# Chapter 4
# Efficient Likelihood Estimation and Related Tests

1. **Maximum likelihood and efficient likelihood estimation**
2. **Likelihood ratio, Wald, and Rao (or score) tests**
3. **Examples**
4. **Consistency of Maximum Likelihood Estimates**
5. **The EM algorithm and related methods**
6. **Nonparametric MLE**
7. **Limit theory for the statistical agnostic:** $P \notin \mathcal{P}$

# Chapter 4

# Efficient Likelihood Estimation and Related Tests

## 1 Maximum likelihood and efficient likelihood estimation

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 $(\mathbb{X}, \mathcal{A})$ with densities $p$ and $q$ with respect to a sigma-finite measure $\mu$ ($\mu = P + Q$ always works). Thus $P(\mathbb{X}) = 1$ and $Q(\mathbb{X}) \le 1$. Then the *Kullback - Leibler information* $K(P,Q)$ is

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

**Lemma 1.1** For a probability measure $P$ and a (sub-)probability measure $Q$, the Kullback-Leibler information $K(P,Q)$ is always well-defined, and

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

**Proof.** Now

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

If $P \ne MQ$, then Jensen's inequality is strict and yields

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

$\square$

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

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

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

**Assumptions:**

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

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

**A2.** $P_\theta$ has density $p_\theta$ with respect to the $\sigma-$finite measure $\mu$ and $X_1, \ldots, X_n$ are i.i.d. $P_{\theta_0} \equiv P_0$.

**Notation:**

$$
\begin{aligned}
L(\theta) &\equiv L_n(\theta) \equiv L(\theta|\underline{X}) \equiv \prod_{i=1}^n p_\theta(X_i)\,, \\
l(\theta) &= l(\theta|\underline{X}) \equiv l_n(\theta) \equiv \log L_n(\theta) = \sum_{i=1}^n \log p_\theta(X_i)\,, \\
l(B) &\equiv l(B|\underline{X}) \equiv l_n(B) = \sup_{\theta \in B} l(\theta|\underline{X})\,.
\end{aligned}
$$

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

**Theorem 1.1** If A0 - A2 hold, then for $\theta \neq \theta_0$

$$
\frac{1}{n} \log\left(\frac{L_n(\theta_0)}{L_n(\theta)}\right) = \frac{1}{n}\sum_{i=1}^n \log \frac{p_{\theta_0}(X_i)}{p_\theta(X_i)} \rightarrow_{a.s.} K(P_{\theta_0}, P_\theta) > 0\,,
$$

and hence

$$
P_{\theta_0}(L_n(\theta_0|\underline{X}) > L_n(\theta|\underline{X})) \rightarrow 1 \qquad \text{as} \qquad n \rightarrow \infty\,.
$$

**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 1.1 and A0. The second assertion is an immediate consequence of the first. □

Theorem 1.1 motivates the following definition.

**Definition 1.2** The value $\widehat{\theta} = \widehat{\theta}_n$ of $\theta$ which maximizes the likelihood $L(\theta|\underline{X})$, if it exists and is unique, is the *maximum likelihood estimator* (MLE) of $\theta$. Thus $L(\widehat{\theta}) = L(\Theta)$ or $\mathbf{l}(\widehat{\theta}_n) = \mathbf{l}(\Theta)$.

**Cautions:**

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

When $\Theta \subset R^d$, the usual approach to finding $\widehat{\theta}_n$ is to solve the *likelihood* (or *score*) equations

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

i.e. $\dot{\mathbf{l}}_{\theta_i}(\theta|\underline{X}) = 0$, $i = 1, \ldots, d$. The solution $\widetilde{\theta}_n$ 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

$$\begin{aligned} \lambda_n &= \frac{L(\Theta)}{L(\theta_0)} = \frac{\sup_{\theta\in\Theta} L(\theta|\underline{X})}{L(\theta_0|\underline{X})} = \frac{L(\widehat{\theta}_n)}{L(\theta_0)}\,, \\ \widetilde{\lambda}_n &= \frac{L(\widetilde{\theta}_n)}{L(\theta_0)}\,. \end{aligned}$$

Write $P_0$, $E_0$ for $P_{\theta_0}$, $E_{\theta_0}$. Here are some more assumptions about the model $\mathcal{P}$ which we will use to treat these estimators and test statistics.

**Assumptions, continued:**

**A3.** $\Theta$ contains an open neighborhood $\Theta_0 \subset R^d$ of $\theta_0$ for which:

- **(i)** For $\mu$ a.e. $x$, $l(\theta|x) \equiv \log p_\theta(x)$ is twice continuously differentiable in $\theta$.
- **(ii)** For a.e. $x$, the third order derivatives exist and $\dddot{\mathbf{l}}_{jkl}(\theta|x)$ satisfy $|\dddot{\mathbf{l}}_{jkl}(\theta|x)| \le M_{jkl}(x)$ for $\theta \in \Theta_0$ for all $1 \le j,k,l \le d$ with $E_0 M_{jkl}(X) < \infty$.

**A4.** **(i)** $E_0\{\dot{\mathbf{l}}_j(\theta_0|X)\} = 0$ for $j = 1, \ldots, d$.

- **(ii)** $E_0\{\dot{\mathbf{l}}_j^2(\theta_0|X)\} < \infty$ for $j = 1, \ldots, d$.
- **(iii)** $I(\theta_0) = (-\,E_0\{\ddot{\mathbf{l}}_{jk}(\theta_0|X)\})$ is positive definite.

Let

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

so that

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

**Theorem 1.2** Suppose that $X_1, \ldots, X_n$ are i.i.d. $P_{\theta_0} \in \mathcal{P}$ with density $p_{\theta_0}$ where $\mathcal{P}$ satisfies A0 - A4. Then:

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

**(ii)** $\widetilde{\theta}_n$ is asymptotically linear with influence function $\widetilde{\mathbf{l}}(\theta_0|x)$. That is,

$$\begin{aligned} \sqrt{n}(\widetilde{\theta}_n - \theta_0) &= I^{-1}(\theta_0) Z_n + o_p(1) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \widetilde{\mathbf{l}}(\theta_0|X_i) + o_p(1) \\ &\to_d I^{-1}(\theta_0) Z \equiv D \sim N_d(0, I^{-1}(\theta_0))\,. \end{aligned}$$

**(iii)**

$$2\log \widetilde{\lambda}_n \to_d Z^T I^{-1}(\theta_0) Z = D^T I(\theta_0) D \sim \chi^2_d \, .$$

**(iv)**

$$W_n \equiv \sqrt{n}(\widetilde{\theta}_n - \theta_0)^T \widehat{I}_n(\widetilde{\theta}_n)\sqrt{n}(\widetilde{\theta}_n - \theta_0) \to_d D^T I(\theta_0) D = Z^T I^{-1}(\theta_0) Z \sim \chi^2_d \, ,$$

where

$$\widehat{I}_n(\widetilde{\theta}_n) = \begin{cases} I(\widetilde{\theta}_n)\, , & \text{or} \\ n^{-1}\sum_{i=1}^n \dot{\mathbf{l}}(\widetilde{\theta}_n|X_i)\dot{\mathbf{l}}(\widetilde{\theta}_n|X_i)^T\, , & \text{or} \\ -n^{-1}\sum_{i=1}^n \ddot{\mathbf{l}}(\widetilde{\theta}_n|X_i)\, . & \end{cases}$$

**(v)**

$$R_n \equiv Z_n^T I^{-1}(\theta_0) Z_n \to Z^T I^{-1}(\theta_0) Z \sim \chi^2_d \, .$$

Here we could replace $I(\theta_0)$ by any of the possibilities for $\widehat{I}_n(\widetilde{\theta}_n)$ given in (iv) and the conclusion continues to hold.

