Chapter 5: Likelihood estimation JAW Ch.4, Sev. Ch. 4

5.1 Maximum likelihood estimation: basics

(i) Suppose:
(a) A0: $X_1, ..., X_n$ **are i.i.d.** P_θ**:** $\theta \in \Theta \subset \Re^k$
(b) A1: Identifiability: $\theta \neq \theta^* \quad \Rightarrow \quad P_\theta \neq P_{\theta^*}$
(c) A2: P_θ **has density** $f(\cdot;\theta)$ **w.r.t** σ**-finite** μ**.**
(d) A3: $A = \{x : f(x;\theta) > 0\}$ **does not depend on** θ**.**

(ii) Given A0,A1,A2:
the ***likelihood*** $L_n(\theta) = L(\theta; X^{(n)}) = f_n(X^{(n)};\theta) = \Pi_{i=1}^n f(X_i;\theta)$
The log-likelihood $\ell_n(\theta) = \log_e L_n(\theta) = \Sigma_1^n \log f(X_i;\theta)$
For set $B \subset \Theta$**,** $\ell_n(B) \equiv \sup_{\theta \in B} \ell_n(\theta)$**.**

(iii) Given A0,A1,A2: the value $\widehat{\theta_n}$ **of** θ **which maximises the likelihood** $L_n(\theta)$**, if it exists and is unique, is the maximum likelihood estimator (MLE) of** θ**. Note** $\ell_n(\Theta) = \ell_n(\widehat{\theta_n})$**.**

(iv) Given A0-A3, and differentiability of $L_n(\theta)$ **the MLE may be found by solving the** ***likelihood equation*** **or** ***score equation*****,** $\bigtriangledown \ell_n(\theta) = 0$**. (However, this equation may have no roots in** Θ**, or multiple roots.)**

(v) Basic properties:
(a) the MLE depends only on the minimal sufficient statistic
(b) MLE's are NOT necessarily unbiased − if consistent, then unbiased in the limit ($\mathrm{E}(T_n) \rightarrow q(\theta)$**, provided** $\mathrm{E}(T_n) < M < \infty$**) (cf TPE "asymptotically unbiased").**
(c) If an unbiased estimator attaining the CRLB exists, it is the MLE.
(d) If $q(\theta)$ **is any 1-1 function of** θ**,** $\widehat{q(\theta)} = q(\hat{\theta})$**.**

5.2 Kullback-Leibler Information (JAW 4.3)

(i) Defn: Let P and Q be probability measures (Q may be sub-prob.meas.) with densities p and q. Then $K(P,Q) \equiv$ $\mathrm{E}_P(\log(p(X)/q(X)))$.

(ii) $K(P,Q)$ is well-defined, and ≥ 0 (possibly ∞), and $=0$ iff $Q = P$. (Proof by Jensen's inequality, or by $\log x \leq (x-1)$.)

(iii) If A0-A3, the SLLN gives, under P_{θ_0}, for $\theta \neq \theta_0$

$$\frac{1}{n}\log\frac{L(\theta_0 : X^{(n)})}{L(\theta; X^{(n)})} \;=\; \frac{1}{n}\sum_1^n \log\frac{P_{\theta_0}(X_i)}{P_\theta(X_i)} \;\rightarrow_{a.s.} K(P_{\theta_0}, P_\theta) \;>\; 0$$

(iv) Thus $P_{\theta_0}(L(\theta_0 : X^{(n)}) \;>\; L(\theta; X^{(n)})) \rightarrow 1$ as $n \rightarrow \infty$. This motivates the definition of $\widehat{\theta_n}$, but

"Likelihood is a pointwise function on Θ"

To proceed we need some metric/uniformity/smoothness w.r.t θ.

(iv) A4: $\Theta \supset \Theta_0$, an open set in $\Re^k$, and for $\theta \in \Theta_0$:
(a) $\ell(\theta; x) \;\equiv\; \log p_\theta(x)$ is twice continuously diffble in θ (μ $(a.e.)$ x)
(b) μ $(a.e.)$ x, third order derivatives exist, with $\frac{\partial^3 \ell}{\partial\theta_j \partial\theta_l \partial\theta_u}$ bounded by $M_{jlu}(x)$ for all $\theta \in \Theta_0$, and $\mathrm{E}_{\theta_0}(M_{jlu}(X)) < \infty$ for all $j, l, u = 1, ..., k$

(v) A5: (a) $\mathrm{E}_{\theta_0}(\bigtriangledown \ell(\theta; X)|_{\theta=\theta_0}) \;=\; 0$.
(b) $\mathrm{E}_{\theta_0}((\bigtriangledown \ell(\theta; X))^t(\bigtriangledown \ell(\theta; X))|_{\theta=\theta_0}) \;=\; \Sigma_{j=1}^k \left(\frac{\partial l}{\partial \theta_j}\right)^2 \;< \infty$.
(c) $I(\theta_0) = -\left(\mathrm{E}_{\theta_0}(\frac{\partial^2 \ell}{\partial\theta_j \partial\theta_l})|_{\theta=\theta_0}\right)$ is positive definite.

