(s)$.

Condition 1^p. For each s, $\ell(\cdot \mid s)$ is a positive continuously differentiable function of θ.

Condition 2^p. Given any $\theta \in \Theta$, we may find an $\varepsilon = \varepsilon(\theta) > 0$ such that

$$\max\{|L_j(\delta \mid s)| : |\delta_i - \theta_i| \le \varepsilon\} \in V_\theta$$

(i.e., the function is square-integrable with respect to P_θ), or at least

$$\frac{\max\{|\ell_j(\delta \mid s)| : |\delta_i - \theta_i| \le \varepsilon\}}{\ell(\theta \mid s)} \in V_\theta.$$

Let $I(\theta) = E_\theta\big(L_i(\theta \mid s) L_j(\theta \mid s)\big)$.

Condition 3^p. For each θ, $I(\theta)$ is positive definite.

12^pE. a. For each θ, $1, L_1(\theta \mid s), \ldots, L_p(\theta \mid s) \subseteq W_\theta$, and $1 \perp L_j(\theta \mid s)$ in V_θ for $j = 1, \ldots, p$.

b. If U_g is non-empty, then g is differentiable and the projection of any $t \in U_g$ to $\operatorname{Span}\{1, L_1, \ldots, L_p\}$ (which is the projection of $\tilde{t}$ to $\operatorname{Span}\{1, L_1, \ldots, L_p\}$) is

$$t^*_{\theta,1} = g(\theta) + h(\theta)I^{-1}(\theta)\big(L_1(\theta \mid s), \ldots, L_p(\theta \mid s)\big)',$$

where $h(\theta) = \operatorname{grad} g(\theta)$.

c. If $t \in U_g$, then $\operatorname{Var}_\theta(t) \geq h(\theta)I^{-1}(\theta)h'(\theta)$ for all $\theta \in \Theta$.

Proof. The proof is left as an exercise for the reader. See the proof in the case $p = 1$ and use Approach 1 rather than Approach 2.

Note also that $g(\theta)$ is a projection of t to $\operatorname{Span}\{1\}$ and that 1 is orthogonal to $L_1, \ldots, L_p$, so that the projection of $t - g(\theta)$ to $\operatorname{Span}\{1, L_1, \ldots, L_p\}$ is the same as its projection to $\operatorname{Span}\{L_1, \ldots, L_p\}$. Thus

$$\begin{aligned}\operatorname{Var}_\theta\big(t - g(\theta)\big) &\geq E_\theta\big(\text{projection of } t - g(\theta) \text{ to } \operatorname{Span}\{L_1, \ldots, L_p\}\big)^2 \\ &= E_\theta\big(hI^{-1}(L_1, \ldots, L_p)'[hI^{-1}(L_1, \ldots, L_p)']'\big) \\ &= E_\theta\big(hI^{-1}(L_1, \ldots, L_p)'(L_1, \ldots, L_p)I^{-1}h'\big) = hI^{-1}h'.\end{aligned}$$

Lecture 23

Note. In the case when Θ is open in $\mathbb{R}^p$, $g : \Theta \to \mathbb{R}'$ is differentiable and conditions 1^p–3^p are satisfied, then, for any estimate t,

$$R_t(\theta) := E_\theta\big(t(s) - g(\theta)\big)^2 \geq \beta_t(\theta)I^{-1}(\theta)\beta_t'(\theta) + \big[b_t(\theta)\big]^2,$$

where $b_t(\theta) := E_\theta(t) - g(\theta)$ and $\beta_t(\theta) := \operatorname{grad} E_\theta(t) = \operatorname{grad} g(\theta) + \operatorname{grad} b_t(\theta)$.

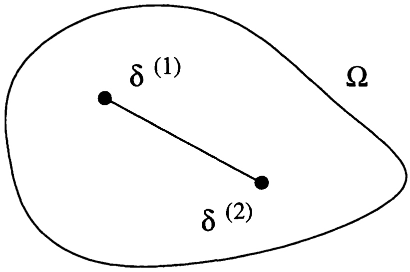
Proof. Let $\gamma(\delta) = E_\delta(t)$, so that $t \in U_\gamma$. Then

$$R_t(\theta) = \operatorname{Var}_\theta(t) + \big[b_t(\theta)\big]^2 \geq \big[\operatorname{grad}\gamma(\theta)\big]I^{-1}(\theta)\big[\operatorname{grad}\gamma(\theta)\big]'$$

by C-R bound. □

This result is useful even in case $p = 1$ – see, for example, the proof of the admissibility of $\hat{\theta}$ in Example 1(a) in Lehmann (1983, *Theory of point estimation*).

