CHAPTER 4.

EFFICIENT LIKELIHOOD ESTIMATION AND RELATED TESTS

4.1

1. MAXIMUM LIKELIHOOD AND EFFICIENT LIKELIHOOD ESTIMATES

We begin with a brief discussion of *Kullback - Leibler information.*

Definition 1.1. Let P be a probability measure and let Q be a sub-probability measure on $(\mathbf{X}, \mathbf{A})$ with densities p and q with respect to a $\sigma -$ finite measure μ ($\mu = P + Q$ always works). Thus $P(\mathbf{X}) = 1$ and $Q(\mathbf{X}) \leq 1$. Then the *Kullback - Leibler information* $K(P,Q)$ is

$$K(P,Q) \equiv E_P\{\log \frac{p(X)}{q(X)}\} . \tag{1}$$

Lemma 1.2. For a probability measure P and a (sub)-probability measure Q, the Kullback - Leibler information $K(P,Q)$ is always well - defined, and

$$K(P,Q) \begin{cases} \in [0,\infty] & \text{always} \\ = 0 & \text{iff } Q = P \end{cases} .$$

Proof. Now

$$K(P,Q) = \begin{cases} \log 1 = 0 & \text{if } P = Q \\ \log M > 0 & \text{if } P = M\,Q\,, \quad M > 1 \end{cases} .$$

If $P \neq M\,Q$, then Jensen's inequality is strict and yields

$$\begin{aligned} K(P,Q) &= E_P(-\log \frac{q(X)}{p(X)}) \\ &> -\log E_P(\frac{q(X)}{p(X)}) = -\log E_Q 1_{[p(X)>0]} \\ &\geq -\log 1 = 0 . \quad \square \end{aligned}$$

Now we need some assumptions and notation. Suppose that the model $\mathbf{P}$ is given by

$$\mathbf{P} = \{P_\theta : \theta \in \Theta\} .$$

We will impose the following as hypotheses about $\mathbf{P}$:

ASSUMPTIONS:

A0. $\theta \neq \theta^*$ implies $P_\theta \neq P_{\theta^*}$.

A1. $A \equiv \{x : p_\theta(x) > 0\}$ does not depend on θ.

A2. P_θ has density p_θ with respect to the $\sigma -$ finite measure μ and $X_1, \cdots, X_n$ are iid $P_{\theta_0} \equiv P_0$.

NOTATION:

$$L(\theta) \equiv L(\theta \mid \underline{X}) \equiv \prod_{i=1}^{n} p_\theta(X_i) ,$$

$$\mathbf{l}(\theta) = \mathbf{l}(\theta \mid \underline{X}) \equiv \sum_{i=1}^{n} \log p_\theta(X_i) ,$$

$$\mathbf{l}(B) \equiv \mathbf{l}(B \mid \underline{X}) = \sup_{\theta \in B} \mathbf{l}(\theta \mid \underline{X}) .$$

Here is a preliminary result which motivates our definition of the maximum likelihood estimator:

Theorem 1.3. If A0 - A2 hold, then for $\theta \neq \theta_0$

$$\frac{1}{n} \log(\frac{L(\theta_0)}{L(\theta)}) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p_{\theta_0}(X_i)}{p_\theta(X_i)} \rightarrow K(P_{\theta_0}, P_\theta) > 0 , \quad P_{\theta_0} \quad \text{a.s.}$$

and hence

$$P_{\theta_0}(L(\theta_0 \mid \underline{X}) > L(\theta \mid \underline{X})) \rightarrow 1 \quad \text{as} \quad n \rightarrow \infty .$$

Proof. The first assertion is just the strong law of large numbers; note that

$$E_{\theta_0} \log \frac{p_{\theta_0}(X)}{p_\theta(X)} = K(P_{\theta_0}, P_\theta) > 0$$

by lemma 1.2 and A0. The second assertion is an immediate consequence of the first. □

Theorem 1.3 motivates the following definition:

Definition 1.4. The value $\hat{\theta} = \hat{\theta}_n$ of θ which maximizes the likelihood $L(\theta \mid \underline{X})$, if it exists and is unique, is the *maximum likelihood estimator* (MLE) of θ. Thus $L(\hat{\theta}_n) = L(\Theta)$ or $\mathbf{l}(\hat{\theta}_n) = \mathbf{l}(\Theta)$.

Cautions:

- $\hat{\theta}_n$ may not exist.
- $\hat{\theta}_n$ may exist, but not be unique.
- Note that the definition depends on the version of the density p_θ which is selected; since this is not unique, different versions of p_θ lead to different MLE's.

The usual approach to finding $\hat{\theta}_n$ is to solve the *likelihood equations*

(2) $$\underline{\dot{\mathbf{l}}}(\theta \mid \underline{X}) \equiv \underline{\dot{\mathbf{l}}}_n(\theta) = \underline{0} ;$$

i.e. $\dot{\mathbf{l}}_{\theta_i}(\theta \mid \underline{X}) = 0$, $i = 1, \cdots, d$. The solution $\tilde{\theta}_n$ say, may not be the MLE, but may yield simply a local maximum of $l(\theta)$.

The *likelihood ratio statistic* for testing $H : \theta = \theta_0$ versus $K : \theta \neq \theta_0$ is

$$\lambda_n = \frac{L(\Theta)}{L(\theta_0)} = \frac{\sup_{\theta \in \Theta} L(\theta \mid \underset{\sim}{X})}{L(\theta_0 \mid \underset{\sim}{X})} = \frac{L(\hat{\theta}_n)}{L(\theta_0)}, \qquad \tilde{\lambda}_n = \frac{L(\tilde{\theta}_n)}{L(\theta_0)}.$$

Write P_0, E_0 for P_{θ_0}, E_{θ_0}. Here are some more assumptions about the model $\mathbf{P}$ which we need to treat these estimators:

ASSUMPTIONS; continued.

A3. Θ contains an open neighborhood $\Theta_0 \subset R^d$ of θ_0 for which:

(i) For μ a.e. x, $\mathbf{l}(\theta \mid x) \equiv \log p_\theta(x)$ is twice continuously differentiable in θ.

(ii) For a.e. x, the third order derivatives exist and $\mathbf{l}_{jkl}(\theta \mid x)$ satisfy $|\mathbf{l}_{jkl}(\theta \mid x)| \leq M_{jkl}(x)$ for $\theta \in \Theta_0$ for all j, k, l with $E_0 M_{jkl}(X) < \infty$.

A4. (i) $E_0\{\dot{\mathbf{l}}_i(\theta_0 \mid X)\} = 0$.

(ii) $E_0\{\dot{\mathbf{l}}_i^2(\theta_0 \mid X)\} < \infty$ for $i = 1, \cdots, d$.

(iii) $I(\theta_0) = (- E_0(\ddot{\mathbf{l}}_{ij}(\theta_0 \mid X)))$ is positive definite.

Let

$$Z_n \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\mathbf{l}}(\theta_0 | X_i) \qquad \text{and} \qquad \tilde{\mathbf{l}}(\theta_0 | X) = I^{-1}(\theta_0) \dot{\mathbf{l}}(\theta_0 | X)$$

so that

$$I^{-1}(\theta_0) Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\mathbf{l}}(\theta_0 | X_i).$$

Theorem 1.5. Suppose that $X_1, \cdots, X_n$ are iid $P_{\theta_0} \in \mathbf{P}$ with density p_{θ_0} where $\mathbf{P}$ satisfies A0 - A4. Then:

(i) With probability converging to 1 there exist solutions $\tilde{\theta}_n$ of the likelihood equations such that $\tilde{\theta}_n \to_p \theta_0$ when $P_0 = P_{\theta_0}$ is true.

(ii) $\sqrt{n}(\tilde{\theta}_n - \theta_0) = I^{-1}(\theta_0) Z_n + o_p(1) \to_d I^{-1}(\theta_0) Z \equiv D \sim N_d(0, I(\theta_0)^{-1})$.

(iii) $2 \log \tilde{\lambda}_n \to_d \chi_d^2$.

(iv) $W_n \equiv \sqrt{n}(\tilde{\theta}_n - \theta_0)^T \hat{I}_n(\tilde{\theta}_n) \sqrt{n}(\tilde{\theta}_n - \theta_0) \to_d D^T I(\theta_0) D \stackrel{d}{=} Z^T I^{-1}(\theta_0) Z$

where

$$\hat{I}_n(\tilde{\theta}_n) \equiv \begin{cases} I(\tilde{\theta}_n) & \text{or} \\ n^{-1} \sum_{i=1}^{n} \dot{\mathbf{l}}(\tilde{\theta}_n \mid X_i) \dot{\mathbf{l}}^T(\tilde{\theta}_n \mid X_i) & \text{or} \\ - n^{-1} \sum_{i=1}^{n} \ddot{\mathbf{l}}(\tilde{\theta}_n) . \end{cases}$$

(v) $R_n \equiv Z_n^T I^{-1}(\theta_0) Z_n \to_d Z^T I^{-1}(\theta_0) Z \sim \chi_d^2$. Here we could replace $I(\theta_0)$ by any of the possibilities for $\hat{I}_n(\tilde{\theta}_n)$ given in (iv) and the conclusion continues to hold.

(vi) The model $\mathbf{P}$ satisfies the LAN condition at θ_0 :

$$\begin{aligned} \mathbf{l}(\theta_0 + n^{-1/2}t) - \mathbf{l}(\theta_0) &= t^T Z_n - \frac{1}{2} t^T I(\theta_0) t + o_{P_0}(1) \\ &\to_d t^T Z - \frac{1}{2} t^T I(\theta_0) t \sim N(-\frac{1}{2}\sigma_0^2, \sigma_0^2) \end{aligned}$$

where $\sigma_0^2 \equiv t^T I(\theta_0) t$.

Corollary 1.6. Suppose that A0 - A4 hold and that $\nu \equiv \nu(P_\theta) = q(\theta)$ is differentiable at $\theta_0 \in \Theta$. Then $\tilde{\nu}_n \equiv q(\tilde{\theta}_n)$ satisfies

$$\sqrt{n}(\tilde{\nu}_n - \nu_0) = n^{-1/2} \sum_{i=1}^n \tilde{\mathbf{l}}_\nu(\theta_0|X_i) + o_p(1) \to_d N(0, \dot{q}^T(\theta_0) I^{-1}(\theta_0) \dot{q}(\theta_0))$$

where $\tilde{\mathbf{l}}_\nu(\theta_0|X_i) = \dot{q}^T(\theta_0) I^{-1}(\theta_0) \dot{\mathbf{l}}(\theta_0|X_i)$ and $\nu_0 \equiv q(\theta_0)$.

If the likelihood equations (2) are difficult to solve or have multiple roots, then it is possible to use a one-step approximation. Suppose that $\overline{\theta}_n$ is a preliminary estimator of θ and set

$$\theta_n \equiv \overline{\theta}_n + \hat{I}_n^{-1}(\overline{\theta}_n)(\frac{1}{n}\dot{\mathbf{l}}(\overline{\theta}_n)) .$$

Theorem 1.7. Suppose that A0 - A4 hold, and that $\overline{\theta}_n$ satisfies $n^{1/4}(\overline{\theta}_n - \theta_0) = o_p(1)$; note that the latter holds if $\sqrt{n}(\overline{\theta}_n - \theta_0) = O_p(1)$. Then

$$\sqrt{n}(\theta_n - \theta_0) = I^{-1}(\theta_0) Z_n + o_p(1) \to_d N_d(0, I^{-1}(\theta_0))$$

where $Z_n \equiv n^{-1/2} \sum_{i=1}^n \dot{\mathbf{l}}(\theta_0 \mid X_i)$.

Proof of theorem 1.5. (i) Existence and consistency: For $a > 0$, let

$$Q_a \equiv \{\theta \in \Theta : |\theta - \theta_0| = a\} .$$

We will show that

(a) $$P_0\{\mathbf{l}(\theta) < \mathbf{l}(\theta_0) \text{ for all } \theta \in Q_a\} \to 1 \quad \text{as } n \to \infty .$$

This implies that L has a local maximum inside Q_a . Since the likelihood equations must be satisfied at a local maximum, it will follow that for any $a > 0$ with probability converging to 1 that the likelihood equations have a solution $\hat{\theta}_n(a)$ within Q_a ; taking the root closest to θ_0 completes the proof.

To prove (a), write

(b) $$\frac{1}{n}(\mathbf{l}(\theta) - \mathbf{l}(\theta_0))$$

$$= \frac{1}{n}(\theta - \theta_0)^T \underline{\dot{\mathrm{l}}}(\theta_0) - \frac{1}{2}(\theta - \theta_0)^T(-\frac{1}{n}\ddot{\mathrm{l}}(\theta_0))(\theta - \theta_0)$$

$$+ \frac{1}{6n}\sum_{j=1}^{d}\sum_{k=1}^{d}\sum_{l=1}^{d}(\theta_j - \theta_{j0})(\theta_k - \theta_{k0})(\theta_l - \theta_{l0})\sum_{i=1}^{n}\gamma_{jkl}(X_i)M_{jkl}(X_i)$$

$$= S_1 + S_2 + S_3$$

where, by A3(ii), $0 \le |\gamma_{jkl}(x)| \le 1$. Furthermore, by A3(ii) and A4,

(c) $$S_1 \to_p 0\,,$$

(d) $$S_2 \to_p -\frac{1}{2}(\theta - \theta_0)^T I(\theta_0)(\theta - \theta_0)\,,$$

where

(e) $$(\theta - \theta_0)^T I(\theta_0)(\theta - \theta_0) \ge \lambda_d|\theta - \theta_0|^2 = \lambda_d a^2$$

and λ_d is the smallest eigenvalue of $I(\theta_0)$ (recall that $\sup_x (x^TAx)/(x^Tx) = \lambda_1$, $\inf_x (x^TAx)/(x^Tx) = \lambda_d$ where $\lambda_1 \ge \cdots \ge \lambda_d > 0$ are the eigenvalues of A symmetric positive definite), and

(f) $$S_3 \to_p \frac{1}{6}\sum_j\sum_k\sum_l(\theta_j - \theta_{j0})(\theta_k - \theta_{k0})(\theta_l - \theta_{l0})\; E\gamma_{jkl}(X)M_{jkl}(X)\,.$$

Thus, for any given $\epsilon, a > 0$, for n sufficiently large with probability larger than $1 - \epsilon$, for all $\theta \in Q_a$,

(g) $$|S_1| < da^3\,,$$

(h) $$S_2 < -\lambda_d a^2/4$$

and

(i) $$|S_3| \le \frac{1}{3}(da)^3\sum_{j,k,l} m_{jkl}$$

where $m_{jkl} \equiv EM_{jkl}(X)$. Hence, combining (f), (g), and (i) yields

(j) $$\begin{aligned}\sup_{\theta\in Q_a}(S_1 + S_2 + S_3) &\le \sup_{\theta\in Q_a}|S_1 + S_3| + \sup_{\theta\in Q_a} S_2\\ &< da^3 + Ba^3 - \frac{\lambda_d}{4}a^2\\ &\le (B + d)a^3 - \frac{\lambda_d}{4}a^2\end{aligned}$$

where $B \equiv 3^{-1}d^3\sum_{j,k,l} m_{jkl}$. The right side of (j) is < 0 if $a < \lambda_d/\{4(B + d)\}$, and hence (a) holds.

(ii) On the set

$$G_n \equiv \{\tilde{\theta}_n \text{ solves } \dot{\mathbf{l}}_n(\tilde{\theta}_n) = 0 \quad \text{and} \quad |\tilde{\theta}_n - \theta_0| < \epsilon\}$$

with $P_0(G_n) \to 1$ as $n \to \infty$ we have

(k) $$0 = \frac{1}{\sqrt{n}} \dot{\mathbf{l}}_n(\tilde{\theta}_n) = \frac{1}{\sqrt{n}} \dot{\mathbf{l}}_n(\theta_0) - \sqrt{n}(\tilde{\theta}_n - \theta_0)^T(-\frac{1}{n} \ddot{\mathbf{l}}_n(\theta_n^*))$$

where $|\theta_n^* - \theta_0| \le |\tilde{\theta}_n - \theta_0|$. Now from A4(i), (ii)

(l) $$Z_n \equiv \frac{1}{\sqrt{n}} \dot{\mathbf{l}}_n(\theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\mathbf{l}}(\theta_0 \mid X_i) \to_d N_d(0, I(\theta_0)) .$$

Furthermore

(m) $$-\frac{1}{n} \ddot{\mathbf{l}}_n(\theta_n^*) = -\frac{1}{n} \ddot{\mathbf{l}}_n(\theta_0) + o_p(1)$$

by using $\tilde{\theta}_n \to_p \theta_0$ and A3(ii) together with Taylor's theorem. Since matrix inversion is continuous (at nonsingular matrices), it follows that the inverse

$$(-\frac{1}{n} \ddot{\mathbf{l}}(\theta_n^*))^{-1}$$

exists with high probability, and satisfies

(n) $$(-\frac{1}{n} \ddot{\mathbf{l}}(\theta_n^*))^{-1} \to_p I(\theta_0)^{-1} .$$

Hence we can use (k) to write, on G_n,

(o) $$\begin{aligned} \sqrt{n}(\tilde{\theta}_n - \theta_0) &= I^{-1}(\theta_0) Z_n + o_p(1) \\ &\to I^{-1}(\theta_0) Z \sim N_d(0, I^{-1}(\theta_0)) . \end{aligned}$$

This proves (ii).