5.3 Consistency of MLE (JAW 4.7, Severini 4.2)

(i) Theorem: Suppose A0-A5. Then, with probability $\to 1$ as $n \to \infty$, $\exists$ solution $\tilde{\theta_n}$ of the likelihood equations s.t. $\tilde{\theta_n} \to_p \theta_0$ when P_{θ_0} is true.

(ii) For $a > 0$ let $Q_a \equiv \{\theta \in \Theta : |\theta - \theta_0| = a\}$. We show below that, provided $Q_a \subset \Theta_0$, then $P_{\theta_0}(\sup_{\theta \in Q_a} \ell(\theta) < \ell(\theta_0)) \to 1$ as $n \to \infty$. Hence there is a local max, which must be root of the likelihood eqn, inside Q_a.
(Note $\sup_{Q_a} \ell(\theta)$ is attained on Q_a by some $\theta \in Q_a$.)

(iii) Define "observed information" $J(\theta_0) = -\left(\frac{\partial^2 \ell}{\partial \theta_i \partial \theta_j}|_{\theta=\theta_0}\right)$.
(Note that $J(\theta)$, unlike $I(\theta)$, is a r.v.)

(iv) $n^{-1}(\ell(\theta; X^{(n)}) - \ell(\theta_0; X^{(n)}))$

$$
\begin{aligned}
&= n^{-1}(\theta - \theta_0)^t \bigtriangledown (\ell(\theta_0)) + \frac{1}{2}(\theta - \theta_0)^t(-n^{-1}J(\theta_0))(\theta - \theta_0) \\
&\quad +(6n)^{-1} \sum_{jlu}(\theta_j - \theta_{0,j})(\theta_l - \theta_{0,l})(\theta_u - \theta_{0,u}) \sum_i \gamma_{jlu}(X_i) M_{jlu}(X_i) \\
&= S_1 + S_2 + S_3
\end{aligned}
$$

where $|\gamma_{jlu}| < 1$ (by A4 (b)).

(v) By A5 and WLLN: $S_1 \to_p 0$, $-2S_2 \to_p (\theta - \theta_0)^t I(\theta_0)(\theta - \theta_0) \geq \lambda_k a^2$ where λ_k is smallest eigenvalue of $I(\theta_0)$.
$S_3 \to_p (1/6) \Sigma_{jlu}(\theta_j - \theta_{0,j})(\theta_l - \theta_{0,l})(\theta_u - \theta_{0,u}) \mathrm{E}_{\theta_0}(\gamma_{jlu}(X_1) M_{jlu}(X_1))$
and $|S_3| \leq (1/3)(ka)^3 \Sigma_{jlu} m_{jlu} \ \equiv \ Ba^3$ as $n \to \infty$.

(vi) For large enough n

$$
\begin{aligned}
\sup_{\theta \in Q_a}(S_1 + S_2 + S_3) &\leq \sup_{\theta \in Q_a}|S_1 + S_3| + \sup_{\theta \in Q_a}(S_2) \\
&< (k + B)a^3 - \lambda_k a^2/4 \\
&< 0 \quad \text{for small enough } a
\end{aligned}
$$

5.4 Aymptotic normality and efficiency of $\tilde{\theta_n}$. (Sev. 4.2.2)

(i) Notation

$Z_n \equiv n^{-\frac{1}{2}} \sum_i \bigtriangledown(\ell(\theta_0; X_i)) \;=\; n^{-\frac{1}{2}} \bigtriangledown \ell_n(\theta_0; X^{(n)})$,
$\tilde{\ell}(\theta_0; X) \equiv I^{-1}(\theta_0) \bigtriangledown \ell(\theta_0; X)$ **so** $n^{-\frac{1}{2}} \sum_{i=1}^n \tilde{\ell}(\theta_0; X_i) = I^{-1}(\theta_0) Z_n$.
Define $G_n(\epsilon) \equiv \{\tilde{\theta_n}; \bigtriangledown \ell_n(\tilde{\theta_n}) = 0, |\tilde{\theta_n} - \theta_0| < \epsilon\}$ **non-empty as** $n \to \infty$, $\forall \epsilon > 0$. **(Also note here** $I(\theta) \equiv I_1(\theta)$.**)**