**(vi)** The model $\mathcal{P}$ satisfies the LAN condition at $\theta_0$:

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

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

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n - \theta_0) = \hat{t}_n &= \text{argmax}\{l_n(\theta_0 + n^{-1/2}t) - l_n(\theta_0)\} \\ &\to_d \text{argmax}\{t^T Z - (1/2) t^T I(\theta_0) t\} = I^{-1}(\theta_0) Z \\ &\sim N_d(0, I^{-1}(\theta_0)). \end{aligned}$$

**Remark 1.1** Note that the asymptotic form of the log-likelihood given in part (vi) of theorem 1.2 is exactly the log-likelihood ratio for a normal mean model $N_d(I(\theta_0)t, I(\theta_0))$. Also note that

$$t^T Z - \frac{1}{2} t^T I(\theta_0) t = \frac{1}{2} Z^T I^{-1}(\theta_0) Z - \frac{1}{2}(t - I^{-1}(\theta_0)Z)^T I(\theta_0)(t - I^{-1}(\theta_0)Z)\, ,$$

which is maximized as a function of $t$ by $\widehat{t} = I^{-1}(\theta_0)Z$ with maximum value $Z^T I^{-1}(\theta_0) Z/2$.

**Corollary 1** Suppose that A0-A4 hold and that $\nu \equiv \nu(P_\theta) = q(\theta)$ is differentiable at $\theta_0 \in \Theta$. Then $\widetilde{\nu}_n \equiv q(\widetilde{\theta}_n)$ satisfies

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

where $\widetilde{\mathbf{l}}_\nu(\theta_0|X_i) = \dot{q}^T(\theta_0) I^{-1}(\theta_0) \dot{\mathbf{l}}(\theta_0|X_i)$ 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 $\theta$ and set

(3) $\quad \check{\theta}_n \equiv \overline{\theta}_n + \widehat{I}_n^{-1}(\overline{\theta}_n)(n^{-1}\dot{\mathbf{l}}(\overline{\theta}_n|\underline{X}))\,.$

The estimator $\check{\theta}_n$ is sometimes called a *one-step* estimator.

**Theorem 1.3** 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

$$\sqrt{n}(\check{\theta}_n - \theta_0) = I^{-1}(\theta_0)Z_n + o_p(1) \rightarrow_d N_d(0, I^{-1}(\theta_0))$$

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

**Proof.** **Theorem 1.2**. (i) Existence and consistency. For $a > 0$, let

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

We will show that

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

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 $\widetilde{\theta}_n(a)$ within $Q_a$; taking the root closest to $\theta_0$ completes the proof.

To prove (a), write

$$\begin{aligned}
\frac{1}{n}(l(\theta) - l(\theta_0)) &= \frac{1}{n}(\theta - \theta_0)^T\underline{\dot{\mathbf{l}}}(\theta_0) - \frac{1}{2}(\theta - \theta_0)^T\left(-\frac{1}{n}\ddot{\mathbf{l}}(\theta_0)\right)(\theta - \theta_0) \\
&\quad + \frac{1}{6n}\sum_{j=1}^d\sum_{k=1}^d\sum_{l=1}^d(\theta_j - \theta_{j0})(\theta_k - \theta_{k0})(\theta_l - \theta_{l0})\sum_{i=1}^n \dddot{\mathbf{l}}_{jkl}\,(\theta_n^*|X_i) \\
\text{(b)} \qquad &= S_1 + S_2 + S_3
\end{aligned}$$

where, by A3(ii), $|\theta_n^* - \theta_0| \le |\theta - \theta_0|$, and by A3(iii), $|\,\dddot{\mathbf{l}}_{jkl}\,(\theta_n^*|X_i)| \le M_{j,k,l}(X_i)$ for $|\theta - \theta_0| = a$ small enough so that $Q_a \subset \Theta_0$. Furthermore, by A3(ii) and A4,

(c) $\quad S_1 \rightarrow_p 0\,,$

(d) $\quad S_2 \rightarrow_p -\dfrac{1}{2}(\theta - \theta_0)^T I(\theta_0)(\theta - \theta_0)\,,$

where

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

and $\lambda_d$ 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 $\lambda_1 \ge \ldots \ge \lambda_d > 0$ are the eigenvalues of $A$ symmetric and positive definite), and

(f) $\quad S_3 = \dfrac{1}{6n}\displaystyle\sum_{j,k,l}(\theta_j - \theta_{j0})(\theta_k - \theta_{k0})(\theta_l - \theta_{l0})\sum_{i=1}^n \dddot{\mathbf{l}}_{jkl}\,(\theta_n^*|X_i)|$

and hence

$$\begin{aligned}
|S_3| &\le \frac{1}{6}\sum_{j,k,l}|\theta_j-\theta_{j0}||\theta_k-\theta_{k0}||\theta_l-\theta_{l0}|\frac{1}{n}\sum_{i=1}^n|\dddot{\mathbf{l}}_{jkl}(\theta_n^*|X_i)| \\
&\le \frac{1}{6}\sum_{j,k,l}|\theta_j-\theta_{j0}||\theta_k-\theta_{k0}||\theta_l-\theta_{l0}|\frac{1}{n}\sum_{i=1}^n M_{j,k,l}(X_i).
\end{aligned}$$

This implies that

$$\begin{aligned}
\sup_{\theta\in Q_a}|S_3| &\le \frac{1}{6}\sum_{j,k,l}a^3\frac{1}{n}\sum_{i=1}^n M_{j,k,l}(X_i) \le \frac{(da)^3}{6}\sum_{j,k,l}\frac{1}{n}\sum_{i=1}^n M_{j,k,l}(X_i) \\
&\to_p \frac{(da)^3}{6}\sum_{j,k,l}m_{j,k,l}, \quad m_{j,k,l}\equiv E_{\theta_0}M_{j,k,l}(X_1).
\end{aligned}$$

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

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

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

and

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

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

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

The right side of (j) is $<0$ if $a < \lambda_d/\{4(B+d)\}$, and hence (a) holds.

On the set

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

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

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

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

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

Furthermore

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

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

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

exists with high probability, and satisfies

(p) $$\left(-\frac{1}{n}\ddot{\mathbf{l}}(\theta_n^*)\right)^{-1} \rightarrow_p I(\theta_0)^{-1}\,.$$

Hence we can use (l) to write, on $G_n$,

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

This proves (ii).