It also follows from (m) that

(p) $$\sqrt{n}(\tilde{\theta}_n - \theta_0)(-\frac{1}{n} \ddot{\mathbf{l}}(\tilde{\theta}))\sqrt{n}(\tilde{\theta}_n - \theta_0) \to_d Z^T I^{-1}(\theta_0) Z \sim \chi_d^2 ,$$

and that, if $I(\theta)$ is continuous at θ_0,

(q) $$\sqrt{n}(\tilde{\theta}_n - \theta_0)^T I(\tilde{\theta}_n)\sqrt{n}(\tilde{\theta}_n - \theta_0) \to_d Z^T I^{-1}(\theta_0) Z \sim \chi_d^2 .$$

(iii) To prove (iii), on the set G_n write

$$\mathbf{l}(\theta_0) = \mathbf{l}(\tilde{\theta}_n) + \dot{\mathbf{l}}^T(\tilde{\theta}_n)(\theta_0 - \tilde{\theta}_n) - \frac{1}{2} n(\theta_0 - \tilde{\theta}_n)^T(-\frac{1}{n} \ddot{\mathbf{l}}(\theta_n^*))(\theta_0 - \tilde{\theta}_n)$$

where $|\theta_n^* - \theta_0| \le |\tilde{\theta}_n - \theta_0|$. Thus

$$\begin{aligned} 2 \log \tilde{\lambda}_n &= 2\{\mathbf{l}(\tilde{\theta}_n) - \mathbf{l}(\theta_0)\} \\ &= 0 + 2\frac{1}{2} \sqrt{n}(\tilde{\theta}_n - \theta_0)^T(-\frac{1}{n} \ddot{\mathbf{l}}(\theta_n^*))\sqrt{n}(\tilde{\theta}_n - \theta_0) \end{aligned}$$

$$= D_n^T I(\theta_0) D_n + o_p(1) \quad \text{where} \quad D_n \equiv \sqrt{n}(\tilde{\theta}_n - \theta_0)$$
$$\to_d D^T I(\theta_0) D \quad \text{where} \quad D \sim N_d(0, I^{-1}(\theta_0))$$
$$\sim \chi_d^2 .$$

Finally, (v) is trivial since everything is evaluated at the fixed point θ_0 . □

Proof of theorem 1.7. First note that

(a)
$$\frac{1}{n}\ddot{\mathbf{l}}_n(\overline{\theta}_n) = \frac{1}{n}\ddot{\mathbf{l}}_n(\theta_0) + \frac{1}{n}\mathbf{l}(\theta_n^*)(\overline{\theta}_n - \theta_0)$$
$$= \frac{1}{n}\ddot{\mathbf{l}}_n(\theta_0) + O_p(1)|\overline{\theta}_n - \theta_0|$$

so that

$$(-\frac{1}{n}\ddot{\mathbf{l}}_n(\overline{\theta}_n))^{-1} = (-\frac{1}{n}\ddot{\mathbf{l}}_n(\theta_0))^{-1} + O_p(1)|\overline{\theta}_n - \theta_0|$$

and

(b)
$$\frac{1}{\sqrt{n}}\dot{\mathbf{l}}_n(\overline{\theta}_n) = \frac{1}{\sqrt{n}}\dot{\mathbf{l}}_n(\theta_0) + \frac{1}{n}\ddot{\mathbf{l}}_n(\theta_0)\sqrt{n}(\overline{\theta}_n - \theta_0)$$
$$+ \frac{1}{2}\sqrt{n}(\overline{\theta}_n - \theta_0)^T(\frac{1}{n}\ddot{\mathbf{l}}(\theta_n^*))(\overline{\theta}_n - \theta_0) .$$

Hence

$$\sqrt{n}(\theta_n - \theta_0) =$$
$$= \sqrt{n}(\overline{\theta}_n - \theta_0) + (-\frac{1}{n}\ddot{\mathbf{l}}_n(\overline{\theta}_n))^{-1}\frac{1}{\sqrt{n}}\dot{\mathbf{l}}_n(\overline{\theta}_n)$$
$$= \sqrt{n}(\overline{\theta}_n - \theta_0)$$
$$+ \{(-\frac{1}{n}\ddot{\mathbf{l}}_n(\theta_0))^{-1} + O_p(1)|\overline{\theta}_n - \theta_0|\}$$
$$\cdot\{Z_n + \frac{1}{n}\ddot{\mathbf{l}}(\theta_0)\sqrt{n}(\overline{\theta}_n - \theta_0) + \frac{1}{2}\sqrt{n}(\overline{\theta}_n - \theta_0)^T(\frac{1}{n}\ddot{\mathbf{l}}(\theta_n^*))(\overline{\theta}_n - \theta_0)\}$$
$$= (-\frac{1}{n}\ddot{\mathbf{l}}_n(\theta_0))^{-1}Z_n + O_p(1)|\overline{\theta}_n - \theta_0|Z_n$$
$$+ O_p(1)\frac{1}{n}\ddot{\mathbf{l}}_n(\theta_0)\sqrt{n}|\overline{\theta}_n - \theta_0|^2$$
$$+ O_p(1)\frac{1}{2}\sqrt{n}(\overline{\theta}_n - \theta_0)^T(\frac{1}{n}\ddot{\mathbf{l}}_n(\theta_n^*))(\overline{\theta}_n - \theta_0)$$
$$= I^{-1}(\theta_0)Z_n + o_p(1) + O_p(1)\sqrt{n}|\overline{\theta}_n - \theta_0|^2$$
$$= I^{-1}(\theta_0)Z_n + o_p(1) + o_p(1) .$$

Here we used

$$|\frac{1}{\sqrt{n}}\mathbf{l}_n(\theta_n^*)(\overline{\theta}_n - \theta_0)(\overline{\theta}_n - \theta_0)|$$

$$= \left| \sum_{k=1}^{d} \sum_{l=1}^{d} \sqrt{n}(\overline{\theta}_{nk} - \theta_{0k})(\overline{\theta}_{nl} - \theta_{0l}) \frac{1}{n} \ddot{\mathbf{l}}_{jkl}(\theta_n^* | \underline{X}) \right|$$

$$\leq d^3 \sqrt{n} |\overline{\theta}_n - \theta_0|^2 \sum_{j=1}^{d} \frac{1}{n} \sum_{i=1}^{n} |\ddot{\mathbf{l}}_{jkl}(\theta_n^* \mid X_i)|$$

$$= O_p(1) \sqrt{n} |\overline{\theta}_n - \theta_0|^2$$

since $|\overline{\theta}_{nk} - \theta_{0k}| \leq |\overline{\theta}_n - \theta_0|$ for $k = 1, \cdots, d$ and $|\underline{x}| \leq d \max_{1 \leq k \leq d} |x_k| \leq d \sum_{k=1}^{d} |x_k|$. □

2. THE WALD, LIKELIHOOD RATIO, AND SCORE (OR RAO) TESTS

Let $\theta_0 \in \Theta$ be fixed. For testing

(1) $$H : \theta = \theta_0 \quad \text{versus} \quad K : \theta \neq \theta_0 \, ,$$

recall the three test statistics

(2) $$2 \log \tilde{\lambda}_n \equiv 2\{\mathbf{l}_n(\tilde{\theta}_n) - \mathbf{l}_n(\theta_0)\} \, ,$$

(3) $$W_n \equiv \sqrt{n}(\tilde{\theta}_n - \theta_0)^T \hat{I}_n(\theta_0) \sqrt{n}(\tilde{\theta}_n - \theta_0) \, ,$$

and

(4) $$R_n \equiv Z_n^T I^{-1}(\theta_0) Z_n$$

where

$$Z_n \equiv \frac{1}{\sqrt{n}} \dot{\mathbf{l}}_n(\theta_0) = \frac{1}{\sqrt{n}} \dot{\mathbf{l}}(\theta_0 \,|\, \underline{X}) \, .$$

Theorem 4.1.5 described the null hypothesis behavior of these statistics; all three $\to_d \chi_d^2$ when $P_0 \equiv P_{\theta_0}$ is true. We now examine their behavior under alternatives, i.e. for $X_1, \cdots, X_n$ iid P_θ with $\theta \neq \theta_0$.

Theorem 2.1. (Fixed alternatives). Suppose that $\theta \neq \theta_0$ and A0 - A4 hold at θ. Then:

(5) $$\frac{1}{n} 2 \log \tilde{\lambda}_n \to_p 2\, K(P_\theta, P_{\theta_0}) = 2\, K(P_{true}, P_{hypothesized}) > 0$$

(6) $$\frac{1}{n} W_n \to_p (\theta - \theta_0)^T I(\theta)(\theta - \theta_0) > 0 \, .$$

(7) If, furthermore,

A5. $$E_\theta |\dot{\mathbf{l}}_i(\theta_0 \,|\, X)| < \infty \, , \quad i = 1, \cdots, d$$

holds, then

$$\frac{1}{n} R_n \to_p E_\theta\{\dot{\mathbf{l}}(\theta_0 \,|\, X)\}^T I^{-1}(\theta_0) E_\theta\{\dot{\mathbf{l}}(\theta_0 \,|\, X)\} > 0$$

if $E_\theta\{\dot{\mathbf{l}}(\theta_0 \,|\, X)\} \neq 0$.

Proof. When $\theta \neq \theta_0$ is really true,

(a) $$\begin{aligned} \frac{2}{n} \log \tilde{\lambda}_n &= \frac{2}{n} \{\mathbf{l}(\tilde{\theta}_n) - \mathbf{l}(\theta_0)\} \\ &= \frac{2}{n} \{\mathbf{l}(\theta) - \mathbf{l}(\theta_0)\} + \frac{2}{n} \{\mathbf{l}(\tilde{\theta}_n) - \mathbf{l}(\theta)\} \\ &= \frac{2}{n} \sum_{i=1}^n \log \frac{p_\theta}{p_{\theta_0}}(X_i) + \frac{2}{n} \{\mathbf{l}(\tilde{\theta}_n) - \mathbf{l}(\theta)\} \end{aligned}$$

$$\to_p \; 2\, E_\theta \log \frac{p_\theta}{p_{\theta_0}}(X) \;+\; 0\,\chi_d^2 \;=\; 2\, K(P_\theta\,, P_{\theta_0})$$

by the WLLN and theorem 4.1.5. Also, by Mann - Wald (continuous mapping),

(b) $$\frac{1}{n} W_n \;=\; (\tilde{\theta}_n \;-\; \theta_0)^T \hat{I}(\tilde{\theta}_n)(\tilde{\theta}_n \;-\; \theta_0)$$
$$\to_p \; (\theta \;-\; \theta_0)^T I(\theta)(\theta \;-\; \theta_0)\,,$$

and, since

(c) $$\frac{1}{\sqrt{n}} Z_n \;=\; \frac{1}{n}\dot{\mathbf{l}}_n(\theta_0) \;=\; \frac{1}{n}\sum_{i=1}^n \dot{\mathbf{l}}(\theta_0 \,|\, X_i) \;\to_p\; E_\theta \dot{\mathbf{l}}(\theta_0 \,|\, X)\,,$$

(d) $$\frac{1}{n} R_n \;\to_p\; E_\theta(\dot{\mathbf{l}}(\theta_0 \,|\, X))^T I^{-1}(\theta_0) E_\theta(\dot{\mathbf{l}}(\theta_0 \,|\, X))\,. \qquad \square$$

Corollary 2.2. (Consistency of the $2\log\lambda_n$, W_n, R_n tests). If assumptions A0 - A5 hold, then the tests are consistent: i.e. if $\theta \neq \theta_0$, then

(8) $$P_\theta(\text{LR test rejects } H) \;=\; P_\theta(2\log\tilde{\lambda}_n \;\geq\; \chi^2_{d,\alpha}) \;\to\; 1\,,$$

(9) $$P_\theta(\text{Wald test rejects } H) \;=\; P_\theta(W_n \;\geq\; \chi^2_{d,\alpha}) \;\to\; 1\,,$$

(10) $$P_\theta(\text{score test rejects } H) \;=\; P_\theta(R_n \;\geq\; \chi^2_{d,\alpha}) \;\to\; 1\,,$$

assuming $E_\theta(\dot{\mathbf{l}}(\theta_0 \,|\, X)) \;\neq\; 0$.

It remains to examine the behavior of these three tests under *local alternatives*, $\theta_n \;=\; \theta_0 \;+\; tn^{-1/2}$ with $t \neq 0$. We first examine Z_n and $\tilde{\theta}_n$ under θ_0 using Le Cam's third lemma 3.3.14.

Theorem 2.3. Suppose that A0 - A4 hold. Then, if $\theta_n \;=\; \theta_0 \;+\; tn^{-1/2}$ is true, under P_{θ_n}

(10) $$\sqrt{n}(\tilde{\theta}_n \;-\; \theta_0) \;\to_d\; D \;+\; t \;\sim\; N_d(\,t\,, I^{-1}(\theta_0))\,;$$

furthermore,

(11) $$Z_n(\theta_0) \;\equiv\; \frac{1}{\sqrt{n}}\dot{\mathbf{l}}(\theta_0 \,|\, \underline{X}) \;\to_d\; Z \;+\; I(\theta_0)t \;\sim\; N_d(I(\theta_0)t\,, I(\theta_0))\,.$$

Hence, we also have under P_{θ_n} ,

(12) $$\sqrt{n}(\tilde{\theta}_n \;-\; \theta_n) \;=\; \sqrt{n}(\tilde{\theta}_n \;-\; \theta_0) \;-\; \sqrt{n}(\theta_n \;-\; \theta_0)$$
$$\to_d\; D \;+\; t \;-\; t \;=\; D \;\sim\; N_d(\,0\,, I^{-1}(\theta_0))\,;$$

i.e. $\tilde{\theta}_n$ is locally regular. Furthermore, and

(13) $$Z_n(\theta_n) \;=\; Z_n(\theta_0) \;-\; (\,-\frac{1}{n}\,\ddot{\mathbf{l}}(\theta_n^*))\sqrt{n}(\theta_n \;-\; \theta_0)$$
$$\to_d\; Z \;+\; I(\theta_0)\underline{t} \;-\; I(\theta_0)\underline{t} \;=\; Z \;\sim\; N_d(\,0\,, I(\theta_0))\,.$$

Proof. From the proof of theorem 4.1.5 we know that $\tilde{\theta}_n$ is asymptotically linear under $P_0 = P_{\theta_0}$:

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\mathbf{l}}_\theta(\theta_0 \mid X_i) + o_p(1)$$

where $\tilde{\mathbf{l}}_\theta(x) \equiv I^{-1}(\theta_0)\dot{\mathbf{l}}_\theta(x)$. Furthermore, it follows from theorem 4.1.5 part (vi) that the log likelihood ratio is asymptotically linear:

$$\log \frac{dP^n_{\theta_n}}{dP^n_{\theta_0}}(\underline{X}) = \mathbf{l}(\theta_n) - \mathbf{l}(\theta_0) = t^T Z_n - \frac{1}{2} t^T I(\theta_0) t + o_p(1) .$$

Let $a \in R^d$. Then with $T_n \equiv a^T \sqrt{n}(\tilde{\theta}_n - \theta_0)$ it follows from the multivariate CLT that

$$\begin{pmatrix} T_n \\ \log \frac{dP^n_{\theta_n}}{dP^n_{\theta_0}} \end{pmatrix} = \begin{pmatrix} a^T \sqrt{n}(\tilde{\theta}_n - \theta_0) \\ \log \frac{dP^n_{\theta_n}}{dP^n_{\theta_0}} \end{pmatrix} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} a^T \tilde{\mathbf{l}}_\theta(\theta_0 \mid X_i) \\ t^T \dot{\mathbf{l}}_\theta(\theta_0 \mid X_i) \end{pmatrix} + \begin{pmatrix} 0 \\ -\sigma^2/2 \end{pmatrix} + o_p($$

$$\to_d N_2(\begin{pmatrix} 0 \\ -\sigma^2/2 \end{pmatrix}, \begin{pmatrix} a^T I^{-1}(\theta_0) a & a^T t \\ a^T t & \sigma^2 \end{pmatrix}) .$$

Thus the hypothesis of Le Cam's third lemma 3.3.14 is satisfied with $c \equiv a^T t$, and we deduce that, under P_{θ_n} ,

$$\begin{pmatrix} a^T \sqrt{n}(\tilde{\theta}_n - \theta_0) \\ \log \frac{dP^n_{\theta_n}}{dP^n_{\theta_0}} \end{pmatrix} \to_d N_2(\begin{pmatrix} a^T t \\ +\sigma^2/2 \end{pmatrix}, \begin{pmatrix} a^T I^{-1}(\theta_0) a & a^T t \\ a^T t & \sigma^2 \end{pmatrix}) .$$

In particular, under P_{θ_n} ,

$$a^T \sqrt{n}(\tilde{\theta}_n - \theta_0) \to_d N(a^T t , a^T I^{-1}(\theta_0) a)$$

and from the Cramér - Wold device this implies that under P_{θ_n}

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \to_d N_d(t , I^{-1}(\theta_0)) .$$

This, in turn, implies that

$$\sqrt{n}(\tilde{\theta}_n - \theta_n) \to_d N_d(0 , I^{-1}(\theta_0))$$

under P_{θ_n} . The proof of the second part is similar, but easier, by taking $T_n \equiv a^T Z_n(\theta_0)$ which is already a linear statistic. □

Corollary 2.4. If A0 - A4 hold,then if $\theta_n = \theta_0 + tn^{-1/2}$, under P_{θ_n}

(14) $2 \log \tilde{\lambda}_n \to_d (D + t)^T I(\theta_0)(D + t) \sim \chi^2_d(\delta)$,

(15) $W_n \to_d (D + t)^T I(\theta_0)(D + t) \sim \chi^2_d(\delta)$,

(16) $R_n \to_d (Z + I(\theta_0)t)^T I^{-1}(\theta_0)(Z + I(\theta_0)t) = (D + t)^T I(\theta_0)(D + t) \sim \chi^2_d(\delta)$
where $\delta \equiv t^T I(\theta_0) t$.