On the distance between θ and δ



Should one use the Euclidean distance d_1? What is really of interest is the "distance" between $P_{\delta^{(1)}}$ and $P_{\delta^{(2)}}$ – given, say, by

$$d_2(\delta^{(1)}, \delta^{(2)}) = \sup_{A \in \mathcal{A}} |P_{\delta^{(1)}}(A) - P_{\delta^{(2)}}(A)| = \frac{1}{2} \int_S \left| \ell(\delta^{(1)} \mid s) - \ell(\delta^{(2)} \mid s) \right| ds$$

or

$$d_3(\delta^{(1)}, \delta^{(2)}) = \int_S \left(\sqrt{\ell(\delta^{(1)} \mid s)} - \sqrt{\ell(\delta^{(2)} \mid s)} \right)^2 d\mu(s).$$

The distance d_3 is used in E. J. G. Pitman (1979, *Some basic theory of statistic inference*). It is related to the Fisher information in the following way:

Suppose that we want to distiguish between $P_{\delta^{(1)}}$ and $P_{\delta^{(2)}}$ on the basis of s. Instead of a hypothesis-testing approach, let us choose a real-valued function $t(s)$. What is the difference between $\delta^{(1)}$ and $\delta^{(2)}$ on the basis of t?

Regard t as an estimate of $g(\delta) := E_\delta(t)$. Then $|g(\delta^{(1)}) - g(\delta^{(2)})|$ might be taken as a measure of the distance between $\delta^{(1)}$ and $\delta^{(2)}$ on the basis of t. It is, however, more plausible to use the standardized versions

$$\frac{1}{\mathrm{SD}_{\delta^{(1)}}(t)} |g(\delta^{(1)}) - g(\delta^{(2)})| \quad \text{and} \quad \frac{1}{\mathrm{SD}_{\delta^{(2)}}(t)} |g(\delta^{(1)}) - g(\delta^{(2)})|,$$

especially if t is approximately normally distributed.

Now choose and fix $\theta \in \Theta$ and restrict δ to a small neighborhood of θ. Then $\mathrm{Var}_\delta(t) \approx \mathrm{Var}_\theta(t)$, and hence the distance (between $\delta^{(1)}$ and $\delta^{(2)}$, on the basis of t) is approximately

$$\frac{|g(\delta^{(1)}) - g(\delta^{(2)})|}{\sqrt{\mathrm{Var}_\theta(t)}} =: d_{t,\theta}(\delta^{(1)}, \delta^{(2)}).$$

Since the distance should be "intrinsic", we should maximize it with respect to t. First, we maximize $d_{t,\theta}$ with respect to t with the expectation function g fixed to get

$$\frac{|g(\delta^{(1)}) - g(\delta^{(2)})|}{\sqrt{(\operatorname{grad} g(\theta)) I^{-1}(\theta) (\operatorname{grad} g(\theta))'}}.$$

With $\delta^{(1)} \to \theta$ and $\delta^{(2)} \to \theta$, this is approximately

$$\frac{|(\delta^{(1)} - \delta^{(2)})[\operatorname{grad} g(\theta)]'|}{\sqrt{(\operatorname{grad} g(\theta)) I^{-1}(\theta) (\operatorname{grad} g(\theta))'}}.$$

Next, maximize the square of this with respect to $h(\theta) = \text{grad}\, g(\theta)$, which then leads to the squared distance

$$D_\theta^2(\delta^{(1)}, \delta^{(2)}) = (\delta^{(2)} - \delta^{(1)})I(\theta)(\delta^{(2)} - \delta^{(1)})'.$$

The distance D_θ is called the LOCAL FISHER METRIC in the vicinity of θ. It is the distance between $P_{\delta^{(1)}}$ and $P_{\delta^{(2)}}$ as measured in standard units for a real-valued statistic of the form $g(\hat{\theta})$, where g is suitably chosen so that $\text{grad}\, g(\theta) = (\delta^{(2)} - \delta^{(1)})I(\theta)$.

