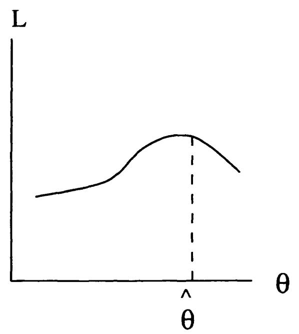
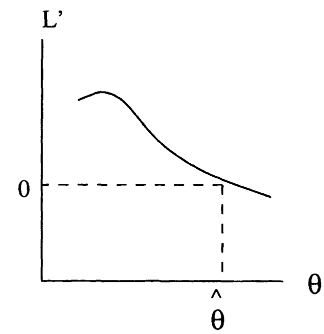
Chapter 7

Lecture 25

Using the score function (or vector)

Assume the usual setting, $(S, \mathcal{A}, P_\theta)$, $\theta \in \Theta \subseteq \mathbb{R}^p$.





First consider the case $p = 1$. Let $u(s)$ be a trial solution of $L'(\theta \mid s) = 0$. Assume that $\hat{\theta} - \theta = O(1/\sqrt{I(\theta)})$ and I is large. (Here θ is the true parameter, $E_\theta(\hat{\theta}) \approx 0$ and $\mathrm{Var}_\theta(\hat{\theta}) \approx 1/I(\theta)$.) Assume that u is not very inaccurate in the sense that, for any θ, $u(s) - \theta = O(1/\sqrt{I(\theta)})$. Then $\hat{\theta} - u = O(1/\sqrt{I(\theta)})$ under θ,

$$0 = L'\big(\hat{\theta}(s) \mid s\big) = L'(u(s) \mid s) + \big(\hat{\theta}(s) - u(s)\big)L''(u(s) \mid s) + O(1/I(\theta))$$

and

$$\hat{\theta}(s) = u(s) + \left(-\frac{1}{L''(u(s) \mid s)}\right)L'(u(s) \mid s) + O(1/I(\theta)).$$

Dropping the last term (order $1/I(\theta)$), we obtain the 'first Newton iterate' for solving $L'(\theta \mid s) = 0$.

Application 1. Let $u^{(0)}(s)$ be a trial solution of $L'(\theta \mid s) = 0$. Let

$$u^{(j+1)}(s) = u^{(j)}(s) + \left(-\frac{1}{L''(u^{(j)}(s) \mid s)}\right)L'(u^{(j)}(s) \mid s).$$

One hopes that $u^{(j)}(s) \to \hat{\theta}(s)$.

A variant of this approach consists in taking

$$u^{(j+1)}(s) = u^{(j)}(s) + \frac{1}{I(u^{(j)}(s))} L'(u^{(j)}(s) \mid s)$$

(since typically $-L''(\theta \mid s)/I(\theta) \approx 1$ if $I(\theta)$ is large).

Suppose we do not think it worthwhile to find $\hat{\theta}$ exactly.

Application 2. Start with a plausible estimate $u(s)$ of θ, and improve it to

$$u^*(s) = u(s) + \left(-\frac{1}{L''(u(s) \mid s)}\right) L'(u(s) \mid s)$$

or

$$u^{**}(s) = u(s) + \frac{1}{I(\theta)} L'(u(s) \mid s).$$

If $u - \theta = O(1/\sqrt{I(\theta)})$ and $E_\theta(u) - \theta = O(1/\sqrt{I(\theta)})$, then the first iterates have the same properties as $\hat{\theta}$, i.e., $u^* - \theta$ and $u^{**} - \theta$ are of order $1/\sqrt{I(\theta)}$ and $\mathrm{Var}_\theta(u^*)$ and $\mathrm{Var}_\theta(i^{**})$ are $b_1(\theta) = 1/I(\theta)$.

The case $p \geq 1$

Let $u(s) = (u_1(s), \ldots, u_p(s)) : S \to \Theta \subseteq \mathbb{R}^p$ be some plausible estimate of θ. Then

$$u^*(s) = u(s) + \{-L_{ij}(u(s) \mid s)\}^{-1} \{\mathrm{grad}\, L(\theta \mid s)\big|_{\theta = u(s)}\}$$

and

$$u^{**}(s) = u(s) + I^{-1}(u(s)) \{\mathrm{grad}\, L(\theta \mid s)\big|_{\theta = u(s)}\}$$

are versions of the first iteration of the Newton-Raphson method for solving $\mathrm{grad}\, L(\theta \mid s) = 0$.