Proof. This follows from theorem 2.3, the Mann - Wald theorem, and the fact that

$$X \sim N_d(\mu, \Sigma) \quad \text{implies} \quad X^T \Sigma^{-1} X \sim \chi^2_d(\delta)$$

with $\delta = \mu^T \Sigma^{-1} \mu$. □

Corollary 2.5. If A0 - A4 hold, then with $T_n = 2 \log \tilde{\lambda}_n$, W_n, or R_n,

(17) $$P_{\theta_n}(T_n \geq \chi^2_{d,\alpha}) \to P(\chi^2_d(\delta) \geq \chi^2_{d,\alpha}) .$$

Three Statistics for Testing a Composite Null Hypothesis

Now consider testing $\theta \in \Theta_0 \equiv \{\theta \in \Theta : \theta_1 = \theta_{10} \}$; i.e.

$$H : \theta_1 = \theta_{10} , \; \theta_2 = \text{anything}$$

while

$$K : \theta = (\theta_1 , \theta_2) \neq (\theta_{10} , \theta_2)$$

where $\theta \equiv (\theta_1 , \theta_2) \in R^m \times R^{d-m} = R^d$. Recall the corresponding partitioning of $I(\theta)$ and $I^{-1}(\theta)$, and the matrices $I_{11\cdot 2}$, $I_{22\cdot 1}$ introduced in section 3.2.

The likelihood ratio, Wald, and Rao (or score) statistics for testing H versus K are:

(18) $$2 \log \lambda_n \qquad \text{with} \qquad \lambda_n \equiv \frac{\sup_{\theta \in \Theta} L(\theta \mid \underline{X})}{\sup_{\theta \in \Theta_0} L(\theta \mid \underline{X})} = \frac{L(\hat{\theta}_n \mid \underline{X})}{L(\hat{\theta}^0_n \mid \underline{X})}$$

(or

(19) $$2 \log \tilde{\lambda}_n \qquad \text{with} \qquad \tilde{\lambda}_n \equiv \frac{L(\tilde{\theta}_n \mid \underline{X})}{L(\tilde{\theta}^0_n \mid \underline{X})}$$

where $\tilde{\theta}_n$, $\tilde{\theta}^0_n$ are consistent solutions of the likelihood equations under K and H respectively);

(20) $$W_n \equiv \sqrt{n}(\hat{\theta}_{n1} - \theta_{10})^T \, \hat{I}_{11\cdot 2} \sqrt{n}(\hat{\theta}_{n1} - \theta_{10}) ,$$

and

(21) $$R_n \equiv Z^T_n(\hat{\theta}^0_n) \hat{I}^{-1}(\hat{\theta}^0_n) Z_n(\hat{\theta}^0_n)$$

where $\hat{\theta}^0_n$ ($\tilde{\theta}^0_n$) is an MLE (ELE) of $\theta \in \Theta_0$.

Now under $H : \theta \in \Theta_0$ we have

(22) $$\sqrt{n}(\tilde{\theta}_{n1} - \theta^0_1) \to_d D_1 \sim N_m(0 , I^{-1}_{11\cdot 2})$$

where

$$D \;=\; \begin{pmatrix} D_1 \\ D_2 \end{pmatrix} \;=\; I^{-1}(\theta)\, Z \;=\; \begin{pmatrix} I_{11\cdot 2}^{-1}(Z_1 \;-\; I_{12}I_{22}^{-1}Z_2) \\ I_{22\cdot 1}^{-1}(Z_2 \;-\; I_{21}I_{11}^{-1}Z_1) \end{pmatrix}$$

and

$$\begin{aligned} Z_n(\tilde\theta_n^0) \;&=\; \begin{pmatrix} Z_{n1}(\tilde\theta_n^0) \\ Z_{n2}(\tilde\theta_n^0) \end{pmatrix} \\ &=\; \begin{pmatrix} Z_{n1}(\theta_0) \;-\; I_{12}(\theta_n^*)\sqrt{n}(\tilde\theta_{n2}^0 \;-\; \theta_2^0) \;+\; o_p(1) \\ 0 \end{pmatrix} \\ &=\; \begin{pmatrix} Z_{n1}(\theta_0) \;-\; I_{12}(\theta_n^*)I_{22}^{-1}Z_{n2}(\theta_0) \;+\; o_p(1) \\ 0 \end{pmatrix} \\ &\to_d\; \begin{pmatrix} Z_1(\theta_0) \;-\; I_{12}(\theta_0)I_{22}^{-1}Z_2(\theta_0) \\ 0 \end{pmatrix} \;\sim\; \begin{pmatrix} N_m(0\,, I_{11\cdot 2}) \\ 0 \end{pmatrix} . \end{aligned} \tag{23}$$

The natural consequences of (22) and (23) for the likelihood ratio, Wald, and Rao statistics are sumarized in the following theorem.

Theorem 2.6. If A0 - A4 hold and $\theta_0 \in \Theta_0$ is true, then

$$\left\{\begin{matrix} 2\log\tilde\lambda_n \\ W_n \\ R_n \end{matrix}\right\} \;\to_d\; D_1^T I_{11\cdot 2} D_1 \;\sim\; \chi_m^2 \;=\; \chi^2_{d\,-\,(d\,-\,m)} \,.$$

Proof. That $W_n \to_d D_1^T I_{11\cdot 2} D_1$ follows from (22) and consistency of $\hat I_{11\cdot 2}$; similarly, $R_n \to_d D_1^T I_{11\cdot 2} D_1$ follows from (23) and $\hat I^{-1}(\hat\theta_n^0) \to_p I^{-1}(\theta_0)$. To prove the claimed convergence of $2\log\tilde\lambda_n$, write

$$\begin{aligned} 2\log\tilde\lambda_n \;&=\; 2\{l_n(\tilde\theta_n) \;-\; l_n(\tilde\theta_n^0)\} \\ &=\; 2\{l_n(\tilde\theta_n) \;-\; l_n(\theta_0) \;-\; (l_n(\tilde\theta_n^0) \;-\; l_n(\theta_0))\} \end{aligned}$$

where

$$2\{l_n(\tilde\theta_n) \;-\; l_n(\theta_0)\} \;\to_d\; D^T I(\theta_0) D \;=\; Z^T I^{-1}(\theta_0) Z \tag{a}$$

by our proof of theorem 4.1.5, and

$$2\{l_n(\tilde\theta_n^0) \;-\; l_n(\theta_0)\} \;\to_d\; Z_2^T I_{22}^{-1}(\theta_0) Z_2 \tag{b}$$

by the proof of theorem 4.1.5. In fact, by the asymptotic linearity of $\tilde\theta_n$ (and $\tilde\theta_n^0$) proved there, the convergence in (a) and (b) holds jointly, and it follows by the continuous mapping (Mann - Wald) theorem and (a) and (b) that

$$\begin{aligned} 2\log\tilde\lambda_n \;&\to_d\; Z^T I^{-1}(\theta_0) Z \;-\; Z_2^T I_{22}^{-1} Z_2 \\ &=\; (Z_1 - I_{12}I_{22}^{-1}Z_2)^T\, I_{11\cdot 2}^{-1}(Z_1 - I_{12}I_{22}^{-1}Z_2) \end{aligned}$$

$$= D_1^T I_{11\cdot 2} D_1$$

where we have used the block inverse form of $I^{-1}(\theta_0)$ given in (3.2.3) and the matrix identity (3.2.15) with the roles of "1" and "2" interchanged. □

Now under local alternatives the situation is as follows:

Theorem 2.7. If A0 - A4 hold, and $\theta_n = \theta_0 + tn^{-1/2}$ with $\theta_0 \in \Theta_0$, then under P_{θ_n}

$$\left\{ \begin{array}{l} 2\log \tilde{\lambda}_n \\ W_n \\ R_n \end{array} \right\} \to_d (D_1 + t_1)^T I_{11\cdot 2}(D_1 + t_1) \sim \chi^2_m(\delta) = \chi^2_{d-(d-m)}(\delta)$$

where $\delta \equiv t_1^T I_{11\cdot 2} t_1$.

3. Selected Examples

Now we consider several examples to illustrate the preceding theory and its limitations.

Examples 3.1. Let $X_1, \cdots, X_n$ be iid $N(\mu, \sigma^2)$; $\theta = (\mu, \sigma^2)$. Consider:

(i) Estimation of θ.

(ii) Testing $H_1 : \mu = 0$ versus $K_1 : \mu \neq 0$.

(iii) Testing $H_2 : \theta = (0, \sigma_0^2) \equiv \theta_0$ versus $K_2 : \theta \neq \theta_0$.

(i) Now the likelihood function is:

$$L(\theta) = (2\pi\sigma^2)^{-n/2} \exp\left(- \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 \right).$$

Thus the MLE is

$$\hat{\theta}_n = \begin{pmatrix} \bar{X} \\ S^2 \end{pmatrix} \quad \text{where} \quad S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$$

and

$$I(\theta) = \begin{pmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} \end{pmatrix}$$

since

$$l(\theta \mid X_1) = - \frac{1}{2\sigma^2} (X_1 - \mu)^2 - \frac{1}{2} \log \sigma^2$$

so that

$$\dot{\mathbf{l}}_\mu(X_1) = \frac{1}{\sigma^2}(X_1 - \mu), \qquad \ddot{\mathbf{l}}_{\mu\mu}(X_1) = -\frac{1}{\sigma^2}$$

$$\dot{\mathbf{l}}_{\sigma^2}(X_1) = \frac{(X_1 - \mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2}; \qquad \ddot{\mathbf{l}}_{\sigma^2\sigma^2}(X_1) = -\frac{(X_1 - \mu)^2}{\sigma^6} + \frac{1}{2\sigma^4}.$$

Now

(1) $$\hat{\theta}_n = (\bar{X}, S^2) \to_{a.s.} (\mu, \sigma^2) = \theta,$$

and

(2) $$\sqrt{n}(\hat{\theta}_n - \theta) \to_d N_2(0, I^{-1}(\theta))$$

with

$$I^{-1}(\theta) = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}.$$

Proof. (1) and (2) follow from either a direct application of the SLLN and the CLT, or from theorem 4.1.2. □

(ii) The likelihood ratio statistic for testing H_1 is

$$\lambda_n \;=\; \frac{L(\hat{\theta}_n)}{L(\hat{\theta}^0_n)} = \frac{L(\bar{X}\,, S^2)}{L(0\,, n^{-1}\sum_{i=1}^n X_i^2)} \;=\; \left\{ \frac{n^{-1}\sum_i X_i^2}{n^{-1}\sum_i (X_i - \bar{X})^2} \right\}^{n/2}$$

and hence

$$2\log\lambda_n \;=\; -\,n\log(1 \;-\; \frac{\bar{X}^2}{\frac{1}{n}\sum X_i^2})\;;$$

note that $-\log(1-x) \;\sim\; x$ for $x \to 0$. The Wald statistic is

$$W_n \;=\; \{\sqrt{n}(\bar{X} - 0)\}\hat{I}_{11\cdot 2}\{\sqrt{n}(\bar{X} - 0)\} \;=\; \frac{n\bar{X}^2}{S^2} \;=\; \{\frac{\sqrt{n}\bar{X}}{S}\}^2\,.$$

Finally, the Rao or score statistic is

$$R_n \;=\; Z_n(\hat{\theta}^0_n)^T I(\hat{\theta}_0)^{-1} Z_n(\hat{\theta}^0_n)$$

$$= \begin{pmatrix} \dfrac{\sqrt{n}\bar{X}}{n^{-1}\sum X_i^2} \\ 0 \end{pmatrix}^T \begin{pmatrix} n^{-1}\sum X_i^2 & 0 \\ 0 & 2(n^{-1}\sum X_i^2)^2 \end{pmatrix} \begin{pmatrix} \dfrac{\sqrt{n}\bar{X}}{n^{-1}\sum X_i^2} \\ 0 \end{pmatrix} \;=\; \{\frac{\sqrt{n}\bar{X}}{\sqrt{n^{-1}\sum X}}$$

If $\theta \;=\; \theta_0 \;=\; (0\,,\sigma^2)$ so H_1 holds, then

$$2\log\lambda_n\,,\quad W_n\,,\quad R_n \quad \to_d \quad \chi^2_1\,.$$

if $\mu \neq 0$ so $\theta \not\in \Theta_1$, then

$$\frac{1}{n}2\log\lambda_n \;\to_p\; -\log(1 \;-\; \frac{\mu^2}{\sigma^2 + \mu^2}) = -\log(\frac{\sigma^2}{\sigma^2+\mu^2}) \;>\; 0\,,$$

$$\frac{1}{n}W_n \;\to_p\; \frac{\mu^2}{\sigma^2} \;>\; 0\,,$$

and

$$\frac{1}{n}R_n \;\to_p\; \frac{\mu^2}{\sigma^2+\mu^2} \;>\; 0\,.$$

(iii) The likelihood ratio statistic λ_n for testing H_2 is

$$\lambda_n \;=\; \frac{L(\bar{X}\,, S^2)}{L(0\,, \sigma_0^2)} \;=\; \frac{(2\pi S^2)^{-n/2}\exp(\,-\,n/2)}{(2\pi\sigma_0^2)^{-n/2}\exp(\,-\sum_i X_i^2/2\sigma_0^2)}$$

$$= (\frac{S^2}{\sigma_0^2})^{-n/2}\exp(\frac{1}{2\sigma_0^2}\sum_i X_i^2 \;-\; n/2)$$

so that

$$2 \log \lambda_n = \frac{1}{\sigma_0^2} \sum_{i=1}^n X_i^2 - n - n \log(\frac{S^2}{\sigma_0^2})$$

$$= n\{ \frac{S^2}{\sigma_0^2} - 1 - \log(\frac{S^2}{\sigma_0^2})\} + \frac{n\overline{X}^2}{\sigma_0^2} .$$

The Wald statistic is

$$W_n = \begin{pmatrix} \sqrt{n}(\overline{X} - 0) \\ \sqrt{n}(S^2 - \sigma_0^2) \end{pmatrix}^T \begin{pmatrix} \frac{1}{S^2} & 0 \\ 0 & \frac{1}{2S^4} \end{pmatrix} \begin{pmatrix} \sqrt{n}(\overline{X} - 0) \\ \sqrt{n}(S^2 - \sigma_0^2) \end{pmatrix}$$

$$= \frac{n\overline{X}^2}{S^2} + \frac{\{\sqrt{n}(S^2 - \sigma_0^2)\}^2}{2S^4} .$$

The Rao or score statistic is

$$R_n = \begin{pmatrix} \sqrt{n}\overline{X}/\sigma_0^2 \\ \frac{1}{2} \frac{\sqrt{n}}{\sigma_0^2} \{ \frac{n^{-1}\sum_i X_i^2}{\sigma_0^2} - 1\} \end{pmatrix}^T \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & 2\sigma_0^4 \end{pmatrix} \begin{pmatrix} \sqrt{n}\overline{X}/\sigma_0^2 \\ \frac{1}{2} \frac{\sqrt{n}}{\sigma_0^2} \{ \frac{n^{-1}\sum_i X_i^2}{\sigma_0^2} - 1\} \end{pmatrix}$$

$$= \frac{n\overline{X}^2}{\sigma_0^2} + \frac{n}{2} \{ \frac{n^{-1}\sum_i X_i^2}{\sigma_0^2} - 1\}^2 .$$

If H_2 holds, then

$$2 \log \lambda_n , \ W_n , \ R_n \ \to_d \ \chi_2^2 .$$

[**Question:** What are the limits in probability of $n^{-1} 2 \log \lambda_n$, $n^{-1} W_n$, $n^{-1} R_n$ under $\theta \neq \theta_0$.]