(ii) Theorem

$$(a) \qquad (n^{\frac{1}{2}}(\tilde{\theta_n} - \theta_0) \;-\; n^{-\frac{1}{2}} \sum_{i=1}^n \tilde{\ell}(\theta_0; X_i)) \;\;\to_p\;\; 0$$

$$(b) \qquad n^{-\frac{1}{2}} \sum_{i=1}^n \tilde{\ell}(\theta_0; X_i) \;\to_d\; I^{-1}(\theta_0) Z \;\equiv\; D \;\sim N_k(0, I^{-1}(\theta_0)).$$

(iii) First (b): by CLT $Z_n \to_d N(0, I(\theta_0))$,
so $n^{-\frac{1}{2}} \sum_{i=1}^n \tilde{\ell}(\theta_0; X_i) = I^{-1}(\theta_0) Z_n \to_d N(0, I^{-1}(\theta_0))$

(iv) On G_n,

$$0 = n^{-\frac{1}{2}} \bigtriangledown \ell_n(\tilde{\theta_n}) \;=\; n^{-\frac{1}{2}} \bigtriangledown \ell_n(\theta_0) - n^{-1} J_n(\theta_n^*) n^{\frac{1}{2}}(\tilde{\theta_n} - \theta_0)$$

where $|\theta_n^* - \theta_0| < |\tilde{\theta_n} - \theta_0|$.
Or $n^{\frac{1}{2}}(\tilde{\theta_n} - \theta_0) = (n^{-1} J_n(\theta_n^*))^{-1} Z_n$ **if** $J_n^{-1}(\theta_n^*)\ \exists$.

(v)$\tilde{\theta_n} \to_p \theta_0$, **so using one-term expansion of 2 nd. deriv, and boundedness of 3 rd.,** $n^{-1}(J_n(\theta_n^*) - J_n(\theta_0)) \to_p 0$. **By continuity,** $(n^{-1} J_n(\theta_n^*))^{-1} \to_p (\mathrm{E}(J_1(\theta_0)))^{-1} \;=\; I^{-1}(\theta_0)$. **(and** $J_n(\theta_n^*)$ **is pos def with prob approaching 1).**
Now $I^{-1}(\theta_0) Z_n = n^{-\frac{1}{2}} \sum_{i=1}^n \tilde{\ell}(\theta_0; X_i)$ **hence (a).**

(vi) Transforming (ii) to $q(\theta)$**: dim(**q**)=** k^*, $1 \leq k^* \leq k$

$$(a) \qquad (n^{\frac{1}{2}}(q(\tilde{\theta_n}) - q(\theta_0)) \;-\; n^{-\frac{1}{2}} \sum_{i=1}^n \tilde{\ell}_q(\theta_0; X_i)) \;\;\to_p\;\; 0$$

$$(b) \qquad n^{-\frac{1}{2}} \sum_{i=1}^n \tilde{\ell}_q(\theta_0; X_i) \;\to_d\; N_{k^*}(0, (\bigtriangledown q(\theta_0))^t I^{-1}(\theta_0)(\bigtriangledown q(\theta_0))).$$

where $\tilde{\ell}_q(\theta_0; X_i) \;=\; (\bigtriangledown q(\theta_0))^t \; I^{-1}(\theta_0)(\bigtriangledown \ell(\theta_0, X_i))$

5.5 Bits and pieces

5.5.1 Estimation of $I(\theta)$: Suppose we need to estimate $I(\theta_0)$, and have A0-A5, as above, so ℓ_n is twice continuously diffble, and expectations $\exists$:

(a) $\tilde{\theta_n} \to_p \theta_0$, so $I(\tilde{\theta_n}) \to_p I(\theta_0)$, but $I()$ can be hard to compute.

(b) $n^{-1}\Sigma_{i=1}^n (\bigtriangledown\ell(\tilde{\theta_n}; X_i))(\bigtriangledown\ell(\tilde{\theta_n}; X_i))^t$ is also a consistent estimator of $I(\theta_0)$, since $\bigtriangledown\ell(\tilde{\theta_n}; X_i) \to_p \bigtriangledown \ell(\theta_0; X_i)$.