Example 1(a). Let $n = 1$, $s \sim N(\theta, \sigma^2)$ and $\theta \in \Theta = (-\infty, \infty)$, where σ^2 is a fixed known quantity. Then $I(\theta) = 1/\sigma^2$ for all θ,

$$D_\theta^2(\delta^{(2)}, \delta^{(1)}) = \frac{(\delta^{(2)} - \delta^{(1)})^2}{\sigma^2}$$

and

$$D = \frac{|\delta^{(2)} - \delta^{(1)}|}{\sigma} = \left|\frac{\text{mean of } P_{\delta^{(1)}} - \text{mean of } P_{\delta^{(2)}}}{\text{common SD}}\right|.$$

If $n > 1$ and $s = (X_1, \ldots, X_n)$ with the X_i iid, then

$$D_\theta(\delta^{(1)}, \delta^{(2)}) = \sqrt{n}\left|\frac{\delta^{(2)} - \delta^{(1)}}{\sigma}\right|.$$

For fixed $\theta \in \mathbb{R}^p$, D_θ is the metric derived from the inner product

$$(u|v)_* := \sum_{i,j} u_i I_{ij}(\theta) v_j = uI(\theta)v',$$

which has been used before. *Exercise* (informal): Look at D_θ in Example 3, $N(\theta_1, \theta_2)$.

Example 4. $Y \in \mathbb{R}^k$ has the $N_k(\theta, \Sigma)$ distribution and density

$$\ell(\theta \mid y) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{k/2}} e^{-\frac{1}{2}(y-\theta)\Sigma^{-1}(y-\theta)'}$$

with respect to Lebesgue measure. With this density, θ and Σ are respectively the mean and covariance matrices of Y. Show that $I(\theta) = \Sigma^{-1}$ for all θ and hence $D_\theta^2(\delta^{(2)}, \delta^{(1)})$ is the fixed square distance $(\delta^{(2)} - \delta^{(1)})\Sigma^{-1}(\delta^{(2)} - \delta^{(1)})$.

Lecture 24

Note. A sufficient condition for 13^p – i.e., the equality $I(\theta) = -\{E_\theta(L_{ij}(\theta \mid s))\}$ – is that, given any $\theta \in \Theta$, we may find an $\varepsilon = \varepsilon(\theta) > 0$ such that

$$\max\{|\ell_{ij}(\delta \mid s)/\ell(\delta \mid s)| : |\delta_i - \theta_i| \le \varepsilon\},$$

or at least

$$\max\{|\ell_{ij}(\delta \mid s)| : |\delta_i - \theta_i| \le \varepsilon\}/\ell(\theta \mid s),$$

be P_θ integrable (for $i, j = 1, \ldots, p$).

Note. The theory extends to estimation of vector-valued functions – for example, if $u(s) = \big(u_1(s), \dots, u_p(s)\big)$ is an unbiased estimate of θ and $\mathrm{Var}_\theta(u_i) < +\infty$ for each $i = 1, \dots, p$ and $\theta \in \Theta$, then $\mathrm{Cov}_\theta(u) - I^{-1}(\theta)$ is positive semidefinite for each $\theta \in \Theta$.

Proof. Fix $a = (a_1, \dots, a_p) \in \mathbb{R}^p$ and define $g(\theta) = \sum_{i=1}^p a_i\theta_i = a\theta'$. Then $t(s) = au'(s)$ is an unbiased estimate of g. Since $\operatorname{grad} g(\theta) = a$, we have

$$\mathrm{Var}_\theta(t) = a\,\mathrm{Cov}_\theta(u)a' \geq aI^{-1}(\theta)a',$$

so that ($a \in \mathbb{R}^p$ having been arbitrary) $\mathrm{Cov}_\theta(u) - I^{-1}(\theta)$ is positive semidefinite. □

Definition. $(S, \mathcal{A}, P_\theta)$, $\theta \in \Theta \subseteq \mathbb{R}^p$ is a (p-parameter) EXPONENTIAL FAMILY with statistic $T = (T_1, \dots, T_p) : S \to \mathbb{R}^p$ if $dP_\theta(s) = \ell(\theta \mid s)d\mu(s)$, where

$$\ell(\theta \mid s) = C(s)e^{B_1(\theta)T_1(s) + \cdots + B_p(\theta)T_p(s) + A(\theta)}.$$

The family is NON-DEGENERATE at a particular $\theta \in \Theta$ if

$$\big\{(B_1(\delta) - B_1(\theta), \dots, B_p(\delta) - B_p(\theta)) : \delta \in \Theta\big\}$$

contains a neighborhood of $0 = (0, \dots, 0)$.