It also follows from (n) that

(r) $$\sqrt{n}(\widetilde{\theta}_n - \theta_0)^T\left(-\frac{1}{n}\ddot{\mathbf{l}}(\widetilde{\theta}_n)\right)\sqrt{n}(\widetilde{\theta}_n - \theta_0) \rightarrow_d Z^T I^{-1}(\theta_0)Z \sim \chi^2_d\,,$$

and that, since $I(\theta)$ is continuous at $\theta_0$,

(s) $$\sqrt{n}(\widetilde{\theta}_n - \theta_0)^T I(\widetilde{\theta}_n)\sqrt{n}(\widetilde{\theta}_n - \theta_0) \rightarrow_d Z^T I^{-1}(\theta_0)Z \sim \chi^2_d\,.$$

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

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

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

$$\begin{aligned}2\log\widetilde{\lambda}_n &= 2\{l(\widetilde{\theta}_n) - l(\theta_0)\} \\ &= 0 + 2\frac{1}{2}\sqrt{n}(\widetilde{\theta}_n - \theta_0)^T\left(-\frac{1}{n}\ddot{\mathbf{l}}(\theta_n^*)\right)\sqrt{n}(\widetilde{\theta}_n - \theta_0) \\ &= D_n^T I(\theta_0)D_n + o_p(1)\,, \qquad \text{with } \ D_n \equiv \sqrt{n}(\widetilde{\theta}_n - \theta_0) \\ &\rightarrow_d D^T I(\theta_0)D \qquad \text{where } \ D \sim N_d(0, I^{-1}(\theta_0)) \\ &\sim \chi^2_d\,.\end{aligned}$$

Finally, (v) is trivial since everything is evaluated at the fixed point $\theta_0$. □

**Proof.** **Theorem 1.3**. First note that

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

so that

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

and

$$\begin{aligned}\text{(b)} \qquad \frac{1}{\sqrt{n}}\dot{\mathbf{l}}_n(\overline{\theta}_n) &= \frac{1}{\sqrt{n}}\dot{\mathbf{l}}_n(\theta_0) + \frac{1}{n}\ddot{\mathbf{l}}_n(\theta_0)\sqrt{n}(\overline{\theta}_n - \theta_0)\\ &\quad + \frac{1}{2}\sqrt{n}(\overline{\theta}_n - \theta_0)^T\left(\frac{1}{n}\dddot{\mathbf{l}}_n(\theta_n^*)\right)(\overline{\theta}_n - \theta_0)\,.\end{aligned}$$

Therefore it follows that

$$\begin{aligned}\sqrt{n}(\check{\theta}_n - \theta_0) &= \sqrt{n}(\overline{\theta}_n - \theta_0) + \left(-\frac{1}{n}\ddot{\mathbf{l}}_n(\overline{\theta}_n)\right)^{-1}\frac{1}{\sqrt{n}}\dot{\mathbf{l}}_n(\overline{\theta}_n)\\ &= \sqrt{n}(\overline{\theta}_n - \theta_0)\\ &\quad + \left\{\left(-\frac{1}{n}\ddot{\mathbf{l}}_n(\theta_0)\right)^{-1} + O_p(1)|\overline{\theta}_n - \theta_0|\right\}\\ &\quad \cdot\left\{Z_n + \frac{1}{n}\ddot{\mathbf{l}}_n(\theta_0)\sqrt{n}(\overline{\theta}_n - \theta_0) + \frac{1}{2}\sqrt{n}(\overline{\theta}_n - \theta_0)^T\left(\frac{1}{n}\dddot{\mathbf{l}}_n(\theta_n^*)\right)(\overline{\theta}_n - \theta_0)\right\}\\ &= \left(-\frac{1}{n}\ddot{\mathbf{l}}_n(\theta_0)\right)^{-1}Z_n + O_p(1)|\overline{\theta}_n - \theta_0|Z_n\\ &\quad + O_p(1)\frac{1}{n}\ddot{\mathbf{l}}_n(\theta_0)\sqrt{n}|\overline{\theta}_n - \theta_0|^2\\ &\quad + O_p(1)\frac{1}{2}\sqrt{n}(\overline{\theta}_n - \theta_0)^T\left(\frac{1}{n}\dddot{\mathbf{l}}_n(\theta_n^*)\right)(\overline{\theta}_n - \theta_0)\\ &= I^{-1}(\theta_0)Z_n + o_p(1) + O_p(1)\sqrt{n}|\overline{\theta}_n - \theta_0|^2\\ &= I^{-1}(\theta_0)Z_n + o_p(1)\,.\end{aligned}$$

Here we used

$$\begin{aligned}&\left|\frac{1}{\sqrt{n}}\dddot{\mathbf{l}}_n(\theta_n^*)(\overline{\theta}_n - \theta_0)(\overline{\theta}_n - \theta_0)\right|\\ &\quad= \left|\sum_{k=1}^d\sum_{l=1}^d\sqrt{n}(\overline{\theta}_{nk} - \theta_{0k})(\overline{\theta}_{nl} - \theta_{0l})\frac{1}{n}\dddot{\mathbf{l}}_{jkl}(\theta_n^*|\underline{X})\right|\\ &\quad\le d^3\sqrt{n}|\overline{\theta}_n - \theta_0|^2\sum_{j=1}^d\frac{1}{n}\sum_{i=1}^n|\dddot{\mathbf{l}}_{jkl}(\theta_n^*|X_i)|\\ &\quad= O_p(1)\sqrt{n}|\overline{\theta}_n - \theta_0|^2\end{aligned}$$

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

**Exercise 1.1** Show that $K(P,Q) \ge 2H^2(P,Q)$.

## 2 The Wald, Likelihood ratio, and Score (or Rao) Tests

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

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

recall the three test statistics

$$(2) \qquad 2\log \widetilde{\lambda}_n \equiv 2\{l_n(\widetilde{\theta}_n) - l_n(\theta_0)\}\,,$$

$$(3) \qquad W_n \equiv \sqrt{n}(\widetilde{\theta}_n - \theta_0)^T \widehat{I}_n(\widetilde{\theta}_n)\sqrt{n}(\widetilde{\theta}_n - \theta_0)\,,$$

and

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

where

$$(5) \qquad Z_n \equiv \frac{1}{\sqrt{n}}\dot{\mathbf{l}}_n(\theta_0) = \frac{1}{\sqrt{n}}\dot{\mathbf{l}}_n(\theta_0|\underline{X})\,.$$

Theorem 1.2 described the null hypothesis behavior of these statistics; all three converge in distribution to $\chi^2_d$ when $P_0 = P_{\theta_0}$ is true. We now examine their behavior under alternatives, i.e. for $X_1, \ldots, X_n$ i.i.d. $P_\theta$ with $\theta \neq \theta_0$.

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

$$(6) \qquad \frac{1}{n}2\log\widetilde{\lambda}_n \to_p 2K(P_\theta, P_{\theta_0}) = 2K(P_{true}, P_{hypothesized}) > 0\,,$$

$$(7) \qquad \frac{1}{n}W_n \to_p (\theta - \theta_0)^T I(\theta)(\theta - \theta_0) > 0\,.$$

If, furthermore,
A5. $E_\theta|\dot{\mathbf{l}}_i(\theta_0|X)| < \infty$ for $i = 1, \ldots, d$, holds, then

$$(8) \qquad \frac{1}{n}R_n \to_p E_\theta\{\dot{\mathbf{l}}(\theta_0|X)\}^T I^{-1}(\theta_0) E_\theta\{\dot{\mathbf{l}}(\theta_0|X)\} > 0$$