Let $||\cdot||$ be the Euclidean norm. If $||u - \theta||$ and $||\hat{\theta} - \theta||$ are of the same order and $E_\theta(\hat{\theta}) \approx \theta$ and $\mathrm{Cov}_\theta(\hat{\theta}(s)) \approx I^{-1}(\theta)$, then u^* and u^{**} also have these properties – i.e., $E_\theta(u^*) \approx \theta$ and $\mathrm{Cov}_\theta(u^*(s)) \approx I^{-1}(\theta)$ (and similarly for u^{**}).

Example 1. $s = (X_1, \ldots, X_n)$, with the X_i iid with density $f(x - \theta)$ for $\theta \in \mathbb{R}^1$.

a. f is the normal density. $I(\theta) = n$, $L'(\theta \mid s) = n(\overline{X} - \theta)$ and $L''(\theta \mid s) = -n$. For any u, the first iteration gives $u^* = \overline{X} = u^{**}$.

b. $f(x) = \frac{1}{2}e^{-|x|}$. We know from the homework that $\hat{\theta}$ is the median of $X_1, \ldots, X_n$. Here L' and I do not exist, but the Chapman-Robbins bound gives $\mathrm{Var}_\theta(t) \geq \frac{1}{n}$ for any unbiased estimate t of g. Show that $\mathrm{Var}_\theta(\hat{\theta}) = \frac{1}{n} + O(\frac{1}{n^2})$. (Note that

$$\mathrm{Var}_\theta(\overline{X}) = \frac{1}{n}\mathrm{Var}_\theta(X_1) = \frac{1}{n}\int \frac{x^2}{2} e^{-|x|} dx = \frac{1}{n}\int_0^\infty x^2 e^{-x} dx = \frac{\Gamma(3)}{n} = \frac{2}{n},$$

so that the variance bound is true for $\overline{X}$.)

c. $f(x) = \frac{1}{\pi}\frac{1}{1+x^2}$. Here $I_1(\theta) = \frac{1}{2}$ and $I(\theta) = \frac{n}{2}$. $\hat{\theta}$ is hard to find (there are many roots of $L'(\theta \mid s) = 0$).

$$L(\theta \mid s) = C - \sum_{i=1}^{n} \log\left[1 + (X_i - \theta)^2\right],$$

where C is a constant, and

$$L'(\theta \mid s) = \sum_{i=1}^{n} \frac{2(X_i - \theta)}{1 + (X_i - \theta)^2}.$$

Let $u(s)$ be the median of $\{X_1, \ldots, X_n\}$; then

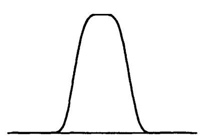
$$u^{**}(s) = u(s) + \frac{4}{n}\sum_{i=1}^{n} \frac{X_i - u(s)}{1 + (X_i - u(s))^2}.$$

Since it is true that $u(s) - \theta$ is $O(1/\sqrt{n})$, we have $E_\theta(u^{**}) \approx \theta$ and $\mathrm{Var}_\theta(u^{**}) \approx \frac{2}{n}$, the information bound.

e. $f(x) = ae^{-bx^4}$, $a, b > 0$, and $\mathrm{Var}(x) = 1$. Here, as in (c) above, it is difficult to find W_θ, and $W_{\theta,1}$ and $W_{\theta,2}$ look awful. $\overline{X}$ is a plausible estimate since $E_\theta(\overline{X}) = \theta$ and $\mathrm{Var}_\theta(\overline{X}) = \frac{1}{n} = O(1/I(\theta))$ $(I(\theta) = n)$.

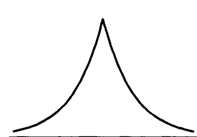
The most important differences among the above four densities are the different tail behaviors:

1(e). Short tail: Here a good estimate gives more weight to the extreme values than to the central values.