Example 3.2. (One-Parameter Exponential family; Lehmann, page 417). Let $p_\theta(x) = \exp(\theta T(x) - A(\theta))$ with respect to μ . Then

$$\dot{\mathbf{l}}_\theta(x) = T(x) - A'(\theta) , \quad I(\theta) = Var_\theta(T(X)) ,$$

and the likelihood equation may be written as

$$\frac{1}{n} \sum_{i=1}^n T(X_i) = A'(\theta) . \tag{3}$$

Now $A'(\theta) = E_\theta T(X)$, and $A''(\theta) = Var_\theta(T(X)) > 0$, so the right side in (*) is strictly increasing in θ . Thus (3) has at most one root $\hat{\theta}_n$, and

$$\sqrt{n}(\hat{\theta}_n - \theta) \to_d N(0, \frac{1}{I(\theta)}) = N(0, \frac{1}{Var_\theta(T(X))}) . \tag{4}$$

Example 3.3. (Multi-parameter Exponential family; Lehmann, TPE, pages 438, 26, 124) Suppose that

$$p_\theta(x) \;=\; \exp[\sum_{j=1}^{d} \eta_i(\theta) T_j(x) \;-\; B(\theta)] h(x)$$

with respect to some dominating measure μ. Here $B : \Theta \to R$ and $\eta_j : \Theta \to R$, $T_j : \mathbf{X} \to R$ for $j = 1, \cdots, d$. Here $\Theta \subset R^k$ for some k. It is frequently convenient to use the η_i as the parameters and write the density in the *canonical form*

$$p_\eta(x) \;\equiv\; p(x;\eta) \;=\; \exp[\sum_{j=1}^{d} \eta_j T_j(x) \;-\; A(\eta)] h(x) \,.$$

The *natural parameter space* Ξ of the family is

$$\Xi \;\equiv\; \{\eta \in R^d : \int \exp[\underline{\eta}^T \underline{T}(x) h(x)\, d\mu(x) \;<\; \infty\} \,.$$

We will assume that the T_j 's are *affinely independent*: i.e. there do not exist constants $a_1, \cdots, a_d$ and $b \in R$ such that $\sum_1^d a_j T_j(x) \;=\; b$ with probability 1.

By Lehmann, TPE, theorem 4.1, page 29, we can differentiate under the expectations to find that the score vector for $\underline{\eta}$ is given by

$$\dot{\mathbf{l}}_{\eta_j}(x) \;=\; T_j(x) \;-\; \frac{\partial}{\partial \eta_j} A(\eta) \,, \qquad j \;=\; 1, \cdots, d$$

and since this has expectation 0 under p_η,

$$0 \;=\; E_\eta(T_j(X)) \;-\; \frac{\partial}{\partial \eta_j} A(\eta) \,, \qquad j \;=\; 1, \cdots, d$$

If the likelihood equations have a solution, it is unique (and is the MLE) since $\mathbf{l}(\eta)$ is a strictly concave function of η since

$$-\frac{\partial^2}{\partial \eta_j \partial \eta_l} \mathbf{l}(\eta) \;=\; \frac{\partial^2}{\partial \eta_j \partial \eta_l} A(\eta) \;=\; Cov_\eta[T_j(X), T_l(X)]$$

is positive definite by the assumption of affine independence of the T_j 's.

Example 3.4. (Cauchy location family). Suppose that $p_\theta(x) \;=\; g(x - \theta)$ with $g(x) \;=\; \pi^{-1}(1 + x^2)^{-1}$, $-\infty < x < \infty$. Then

$$\dot{\mathbf{l}}_\theta(X) \;=\; -\frac{g'}{g}(X - \theta) \;=\; \frac{2(X - \theta)}{1 + (X - \theta)^2} \,,$$

$I(\theta) \;=\; 1/2$, and the likelihood equation becomes

$$\dot{\mathbf{l}}_\theta(\theta \,|\, \underline{X}) \;=\; \sum_{i=1}^{n} \dot{\mathbf{l}}_\theta(X_i) \;=\; 2 \sum_{i=1}^{n} \frac{X_i - \theta}{1 + (X_i - \theta)^2} \;=\; 0$$

if and only if

$$0 \;=\; \sum_{i=1}^{n} (X_i - \theta) \prod_{j \neq i} \{1 \;+\; (X_j - \theta)^2\} \;\;=$$

a polynomial in θ of degree $2(n-1)+1 = 2n-1$ which could have as many as $2n-1$ roots. Let $\bar{\theta}_n \;\equiv\; \text{median}(X_i) \;\equiv\; IF_n^{-1}(1/2)$. Then

$$\sqrt{n}(\bar{\theta}_n \;-\; \theta) \;\to_d\; N(0\,,\pi^2/4)$$

and the one - step adjustment (or "method of scoring") estimator is

$$\theta_n \;=\; \bar{\theta}_n \;+\; \hat{I}(\bar{\theta}_n)^{-1}(\frac{1}{n}\sum_{i=1}^{n} \dot{\mathbf{l}}_\theta(X_i;\bar{\theta}_n))$$

$$=\; \bar{\theta}_n \;+\; \frac{4}{n}\sum_{i=1}^{n} \frac{X_i - \bar{\theta}_n}{1 \;+\; (X_i - \bar{\theta}_n)^2} \;.$$

Remark: Let r_n denote the (random) number of roots of the likelihood equation. Then with probability one the roots are simple, and they alternately correspond to local minima and local maxima of the likelihood. Thus r_n is odd, there are $(r_n+1)/2$ local maxima, $(r_n-1)/2$ local minima, and one global maxima. Thus the number of roots corresponding to "false" maxima is $(r_n-1)/2$. Reeds (*Ann. Statist.* **13** (1985), 775 - 784) shows that $(r_n - 1)/2 \;\to_d\; Poisson(1/\pi)$, so that as $n \;\to\; \infty$ we can expect to see relatively few roots corresponding to local maxima which are not global maxima.

Example 3.5. (Normal mixture models; see TPE, page 442, example 5.6) Suppose that

$$p_\theta(x) \;=\; pN(\mu\,,\sigma^2) \;+\; (1-p)N(\nu\,,\tau^2)$$

$$=\; p\,\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{1}{2}\,\frac{(x-\mu)^2}{\sigma^2}) \;+\; (1-p)\,\frac{1}{\sqrt{2\pi\tau^2}}\exp(-\frac{1}{2}\,\frac{(x-\nu)^2}{\tau^2})$$

with $\theta \;=\; (p\,,\mu\,,\sigma\,,\nu\,,\tau)$. The simpler case of $\theta \;=\; (\frac{1}{2}\,,\mu\,,\sigma\,,0\,,1) \;\equiv\; \theta_0$ will illustrate the phenomena we wish to illustrate here. When $\theta = \theta_0$ we may reparametrize by $\theta \;=\; (\mu\,,\sigma) \;\in\; R \times R^+ \;=\; \Theta$, and

$$p_\theta(x) \;=\; \frac{1}{2}\,\phi(x) \;+\; \frac{1}{2}\,\frac{1}{\sigma}\,\phi(\frac{x-\mu}{\sigma})\;.$$

Then, if $X_1\,,\cdots,X_n$ are iid p_θ ,

$$\sup_{\theta\,\in\,\Theta} L(\theta\,|\,\underline{X}) \;=\; \infty \qquad a.\,s.$$

To see this, take $\mu \;=$ any X_i say; then let $\sigma \;\to\; 0$. Thus MLE's do not exist.

However, A0 - A4 hold and hence there exists a consistent, asymptotically efficient sequence of roots of the likelihood equations; alternatively, one - step estimators starting with moment estimators are also asymptotically efficient.

Example 3.6. (An inconsistent MLE; Ferguson, JASA 1982, 831-834). Suppose that

$$p_\theta(x) \;=\; \theta \frac{1}{2} 1_{[-1,1]}(x) \;+\; \frac{1-\theta}{\delta(\theta)}(1 \;-\; \frac{|x-\theta|}{\delta(\theta)}) 1_{A(\theta)}(x)$$

where $\theta \in \Theta \equiv [-1,1]$; $\delta \equiv \delta(\theta)$ is decreasing and continuous with $\delta(0) = 1$ and $0 < \delta(\theta) \le 1-\theta$ for $0 < \theta < 1$; and $A(\theta) = [\theta - \delta(\theta), \theta + \delta(\theta)]$.

Note that $p_0(x) = (1-|x|)1_{[-1,1]}(x) \equiv$ traingular density, while $p_1(x) = 2^{-1}1_{[-1,1]}(x) =$ uniform density.

Note that A0 - A2 hold, and, in addition, $p_\theta(x)$ is continuous in θ for all x. Thus a MLE exists for all $n \ge 1$ since a continuous function on a compact set $[0,1]$ achieves its maximum on that set.

Proposition 3.7. (Ferguson). Let $\hat\theta_n \equiv$ a MLE of θ for sample size n. If $\delta(\theta) \to 0$ rapidly enough as $\theta \to 1$, then $\hat\theta_n \to 1$ a.s. P_θ as $n \to \infty$ for every $\theta \in [0,1]$. [In fact, the function $\delta(\theta) = (1-\theta)\exp(-(1-\theta)^{-4}) + 1$ works; for this choice of $\delta(\theta)$ it follows that $n^{1/4}(1 - M_n) \to_{a.s.} 0$ where $M_n = \max\{X_1, \cdots, X_n\}$.]

Other examples of inconsistent MLE's are given by Le Cam (1990), and by Boyles, Marshall, and Proschan (1985). Here is another situation in which the MLE fails, though for somewhat different reasons.

Example 3.8. (Another inconsistent MLE; Neyman and Scott, 1948; TPE example 6.7, page 450).

Let $(X_i, Y_i) \sim N_2((\mu_i, \mu_i), \sigma^2 I)$, $i = 1, \cdots, n$. Consider estimation of σ^2. Now $Z_i \equiv X_i - Y_i \sim N(0, 2\sigma^2)$, and therefore

$$\frac{1}{2n}\sum_{i=1}^n Z_i^2 \;\sim\; \sigma^2 \frac{\chi_n^2}{n}$$

is unbiased and consistent. However

$$L(\theta \mid \underline{X}, \underline{Y}) \;=\; (2\pi\sigma^2)^{-n/2}\exp(-\frac{1}{2\sigma^2}\sum_{i=1}^n \{(X_i-\mu_i)^2 \;+\; (Y_i-\mu_i)^2\})$$

and therefore the MLE of $\mu_i = (X_i + Y_i)/2$ and hence the MLE of σ^2 is

$$\hat\sigma^2 \;=\; \frac{1}{4}\frac{1}{n}\sum_{i=1}^n Z_i^2 \;=\; \frac{1}{2n}\{\sum_{i=1}^n (X_i - \hat\mu_i)^2 \;+\; \sum_{i=1}^n (Y_i - \hat\mu_i)^2\}.$$

Thus $\hat\sigma^2 \to_{a.s.} \frac{1}{2}\sigma^2$ as $n \to \infty$.

What went wrong? Note that the dimensionality of the parameter space increases here with n, so the theory we have developed so far does not apply. One way out of this fix was proposed by Kiefer and Wolfowitz (1956) in their paper on "Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters," *Ann. Math. Statist.* **27**, 887 - 906.

Example 3.9. (Testing in a bivariate Poisson model). Suppose that $U \sim Poisson(\mu)$; $V \sim Poisson(\lambda)$; and $W \sim Poisson(\psi)$ are all independent, and let

$$X \equiv U + W , \quad Y \equiv V + W .$$

Then $X \sim Poisson(\mu + \psi)$, $Y \sim Poisson(\lambda + \psi)$, and jointly

$$(X , Y) \sim \text{bivariate Poisson}(\mu , \lambda , \psi) :$$

for positive integers x and y

$$P_\theta(X = x , Y = y ; \mu , \lambda , \psi) = \ = \sum_{w=0}^{x \wedge y} \frac{\mu^{x-w} \lambda^{y-w} \psi^w}{w!(x-w)!(y-w)!} e^{-(\lambda + \mu + \psi)}$$

where $\theta = (\mu , \lambda , \psi)$. Thus under $\psi = 0$, X and Y are independent ($\psi = 0$ implies $W = 0$ a.s. so $X = U$ and $Y = V$ a.s.).

Consider testing $H : \psi = 0$ (independence) versus $K : \psi > 0$ based on a sample of n iid pairs from P_θ. Note that maximum likelihood estimation of $\theta = (\mu , \lambda , \psi)$ for general θ is not at all simple, so both the Wald and LR statistics must be calculated numerically. However, for $\theta \in \Theta_0 \equiv \{\theta \in \Theta : \psi = 0\}$, X and Y are just independent Poisson rv's and some calculaton shows that the scores, for any point $\theta_0 = (\lambda , \mu , 0) \in \Theta_0$ are given by

$$\dot{\mathbf{l}}_\mu(X, Y; \theta_0) = -1 + \frac{X}{\mu} ,$$

$$\dot{\mathbf{l}}_\lambda(X, Y; \theta_0) = -1 + \frac{Y}{\lambda} ,$$

and

$$\dot{\mathbf{l}}_\psi(X, Y; \theta_0) = -1 + \frac{XY}{\mu \lambda} .$$

Hence, under H the MLE of θ is $\hat{\theta}_n^0 = (\overline{X} , \overline{Y} , 0)$, and the Rao or scores statistic for testing H is

$$R_n = n \left(\frac{n^{-1} \sum_i X_i Y_i}{\overline{X}\,\overline{Y}} - 1 \right)^2 \overline{X}\,\overline{Y} = n \frac{\{n^{-1} \sum_i X_i Y_i - \overline{X}\,\overline{Y}\}^2}{\overline{X}\,\overline{Y}}$$

since

$$I(\theta_0) \;=\; \begin{pmatrix} 1/\mu & 0 & 1/\mu \\ 0 & 1/\lambda & 1/\lambda \\ 1/\mu & 1/\lambda & 1/\mu + 1/\lambda + 1/(\mu\lambda) \end{pmatrix}$$

so that $I_{22\cdot 1} \;=\; 1/(\mu\,\lambda)$.

$$I_{22\cdot 1} \;=\; 1/(\mu\,\lambda)\ .$$

Note that R_n is a close relative of the classical correlaton coefficient, and in fact, under H we have $R_n \;=\; nr_n^2 \;+\; o_p(1)$.

Note that the parameter space for this model is $\Theta \;=\; \{(\mu\,,\lambda\,,\psi):\ \mu > 0\,,\ \lambda > 0\,,\ \psi \geq 0\}$, which is not closed. Moreover the points in the null hypothesis $\psi = 0$ are on a boundary of the parameter space. Thus the theory we have developed does not quite apply. Nonetheless, under H we have $R_n \;\rightarrow_d\; \chi_1^2$. [Why is this so?]