(c) Often easiest is to use the second derivatives:

$$\left(-n^{-1}\sum_{i=1}^n \frac{\partial^2 \ell(\theta; X_i)}{\partial\theta_j \partial\theta_l}\right)|_{\theta=\tilde{\theta_n}} \quad = \quad \left(n^{-1} J_n(\tilde{\theta_n})\right)$$

is also a consistent estimator of $I(\theta_0)$.

(d) If CRLB attained: $\bigtriangledown\ell_n(\theta) = nI(\theta)(\widehat{\theta_n} - \theta)$
Hence (differentiating), $n^{-1}J_n(\widehat{\theta_n}) = I(\widehat{\theta_n})$.

5.5.2 The one-step estimator

We want to solve $\bigtriangledown\ell_n(\theta; X^{(n)}) = 0$. This can be hard. Suppose we have a preliminary estimator $\overline{\theta_n}$. Then we can do one-step Newton-Raphson:

$$0 = \bigtriangledown\ell_n(\theta; X^{(n)}) \;\approx\; \bigtriangledown\ell_n(\overline{\theta_n}; X^{(n)}) + \left(\frac{\partial^2 \ell_n}{\partial\theta_j \partial\theta_l}\right)|_{\theta=\overline{\theta_n}}(\theta - \overline{\theta_n})$$

Thus, replacing the second derivatives by some consistent estimator $-\hat{I}$ from (a),(b) or (c) above, new θ_n^* is

$$\theta_n^* \;=\; \overline{\theta_n} \;+\; (nI(\widehat{\overline{\theta_n}}))^{-1} \bigtriangledown \ell_n(\overline{\theta_n}; X^{(n)})$$

If $n^{1/4}(\overline{\theta_n} - \theta_0) \to_p 0$ then θ_n^* satisfies same Theorem 5.4(ii) (a) and (b) as $\tilde{\theta_n}$ –see JAW 4.7.

5.6 Appendix: summary of notation

Notation, definition	**description, result, etc.**
$X^{(n)} = (X_1, ..., X_n)$	**sample of i.i.d.** X_i **from pdf** f
$\ell_n(\theta; X^{(n)}) = \Sigma_1^n \log f(X_i;\theta)$	**log-likelihood function.**
$\bigtriangledown \ell(\theta; X_i) = (\frac{\partial \log f(X_i;\theta)}{\partial \theta_j})$	**contribution of** X_i **to score**
$\bigtriangledown \ell_n(\theta) = \Sigma_1^n \bigtriangledown \ell(\theta; X_i)$	**the score function: deriv. of** ℓ_n
$I(\theta) = \mathrm{E}(-\frac{\partial^2 \log f(X_i;\theta)}{\partial \theta_j \partial \theta_l})$	$I_n(\theta) = nI(\theta)$**: Fisher Information**
$J_n(\theta; X^{(n)}) = (-\frac{\partial^2 \ell_n}{\partial \theta_j \partial \theta_l})$	**observed information**
	$n^{-1} J_n(\theta) \rightarrow_p I(\theta)$
$\theta_0 \in \Theta \subset \Re^k$	**true value of** θ
$\bigtriangledown \ell_n(\theta) = 0$	**the likelihood equation**
$\widehat{\theta_n}$**; MLE**	**maximizes** $\ell_n(\theta)$ **in** Θ
$\tilde{\theta}_n$	**root of the likelihood eqn**

$Z_n \equiv n^{-\frac{1}{2}} \bigtriangledown \ell_n(\theta_0; X^{(n)})$	$Z_n \rightarrow_d N(0, I(\theta_0))$
$D_n = I^{-1}(\theta_0) Z_n$	$D_n \rightarrow_d D \sim N_k(0, I^{-1}(\theta_0))$
$\tilde{\ell}(\theta_0; X) \equiv I^{-1}(\theta_0) \bigtriangledown \ell(\theta_0; X)$	$n^{-\frac{1}{2}} \Sigma_{i=1}^n \tilde{\ell}(\theta_0; X_i) = I^{-1}(\theta_0) Z_n$
$\Delta_n \equiv n^{\frac{1}{2}}(\tilde{\theta}_n - \theta_0)$	$(\Delta_n - D_n) \ \rightarrow_p \ 0$
$\tilde{\ell}_n(\theta, X^{(n)}) = \Sigma_{i=1}^n \tilde{\ell}(\theta; X_i)$	**influence fn for** θ
$\tilde{\ell}_q(\theta; X^{(n)}) \ = \ (\bigtriangledown q(\theta))^t \tilde{\ell}_n(\theta; X^{(n)})$	**influence fn for** q**.**