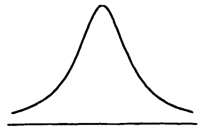


1(a). Normal: Here the best estimate $\overline{X}$ gives equal weight to all observations.

1(b). Double exponential: Here the best estimate, the median, gives weights concentrated in the middle.



1(c). CAUCHY: Here the optimal estimate(s) is (are) unknown.



Lecture 26

Continuing Example 1(e)

$$\ell(\theta \mid s) = a^n e^{-b\sum_{i=1}^n (X_i-\theta)^4} = \varphi(s)e^{-b[-4\theta\sum X_i^3+6\theta^2\sum X_i^2-4\theta^3\sum X_i]+A(\theta)}$$
$$= \varphi(s)e^{B_1(\theta)m_3'+B_2(\theta)m_2'+B_1(\theta)m_1'+A(\theta)},$$

where $m_j' = \frac{1}{n}\sum_{i=1}^n X_i^j$ for $j = 1, 2, 3$ (notice that $m_1' = \overline{X}$). This is not a three-parameter exponential family but a curved exponential family; but (m_1', m_2', m_3') is equivalent to $(\overline{X}, m_2, m_3)$, where $m_j = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^j$ for $j = 2, 3$, which is the minimal sufficient statistic – i.e., $(\overline{X}, m_2, m_3)$ is an adequate summary of data (for any statistical purpose) and nothing less will do. (In Example 1(a), $\overline{X}$ is the minimal sufficient statistic, and, in Example 3, $(\overline{X}, m_2)$ is the minimal sufficient statistic.)

$L'(\theta) = 4b\sum_{i=1}^n (X_i - \theta)^3$. Let $\hat{\theta} = \overline{X} + zm_2^{1/2}$. Since $L'(\hat{\theta}) = 0$, we have

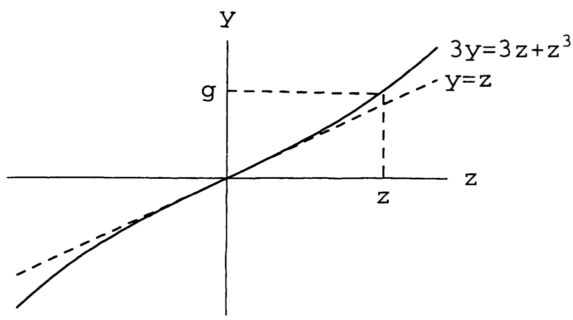
$$z + \frac{1}{3}z^3 = \frac{1}{3}\gamma_1,$$

where $\gamma_1 = m_3/m_2^{3/2}$ is the sample coefficient of kurtosis.

There are several approaches to getting $\hat{\theta}$:

Approach 3. Get an explicit form of z from the equation in z above, and substitute it into the expression for $\hat{\theta}$ in terms of z.

Approach 4. The graphic method:



(In the picture above, $g = \gamma_1/3$.) $0 < z < \frac{1}{3}\gamma_1$ if $\gamma_1 > 0$ and $\frac{1}{3}\gamma_1 < z < 0$ if $\gamma_1 < 0$; so a solution is

$$z = \frac{1}{3}\gamma_1 - \frac{\eta}{27}\gamma_1^3,$$

where $0 < \eta < 1$.

Approach 5.

$$\hat{\theta} \approx \overline{X} + \frac{1}{3}\gamma_1 m_2^{1/2} = \overline{X} + \frac{1}{3}\frac{m_3}{m_2}.$$

Note that, if n is large, then $m_2 \approx 1$ (since $\mathrm{Var}_\theta(X_1) = 1$) and so $\hat{\theta} \approx \overline{X} + \frac{1}{3}m_3$ (so outliers are given more weight than given by $\overline{X}$). Here $I(\theta) = n$ and $\mathrm{Var}_\theta(\overline{X}) = \frac{1}{n} = O\big(\frac{1}{I(\theta)}\big)$, so $\overline{X}$ is an acceptable starting value for approximating $\hat{\theta}$.