Example 3.10. (The Multinomial Distribution). Suppose that $\underline{X}_1\,,\cdots,\underline{X}_n$ are i.i.d. $Mult_k(1\,,\underline{p})$. Then

$$\underline{N}_n \;\equiv\; \sum_{i=1}^{n} \underline{X}_i \;\sim\; Mult_k(n\,,\underline{p})\ .$$

Then the log likelihood is

$$\begin{aligned} \mathbf{l}(\underline{p}\,|\,\underline{X}) \;&=\; \log\left\{\prod_{i=1}^{n} \frac{1!}{X_{i1}!\cdots X_{ik}!}\, p_1^{X_{i1}}\cdots p_k^{X_{ik}}\right\} \\ &=\; \sum_{j=1}^{k} N_j \log p_j \;+\; \sum_{i=1}^{n} \log\left(\frac{1!}{X_{i1}!\cdots X_{ik}!}\right)\,, \end{aligned}$$

and the *constrained log - likelihood* is

$$\mathbf{l}(\underline{p},\lambda\,|\,\underline{X}) \;=\; \log\left\{\prod_{i=1}^{n} \frac{1!}{X_{i1}!\cdots X_{ik}!}\, p_1^{X_{i1}}\cdots p_k^{X_{ik}}\right\} \;+\; \lambda\Big(\sum_{j=1}^{k} p_j \;-\; 1\Big)\ .$$

Thus the likelihood equations are

$$0 \;=\; \dot{\mathbf{l}}_{p_j}(\underline{p}\,,\lambda\,|\,\underline{X}) \;=\; \frac{N_j}{p_j} \;+\; \lambda\,, \qquad j = 1\,,\cdots,k$$

and

$$0 \;=\; \sum_{j=1}^{k} p_j \;-\; 1\ .$$

The solution of the first set of k equations yields

$$\hat{p}_j \;=\; -\,\frac{N_j}{\lambda}\,, \qquad j \;=\; 1\,,\cdots,k\ .$$

But to satisfy the constraint we must have

$$1 = \sum_{j=1}^{k} \hat{p}_j = -\frac{n}{\lambda} ,$$

or $\lambda = -n$, and thus the MLE of $\underline{p}$ is

$$\underline{\hat{p}} = \underline{N}_n / n .$$

The score vector (for $n = 1$) becomes

$$\underline{\dot{\mathbf{l}}}(p \mid \underline{X}_1) = (\frac{X_{1j}}{p_j} - 1) = (\frac{X_{1j} - p_j}{p_j})_{j=1,\cdots,k} .$$

Thus the information matrix is

$$\begin{aligned} I(p) &= E(\underline{\dot{\mathbf{l}}}(p \mid \underline{X}_1)\underline{\dot{\mathbf{l}}}^T(p \mid \underline{X}_1)) = diag(1/p_j)(diag(p_j) - \underline{p}\underline{p}^T)diag(1/p_j) \\ &= diag(1/p_j) - \underline{1}\underline{1}^T . \end{aligned}$$

Since $I(p)\underline{p} = \underline{1} - \underline{1} = \underline{0}$, this matrix is *singular*. But note that it has a generalized inverse given by $I^-(p) = diag(p_j) - \underline{p}\underline{p}^T$:

$$I(p)I^-(p)I(p) = I(p) .$$

In fact, we know by direct calculations and the multivariate CLT that

$$\sqrt{n}(\underline{\hat{p}}_n - \underline{p}) \to_d N_k(0, I^-(p)) .$$

Note that the natural Rao statistic is in fact the usual Chi-square statistic:

$$\underline{Z}_n^T(\underline{p}_0)I^-(\underline{p}_0)\underline{Z}_n(\underline{p}_0) = \sum_{j=1}^{k} \frac{(N_j - np_{j0})^2}{np_{j0}}$$

where

$$\underline{Z}_n^T(p_0) = \frac{1}{\sqrt{n}}(\frac{N_1 - np_{10}}{p_{10}}, \cdots, \frac{N_k - np_{k0}}{p_{k0}})$$

and

$$I^-(\underline{p}_0) = diag(p_{j0}) - \underline{p}_0\underline{p}_0^T .$$

The Wald statistic is given by

$$\sqrt{n}(\underline{\hat{p}}_n - \underline{p}_0)^T \hat{I}(\underline{\hat{p}})\sqrt{n}(\underline{\hat{p}} - \underline{p}_0) = \sum_{j=1}^{k} \frac{(N_j - np_{j0})^2}{n\hat{p}_j} .$$

4. Consistency of Maximum Likelihood Estimates

Our goal here will be to prove strong consistency under somewhat restrictive hypotheses. The method we use here goes back to Wald (1949).

Let $0 < a < 1$ be fixed (ordinarily at a value close to 1). Then

$$\hat{\Theta}_n \equiv \{\theta^* \in \Theta : L_n(\theta^*) \geq aL_n(\Theta)\}$$

is called the set of *approximate MLE's.*

Theorem 4.1. (Wald). Suppose that:

(C1) $0 \leq E_{\theta_0}[l(\Theta \mid X) - l(\theta_0 \mid X)] < \infty$.

(C2) Θ is compact.

(C3) $P_{\theta_0}(\{x : p_\theta(x)$ is continuous in $\theta\}) = 1$ (or (C3') below holds).

Then $\hat{\Theta}_n \rightarrow_{a.s.} \theta_0$ with respect to P_{θ_0} ; i.e.

$$P_{\theta_0}(\lim_{n\to\infty} \sup_{\theta^* \in \hat{\Theta}_n} |\theta^* - \theta_0| = 0) = 1 .$$

Proof. Let $\theta \neq \theta_0$ where θ_0 is true. Then

$$(a) \quad -\infty \leq E_{\theta_0}[\mathbf{l}(\theta \mid X) - \mathbf{l}(\theta_0 \mid X)] = -K(P_{\theta_0}, P_\theta) < 0$$

by lemma 4.1.1 since $P_\theta \neq P_{\theta_0}$.

We also have

(b) $$\lim_{t\downarrow 0} E_{\theta_0}[\mathbf{l}(S_t(\theta)|X) - \mathbf{l}(\theta_0 \mid X)] = -K(P_{\theta_0}, P_\theta)$$

where $S_t(\theta)$ denotes the sphere about θ of radius t . To see this we note that the monotone convergence theorem yields $Ef_t \downarrow Ef_0$ when $0 \leq f - f_t \uparrow f - f_0$ with $f \geq 0$ having $Ef < \infty$. This applies with

(C3') $$f_t \equiv \mathbf{l}(S_t(\theta)\,|X) - \mathbf{l}(\theta_0 \,|\, X) \downarrow f_0 \equiv \mathbf{l}(\theta\,|X) - \mathbf{l}(\theta_0\,|X) \quad \text{a.s. } P_{\theta_0}$$

by (C3)and with $f_t \leq f \equiv \mathbf{l}(\Theta \mid X) - \mathbf{l}(\theta_0 \mid X)$ where $E_{\theta_0}(f) < \infty$ by (C1).

Combining (a) and (b) we see that there exists a number $t(\theta)$ such that

(c) $$E_{\theta_0}[\mathbf{l}(S_{t(\theta)}(\theta) \mid X) - \mathbf{l}(\theta_0 \mid X)] = \text{some } b(\theta) < 0 .$$

Let $\epsilon > 0$. Since $\Theta - S_\epsilon(\theta_0)$ is compact, it is covered by a finite number of spheres $S_i \equiv S_{t(\theta_i)}(\theta_i)$ having $1 \leq i \leq \text{some } M$. For each fixed i (using $\sup \sum \leq \sum \sup$)

(d) $$\mathbf{l}(S_i) - \mathbf{l}(\theta_0) \leq \sum_{j=1}^{n} [\mathbf{l}(S_i \mid X_j) - \mathbf{l}(\theta_0 \mid X_j)] \equiv \sum_{j=1}^{n} Z_{ij}$$

where the Z_{ij} 's are i.i.d. with $E_{\theta_0} Z_{ij} = b(\theta_i) < 0$. Thus for each i we have

$\sum_{j=1}^{n} Z_{ij} \to_{a.s.} -\infty$ w.r.t. P_{θ_0} by the strong law of large numbers (SLLN). hence for a.e. ω we have for $n \geq$ some N_ω that

(e) $\qquad \mathbf{l}(S_i) - \mathbf{l}(\theta_0) < \log a \qquad$ for all $\quad 1 \leq i \leq M$.

But if $\theta \in \hat{\Theta}_n$, then

$$L(\theta) \geq a\,L(\Theta) \geq a\,L(\theta_0)\,,$$

so that

(f) $\qquad \mathbf{l}(\theta) \geq \log a + \mathbf{l}(\theta_0) \qquad$ for $\quad \theta \in \hat{\Theta}_n$.

Thus for all $n \geq N_\omega$ we have $\theta \notin \cup_{i=1}^{M} S_i \supset \Theta - S_\epsilon(\theta_0)$; that is

$$\theta \in S_\epsilon(\theta_0) \qquad \text{for all} \quad n \geq N_\omega\,.$$

Thus $\hat{\Theta}_n \subset S_\epsilon(\theta_0)$ for all $n \geq N_\omega$. Note that our proof did not depend on the dimensionality of Θ , just its compactness. □

The set Θ is *compactible* if $\Theta \subset$ some compact $\bar{\Theta}$ for which the basic model properties hold (except we allow the new P_θ 's to be sub-probability distributions) and if $L(\bar{\Theta}) = L(\Theta)$ a.e. P_θ .

Theorem 4.2. If Θ is compactible by $\bar{\Theta}$ and (C1), (C2) and (C3) hold for $\bar{\Theta}$, then $\hat{\Theta}_n \to_{a.s.} \theta_0$ with respect to P_{θ_0} .

Example 4.3. (Cauchy) Let X be Cauchy with density $p_\theta(x) = \pi^{-1}[1 + (x-\theta)^2]^{-1}$. Clearly the basic model properties hold. also (i) holds since

$$L(\Theta \mid X) = \sup_\theta L(\theta \mid X) = \sup_\theta\{-\log\pi - \log(1 + (X-\theta)^2) = -\log\pi\,,$$

so that

$$\begin{aligned} E_\theta\{L(\Theta \mid X) - L(\theta \mid X)\} &= E_\theta\{\log(1 + (X-\theta)^2\} \\ &= \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{\log(1+x^2)}{1+x^2}\,dx < \infty\,. \end{aligned}$$

But (ii) clearly fails, so we must compactify Θ . Replace Θ by $\bar{\Theta} \equiv$ the unit circle making the identification shown and adding "the north pole"; and let the north pole correspond to the density function identically zero, so that (5) holds (the fact the added density is not a probability density does not disrupt our proof since (C3') does hold). The reparametrized Θ is now compactible by $\bar{\Theta}$ and $\hat{\Theta}_n \to_{a.s.} \theta_0$ a.s. P_{θ_0} .

Example 4.4. (Uniform$(0,\theta)$). Let X be Uniform$(0,\theta)$ with $\Theta = (0,\infty)$. Add point mass at 0 corresponding to $\theta = 0$. Add the zero measure corresponding to $\theta = \infty$. Let $[0,\infty]$ correspond to the points of the semicircle as in example 4.3. Let $\mu = \delta_0 + Lebesgue$. Then theorem 4.2 gives $\hat{\Theta}_n \rightarrow_{a.s.} \theta_0$ a.s. P_{θ_0} .

Example 4.5. (How Wald's theorem can fail when (C3') fails.) Suppose that the sample space is $\{0,1\}$, and let

$$p_\theta(1) = 1 - p_\theta(0) = \begin{cases} \theta & \text{for } \theta \text{ rational} \\ 1-\theta & \text{for } \theta \text{ irrational}. \end{cases}$$

Then the likelihood is given by

$$L(\theta \mid \underline{X}) = \begin{cases} \theta^{\sum_1^n X_i}(1-\theta)^{n-\sum_1^n X_i} & \text{for } \theta \text{ rational} \\ (1-\theta)^{\sum_1^n X_i}\theta^{n-\sum_1^n X_i} & \text{for } \theta \text{ irrational}. \end{cases}$$

The MLE is $\hat{\theta}_n = \bar{X}$ (corresponding to θ rational) since $L(\theta \mid \underline{X})$ approaches the same maximum over irrational θ's as θ approaches $1-\bar{X}$, but it does not achieve it since $1-\bar{X}$ is a rational number. Thus

$$\hat{\theta}_n \rightarrow_{a.s.} \begin{cases} \theta & \text{if } \theta \text{ is rational} \\ 1-\theta & \text{if } \theta \text{ is irrational}. \end{cases}$$

Thus $\hat{\theta}_n$ is not consistent for irrational θ's. Note that condition (C3) does not hold here. Neither does (C3'). □

Existence of MLE's

Notation:

- $\dot{\mathbf{l}}_i(\theta \mid \underline{X}) = \dfrac{\partial}{\partial\theta_i}\mathbf{l}(\theta \mid \underline{X})$ if it exists, $1 \le i \le d$.
- The *likelihood equations* are $\dot{\mathbf{l}}_i(\theta \mid \underline{X}) = 0$ for $1 \le i \le d$.
- $\hat{S}_n \equiv \{\theta \in \Theta : \dot{\mathbf{l}}_i(\theta \mid \underline{X}) = 0 , \ 1 \le i \le d\}$.
  $= \{\text{all solutions of the likelihood equations}\}$.

- $\tilde{S}_n^* \equiv \{\theta \in S_n : \mathbf{l}(\theta \mid \underline{X}) = \mathbf{l}(\tilde{S}_n \mid \underline{X})$,
  which may be void even if $\tilde{S}_n$ is ot void and which need not contain the MLE.
- $\hat{S}_n \equiv \{\theta \in \Theta : \mathbf{l}(\theta \mid \underline{X}) = \mathbf{l}(\Theta \mid \underline{X})\} \equiv \{$all MLE's$\}$.

Theorem x.xx. (When is the MLE a solution of the likelihood equations). Suppose that

(i) Θ is open.

(ii) $\mathbf{l}(\theta \mid X_1)$ is continuous and $\dot{\mathbf{l}}_i(\theta \mid X_1)$ exists for $1 \le i \le d$, for all $\theta \in \Theta$ with probability 1 .

(iii) For all θ, w.p. 1 , we have $\mathbf{l}(\Theta \mid \underline{X}) < \infty$.

Then $\hat{S}_n \subset \tilde{S}_n^* \subset \tilde{S}_n$ for all n w.p. 1 .

Proof. Now $\tilde{S}_n^* \subset \tilde{S}_n$ by definition. Let $\hat{\theta} \in \hat{S}_n$. Then w.p. 1

$$
\begin{aligned}
\mathbf{l}(\Theta \mid \underline{X}) &= \text{a finite maximum by (iii)} \\
&= \mathbf{l}(\hat{\theta} \mid \underline{X}) \qquad \text{since } \hat{\theta} \in \hat{S}_n \\
&= \text{the sum of } n \text{ continuous functions} \\
&= \sum_{j=1}^{n} \mathbf{l}(\hat{\theta} \mid X_j) ,
\end{aligned}
$$

and since $\dot{\mathbf{l}}_i(\hat{\theta} \mid \underline{X})$ exists for $1 \le i \le d$ with Θ open we have

$$\dot{\mathbf{l}}_i(\hat{\theta} \mid \underline{X}) = 0 \qquad \text{for } 1 \le i \le d ;$$

that is $\hat{\theta} \in \tilde{S}_n$. That $\hat{\theta} \in \tilde{S}_n^*$ follows from

$$\mathbf{l}(\hat{\theta} \mid \underline{X}) = \mathbf{l}(\Theta \mid \underline{X}) \ge \mathbf{l}(\tilde{S}_n \mid \underline{X}) . \qquad \square$$

4.5

5. The EM algorithm

In this section our goal is to introduce one quite general scheme for maximizing likelihoods, the *EM - algorithm.* The following examples will be helpful before trying to formalize the general method.

Example 5.1. (Multinomial) Suppose that $\underline{Y} \sim Mult_4(n = 197, \underline{p})$ where

$$\underline{p} = (\frac{1}{2} + \frac{1}{4}\theta, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{1}{4}\theta), \qquad 0 < \theta < 1 .$$

Thus

$$p_\theta(\underline{y}) = \frac{n!}{y_1! y_2! y_3! y_4!}(\frac{1}{2} + \frac{\theta}{4})^{y_1}(\frac{1}{4}(1-\theta))^{y_2}(\frac{1}{4}(1-\theta))^{y_3}(\frac{1}{4}\theta)^{y_4} ,$$

$$l(\theta \mid \underline{Y}) = Y_1 \log(\frac{1}{2} + \frac{\theta}{4}) + (Y_2 + Y_3)\log(1-\theta) + Y_4 \log\theta + const. ,$$

$$\dot{\mathbf{l}}_\theta(\underline{Y}) = Y_1 \frac{1/4}{1/2 + \theta/4} - (Y_2 + Y_3)\frac{1}{1-\theta} + Y_4 \frac{1}{\theta} ,$$

$$\ddot{\mathbf{l}}_{\theta\theta}(\underline{Y}) = -Y_1 \frac{(1/4)^2}{(1/2 + \theta/4)^2} - (Y_2 + Y_3)\frac{1}{(1-\theta)^2} - Y_4 \frac{1}{\theta^2} ,$$

and

$$I(\theta) = \frac{(1/4)^2}{1/2 + \theta/4} + \frac{1/2}{(1-\theta)} + \frac{1/4}{\theta} .$$

The one - step estimator of θ is then given by

$$\theta_n = \overline{\theta}_n + I(\overline{\theta}_n)^{-1}\frac{1}{n}\dot{\mathbf{l}}(\overline{\theta}_n) .$$

Thus, if $\underline{Y} = (125, 18, 20, 34)$ is observed, and $\overline{\theta}_n = 4Y_4/n = (4 \cdot 34)/197 = .6904$, then a one - step estimator is

$$\theta_n = .6904 + \frac{1}{2.07}\frac{1}{197}(-27.03) = .6904 - .0663 = .6241 .$$

Note that solving the likelihood equation $\dot{\mathbf{l}}_\theta(\underline{Y}) = 0$ involves solving a cubic equation.