We assume non-degeneracy at each $\theta \in \Theta$.

Exercise: Check that Example 1(a) is a non-degenerate exponential family with $p = 1$, with $T_1 = \overline{X}$ if $\Theta = \mathbb{R}^1$; Example 2(a) is a non-degenerate exponential family with $p = 1$, $T_1 = \overline{X}$ and $\Theta = (0, 1)$; Example 2(b) is a non-degenerate exponential family with $p = 1$, $T_1 = N$ and $\Theta = (0, 1)$; Example 3 is a two-parameter non-degenerate exponential family with $T_1 = \sum X_i$, $T_2 = \sum X_i^2$ and

$$\Theta = \{(\mu, \tau) : -\infty < \mu < +\infty \text{ and } 0 < \tau < +\infty\};$$

and Example 4 is a k-parameter exponential family with $T = \sum y_i = (T_1, \dots, T_k)$.

15^p.
a. For each $\theta \in \Theta$, W_θ is the space of all Borel functions of $T = (T_1, \dots, T_p)$ which are in V_θ.

b. $C = \bigcap_{\theta \in \Theta} W_\theta$ is the class of all UMVUE – i.e., the class of all Borel functions of T which are in $L^2(P_\theta)$ for all $\theta \in \Theta$.

c. For any g such that U_g is non-empty, there exists an essentially unique estimate $\tilde{t} = \tilde{t}(T) \in C \cap U_g$.

d. $\tilde{t} = E_\theta(t \mid T)$ for all $t \in U_g$ and $\theta \in \Theta$.

e. For all $A \subseteq S$, $E_\theta(I_A \mid T) = P_\theta(A \mid T)$ (essentially) is the same for each $\theta \in \Theta$, i.e., T is a sufficient statistic.

f. T is a complete statistic.

Proof.

a. Choose $\theta \in \Theta$ and write $\xi_i = B_i(\delta) - B_i(\theta)$. Then

$$\Omega_{\delta,\theta} = e^{\sum_{i=1}^{p} \xi_i T_i(s) - K_\theta(\xi_1,\ldots,\xi_p)},$$

where $K_\theta(\xi_1, \ldots, \xi_p) = \log E_\theta(e^{\sum \xi_i T_i(s)})$ is the cumulant generating function of T at $(\xi_1, \ldots, \xi_p)$ under P_θ. Non-degeneracy means that

$$K_\theta(\xi_1, \ldots, \xi_p) < +\infty$$

for $(\xi_1, \ldots, \xi_p)$ in a neighborhood of 0, and hence W_θ contains all functions $e^{\sum \xi_i T_i}$ for $(\xi_1, \ldots, \xi_p)$ in a neighborhood of 0. By differentiation, we find that W_θ contains all polynomials in $T_1, \ldots, T_p$, so W_θ contains all Borel functions of T which belong to V_θ.

On the other hand, since each $\Omega_{\delta,\theta}$ is a Borel function of T, every function in W_θ is such; so (a) is proved.

b. This follows from (a) and (9).

c. This follows from (a) and (8).

d. This follows from (a) and (8) and the fact that, if W is the space of all functions of T, projection to W is the conditional expectation given T.

e. This follows from (c) and (d) by letting $g(\theta) = P_\theta(A)$.

f. Suppose $E_\theta h = 0$ and $E_\theta h^2 < +\infty$ for all $\theta \in \Theta$. Then $h(T)$ is the UMVUE of $g(\theta) = 0$; but 0 is an unbiased estimate of this g, so $\mathrm{Var}_\theta\, h = 0$ for all $\theta \in \Theta$ and hence $P_\theta(h = 0) = 1$ for all $\theta \in \Theta$. □