Approach 6. $u^* = \overline{X} + \frac{1}{3}\frac{m_3}{m_2}$ and $u^{**} = \overline{X} + \frac{1}{3}m_3$ (please check). It is not easy to find the exact properties of $\hat{\theta}$, u^* and u^{**}, but u^{**} is the easiest to examine.

Homework 5

1. Show that $E_\theta(u^{**}) = \theta$ and

$$\mathrm{Var}_\theta(u^{**}) = b_1(\theta) + O\left(\frac{1}{n^2}\right) = \frac{1}{12bn} + O\left(\frac{1}{n^2}\right) = \frac{1}{1.37n} + O\left(\frac{1}{n^2}\right)$$

(so that $E_\theta(m^3) = 0$ and $\mathrm{Cov}_\theta(\overline{X}, m_3) < 0$).

Since m_3 is a function of the (minimal) sufficient statistic $T(s) = (\overline{X}, m_2, m_3)$, this statistic is not complete. Since $\mathrm{Cov}_\theta(\overline{X}, m_3) \neq 0$ (m_3 is an unbiased estimate of 0), we know that $\overline{X}$ is not even locally MVUE. (See Kendall and Stuart, vol. I, for "standard error of moments". A good reference to the use of the score function in general is C. R. Rao's *Linear Statistical Inference.*)

Example 5. Our state space is $\{1, 2\}$ and the transition probability matrix is

$$\begin{pmatrix} \theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22} \end{pmatrix} = \begin{pmatrix} \theta_1 & 1-\theta_1 \\ 1-\theta_2 & \theta_2 \end{pmatrix}.$$

Suppose first that $\Theta = (0,1) \times (0,1)$ and that a Markoff chain with transition probability matrix as above starts at '1' and is observed for n one-step transitions. Thus $s = (X_0, X_1, \ldots, X_n)$, where $X_0 \equiv 1$, and

$$\ell(\theta \mid s) = \prod_{i,j=1,2} \theta_{ij}^{f_{ij}(s)} = \theta_1^{f_{11}(s)}(1-\theta_1)^{f_{12}(s)}\theta_2^{f_{22}(s)}(1-\theta_2)^{f_{21}(s)},$$

where $f_{ij}(s)$ is the number of one-step transitions from i to j in s. Since $f_{11}+f_{12}+f_{22}+f_{21} = n$, we have a three-dimensional minimal sufficient statistic and two parameters. If $f_{21} + f_{22} > 0$, $f_{11} > 0$ and $f_{22} > 0$, then we have (noticing that $f_{11} + f_{12} > 0$) $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2)$, where $\hat{\theta}_1 = \frac{f_{11}}{f_{11}+f_{12}}$ and $\hat{\theta}_2 = \frac{f_{22}}{f_{21}+f_{22}}$. Since

$$L_1 = \frac{f_{11}}{\theta_1} - \frac{f_{12}}{1-\theta_1}, \quad L_2 = \frac{f_{22}}{\theta_2} - \frac{f_{21}}{1-\theta_2}, \quad L_{11} = -\frac{f_{11}}{\theta_1^2} - \frac{f_{12}}{(1-\theta_1)^2}, \ldots,$$

we have $E_\theta(L_1) = 0 = E_\theta(L_2)$ (since $(1-\theta_1)E_\theta(f_{11}) = \theta_1 E_\theta(f_{12})$, etc.) and

$$I(\theta) = \begin{pmatrix} E_\theta(f_{11}/\theta_1^2 + f_{12}/(1-\theta_1)^2) & 0 \\ 0 & E_\theta(f_{21}/(1-\theta_2)^2 + f_{22}/\theta_2^2) \end{pmatrix}.$$

It is known that

$$E_\theta(f_{ij}) = n\pi_i(\theta)\theta_{ij} + o(n) \quad \text{as } n \to \infty$$

where $\pi_1(\theta)$ and $\pi_2(\theta)$ are the stationary distrtibution over $\{1, 2\}$ and

$$\pi_1(\theta) = \frac{1-\theta_2}{2-(\theta_1+\theta_2)} \qquad \text{and} \qquad \pi_2(\theta) = \frac{1-\theta_1}{2-(\theta_1+\theta_2)},$$

so

$$I(\theta) = n\begin{pmatrix} \pi_1(\theta)/\theta_1(1-\theta_1) & 0 \\ 0 & \pi_2(\theta)/\theta_2(1-\theta_2) \end{pmatrix} + o(n).$$