Another approach to maximizing $l(\theta \mid \underline{Y})$ is via the E-M algorithm: suppose the "complete data" is

$$\underline{X} \sim Mult_5(n, \underline{p}) \quad \text{with} \quad \underline{p} = (\frac{1}{2}, \frac{\theta}{4}, \frac{1}{4}(1-\theta), \frac{1}{4}(1-\theta), \frac{\theta}{4}) ,$$

so that

$$p_\theta(\underline{x}) = \frac{n!}{\prod_{i=1}^5 x_i!}(\frac{1}{2})^{x_1}(\frac{\theta}{4})^{x_2 + x_5}(\frac{1}{4}(1-\theta))^{x_3 + x_4} ,$$

and the "incomplete data" $\underline{Y}$ is given in terms of the "complete data" $\underline{X}$ by

$$\underline{Y} = (X_1 + X_2, X_3, X_4, X_5) .$$

Then the "E - step" of the algorithm is to estimate $\underline{X}$ given $\underline{Y}$ (and θ):

$$\text{(a)} \qquad E(\underline{X} \mid \underline{Y}) = (Y_1 \frac{1/2}{1/2 + \theta/4}, Y_1 \frac{\theta/4}{1/2 + \theta/4}, Y_2, Y_3, Y_4) ,$$

so we set

$$\underline{\hat{X}}^{(p)} \equiv (Y_1 \frac{1/2}{1/2 + \theta^{(p)}/4}, Y_1 \frac{\theta^{(p)}/4}{1/2 + \theta^{(p)}/4}, Y_2, Y_3, Y_4) .$$

The "M - step" of the algorithm is to maximize the complete data likelihood with X replaced by our estimate of $E(\underline{X} \mid \underline{Y})$: it is easily seen that

$$\dot{\mathbf{l}}_\theta(\underline{X}) = (X_2 + X_5)\frac{1}{\theta} - \frac{X_3 + X_4}{1 - \theta} = 0$$

yields

$$\hat{\theta} = \frac{X_2 + X_5}{X_2 + X_5 + X_3 + X_4} .$$

Thus we take

$$\text{(b)} \qquad \hat{\theta}^{(p+1)} = \frac{\hat{X}_2^{(p)} + Y_4}{\hat{X}_2^{(p)} + Y_4 + Y_2 + Y_3} ,$$

and alternate between (a) and (b) with a "reasonable" guess $\hat{\theta}_n^0$ of θ; say $\hat{\theta}_n^0 \equiv 1/2$. This yields

p	$\hat{\theta}^{(p)}$	$\hat{\theta}^{(p)} - \theta^*$	$\frac{(\hat{\theta}^{(p+1)} - \theta^*)}{(\hat{\theta}^{(p)} - \theta^*)}$
0	.500000000	.126821498	.1465
1	.608247423	.018574075	.1346
2	.624321051	.002500447	.1330
3	.626488879	.000332619	.1328
4	.626777323	.000044176	.1328
5	.626815632	.000005866	.1328
6	.626820719	.000000779	
7	.626821395	.000000104	
8	.626821484	.000000014	

The exact root of the likelihood equation which maximizes the likelihood is $\hat{\theta}_n = .62682149\ldots$.

Example 5.2. (Exponential mixtures model). Suppose that $Y \sim p_\theta(y)$ where

$$p_\theta(y) = \{p\lambda e^{-\lambda y} + (1-p)\mu e^{-\mu y}\}1_{(0,\infty)}(y) ,$$

and $\theta = (p, \lambda, \mu) \in (0,1) \times R^{+2}$. Consider estimation of θ based on $Y_1, \cdots, Y_n$ iid $p_\theta(y)$. The scores for θ based on Y are

$$\dot{\mathbf{l}}_p(Y) = \frac{\lambda e^{-\lambda Y} - \mu e^{-\mu Y}}{p_\theta(Y)} ,$$

$$\dot{\mathbf{l}}_\lambda(Y) = \frac{pe^{-\lambda Y}(1 - \lambda Y)}{p_\theta(Y)} ,$$

and

$$\dot{\mathbf{l}}_\mu(Y) = \frac{(1-p)e^{-\mu Y}(1 - \mu Y)}{p_\theta(Y)} .$$

It is easily seen that $E_\theta \dot{\mathbf{l}}_p^2(Y) < \infty$, $E_\theta \dot{\mathbf{l}}_\lambda^2(Y) < \infty$, $E_\theta \dot{\mathbf{l}}_\mu^2(Y) < \infty$; and moreover that the information matrix $I(\theta)$ is nonsingular if $\lambda \neq \mu$. However, the likelihood equations are complicated and do not have "closed form" solutions.

Thus we will take an approach based on the EM algorithm. The natural complete data for this problem is $X = (Y, \delta) \sim p_\theta(x)$ where

$$p_\theta(x) = p_\theta(y, \delta) = (p\lambda e^{-\lambda y})^\delta ((1-p)\mu e^{-\mu y})^{1-\delta} , \quad y > 0 , \ \delta \in \{0,1\} .$$

Thus the "incomplete data" Y is just the first coordinate of X. Furthermore maximum likelihood estimation of θ in the complete data problem is easy. Now

$$l(\theta \,|\, X) = \delta \log p + (1-\delta)\log(1-p) + \delta(\log\lambda - \lambda Y) + (1-\delta)(\log\mu - \mu Y) ,$$

so that

$$\dot{\mathbf{l}}_p(X) = \frac{\delta}{p} - \frac{(1-\delta)}{1-p} , \qquad \hat{p} = \frac{1}{n}\sum_{i=1}^n \delta_i ,$$

$$\dot{\mathbf{l}}_\lambda(X) = \delta(\frac{1}{\lambda} - Y) , \qquad \frac{1}{\hat{\lambda}} = \frac{\sum_i Y_i\delta_i}{\sum_i \delta_i} ,$$

$$\dot{\mathbf{l}}_\mu(X) = (1-\delta)(\frac{1}{\mu} - Y) , \qquad \frac{1}{\hat{\mu}} = \frac{\sum_i Y_i(1-\delta_i)}{\sum_i (1-\delta_i)} .$$

This gives the "M - step" of the E-M algorithm. To find the E - step, we compute

$$E(\delta \,|\, Y) = \frac{p\lambda e^{-\lambda Y}}{p\lambda e^{-\lambda Y} + (1-p)\mu e^{-\mu Y}} \equiv \hat{\delta}(Y) \equiv p(Y;\theta)$$

since $(\delta \mid Y) \sim \text{Bernoulli}(p(Y;\theta))$. Thus the E - M algorithm becomes: $\hat{\theta}^{(m+1)} = (\hat{p}^{(m+1)}, \hat{\lambda}^{(m+1)}, \hat{\mu}^{(m+1)})$ where

$$\hat{p}^{(m+1)} = \frac{1}{n}\sum_{i=1}^{n} \hat{\delta}_i^{(m)}, \qquad \delta_i^{(m)} \equiv \hat{\delta}(Y_i;\hat{\theta}^{(m)}),$$

$$\frac{1}{\hat{\lambda}^{(m+1)}} \equiv \frac{\sum_i Y_i \hat{\delta}_i^{(m)}}{\sum_i \hat{\delta}_i^{(m)}},$$

and

$$\frac{1}{\hat{\mu}^{(m+1)}} \equiv \frac{\sum_i Y_i (1 - \hat{\delta}_i^{(m))}}{\sum_i (1 - \hat{\delta}_i^{(m)})}.$$

Example 5.3. (Bivariate exponential mixtures model). Consider the following model: suppose that $X = (S,T,Z)$ with density

$$p_\theta(x) = p_\theta(s,t,z) = \lambda z e^{-\lambda z s} \mu z e^{-\mu z t} \frac{z^{\eta-1}}{\Gamma(\eta)} e^{-z} 1_{(0,\infty)}(s) 1_{(0,\infty)}(t) 1_{(0,\infty)}(z)$$

where $\theta = (\lambda, \mu, \eta) \in R^{+3}$ is the "complete data," and $Y = (S,T)$ has the (mixture) density

$$p_\theta(y) = p_\theta(s,t) = \frac{\lambda\mu\eta(\eta+1)}{[1 + \lambda s + \mu t]^{\eta+2}} 1_{[0,\infty)}(s) 1_{[0,\infty)}(t).$$

It is clear that the score equations for θ based on iid copies of $Y = (S,T)$ do not have closed form solutions. However, the score equations for θ based on iid copies of $X = (S,T,Z)$ are easily solved: since

$$\begin{aligned} l(\theta) &= \log p_\theta(X) \\ &= \log\lambda - \lambda Z S + \log\mu - \mu Z T + (\eta - 1)\log Z - \log\Gamma(\eta) + \text{const.}, \end{aligned}$$

we have

$$\dot{\mathbf{l}}_\lambda(X) = \frac{1}{\lambda} - Z\,S, \qquad \dot{\mathbf{l}}_\mu(X) = \frac{1}{\mu} - Z\,T,$$

and

$$\dot{\mathbf{l}}_\eta(X) = \log Z - \psi(\eta)$$

where $\psi = (\log\Gamma)'$ is the digamma function. Thus, for the complete data,

$$\frac{1}{\hat{\lambda}} = \overline{ZS} \equiv \frac{1}{n}\sum_{i=1}^{n} Z_i S_i,$$

$$\frac{1}{\hat{\mu}} = \overline{ZT} \equiv \frac{1}{n}\sum_{i=1}^{n} Z_i T_i,$$

and

$$\hat{\eta} \;=\; \psi^{-1}(\overline{\log Z}) \;=\; \psi^{-1}(\frac{1}{n}\sum_{i=1}^{n} \log Z_i) \;.$$

This gives the basic "M - step" for an E - M algorithm. Moreover, it is easy to compute

$$p_\theta(z|s,t) \;=\; \frac{p_\theta(s,t,z)}{p_\theta(s,t)} \;=\; (1 + \lambda s + \mu t)^{\eta + 2} \frac{z^{\eta + 2 - 1}}{\Gamma(\eta+2)} \, e^{-(1+\lambda s + \mu t)z} 1_{(0,\infty)}(z) \;;$$

i.e. $(Z \mid S,T) \sim \Gamma(\eta + 2 \,, c_\theta(S,T))$ where

$$c_\theta(S,T) \;\equiv\; 1 \;+\; \lambda S \;+\; \mu T \;.$$

Hence

$$E(Z \mid S,T) \;=\; \frac{\eta + 2}{c_\theta(S,T)} \;\equiv\; \hat{Z}(S,T\,;\theta) \;.$$

Thus, starting from some initial value of θ, say $\theta^{(0)} \equiv (1,1,1)$, we can use an E-M algorithm to maximize the likelihood of the Y_i's:

$$\frac{1}{\hat{\lambda}^{(m+1)}} \;\equiv\; \frac{1}{n}\sum_{i=1}^{n} \hat{Z}_i^{(m)} S_i \;,$$

$$\frac{1}{\hat{\mu}^{(m+1)}} \;\equiv\; \frac{1}{n}\sum_{i=1}^{n} \hat{Z}_i^{(m)} T_i \;,$$

and

$$\hat{\eta}^{(m+1)} \;=\; \psi^{-1}(\frac{1}{n}\sum_{i=1}^{n} \log \hat{Z}_i^{(m)})$$

where $\hat{Z}_i^{(m)} \equiv \hat{Z}(S_i\,, T_i\,; \hat{\theta}^{(m)})$. Computation of ψ^{-1} is the only difficult part of this algorithm; once we have a good table for ψ, this is straightforward. See Abramowitz and Stegun (1965), pages 267 - 293.

6. Nonparametric Maximum Likelihood Estimation

The maximum likelihood method has also been used successfully in a variety of nonparametric problems. As in section 5, we will begin with several examples.

Example 6.1. Let $X_1, \cdots, X_n$ be iid $\mathbf{X}$ − valued random variables with common probability distribution (measure) P on $(\mathbf{X}, \mathbf{A})$. For an arbitrary probability distribution (measure) Q on $(\mathbf{X}, \mathbf{A})$, let $Q(\{X_i\}) \equiv q_i$. If there are no ties in the X_i's, then

$$L(Q \mid \underline{X}) = \prod_{i=1}^{n} q_i \equiv L(\underline{q} \mid \underline{X}) .$$

Consider maximizing this likelihood as a function of $\underline{q}$. By Jensen's inequality and concavity of $\log x$,

$$\frac{1}{n} \sum_{i=1}^{n} \log q_i \le \log(\overline{q}) ,$$

or

$$\{ \prod_{i=1}^{n} q_i \}^{1/n} \le \overline{q}$$

with equality iff $q_1 = \cdots = q_n = \overline{q}$. Since $\sum q_i \le 1$, $\overline{q} \le 1/n$. Thus

$$L(\underline{q} \mid \underline{X}) = \prod_{i=1}^{n} q_i \le \overline{q}^n \le (\frac{1}{n})^n$$

with equality iff $\sum q_i = 1$ and $q_1 = \cdots = q_n = 1/n$. Thus the MLE of P is $I\!P_n$ with $I\!P_n(\{X_i\}) = 1/n$, or

$$I\!P_n(A) = \frac{1}{n} \#\{ i \le n : X_i \in A \} = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i}(A) .$$

This extends easily to the case of ties (homework!). Thus we have proved:

Theorem 6.2. The nonparametric maximum likelihood estimate of a completely arbitrary distribution P on any measurable space based on iid data is the empirical measure $I\!P_n = n^{-1} \sum_{i=1}^{n} \delta_{X_i}$.

In particular, if $\mathbf{X} = R$, $I\!F_n(x) \equiv n^{-1} \sum_{i=1}^{n} 1_{[X_i \le x]}$ is the MLE of F. Recall from chapter 2 that Donsker's theorem yields

$$\sqrt{n}(I\!F_n - F) \stackrel{d}{=} I\!U_n(F) \Longrightarrow I\!U(F)$$

where $\mathbb{U}_n$ is the empirical process of iid $U(0,1)$ rv's and $\mathbb{U}$ is a Brownian bridge process.

Example 6.3. (Censored data). Suppose that $X_1, \cdots, X_n$ are iid F, and $Y_1, \cdots, Y_n$ are iid G. Think of the X's as survival times and the Y's as censoring times. Suppose that we can observe only the iid pairs $(Z_1, \delta_1), \cdots, (Z_n, \delta_n)$ where $Z_i \equiv X_i \wedge Y_i$ and $\delta_i \equiv 1[X_i \le Y_i]$. Thus the joint distribution of (Z, δ) is given by

$$H^{uc}(z) \equiv P(Z \le z, \delta = 1) = \int_{[0,z]} (1 - G(x-))\, dF(x)$$

where $G(x-) \equiv \lim_{y \uparrow x} G(y)$, and

$$H^{c}(z) \equiv P(Z \le z, \delta = 0) = \int_{[0,z]} (1 - F(y))\, dG(y) .$$

Furthermore, the survival function $1 - H(z) = P(Z > z)$ is given by

$$1 - H(z) = (1 - F(z))(1 - G(z)) .$$

Now suppose for simplicity that $Z_{n:1} \le \cdots \le Z_{n:n}$ are all distinct. Let $\delta_{n:1}, \cdots, \delta_{n:n}$ be the corresponding δ's. If we let $p_i \equiv F(Z_{n:i}) = F(Z_{n:i}) - F(Z_{n:i}-) = \Delta F(Z_{n:i})$ and $q_i \equiv G(Z_{n:i}) = G(Z_{n:i}) - G(Z_{n:i}-) = \Delta G(Z_{n:i})$, $i = 1, \cdots, n$, $p_{n+1} = 1 - F(Z_{n:n}) = 1 - \sum_{j=1}^{n} p_j$, then a nonparametric likelihood for the censored data model is:

$$\prod_{i=1}^{n} p_i^{\delta_{n:i}} (\sum_{j=i}^{n} q_j)^{\delta_{n:i}} q_i^{1-\delta_{n:i}} (\sum_{j=i+1}^{n+1} p_j)^{1-\delta_{n:i}} = \prod_{i=1}^{n} p_i^{\delta_{n:i}} (\sum_{j=i+1}^{n} p_j)^{1-\delta_{n:i}} \times B$$

where B depends only on G. Thus we can find the nonparametric maximum likelihood estimator of F by maximizing the first term over the p_i's. See the homework set!