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

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

$$\begin{aligned}
(a) \qquad \frac{2}{n}\log\widetilde{\lambda}_n &= \frac{2}{n}\{l(\widetilde{\theta}_n) - l(\theta_0)\} \\
&= \frac{2}{n}\{l(\theta) - l(\theta_0)\} + \frac{2}{n}\{l(\widetilde{\theta}_n) - l(\theta)\} \\
&= \frac{2}{n}\sum_{i=1}^n \log\frac{p_\theta}{p_{\theta_0}}(X_i) + \frac{2}{n}\{l(\widetilde{\theta}_n) - l(\theta)\} \\
&\to_p 2E_\theta\left\{\log\frac{p_\theta}{p_{\theta_0}}(X)\right\} + 0\cdot\chi^2_d = 2K(P_\theta, P_{\theta_0})
\end{aligned}$$

by the WLLN and Theorem 1.2. Also, by the Mann-Wald (or continuous mapping) theorem,

$$\text{(b)} \qquad \frac{1}{n}W_n = (\widetilde{\theta}_n - \theta_0)^T \widehat{I}_n(\widetilde{\theta}_n)(\widetilde{\theta}_n - \theta_0) \rightarrow_p (\theta - \theta_0)^T I(\theta)(\theta - \theta_0)\,,$$

and, since

$$\text{(c)} \qquad \frac{1}{\sqrt{n}}Z_n = \frac{1}{n}\dot{\mathrm{l}}_n(\theta_0) = \frac{1}{n}\sum_{i=1}^n \dot{\mathrm{l}}(\theta_0|X_i) \rightarrow_p E_\theta\{\dot{\mathrm{l}}(\theta_0|X)\}\,,$$

it follows that

$$\text{(d)} \qquad \frac{1}{n}R_n \rightarrow_p E_\theta\{\dot{\mathrm{l}}(\theta_0|X)\}^T I^{-1}(\theta_0) E_\theta\{\dot{\mathrm{l}}(\theta_0|X)\}\,.$$

□

**Corollary 1** (Consistency of the likelihood ratio, Wald, and score tests). If Assumptions A0-A5 hold, then the tests are consistent: i.e. if $\theta \neq \theta_0$, then

$$\text{(9)} \qquad P_\theta(\text{LR test rejects } H) = P_\theta(2\log\widetilde{\lambda}_n \geq \chi^2_{d,\alpha}) \rightarrow 1\,,$$

$$\text{(10)} \qquad P_\theta(\text{Wald test rejects } H) = P_\theta(W_n \geq \chi^2_{d,\alpha}) \rightarrow 1\,,$$

$$\text{(11)} \qquad P_\theta(\text{score test rejects } H) = P_\theta(R_n \geq \chi^2_{d,\alpha}) \rightarrow 1\,,$$

assuming that $E_\theta\{\dot{\mathrm{l}}(\theta_0|X)\} \neq 0$.

It remains to examine the behavior these three tests under *local alternatives*, $\theta_n = \theta_0 + tn^{-1/2}$ with $t \neq 0$. We first examine $Z_n$ and $\widetilde{\theta}_n$ under $\theta_0$ using Le Cam's third lemma 3.3.4.

**Theorem 2.2** Suppose that A0-A4 hold. Then, if $\theta_n = \theta_0 + tn^{-1/2}$ is true, then under $P_{\theta_n}$

$$\text{(12)} \qquad \sqrt{n}(\widetilde{\theta}_n - \theta_0) \rightarrow_d D + t \sim N_d(t, I^{-1}(\theta_0))\,;$$

furthermore,

$$\text{(13)} \qquad Z_n(\theta_0) \equiv \frac{1}{\sqrt{n}}\dot{\mathrm{l}}(\theta_0|\underline{X}) \rightarrow_d Z + I(\theta_0)t \sim N_d(I(\theta_0)t, I(\theta_0))\,.$$

Hence we also have under $P_{\theta_n}$,

$$\text{(14)} \qquad \begin{aligned} \sqrt{n}(\widetilde{\theta}_n - \theta_n) &= \sqrt{n}(\widetilde{\theta}_n - \theta_0) - \sqrt{n}(\theta_n - \theta_0) \\ &\rightarrow_d D + t - t = D \sim N_d(0, I^{-1}(\theta_0))\,; \end{aligned}$$

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

$$\text{(15)} \qquad \begin{aligned} Z_n(\theta_n) &= Z_n(\theta_0) - \left(-\frac{1}{n}\ddot{\mathrm{l}}_n(\theta_n^*)\right)\sqrt{n}(\theta_n - \theta_0) \\ &\rightarrow_d Z + I(\theta_0)t - I(\theta_0)t = Z \sim N_d(0, I(\theta_0))\,. \end{aligned}$$

**Proof.** From the proof of theorem 1.2 we know that $\widetilde{\theta}_n$ is asymptotically linear under $P_0 = P_{\theta_0}$:

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

where $\tilde{\mathbf{l}}_\theta(x) \equiv I^{-1}(\theta_0)\dot{\mathbf{l}}_\theta(x)$. Furthermore, it follows from theorem 1.2 part (vi) that the log likelihood ratio is asymptotically linear:

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

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

$$\begin{aligned}
\begin{pmatrix} T_n \\ \log \frac{dP^n_{\theta_n}}{dP^n_{\theta_0}} \end{pmatrix} &= \begin{pmatrix} a^T\sqrt{n}(\widetilde{\theta}_n - \theta_0) \\ \log \frac{dP^n_{\theta_n}}{dP^n_{\theta_0}} \end{pmatrix} \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} a^T \widetilde{\mathbf{l}}_\theta(\theta_0|X_i) \\ t^T \dot{\mathbf{l}}_\theta(\theta_0|X_i) \end{pmatrix} + \begin{pmatrix} 0 \\ -\sigma^2/2 \end{pmatrix} + o_p(1) \\
&\to_d N_2\left( \begin{pmatrix} 0 \\ -\sigma^2/2 \end{pmatrix}, \begin{pmatrix} a^T I^{-1}(\theta_0) a & a^T t \\ a^T t & \sigma^2 \end{pmatrix} \right).
\end{aligned}$$

Thus the hypothesis of Le Cam's third lemma 3.3.4 is satisfied with $c = a^T t$, and we deduce that, under $P_{\theta_n}$,

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

In particular, under $P_{\theta_n}$,

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

and by the Cramér - Wold device this implies that under $P_{\theta_n}$

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

This, in turn, implies that

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

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

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

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

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

where $\delta = t^T I(\theta_0) t$.

**Proof.** This follows from theorem 2.2, the Mann - Wald theorem, and the fact that

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

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

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

$$(19) \qquad P_{\theta_n}(T_n \ge \chi^2_{d,\alpha}) \to P(\chi^2_d(\delta) \ge \chi^2_{d,\alpha}).$$

**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.

$$H: \ \theta_1 = \theta_{10}, \ \theta_2 = \text{ anything} \qquad \text{versus} \qquad K: \theta = (\theta_1, \theta_2) \ne (\theta_{10}, \theta_2)$$

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

$$(20) \qquad 2\log\lambda_n \qquad \text{with} \qquad \lambda_n \equiv \frac{\sup_{\theta\in\Theta} L(\theta|\underline{X})}{\sup_{\theta\in\Theta^0} L(\theta|\underline{X})} = \frac{L(\widehat{\theta}_n|\underline{X})}{L(\widehat{\theta}^0_n|\underline{X})}$$

(or

$$(21) \qquad 2\log\widetilde{\lambda}_n \qquad \text{with} \qquad \widetilde{\lambda}_n \equiv \frac{L(\widetilde{\theta}_n|\underline{X})}{L(\widetilde{\theta}^0_n|\underline{X})}$$

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

$$(22) \qquad W_n \equiv \sqrt{n}(\widetilde{\theta}_{n1} - \theta_{10})^T \widehat{I}_{11\cdot 2}\sqrt{n}(\widetilde{\theta}_{n1} - \theta_{10}),$$

and

$$(23) \qquad R_n \equiv Z_n^T(\widehat{\theta}^0_n) I^{-1}(\widehat{\theta}^0_n) Z_n(\widehat{\theta}^0_n)$$