The information bound for the variances of estimates of θ_1 is $\frac{\theta_1(1-\theta_1)}{n\pi_1(\theta)}$ (and similarly for θ_2). Is $\mathrm{Var}_\theta(\hat{\theta}_1) \approx \frac{\theta_1(1-\theta_1)}{n\pi_1(\theta)}$? It can be shown (though not easily) that $\mathrm{Var}_\theta(\hat{\theta}_1) = b_1(\theta) + o(1/n)$ as $n \to \infty$, where $b_1(\theta)$ is the C-R bound.

Lecture 27

In Example 5, Θ is an open unit square consisting of points (θ_1, θ_2). Let $\hat{\theta}_1 = \frac{f_{11}}{f_{11}+f_{12}}$ and $\hat{\theta}_2 = \frac{f_{22}}{f_{21}+f_{22}}$ if $f_{ij} > 0$ for all i, j. Otherwise, let $\hat{\theta}_2$ be arbitrary – say $\frac{1}{2}$, for convenience. It can be shown that

$$P_\theta(f_{ij} > 0 \ \forall i, j) \geq 1 - \left[p(\theta)\right]^n$$

for all sufficiently large n and some fixed $0 < p(\theta) < 1$. Hence we can ignore the case $f_{ij} = 0$ in the computation of $E_\theta(\hat{\theta})$ and $\mathrm{Var}_\theta(\hat{\theta})$.

Suppose we know that $\theta_2 = k\theta_1$ for some $0 < k < \infty$; then now $\Theta = \{\theta_1 : 0 < \theta_1 < 1/k\}$ and

$$L \propto f_{11}\log\theta_1 + f_{12}\log(1-\theta_1) + f_{21}\log(1-k\theta_1) + f_{22}\log k\theta_1.$$

Exercise: Show that I in the present case is greater than I in the previous case, for sufficiently large n. (Recall that $E_\theta(f_{ij}) = n\pi_i(\theta)\theta_{ij} + o(n)$.)

The equation for $\hat{\theta}_1$ is now a cubic. We can solve it explicitly, or we can approximate it by $v = u^*$ or u^{**}, with $u = \frac{f_{11}}{f_{11}+f_{12}}$ (say). Then we have $E_\theta(v) = \theta_1 + o(1)$ and $\mathrm{Var}_\theta(v) = 1/(n \cdot \text{present } I) + o(1)$.

A special case of the above is when $\theta_1 = \theta_2$ – i.e., $k = 1$ – so that

$$\ell(\theta \mid s) = \theta_1^{f_{11}(s)+f_{22}(s)}(1-\theta_1)^{f_{12}(s)+f_{21}(s)} = \theta_1^{y(s)}(1-\theta_1)^{n-y(s)},$$

where of course $y = f_{11} + f_{22}$. It turns out that y is a $B(n, \theta_1)$ variable, so that $\hat{\theta}_1 = y/n$ satisfies $\mathrm{Var}_\theta(\hat{\theta}_1) = \frac{1}{n}\theta_1(1-\theta_1)$. This is the new I^{-1}.

Example 6. $X_i \sim N(0, 1)$, $\Theta = (0, 1)$.

a. $\mathrm{Cov}_\theta(X_i, X_j) = \theta^{j-i}$ for all $i < j$.

$$u_1 = \frac{X_1X_2 + X_2X_3 + \cdots + X_{n-1}X_n}{n-1}$$

is unbiased for θ and

$$u_2 = \frac{X_1X_3 + X_2X_4 + \cdots + X_{n-2}X_n}{n-2}$$

is unbiased for θ^2. $u_1 + k\sqrt{u_2}$ is an estimate of θ; what are its properties?

b. $\mathrm{Cov}_\theta(X_i, X_j) = \theta$ for all $i \neq j$.