Here we will take a different approach by using example 1 as follows. If F has density f, then the hazard function λ is given by $\lambda(t) = f(t)/(1 - F(t))$, and the cumulative hazard function Λ is

$$\Lambda(t) = \int_0^t \lambda(s)\, ds = \int_0^t \frac{f(s)}{1 - F(s)}\, ds = \int_0^t \frac{1}{1 - F(s)}\, dF(s) .$$

For an arbitrary df F, it turns out that the "right" way to define Λ, the cumulative hazard function corresponding to F, is:

(1) $$\Lambda(t) \equiv \int_{[0,t]} \frac{1}{1 - F(s-)}\, dF(s) .$$

Note that we can write Λ as

$$\Lambda(t) \equiv \int_{[0,t]} \frac{(1 - G(s-))}{(1 - G(s-))(1 - F(s-))} dF(s) = \int_{[0,t]} \frac{1}{1 - H(s-)} dH^{uc}(s) .$$

Note that we can estimate both H^{uc} and H from the Z, δ observations:

$$I\!\!H_n^{uc}(z) = \frac{1}{n}\sum_{i=1}^{n} 1_{[Z_i \le z, \delta_i = 1]} , \qquad I\!\!H_n(z) = \frac{1}{n}\sum_{i=1}^{n} 1_{[Z_i \le z]} .$$

Thus a natural "nonparametric maximum likelihood" estimator of Λ is $\hat{\Lambda}_n$ given by

(2) $$\hat{\Lambda}_n(t) = \int_0^t \frac{1}{1 - I\!\!H_n(s-)} dI\!\!H_n^{uc}(s) .$$

It remains only to invert the relationship (1) to obtain an estimator of F. To do this we need one more piece of notation: for any nondecreasing right-continuous function A, we define the continuous part A_c of A by

$$A_c(t) \equiv A(t) - \sum_{s \le t} \Delta A(s) , \quad \Delta A(s) \equiv A(s) - A(s-) .$$

Proposition 6.4. Suppose that Λ is the cumulative hazard function corresponding to an arbitrary df F as defined by (1). Then

(3) $$1 - F(t) = exp(- \Lambda_c(t)) \prod_{s \le t} (1 - \Delta\Lambda(s)) \equiv \prod_{s \le t} (1 - d\Lambda(s)) .$$

Proof. In the case of a continuous df F, Λ is also continuous, $\Lambda = \Lambda_c$, and we calculate $\Lambda(t) = -\log(1 - F(t))$ so that (3) holds.

In the case of a purely discrete df F the cumulative hazard function Λ is also discrete so that $\Lambda_c \equiv 0$, and

$$1 - \Delta\Lambda(s) = 1 - \frac{\Delta F(s)}{1 - F(s-)} = \frac{1 - F(s)}{1 - F(s-)} .$$

Thus

$$\begin{aligned}
\prod_{s \le t} (1 - \Delta\Lambda(s)) &= \frac{1 - F(s_1)}{1 - F(s_1-)} \times \frac{1 - F(s_2)}{1 - F(s_2-)} \times \cdots \times \frac{1 - F(s_k)}{1 - F(s_k-)} \\
&= \frac{1 - F(s_1)}{1} \times \frac{1 - F(s_2)}{1 - F(s_1)} \times \cdots \times \frac{1 - F(t)}{1 - F(s_{k-1})} \\
&= 1 - F(t)
\end{aligned}$$

where s_j, $j = 1, \cdots, k$ are the points of jump of F which are less than or equal to t. Hence (3) also holds in this case. For a complete proof of the general case, which relies on rewriting (1) as

$$F(t) = \int_0^t (1 - F(s-))\, d\Lambda(s)$$

and hence

$$1 - F(t) = 1 - \int_0^t (1 - F(s-))\, d\Lambda(s)\ ,$$

see e.g. Liptser and Shiryayev (1978), lemma 18.8, page 255. For a still more general (Doleans - Dade) formula which is valid for martingales, see Shorack and Wellner (1986), page 897. □

The right side of (3) is called the *product integral*; see Gill and Johansen (1990) for a survey. It follows from proposition 1 that the nonparametric maximum likelihood estimator of F in the case of censored data is the *product limit estimator* $\hat{\mathbb{F}}_n$ given by

$$(4) \qquad 1 - \hat{\mathbb{F}}_n(t) = \prod_{s \le t} (1 - \Delta\hat{\Lambda}_n(s))$$

$$= \prod_{i\,:\,Z_{n:i} \le t} (1 - \frac{\delta_{n:i}}{n - i + 1}) \qquad \text{if there are no ties .}$$

This estimator was found by Kaplan and Meier (1958). Breslow and Crowley (1974) proved, using empirical process theory, that

$$(5) \qquad \sqrt{n}(\hat{\Lambda}_n - \Lambda) \Longrightarrow \mathbb{B}(C)\ , \qquad \text{in} \quad D[0,\tau] \qquad \tau < \tau_H\ ;$$

and hence that

$$(6) \qquad \sqrt{n}(\hat{\mathbb{F}}_n - F) \Longrightarrow (1 - F)\,\mathbb{B}(C) \overset{d}{=} \frac{1 - F}{1 - K}\,\mathbb{U}(K) \qquad \text{in} \quad D[0,\tau] \qquad \tau < \tau_H$$

where $\mathbb{B}$ denotes standard brownian motion, $\mathbb{U}$ denotes standard brownian bridge, and

$$C(t) \equiv \int_0^t \frac{1}{(1 - H_-)^2}\, dH^{uc}\ , \qquad K(t) \equiv \frac{C(t)}{1 + C(t)}\ .$$

Note that when there is no censoring $K = F$ and the limit process in (6) becomes just $\mathbb{U}(F)$, the limit process of the usual empirical process. Martingale methods for proving the convergences (5) and (6) are due to Gill (1980), (1983). See Shorack and Wellner (1986), chapter 7.

Example 6.5. (Cox's proportional hazards model and profile likelihood). Suppose that T is a survival time and Z is a real - valued covariate or predictor. Further, suppose that $(T \mid Z)$ has conditional hazard function

$$\lambda(t \mid z) = e^{\theta z}\lambda(t).$$

Here $\theta \in R$ and λ is an (unknown) baseline hazard function. Thus

$$\Lambda(t \mid z) \; = \; e^{\theta z} \Lambda(t) \, ,$$

and

$$\overline{F}(t \mid z) \; = \; \overline{F}(t)^{e^{\theta z}} \, ,$$

or,

$$f(t \mid z) \; = \; e^{\theta z} \overline{F}(t)^{e^{\theta z}} \lambda(t)$$

If we assume that Z has density h, then

$$p(t \, , z) \; = \; p(t, z \, ; \theta \, , \lambda) \; = \; e^{\theta z} \overline{F}(t)^{e^{\theta z}} \lambda(t) \, h(z) \; .$$

Hence

$$\log p(T, Z; \theta, \lambda) \; = \; \theta Z \; - \; e^{\theta Z} \Lambda(T) \; + \; \log \lambda(T) \; + \; \log h(Z) \; .$$

Suppose that $(T_1 \, , Z_1) \, , \cdots , (T_n \, , Z_n)$ are iid p. Assume that $0 < T_{(1)} < \cdots < T_{(n)}$ are the ordered T_i's and $Z_{(1)} , \cdots , Z_{(n)}$ are the corresponding Z_i's. Then, letting $\lambda_i \equiv \Delta\Lambda(T_{(i)})$, $\Lambda(T_{(i)}) = \sum_{j \leq i} \lambda_j$, and $h_i \equiv H(\{Z_{(i)}\})$. a natural nonparametric loglikelihood is given by

$$\begin{aligned} l(\lambda \, , \theta \, , h \vert \, \underline{X}) \; &= \; \sum_{i=1}^{n} \{ \theta Z_{(i)} \; - \; e^{\theta Z_{(i)}} \sum_{j \leq i} \lambda_j \; + \; \log \lambda_i \; + \; \log h_i \} \\ &= \; \sum_{i=1}^{n} \theta Z_{(i)} \; + \; \sum_{i=1}^{n} \{ \log \lambda_i \; - \; \lambda_i \sum_{j \geq i} e^{\theta Z_{(j)}} \} \; + \; \sum_{i=1}^{n} \log h_i \; . \end{aligned}$$

Maximizing this with respect to λ_i and h_i (subject to the constraint $\sum_{i=1}^{n} h_i = 1$) yields

$$\hat{\lambda}_i \; = \; \frac{1}{\sum_{j \geq i} \exp(\theta Z_{(j)})} \, , \qquad \hat{h}_i \; = \; 1/n \; .$$

Thus the profile likelihood for θ is given by

$$\tilde{l}(\theta \mid \underline{X}) \; = \; \log \left\{ \prod_{i=1}^{n} \frac{\exp(\theta Z_{(i)})}{\sum_{j \geq i} \exp(\theta Z_{(j)})} \, \frac{1}{(ne)^n} \right\} .$$

The first factor here is (the log of) *Cox's partial likelihood* for θ; Cox (1972) derived this by other means. Maximizing it over θ yields Cox's partial likelihood estimator of θ -- which is in fact the maximum (nonparametric or semiparametric) profile likelihood estimator. Let $\hat{\theta}$ denote Cox's partial likelihood estimator of θ. It turns out that this estimator is (asymptotically) efficient; this was proved by Efron (1977) and Begun, Hall, Huang, and Wellner (1983). Furthermore the natural cumulative hazard function estimator is just

$$\hat{\Lambda}_n(t) = \sum_{T_{(i)} \le t} \frac{1}{\sum_{j \ge i} \exp(\hat{\theta}\, Z_{(j)})} = \int_0^t \frac{1}{\mathbb{Y}_n(s, \hat{\theta})}\, d\mathbb{H}_n(s)$$

where $\mathbb{H}_n(t) = n^{-1}\sum_{i=1}^n 1_{[T_i \le t]}$, $\mathbb{Y}_n(t, \theta) \equiv n^{-1}\sum_{i=1}^n 1_{[T_i \ge t]} \exp(\theta Z_i)$. This estimator was suggested by Breslow (1972), (1974). It is also asymptotically efficient. See Begun, Hall, Huang, and Wellner (1983) and Bickel, Klaassen, Ritov, and Wellner (1993).

Example 6.6. (Estimation of a concave distribution function). Suppose that the model $\mathbf{P}$ is all probability distributions P on $R^+ = [0, \infty)$ with corresponing df's F which are concave. It follows that the df F corresponding to $P \in \mathbf{P}$ has a density f and that f is nonincreasing. It was shown by Grenander (1956) that if $X_1, \cdots, X_n$ are iid $P \in \mathbf{P}$ with df F, then the MLE of F over $P \in \mathbf{P}$ is the least concave majorant $\hat{\mathbb{F}}_n$ of $\mathbb{F}_n$; and thus the MLE $\hat{f}_n$ of f is given by the slope of $\hat{\mathbb{F}}_n$. See Barlow, et al. (1972) for this and related results. It was shown by Kiefer and Wolfowitz (1976) that

$$\sqrt{n}(\hat{\mathbb{F}}_n - F) \Longrightarrow \mathbb{U}(F) ,$$

and this phenomena of no improvement or reduction in asymptotic variance even though the model $\mathbf{P}$ is not $\mathbf{M} \equiv \{\text{all } P \text{ on } R^+\}$ is explained by Millar (1979). Prakasa Rao (1969) showed that if $f(t) > 0$, then

$$n^{1/3}(\hat{f}_n(t) - f(t)) \to_d 2\,|\frac{1}{2} f(t) f'(t)|^{1/3} Z$$

where Z is the location of the maximum of the process $\{B(t) - t^2 : t \in R\}$ where $B(t)$ is standard two-sided Brownian motion starting from 0 ; his proof has been greatly simplified and the limit distribution examined in detail by Groeneboom (1984), (1989).

Open problem: What is the limit distribution of $n^{1/3}(\hat{f}_n(t) - f(t))$ where $\hat{f}_n$ is the slope of the least concave majorant $\hat{\mathbb{F}}_n^{concave}$ of the product limit estimator $\hat{\mathbb{F}}_n$ of example 5.2?

Example 6.7. (Interval censoring). Suppose, as in example 6.2, that $X_1, \cdots, X_n$ are iid F and that $Y_1, \cdots, Y_n$ are iid G, but now suppose that we only observe (Z_i, δ_i) , $i = 1, \cdots, n$ where $Z_i = Y_i$, $\delta_i = 1_{[X_i \le Y_i]}$. Again the goal is to estimate F (even though we never observe an X directly, but only the indicators δ). If G has density g , then the density $p(z, \delta)$ of the iid (Z, δ) pairs is

$$p(z, \delta) = F(z)^{\delta}(1 - F(z))^{1-\delta} g(z) .$$

If we suppose that there are no ties in the Z's and write $P_i \equiv F(Z_{n:i})$, $q_i \equiv G(\{Z_{n:i}\})$, then a nonparametric likelihood for the data is

$$L(\underline{P}, \underline{q} \mid \underline{Z}, \underline{\delta}) = \prod_{i=1}^{n} P_i^{\delta_{n:i}} (1 - P_i)^{1-\delta_{n:i}} q_i ,$$

or

$$l(\underline{P}, \underline{q}) = \log L = \sum_{i=1}^{n} \{\delta_{n:i} \log(P_i) + (1 - \delta_{n:i}) \log(1 - P_i) + q_i\} ,$$

and we want to maximize this subject to the order restrictions $0 \le P_1 \le \cdots \le P_n \le 1$. Groeneboom and Wellner (1992) show that the NPMLE can be described as follows:

(i) Plot the points $(i, \sum_{j \le i} \delta_{n:j})$, $i = 1, \cdots, n$. This is called the cumulative sum diagram.

(ii) Form $H^*(t)$, the greatest convex minorant of the cumulative sum diagram in (i).

(iii) Let $\hat{P}_i \equiv$ the left derivative of H^* at i, $i = 1, \cdots, n$.

Then $\underline{\hat{P}} = (\hat{P}_1, \cdots, \hat{P}_n)$ is the unique vector maximizing $l(\underline{P}, q)$.

We define $I\hat{F}_n$ to be the piecewise constant function whcih equals $\hat{P}_i$ on the interval $[Z_{n:i}, Z_{n:i+1})$. Groeneboom and Wellner (1992) show that if $f(t), g(t) > 0$, then

$$n^{1/3}(I\hat{F}_n(t) - F(t)) \rightarrow_d 2(\frac{1}{2} F(t)(1 - F(t)) \frac{f(t)}{g(t)})^{1/3} Z$$

where Z is the location of the maximum of the process $\{B(t) - t^2 : t \in R\}$ as in example 6.6.

For further discussion of the definition of nonparametric maximum likelihood estimators, see Kiefer and Wolfowitz (1956) and Scholz (1980). For an application to a mixture model, see Jewell (1982). Othere successful applications of nonparametric maximum likelihood include the work of Vardi (1985) and Gill, Vardi, and Wellner (1988) on biased sampling models, and Bailey (1984) on the Cox regression model. Nonparametric maximum likelihood estimators may be inconsistent; see. e.g Boyles, Marshall, and Proschan (1985), and Barlow, Bartholomew, Bremner, and Brunk (1972), pages 257 - 258. As yet there is no unified theory for nonparametric maximum likelihood estimates.

7. Limit theory for the statistical agnostic: $P \notin \mathbf{P}$

In the preceding sections we have studied the limit behavior the the MLE $\hat{\theta}_n$ (or ELE $\tilde{\theta}_n$) under the assumption that the model $\mathbf{P}$ is true; i.e. assuming that the data $X_1, \cdots, X_n$ were governed by a probability distribution $P_\theta \in \mathbf{P}$. Frequently however we are in the position of not being at all sure that the true P is an element of the model $\mathbf{P}$, and it is natural to ask about the asymptotic behavior of $\hat{\theta}_n$ (or $\tilde{\theta}_n$) when, in fact, $P \notin \mathbf{P}$. This point of view is implicit in the robustness literature, and especially in the work of Huber (1964), (1967) and White (1982).

We begin here with a heuristic and rather informal treatment which will then be made rigorous using additional (convexity) hypotheses. For related results, see Pollard (1985), Pakes and Pollard (1989), Bickel, Klaassen, Ritov, and Wellner (1993), Appendix A.10 and sections 7.2 - 7.4.

Heuristics for Maximum Likelihood

Suppose (temporarily) that $X_1, \cdots, X_n$ are i.i.d. P on $(\mathbf{X}, \mathbf{A})$, and that

$$\rho(x;\theta) \equiv \log p(x;\theta), \quad x \in \mathbf{X}, \quad \theta \in \Theta \subset R^d$$

is twice continuously differentiable in θ for $P-$a.e. x. [We do *not* assume that $P \in \{P_\theta : dP_\theta/d\mu = p_\theta, \ \theta \in \Theta\}$.] Let

$$\psi(x;\theta) \equiv \nabla_\theta \rho(x;\theta),$$

and suppose that $E\,\rho(X_1;\theta) < \infty$ and $E\,|\psi(X_1,\theta)|^2 < \infty$ for all $\theta \in \Theta$.