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

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

$$(24) \qquad \sqrt{n}(\widetilde{\theta}_{n1} - \theta_{10}) \to_d D_1 \sim N_m(0, I^{-1}_{11\cdot 2})$$

where

$$D = \begin{pmatrix} D_1 \\ D_2 \end{pmatrix} = I^{-1}(\theta_0) Z = \begin{pmatrix} I^{-1}_{11\cdot 2}(Z_1 - I_{12} I^{-1}_{22} Z_2) \\ I^{-1}_{22\cdot 1}(Z_2 - I_{21} I^{-1}_{11} Z_1) \end{pmatrix}$$

and

$$\begin{aligned} Z_n(\widetilde{\theta}^0_n) &= \begin{pmatrix} Z_{n1}(\widetilde{\theta}^0_n) \\ Z_{n2}(\widetilde{\theta}^0_n) \end{pmatrix} \\ &= \begin{pmatrix} Z_{n1}(\theta_0) - I_{12}(\theta^*_n)\sqrt{n}(\widetilde{\theta}^0_{n2} - \theta_{02}) + o_p(1) \\ 0 \end{pmatrix} \\ &= \begin{pmatrix} Z_{n1}(\theta_0) - I_{12}(\theta_0) I^{-1}_{22} Z_{n2}(\theta_0) + o_p(1) \\ 0 \end{pmatrix} \\ (25) \qquad &\to_d \begin{pmatrix} Z_1(\theta_0) - I_{12}(\theta_0) I^{-1}_{22} Z_2(\theta_0) \\ 0 \end{pmatrix} \sim \begin{pmatrix} N_m(0, I_{11\cdot 2}) \\ 0 \end{pmatrix}. \end{aligned}$$

The natural consequences of (24) and (25) for the likelihood ratio, Wald, and Rao statistics are summarized in the following theorem.

**Theorem 2.3** (Likelihood ratio, Wald, Rao statistics for composite null under null). If A0 - A4 hold and $\theta_0 \in \Theta_0$ is true, then

$$\left\{ \begin{array}{c} 2\log\widetilde{\lambda}_n \\ W_n \\ R_n \end{array} \right\} \rightarrow_d D_1^T I_{11\cdot 2} D_1 \sim \chi_m^2 = \chi_{d-(d-m)}^2 \,.$$

**Proof.** That $W_n \rightarrow_d D_1^T I_{11\cdot 2} D_1$ follows from (24) and consistency of $\widehat{I}_{11\cdot 2}$. Similarly, $R_n \rightarrow_d D_1^T I_{11\cdot 2} D_1$ follows from (25) and $\widehat{I}_n^{-1}(\widehat{\theta}_n^0) \rightarrow_p I^{-1}(\theta_0)$. To prove the claimed convergence of $2\log\widetilde{\lambda}_n$, write

$$\begin{aligned} 2\log\widetilde{\lambda}_n &= 2\{l_n(\widetilde{\theta}_n) - l_n(\widetilde{\theta}_n^0) \\ &= 2\{l_n(\widetilde{\theta}_n) - l_n(\theta_0) - (l_n(\widetilde{\theta}_n^0) - l_n(\theta_0))\} \end{aligned}$$

where

(a) $\quad 2\{l_n(\widetilde{\theta}_n) - l_n(\theta_0)\} \rightarrow_d D^T I(\theta_0) D = Z^T I^{-1}(\theta_0) Z$

by our proof of theorem 1.2, and

(b) $\quad 2\{l_n(\widetilde{\theta}_n^0) - l_n(\theta_0)\} \rightarrow_d Z_2^T I_{22}^{-1}(\theta_0) Z_2 \,,$

again by the proof of theorem 1.2. In fact, by the asymptotic linearity of $\widetilde{\theta}_n$ (and $\widetilde{\theta}_n^0$) proved there, the convergences in (a) and (b) imply that

$$\begin{aligned} 2\log\widetilde{\lambda}_n &\rightarrow_d Z^T I^{-1}(\theta_0) Z - Z_2^T I_{22}^{-1} Z_2 \\ &= (Z_1 - I_{12} I_{22}^{-1} Z_2)^T I_{11\cdot 2}^{-1} (Z_1 - I_{12} I_{22}^{-1} Z_2) \\ &= D_1^T I_{11\cdot 2} D_1 \end{aligned}$$

where we have used the block inverse form of $I^{-1}(\theta_0)$ given in (3.2.x) and the matrix identity (3.2.15) with the roles of “1” and “2” interchanged. $\square$

Now under local alternatives the situation is as follows:

**Theorem 2.4** If A0 - A4 hold, and $\theta_n = \theta_0 + t n^{-1/2}$ with $\theta_0 \in \Theta_0$, then under $P_{\theta_n}$

$$\left\{ \begin{array}{c} 2\log\widetilde{\lambda}_n \\ W_n \\ R_n \end{array} \right\} \rightarrow_d (D_1 + t_1)^T I_{11\cdot 2} (D_1 + t_1) \sim \chi_m^2(\delta) = \chi_{d-(d-m)}^2(\delta)$$

whre $\delta = t_1^T I_{11\cdot 2} t_1$.

## 3 Selected Examples

Now we consider several example to illustrate the theory of the preceding two sections and its limitations.

**Example 3.1** Let $X_1, \ldots, X_n$ be i.i.d. $N(\mu, \sigma^2)$; $\theta = (\mu, \sigma^2) \in \Theta = R \times R^+$. Consider

**(i)** Estimation of $\theta$.

**(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 of $\theta$ is

$$\widehat{\theta}_n = (\overline{X}_n, S^2) \qquad \text{where} \qquad S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 ,$$

and

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

since

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

so that

$$\begin{aligned} \dot{\mathrm{l}}_\mu(X_1) &= \frac{1}{\sigma^2}(X_1 - \mu) , & \ddot{\mathrm{l}}_{\mu\mu}(X_1) &= -\frac{1}{\sigma^2} , \\ \dot{\mathrm{l}}_{\sigma^2}(X_1) &= \frac{(X_1 - \mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2} , & \ddot{\mathrm{l}}_{\sigma^2\sigma^2}(X_1) &= -\frac{(X_1 - \mu)^2}{\sigma^6} + \frac{1}{2\sigma^4} . \end{aligned}$$

Now

(1) $\qquad \widehat{\theta}_n = (\overline{X}_n, S_n^2) \to_{a.s.} (\mu, \sigma^2) = \theta ,$

and

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

with

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

The almost sure consistency stated in (1) and the limiting distribution in (2) follow from direct application of the strong law of large numbers (proposition 2.2.2) and the central limit theorem (proposition 2.2.3 ) respectively, after easy manipulations and Slutsky's theorem. Alternatively, the in probability version of (1) and the limiting distribution in (2) follow from theorem 1.2 assuming that the normal model for the $X_i$'s holds.