In both cases $(X_1, \ldots, X_n)$ is from a stationary sequence. What is $I(\theta)$ in 6(a) and 6(b)? What estimate(s) t ($t = \hat{\theta}$? $t = u^*$? $t = u^{**}$?) has (have) the property that $E_\theta(t) \approx \theta$ and $\mathrm{Var}_\theta(t) \approx I^{-1}(\theta)$ for large n?

In 6(a), find $|C|$ and C^{-1}, where

$$C = \mathrm{Cov}_\theta(s) = \begin{pmatrix} 1 & \theta & \cdots & \theta^{n-1} \\ \theta & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \theta \\ \theta^{n-1} & \cdots & \theta & 1 \end{pmatrix}.$$

(C^{-1} is tridiagonal.) In 6(b), find $|D|$ and D^{-1}, where

$$D = \begin{pmatrix} 1 & \theta & \cdots & \theta \\ \theta & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \theta \\ \theta & \cdots & \theta & 1 \end{pmatrix}.$$

($D = (1-\theta)I + \theta u$, where $u = \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix}$, so $D^{-1} = \alpha I + \beta u$.)

Homework 5

2. (Optional) Answer the questions in Example 6.

A review of the preceding heuristics

Suppose θ is real.

i. CONSISTENCY: $\hat{\theta}$ is close to the true θ.

ii. $E_\theta(\hat{\theta}) \approx \theta$; in fact, $E_\theta(\hat{\theta}) = \theta + O(1/\sqrt{I(\theta)})$.

iii. $\mathrm{Var}_\theta(\hat{\theta}) = \frac{1}{I(\theta)} + o\left(\frac{1}{I(\theta)}\right)$.

If u is any estimate such that $u = \theta + O\left(\frac{1}{\sqrt{I(\theta)}}\right)$, then u^*, u^{**} etc. also have properties (ii) and (iii).

Consistency is difficult even today. Assuming that $\hat{\theta}$ exists and is consistent, then (ii) and (iii) remain difficult, but one can say that $\hat{\theta}$, u^*, u^{**}, etc. are $\approx N\big(\theta, 1/I(\theta)\big)$ where $I(\theta)$ is large.

Theorem (on consistency). *Let X_i be iid. $\ell(\theta \mid X_i)$ depends on $\theta \in \Theta = (a, b)$ with $-\infty \le a < b \le +\infty$, and $\ell(\theta \mid s) = \prod_{i=1}^n \ell(\theta \mid X_i)$.*

Condition 1. For all s, $\ell(\cdot \mid s)$ is continuous.

Let $\hat{\theta}_n : S \to \Theta$ be some function; $\hat{\theta}$ is an ML estimate $\Leftrightarrow$ $\hat{\theta}$ is measurable and

$$\ell\big(\hat{\theta}(s) \mid s\big) = \sup_{\delta \in \Theta} \ell(\delta \mid s)$$

whenever the supremum exists.

Condition 2. $\lim_{\theta \to a} \ell(\theta \mid X_1)$ and $\lim_{\theta \to b} \ell(\theta \mid X_1)$ exist a.e. with respect to the dominating measure for X_1; denote these limits by $\ell(a \mid X_1)$ and $\ell(b \mid X_1)$.

Condition 3. If $\theta \in \Theta$, then

$$\{x_1 : \ell(\theta \mid x_1) \ne \ell(a \mid x_1)\}$$

and

$$\{x_1 : \ell(\theta \mid x_1) \ne \ell(b \mid x_1)\}$$

have positive measures (with respect to the dominating measure for X_1). For any $\theta, \delta \in \overline{\Theta}$ with $\theta \ne \delta$,

$$\{x_1 : \ell(\theta \mid x_1) \ne \ell(\delta \mid x_1)\}$$

has positive measure.

1 (LeCam). Condition 1 implies that an ML estimate exists.

2 (Wald). Conditions 1–3 imply that, for all $\theta \in \Theta$, with probability 1,

1. $\hat{\theta}_n$ actually maximizes the likelihood for all sufficiently large n.

2. $\lim_{n \to \infty} \hat{\theta}_n = \theta$.

Note. The proof of (2) depends on the fact that $[a, b]$ is compact. There are difficulties in extending the proof to, say, $\Theta \subseteq \mathbb{R}^p$, because it is difficult to find a suitable compactification of Θ.