Suppose that P has density p with respect to a measure μ which also dominates all P_θ, $\theta \in \Theta$. Then the maximum likelihood estimator maximizes

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \rho(X_i;\theta) \ \to_{a.s.}\ E_P\rho(X_1;\theta) &= E_P \log p(X_1;\theta) \\ &= E_P \log p(X_1) - E_P \log \frac{p(X_1)}{p(X_1,\theta)} \\ &= E_P \log p(X_1) - K(P, P_\theta).\end{aligned}$$

Since $K(P,P_\theta) \geq 0$, the last quantity is maximized by choosing θ to make $K(P,P_\theta)$ as small as possible:

$$\begin{aligned}\sup_\theta \{E_P \log p(X_1) - K(P,P_\theta)\} &= E_P \log p(X_1) - \inf_\theta K(P,P_\theta) \\ &= E_P \log p(X_1) - K(P, P_{\theta_0})\end{aligned}$$

if we suppose that the infimum is achieved at $\theta_0 \equiv \theta_0(P)$. Thus it is natural to expect that (under reasonable additional conditions)

$$\hat{\theta}_n = \text{argmax}(\frac{1}{n}\sum_{i=1}^{n} \rho(X_i;\theta))$$

$$\to_p \text{argmax}\{E_P \log p(X_1) - K(P, P_\theta)\} = \theta_0(P) .$$

What about a central limit theorem? First note that since

$$\theta_0(P) \quad \text{maximizes} \quad E_P \rho(X_1;\theta) = P\,\rho(X_1;\theta) = \int \rho(x;\theta)\,dP(x)$$

and

$$\hat{\theta}_n \quad \text{maximizes} \quad \frac{1}{n}\sum_{i=1}^{n}\rho(X_i;\theta) = I\!P_n\,\rho(x;\theta) ,$$

we expect that

$$0 = \nabla_\theta\, E_P\, \rho(X_1;\theta)|_{\theta=\theta_0} = E_P\,\psi(X_1;\theta_0)$$

and

$$0 = \nabla_\theta\, I\!P_n\,\rho(x;\theta)|_{\theta=\hat{\theta}_n} = I\!P_n\,\psi(x;\hat{\theta}_n) = \frac{1}{n}\sum_{i=1}^{n}\psi(X_i,\hat{\theta}_n) .$$

Therefore, by Taylor expansion of $I\!P_n\psi(x,\theta)$ about θ_0, it follows that

(1)
$$0 = \Psi_n(\hat{\theta}_n) \equiv \frac{1}{n}\sum_{i=1}^{n}\psi(X_i;\hat{\theta}_n)$$

$$= \Psi_n(\theta_0) + \dot{\Psi}_n(\theta_n^*)(\hat{\theta}_n - \theta_0)$$

where

$$\sqrt{n}\Psi_n(\theta_0) = \sqrt{n}I\!P_n\psi(\cdot;\theta_0)$$

$$= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(X_i;\theta_0) \to_d Z\!\!Z \sim N_d(0,K)$$

with

$$K \equiv E_P\,\psi(X_1;\theta_0)\,\psi^T(X_1;\theta_0) .$$

We also have

$$\dot{\Psi}_n(\theta_0) = I\!P_n\dot{\psi}(x,\theta_0) \to_{a.s.} E_P\dot{\psi}(X_1,\theta_0) ,$$

and hence we also expect to be able to show that

$$\dot{\Psi}_n(\theta_n^*) \to_p E_P\dot{\psi}(X_1,\theta_0) \equiv J \qquad d\times d .$$

Therefore if J is nonsingular we conclude from (1) that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d -J^{-1}Z\!\!Z \sim N_d(0, J^{-1}K(J^{-1})') .$$

Note that if, in fact, $P \in \mathbf{P}$, then

$$K = -J = I_\theta ,$$

and the asymptotic variance - covariance matrix $J^{-1}K(J^{-1})'$ reduces to just the classical and familiar inverse information matrix I_θ^{-1} .

Maximum Likelihood Estimators when the Likelihood is Convex

In the following we will make the preceeding heuristic rigorous under convexity hypothesis on the likelihood function(s). The following two lemmas are basic to our treatment.

Lemma 7.1. Suppose that A_n is a sequence of random convex functions defined on an open convex set $\Theta \subset R^d$ which converges in probability to some (fixed, nonrandom) function A for each fixed $\theta \in \Theta$; $A_n(\theta) \rightarrow_p A(\theta)$ for $\theta \in \Theta$. Then A_n converges to A uniformly over compact subsets K of Θ :

$$\sup_{\theta \in K} |A_n(\theta) - A(\theta)| \rightarrow_p 0 .$$

Proof. The following proof is from Andersen and Gill (1982), page 1116; a different proof is given by Pollard (1991), section 6.

Convexity of A is obvious. Let $\theta_1 , \theta_2 , \cdots$ be a countable dense set of points in Θ . By a diagonalization argument, there exists a subsequence $\{n_k\}$ such that

$$A_{n_k}(\theta_j) \rightarrow_{a.s.} A(\theta_j) \quad \text{for each} \quad j = 1, 2, \cdots .$$

By Rockafellar (1970), theorem 10.8 it follows that

$$\sup_{\theta \in K} |A_{n_k}(\theta) - A(\theta)| \rightarrow_{a.s.} 0$$

along this subsequence. Since this same argument yields, for any initially given subsequence $\{n'\}$, a further subsequence $\{n''\}$ for which $||A_{n''} - A||_K \rightarrow_{a.s.} 0$, it follows that $||A_n - A||_K \rightarrow_p 0$ for the whole sequence. □

Lemma 7.2. (Convergence of argmin's of convex functions.) Suppose that A_n is convex for each n ; let

$$\alpha_n \equiv \operatorname*{argmin}_\theta A_n(\theta) \qquad \text{and} \qquad \beta_n \equiv \operatorname*{argmin}_\theta B_n(\theta) .$$

Then

$$P(|\alpha_n - \beta_n| \geq \delta) \leq P(\Delta_n(\delta) \geq \frac{1}{2} f_n(\delta))$$

where

$$\Delta_n(\delta) \equiv \sup_{|\theta - \beta_n| \leq \delta} |A_n(\theta) - B_n(\theta)|$$

and

$$f_n(\delta) \;\equiv\; \inf_{|\theta - \beta_n| = \delta} B_n(\theta) \;-\; B_n(\beta_n)\,.$$

Proof. Let $N(\theta_0;\delta) \equiv \{\theta \in R^d : |\theta - \theta_0| \le \delta\}$, the closed ball centered at θ_0 with radius δ . Let $\theta \in N(\beta_n, \delta)^c$, so that

$$\theta \;=\; \beta_n \;+\; lu \qquad \text{for} \quad u \in S^{d-1}, \quad l > \delta\,.$$

Convexity of A_n implies that

$$(1 - \frac{\delta}{l})\, A_n(\beta_n) \;+\; \frac{\delta}{l}\, A_n(\theta) \;\ge\; A_n((1-\delta/l)\beta_n + (\delta/l)s) \;=\; A_n(\beta_n + \delta u)\,.$$

Writing $A_n \equiv B_n + r_n$, we find that

$$\begin{aligned}\frac{\delta}{l}\,\{A_n(\theta) \;-\; A_n(\beta_n)\} &\ge A_n(\beta_n + \delta u) \;-\; A_n(\beta_n)\\ &= B_n(\beta_n + \delta u) \;-\; B_n(\beta_n)\\ &\qquad + \; r_n(\beta_n + \delta u) \;-\; r_n(\beta_n)\\ &\ge f_n(\delta) \;-\; 2\Delta_n(\delta)\,.\end{aligned}$$

If $2\,\Delta_n(\delta) < f_n(\delta)$, then $A_n(\theta) > A_n(\beta_n)$ for θ outside the $\delta -$ ball, which means that the minimizer α_n must be inside this ball. This proves the stated inequality by taking complements . □

If $||A_n - B_n||_K \;=\; o_p(1)$ and $\beta_n \;=\; O_p(1)$, then $\Delta_n(\delta) \to_p 0$ by an easy argument. Hence $\alpha_n - \beta_n \to_p 0$ provided that $1/f_n(\delta) \;=\; O_p(1)$ for each fixed $\delta > 0$. This latter condition says that B_n shouldn't flatten out around its minimum as $n \to \infty$.

Here is a very useful corollary of lemmas 7.1 and 7.2 for the special case in which B_n is a quadratic form.

Corollary 7.3. Suppose that $A_n(\theta)$ is convex in θ and that

$$\begin{aligned}A_n(\theta) &= \frac{1}{2}\,\theta^T V\theta \;+\; U_n^T\theta \;+\; C_n \;+\; r_n(\theta)\\ &\equiv B_n(\theta) \;+\; r_n(\theta)\end{aligned}$$

where V is symmetric and positive definite, $U_n \;=\; O_p(1)$, C_n is arbitrary, and $r_n(\theta) \;=\; o_p(1)$ for each θ . Then

$$\alpha_n \;\equiv\; \text{argmin}(A_n) \;=\; \beta_n \;+\; o_p(1)$$

where

$$\beta_n \;=\; -\,V^{-1}U_n \;=\; \text{argmin}(B_n)\,.$$

Furthermore, if $U_n \to_d U$, then $\alpha_n \to_d -V^{-1}U \;=\; \text{argmin}\,(\theta^T V\theta \;+\; U^T\theta)$.

Proof. $A_n(\theta) - U_n^T\theta - C_n$ is convex and converges in probablity to $\theta^T V\theta/2$ for each fixed θ. By lemma 7.1, this convergence is uniform on compact subsets of Θ.

Let $\Delta_n(\delta)$ be the supremum of $|r_n(\theta)|$ over $N(\beta_n, \delta) = \{\theta : |\theta - \beta_n| \le \delta\}$. Then by lemma 7.2

$$\alpha_n = -V^{-1}U_n + \epsilon_n$$

where

$$P(|\epsilon_n| > \delta) \le P(\Delta_n(\delta) \ge \frac{1}{2}k\delta^2)$$

where k is the smallest eigenvalue of V. This holds since

$$\begin{aligned} B_n(\theta) - B_n(\beta_n) &= \frac{1}{2}\theta^T V\theta + U_n^T\theta + \frac{1}{2}\beta_n^T V\beta_n \\ &= \frac{1}{2}(\theta - \beta_n)^T V(\theta - \beta_n) \ge \frac{1}{2}k\delta^2 \end{aligned}$$

for $|\theta - \beta_n| = \delta$. Since $\beta_n = O_p(1)$ it follows that $\Delta_n(\delta) \to_p 0$, and the result follows. □

Here is a useful slight extension:

Corollary 7.4. Suppose that $A_n(\theta)$ is convex in θ and that

$$\begin{aligned} A_n(\theta) &= \frac{1}{2}\theta^T V_n\theta + U_n^T\theta + C_n + r_n(\theta) \\ &\equiv B_n(\theta) + r_n(\theta) \end{aligned}$$

where $V_n \to_p V$, V is symmetric and positive definite, $U_n = O_p(1)$, C_n is arbitrary, and $r_n(\theta) = o_p(1)$ for each θ. Then the conclusions of corollary 7.3 continue to hold.

Proof. By writing $V_n = V + \gamma_n$, the remainder γ_n can be absorbed into $r_n(\theta)$, and the result follows from corollary 7.3. □

Now we are ready for several fairly general theorems for the case of i.i.d. observations and estimators defined in terms of a convex function $h(x;\theta)$; for maximum likelihood estimators we will take $h(x;\theta) \equiv -\log p(x;\theta) = -\rho(x;\theta)$.

Theorem 7.5. (Consistency). Suppose that

(i) $X_1, \cdots, X_n$ are i.i.d. P.

(ii) $h : \mathbf{X} \times \Theta \to R$, $h(x,\theta)$ is convex in $\theta \in \Theta \subset R^d$ and $E_P|h(X_1,\theta)| < \infty$ for each $\theta \in \Theta$.

(iii) $H(\theta) \equiv P\,h(x,\theta) = \int h(x,\theta)\,dP(x)$ is uniquely minimized by $\theta_0 = \theta_0(P)$.

Then with $\mathbb{H}_n(\theta) \equiv \mathbb{P}_n h(x,\theta)$,

$$\hat{\theta}_n \equiv \mathrm{argmin}\,\mathbb{H}_n(\theta) \to_p \theta_0 .$$

Proof. By (ii) and the strong law of large numbers $\mathbb{H}_n(\theta) \to_p H(\theta)$ for each fixed $\theta \in \Theta$. Hence by lemma 7.1, $||\mathbb{H}_n - H||_K \to_p 0$ for any compact set $K \subset \Theta$, and this holds in particular with the choice $K = N(\theta_0, \delta)$:

$$\Delta_n(\delta) \equiv ||\mathbb{H}_n - H||_{N(\theta_0,\delta)} \to_p 0 .$$

Since (iii) implies that

$$f(\delta) \equiv \inf\{H(\theta) - H(\theta_0) : |\theta - \theta_0| = \delta\} > 0 ,$$

it follows from lemma 7.2 that

$$P(|\hat{\theta}_n - \theta_0| > \delta) \le P(\Delta_n(\delta) > \frac{1}{2} f(\delta)) \to 0 .$$

□

Theorem 7.6. (Asymptotic normality). Suppose that:

(i) $X_1, \cdots, X_n$ are i.i.d. P.

(ii) $h : \mathbf{X} \times \Theta \to R$, $h(x,\theta)$ is convex in $\theta \in \Theta \subset R^d$ and
   (iia) $\theta_0 \equiv \theta_0(P)$ minimizes $E_P h(X_1,\theta) = P\,h(x,\theta)$;
   (iib) $\hat{\theta}_n$ minimizes $\mathbb{P}_n h(x,\theta)$.

(iii) If $R(x;t) \equiv h(x,\theta_0 + t) - h(x,\theta_0) - \dot{h}(x,\theta_0)^T t$ for some vector of functions $\dot{h}(x,\theta_0)$ with $E_P \dot{h}(X_1,\theta_0) = 0$ and $E_P|\dot{h}(X_1,\theta_0)|^2 < \infty$, then

$$E_P\{h(X,\theta_0 + t) - h(X,\theta_0)\} = E_P R(X;t) = \frac{1}{2} t^T J t + o(|t|^2)$$

   where J is a $d \times d$ positive definite matrix.

(iv) $Var_P(R(X;t)) = o(|t|^2)$ as $t \to 0$.

Then $\hat{\theta}_n$ is a $\sqrt{n}-$ consistent estimator of θ_0 and

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n - \theta_0) &= -J^{-1}\sqrt{n}\,\mathbb{P}_n \dot{h}(x,\theta_0) + o_p(1) \\ &= -J^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n \dot{h}(X_i,\theta_0) + o_p(1) \\ &\to_d N_d(0, J^{-1}K(J^{-1})')\end{aligned}$$

where $K = E_P\{\dot{h}(X,\theta_0)\dot{h}^T(X,\theta_0)\}$.

Proof. Consider the convex function

$$A_n(t) \;\equiv\; nIP_n\{h(x,\theta_0+t/\sqrt{n}) \;-\; h(x,\theta_0)\}\;;$$

by (ii) it is minimized by $\sqrt{n}(\hat{\theta}_n-\theta_0)$. Note that with R as defined in (iv),

$$nE_PR(X;t/\sqrt{n}) \;=\; \frac{1}{2}\,t^TJt \;+\; r_{n,0}(t)$$

where $r_{n,0}(t) \;=\; no(|t|^2/n) \;\to\; 0$ for fixed t . Therefore

$$\begin{aligned} A_n(t) &= \sum_{i=1}^{n}\{\dot{h}(X_i,\theta_0)^T\frac{t}{\sqrt{n}} \;+\; R(X_i;\frac{t}{\sqrt{n}}) \;-\; E_PR(X_1;\frac{t}{\sqrt{n}})\} \;+\; n\,E_PR(X_1;t/\sqrt{n}) \\ &= U_n^Tt \;+\; \frac{1}{2}t^TJt \;+\; r_{n,0}(t) \;+\; r_n(t) \end{aligned}$$

where

$$\begin{aligned} U_n &\equiv \sqrt{n}IP_n\dot{h}(x,\theta_0) \;\to_d\; N_d(0,K)\,, \\ r_n(t) &\equiv nIP_n\{R(x;t/\sqrt{n}) \;-\; E_PR(X_1,t/\sqrt{n})\}\,. \end{aligned}$$

Now $r_n(t) \to_p 0$ for each t since its mean is 0 and its variance is $n\,Var\,R(X_1;t/\sqrt{n}) \;=\; no(|t|^2/n) \;\to\; 0$ for each fixed t . This together with corollary 7.3 yields the conclusion. □

Note that

$$Var_PR(X;t) \;=\; E_P\{R(X_1;t)^2\} \;+\; O(|t|^4)\,,$$

so we might as well work with second moments. Note that the differentiability hypothesis (iii) is applied to the function

$$H(t,P) \;\equiv\; E_Ph(X;\theta_0+t)$$

obtained by averaging over the distribution P of X , a smoothing that can eliminate troublesome pointwise behavior of h (and R).

Here are some examples.

Example 7.6. Suppose that $h(x,\theta) \;=\; -\,\rho(x,\theta) \;\equiv\; -\log g(x-\theta)$ where g is the standard logistic density function

$$g(x) \;=\; \frac{e^{-x}}{(1+e^{-x})^2}\,, \quad -\infty < x < \infty\,.$$

It is easy to check that

$$\rho(x,\theta) \;=\; x-\theta+2\log(1+e^{-(x-\theta)})$$

is concave in θ for every x .