(ii). The likelihood ratio statistic for testing $H_1$ is

$$\lambda_n = \frac{L(\widehat{\theta}_n)}{L(\widehat{\theta}_n^0)} = \frac{L(\overline{X}, S^2)}{L(0, n^{-1}\sum_1^n X_i^2)} = \left\{ \frac{n^{-1}\sum_1^n X_i^2}{n^{-1}\sum_1^n (X_i - \overline{X})^2} \right\}^{n/2},$$

and hence

$$2\log\lambda_n = -n\log\left(1 - \frac{\overline{X}^2}{n^{-1}\sum_1^n X_i^2}\right);$$

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

$$W_n = \{\sqrt{n}(\overline{X} - 0)\}\widehat{I}_{11\cdot 2}\{\sqrt{n}(\overline{X} - 0)\} = \frac{n\overline{X}^2}{S^2} = \left\{ \frac{\sqrt{n}\,\overline{X}}{S} \right\}^2.$$

Finally, the Rao or score statistic is

$$\begin{aligned} R_n &= Z_n(\widehat{\theta}_n^0)^T I(\widehat{\theta}_n^0)^{-1} Z_n(\widehat{\theta}_n^0) \\ &= \begin{pmatrix} \frac{\sqrt{n}\overline{X}_n}{n^{-1}\sum_1^n X_i^2} \\ 0 \end{pmatrix}^T \begin{pmatrix} n^{-1}\sum_1^n X_i^2 & 0 \\ 0 & 2(n^{-1}\sum_1^n X_i^2)^2 \end{pmatrix} \begin{pmatrix} \frac{\sqrt{n}\overline{X}_n}{n^{-1}\sum_1^n X_i^2} \\ 0 \end{pmatrix}^T \\ &= \left\{ \frac{\sqrt{n}\overline{X}_n}{\sqrt{n^{-1}\sum_1^n X_i^2}} \right\}^2. \end{aligned}$$

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

$$2\log\lambda_n, \qquad W_n, \qquad R_n \to_d \chi_1^2.$$

If $\mu \neq 0$, so $\theta \notin \Theta_1$, then

$$\begin{aligned} &\frac{1}{n} 2\log\lambda_n \to_p -\log\left(1 - \frac{\mu^2}{\sigma^2 + \mu^2}\right) = -\log\left(\frac{\sigma^2}{\sigma^2 + \mu^2}\right) > 0, \\ &\frac{1}{n} W_n \to_p \frac{\mu^2}{\sigma^2} > 0, \qquad \text{and} \\ &\frac{1}{n} R_n \to_p \frac{\mu^2}{\sigma^2 + \mu^2} > 0. \end{aligned}$$

(iii) The likelihood ratio statistic $\lambda_n$ for testing $H_2$ is

$$\begin{aligned} \lambda_n = \frac{L(\overline{X}, S^2)}{L(0, \sigma_0^2)} &= \frac{(2\pi S^2)^{-n/2}\exp(-n/2)}{(2\pi\sigma_0^2)^{-n/2}\exp(-\sum_1^n X_i^2/2\sigma_0^2)} \\ &= \left(\frac{S^2}{\sigma_0^2}\right)^{-n/2} \exp\left(\frac{1}{2\sigma_0^2}\sum_1^n X_i^2 - n/2\right) \end{aligned}$$

so that

$$\begin{aligned} 2\log\lambda_n &= \frac{1}{\sigma_0^2}\sum_1^n X_i^2 - n - n\log\left(\frac{S^2}{\sigma_0^2}\right) \\ &= n\left\{\frac{S^2}{\sigma_0^2} - 1 - \log\left(\frac{S^2}{\sigma_0^2}\right)\right\} + \frac{n\overline{X}^2}{\sigma_0^2}. \end{aligned}$$

The Wald statistic is

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

The Rao or score statistic is given by

$$\begin{aligned} R_n &= \begin{pmatrix} \sqrt{n}\overline{X}/\sigma_0^2 \\ \frac{1}{2}\frac{\sqrt{n}}{\sigma_0^2}\left\{\frac{n^{-1}\sum_1^n X_i^2}{\sigma_0^2} - 1\right\} \end{pmatrix}^T \begin{pmatrix} \sigma_0^2 & 0 \\ 0 & 2\sigma_0^4 \end{pmatrix} \begin{pmatrix} \sqrt{n}\overline{X}/\sigma_0^2 \\ \frac{1}{2}\frac{\sqrt{n}}{\sigma_0^2}\left\{\frac{n^{-1}\sum_1^n X_i^2}{\sigma_0^2} - 1\right\} \end{pmatrix} \\ &= \frac{n\overline{X}^2}{\sigma_0^2} + \frac{n}{2}\left\{\frac{n^{-1}\sum_1^n X_i^2}{\sigma_0^2} - 1\right\}^2. \end{aligned}$$

If $H_2$ holds, then

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

**Exercise 3.1** What are the limits in probability of $n^{-1}2\log\lambda_n$, $n^{-1}W_n$, and $n^{-1}R_n$ under $\theta \neq \theta_0$?

**Exercise 3.2** In the context of Example 3.1, what are the likelihood ratio, Wald, and score statistics for testing $H_3:\ \sigma^2 = \sigma_0^2$ versus $K_3:\ \sigma^2 \neq \sigma_0^2$?

**Example 3.2** (One parameter exponential family). Suppose that $p_\theta(x) = \exp(\theta T(x) - A(\theta))$ with respect to $\mu$. Then

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

and the likelihood equation may be written as

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

Now $A'(\theta) = E_\theta\{T(X)\}$, and $A''(\theta) = Var_\theta(T(X)) > 0$, so the right side in (4) is strictly increasing in $\theta$. Thus (4) has at most one root $\widehat{\theta}_n$, and

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

**Example 3.3** (Multi-parameter exponential family; Lehmann and Cassella, TPE, pages 23 - 32). Suppose that

$$p_\theta(x) = \exp\left\{\sum_{j=1}^d \eta_j(\theta)T_j(x) - B(\theta)\right\} h(x)$$

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

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

The *natural parameter space* $\Xi$ 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 $a_1, \ldots, a_d$ and $b \in R$ such that $\sum_{j=1}^d a_j T_j(x) = b$ with probability 1.

By Lehmann and Casella, TPE, theorem 5.8, page 27, we can differentiate under the expectations to find that the score vector for $\underline{\eta}$ is given by

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

and since this has expectation 0 under $p_\eta$,

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

If the likelihood equations have a solution, it is unique (and is the MLE) since $l(\eta)$ is a strictly concave function of $\eta$ in view of the fact that $l(\eta)$ has Hessian (times minus one)

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

which 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 $g(x) = \pi^{-1}(1+x^2)^{-1}$, $x \in R$. 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

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

where the right side is a polynomial in $\theta$ of degree $2(n-1)+1 = 2n-1$ which could have as many as $2n-1$ roots. Let $\overline{\theta}_n \equiv \text{median}(X_i) = \mathbb{F}_n^{-1}(1/2)$. Then

$$\sqrt{n}(\overline{\theta}_n - \theta) \rightarrow_d N(0, \pi^2/4) ,$$

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

$$\begin{aligned} \check{\theta}_n &= \overline{\theta}_n + \widehat{I}(\overline{\theta}_n)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \dot{\mathbf{l}}_\theta(X_i; \overline{\theta}_n)\right) \\ &= \overline{\theta}_n + \frac{4}{n}\sum_{i=1}^n \frac{X_i - \overline{\theta}_n}{1+(X_i-\overline{\theta}_n)^2} . \end{aligned}$$

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 maximum. Thus the number of roots corresponding to "false" maxima is $(r_n - 1)/2$. Reeds (1985) 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. In fact,

$$P((r_n - 1)/2 \geq 1) \to P(Poisson(1/\pi) \geq 1) = 1 - e^{-1/\pi} = .272623....$$

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

$$\begin{aligned} p_\theta(x) &= pN(\mu, \sigma^2) + (1-p)N(\nu, \tau^2) \\ &= p\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) + (1-p)\frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{(x-\nu)^2}{2\tau^2}\right) \end{aligned}$$

with $\theta = (p, \mu, \sigma, \nu, \tau)$. The simpler case of $\theta = (1/2, \mu, \sigma, 0, 1)$ will illustrate the phenomena we want to illustrate here. When $\theta = \theta_0$ we may reparametrize by $\theta = (\mu, \sigma) \in R \times R^+ = \Theta$, and the density can be written as

$$(5) \qquad p_\theta(x) = \frac{1}{2}\phi(x) + \frac{1}{2}\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right).$$

Then, if $X_1, \ldots, X_n$ are i.i.d. $p_\theta$,

$$\sup_{\theta \in \Theta} L(\theta | \underline{X}) = \infty \qquad \text{almost surely.}$$

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

However, A0 - A4 hold for this model, 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). Suppose that

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

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

Note that $p_0(x) = (1 - |x|)1_{[-1,1]}(x) \equiv$ triangular 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 $\theta$ for all $x$. Thus a MLE exists for all $n \geq 1$ since a continuous function on a compact set $[0,1]$ achieves its maximum on that set.

**Proposition 3.1** (Ferguson). Let $\widehat{\theta}_n \equiv$ an MLE of $\theta$ for sample size $n$. If $\delta(\theta) \to 0$ rapidly enough as $\theta \to 1$, then $\widehat{\theta}_n \to 1$ a.s. $P_\theta$ 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 $M_n = \max\{X_1, \ldots, X_n\}$.

The details of this example are written out in Ferguson's *A Course in Large Sample Theory*, pages 116 - 117, and will treated in more detail in Section 4. The hypothesis that is violated in this example is that the family of log-likelihoods fails to have an integrable envelope.

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 3.7** (Neyman-Scott). Suppose that $(X_i, Y_i) \sim N_2((\mu_i, \mu_i), \sigma^2 I)$ for $i = 1, \ldots, n$. Consider estimation of $\sigma^2$. 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 an unbiased and consistent estimator of $\sigma^2$. However

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

and therefore the MLE of $\mu_i$ is $\widehat{\mu}_i = (X_i + Y_2)/2$ and it follows that the MLE of $\sigma^2$ is

$$\widehat{\sigma}^2 = \frac{1}{4n}\sum_{i=1}^{n} Z_i^2 = \frac{1}{2n}\left\{\sum_{i=1}^{n}(X_i - \widehat{\mu}_i)^2 + \sum_{i=1}^{n}(Y_i - \widehat{\mu}_i)^2\right\}.$$

Thus $\widehat{\sigma}_n^2 \rightarrow_{a.s.} \sigma^2/2$ as $n \rightarrow \infty$. What went wrong? note that the dimensionality of the parameter space for this model is $n+1$ which increases with $n$, so the theory we have developed so far does not apply. One way out of this difficulty 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. For results on asymptotic normality of the MLE's in the semiparametric mixture models proposed by Kiefer and Wolfowitz (1956), see Van der Vaart (1996), "Efficient maximum likelihood estimation in semiparametric mixture models", *Ann. Statist.* **24**, 862-878.

**Example 3.8** (Bivariate Poisson model). Suppose that $U \sim$ Poisson$(\mu)$, $V \sim$ Poisson$(\lambda)$, and $W \sim$ Poisson$(\psi)$ are all independent, and let

$$X \equiv U + W\,, \qquad Y = V + W\,.$$

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. and then $X = U$ and $Y = V$ a.s.).

Consider testing $H : \psi = 0$ (independence) versus $K : \psi > 0$ based on a sample of $n$ i.i.d. pairs from $P_\theta$. Note that maximum likelihood estimation of $\theta = (\mu, \lambda, \psi)$ for general $\theta$ 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 calculation shows that the scores, for any point $\theta_0 = (\lambda, \mu, 0) \in \Theta_0$ are given by

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

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

$$\begin{aligned}
R_n &= n\left\{\frac{n^{-1}\sum_{i=1}^n X_i Y_i}{\overline{X}_n \overline{Y}_n} - 1\right\}^2 \overline{X}_n \overline{Y}_n \\
&= n\frac{\left\{n^{-1}\sum_{i=1}^n X_i Y_i - \overline{X}_n \overline{Y}_n\right\}^2}{\overline{X}_n \overline{Y}_n}\,.
\end{aligned}$$

since

$$I(\theta_0) = \left(\begin{array}{ccc} 1/\mu & 0 & 1/\mu \\ 0 & 1/\lambda & 1/\lambda \\ 1/\mu & 1/\lambda & 1/\mu + 1/\lambda + 1/(\mu\lambda) \end{array}\right),$$

so that

$$\begin{aligned}
I_{22\cdot 1} &= I_{22} - I_{21} I_{11}^{-1} I_{12} \\
&= 1/\mu + 1/\lambda + 1/(\mu\lambda) - (1/\mu, 1/\lambda)\left(\begin{array}{cc} \mu & 0 \\ 0 & \lambda \end{array}\right)\left(\begin{array}{c} 1/\mu \\ 1/\lambda \end{array}\right) \\
&= 1/(\mu\lambda)\,.
\end{aligned}$$

Note that $R_n$ is a close relative of the classical correlation coefficient, and in fact, under $H$ it follows fairly easily that $R_n = n r_n^2 + o_p(1)$.

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. Nevertheless, under $H$ we have $R_n \to_d \chi_1^2$.

**Example 3.9** (The Multinomial Distribution). Suppose that $\underline{X}_1, \ldots, \underline{X}_n$ are i.i.d. $\mathrm{Mult}_k(1, \underline{p})$. Then

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

Then the log likelihood is

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

and the *constrained log likelihood* is

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

Thus the likelihood equations are

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

and

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

The solution of the first set of $k$ equations yields

$$\widehat{p}_j = -\frac{N_j}{\lambda}\,, \qquad j = 1, \ldots, k\,.$$

But to satisfy the constraint we must have

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

or $\lambda = -n$, and thus the MLE of $\underline{p}$ is $\underline{\widehat{p}} = \underline{N}_n/n$. The score vector (for $n = 1$) becomes

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

Thus the information matrix is

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

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 $I^-(p) = \text{diag}(p_j) - \underline{p}\underline{p}^T$:

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

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

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

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

$$\underline{Z}_n(\theta_0) = \frac{1}{\sqrt{n}} \left( \frac{N_1 - np_{10}}{p_{10}}, \ldots, \frac{N_k - np_{k0}}{p_{k0}} \right)^T$$

and

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

The Wald statistic is given by

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

For more on problems involving singular information matrices, see Rotnitzky, Cox, Bottai, and Robins (2000).

# 4 Consistency of Maximum Likelihood Estimators

**Some Uniform Strong Laws of Large Numbers**

Suppose that:

**A.** $X, X_1, \ldots, X_n$ are i.i.d. $P$ on the measurable space $(\mathcal{X}, \mathcal{A})$.

**B.** For each $\theta \in \Theta$, $f(x, \theta)$ is a measurable, real-valued function of $x$, $f(\cdot, \theta) \in L_1(P)$.

Let $\mathcal{F} = \{f(\cdot, \theta) : \theta \in \Theta\}$. Since $f(\cdot, \theta) \in L_1(P)$ for each $\theta$,

$$g(\theta) \equiv Ef(X, \theta) = \int f(x, \theta) dP(x) \equiv Pf(\cdot, \theta)$$

exists and is finite. Moreover, by the strong law of large numbers,

$$\begin{aligned} \mathbb{P}_n f(\cdot, \theta) &\equiv \int f(x, \theta) d\mathbb{P}_n(x) = \frac{1}{n} \sum_{i=1}^{n} f(X_i, \theta) \\ &\to_{a.s.} Ef(X, \theta) = Pf(\cdot, \theta) = g(\theta). \end{aligned} \tag{1}$$

It is often useful and important to strengthen (1) to hold uniformly in $\theta \in \Theta$:

$$\sup_{\theta \in \Theta} |\mathbb{P}_n f(\cdot, \theta) - Pf(\cdot, \theta)| \to_{a.s.} 0\,. \tag{2}$$

Note that the left side in (2) is equal to

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} \equiv \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf|\,.$$

Here is how (2) can be used: suppose that we have a sequence $\widehat{\theta}_n$ of estimators, possibly dependent on $X_1, \ldots, X_n$, such that $\widehat{\theta}_n \to_{a.s.} \theta_0$. Suppose that $g(\theta)$ is continuous at $\theta_0$. We would like to conclude that

$$\mathbb{P}_n f(\cdot, \widehat{\theta}_n) = \frac{1}{n} \sum_{i=1}^{n} f(X_i, \widehat{\theta}_n) \to_{a.s.} g(\theta_0). \tag{3}$$

The convergence (3) *does not follow* from (1); but (3) *does follow* from (2):

$$\begin{aligned} \left|\mathbb{P}_n f(\cdot, \widehat{\theta}_n) - g(\theta_0)\right| &\leq \left|\mathbb{P}_n f(\cdot, \widehat{\theta}_n) - g(\widehat{\theta}_n)\right| + \left|g(\widehat{\theta}_n) - g(\theta_0)\right| \\ &\leq \sup_{\theta \in \Theta}\left|\mathbb{P}_n f(\cdot, \theta) - g(\theta)\right| + \left|g(\widehat{\theta}_n) - g(\theta_0)\right| \\ &= \|\mathbb{P}_n - P\|_{\mathcal{F}} + \left|g(\widehat{\theta}_n) - g(\theta_0)\right| \\ &\to_{a.s.} 0 + 0 = 0\,. \end{aligned}$$

The following theorems, due to Le Cam, give conditions on $f$ and $P$ under which (2) holds. The first theorem is a prototype for what are now known in empirical process theory as "Glivenko-Cantelli theorems".

**Theorem 4.1** Suppose that:
(a) $\Theta$ is compact.
(b) $f(x,\cdot)$ is continuous in $\theta$ for all $x$.
(c) There exists a function $F(x)$ such that $EF(X) < \infty$ and $|f(x,\theta)| \le F(x)$ for all $x \in \mathcal{X}$, $\theta \in \Theta$.

Then (2) holds; i.e.

$$\sup_{\theta\in\Theta} |\mathbb{P}_n f(\cdot,\theta) - Pf(\cdot,\theta)| \to_{a.s.} 0\,.$$

The second theorem is a "one-sided" version of theorem 4.1 which is useful for the theory of maximum likelihood estimation.

**Theorem 4.2** Suppose that:
(a) $\Theta$ is compact.
(b) $f(x,\cdot)$ is upper semicontinuous in $\theta$ for all $x$; i.e. $\limsup_{n\to\infty} f(x,\theta_n) \le f(x,\theta)$ for all $\theta_n \to \theta$ and all $\theta \in \Theta$.
(c) There exists a function $F(x)$ such that $EF(X) < \infty$ and $f(x,\theta) \le F(x)$ for all $x \in \mathcal{X}$, $\theta \in \Theta$.
(d) For all $\theta$ and all sufficiently small $\rho > 0$

$$\sup_{|\theta'-\theta|<\rho} f(x,\theta')$$

is measurable in $x$.

Then

$$\text{limsup}_{n\to\infty} \sup_{\theta\in\Theta} \mathbb{P}_n f(\cdot,\theta) \le_{a.s.} \sup_{\theta\in\Theta} Pf(\cdot,\theta) = \sup_{\theta\in\Theta} g(\theta).$$

We proceed by first proving Theorem 4.2. Then Theorem 4.1 will follow as a consequence of Theorem 4.2.

**Proof.** **Theorem 4.2.** Let

$$\psi(x,\theta,\rho) \equiv \sup_{|\theta'-\theta|<\rho} f(x,\theta').$$

Then $\psi$ is measurable (for $\rho$ sufficiently small), bounded by an integrable function $F$, and

$$\psi(x,\theta,\rho) \searrow f(x,\theta) \qquad \text{as} \qquad \rho \searrow 0 \qquad \text{by} \ \ (b).$$

Thus by the monotone convergence theorem

$$\int \psi(x,\theta,\rho)dP(x) \searrow \int f(x,\theta)dP(x) = g(\theta).$$

Let $\epsilon > 0$. For each $\theta$, find $\rho_\theta$ so that

$$\int \psi(x,\theta,\rho_\theta)dP(x) < g(\theta) + \epsilon\,.$$

The spheres

$$S(\theta,\rho_\theta) = \{\theta' : |\theta'-\theta| < \rho_\theta\}$$

cover $\Theta$, so by (a) there exists a finite sub cover: $\Theta \subset \cup_{j=1}^m S(\theta_j, \rho_{\theta_j})$. for each $\theta \in \Theta$ there is some $j$, $1 \le j \le m$, such that $\theta \in S(\theta_j, \rho_{\theta_j})$; hence from the definition of $\psi$ it follows that

$$f(x,\theta) \le \psi(x, \theta_j, \rho_{\theta_j})$$

for all $x$. Therefore

$$\mathbb{P}_n f(\cdot,\theta) \le \mathbb{P}_n \psi(\cdot, \theta_j, \rho_{\theta_j}),$$

and hence

$$\begin{aligned}
\sup_{\theta\in\Theta} \mathbb{P}_n f(\cdot,\theta) &\le \sup_{1\le j\le m} \mathbb{P}_n \psi(\cdot, \theta_j, \rho_{\theta_j}) \\
&\to_{a.s.} \sup_{1\le j\le m} P\psi(\cdot, \theta_j, \rho_{\theta_j}) \\
&\le \sup_{1\le j\le m} g(\theta_j) + \epsilon \\
&\le \sup_{\theta\in\Theta} g(\theta) + \epsilon\,.
\end{aligned}$$

We conclude that

$$\text{limsup}_{n\to\infty} \sup_{\theta\in\Theta} \mathbb{P}_n f(\cdot,\theta) \le_{a.s.} \sup_{\theta\in\Theta} g(\theta) + \epsilon\,.$$

Letting $\epsilon \downarrow 0$ completes the proof. $\square$

**Proof.** **Theorem 4.1.** Since $f$ is continuous in $\theta$, condition (d) of Theorem 2 is satisfied: for any countable set $D$ dense in $\{\theta' : |\theta' - \theta| < \rho\}$,

$$\sup_{|\theta'-\theta|<\rho} f(x,\theta') = \sup_{\theta'\in D} f(x,\theta')$$

where the right side is measurable since it is a countable supremum of measurable functions. Furthermore, $g(\theta)$ is continuous in $\theta$:

$$g(\theta) = \lim_{\theta'\to\theta} g(\theta') = \lim_{\theta'\to\theta} \int f(x,\theta') dP(x) = \int f(x,\theta) dP(x)$$

by the dominated convergence theorem. Now Theorem 4.1 follows from Theorem 4.2 applied to the functions $h(x,\theta) \equiv f(x,\theta) - g(\theta)$ and $-h(x,\theta)$: by Theorem 4.2 applied to $\{h(x,\theta) : \theta \in \Theta\}$,

$$\text{limsup}_{n\to\infty} \sup_{\theta\in\Theta} (\mathbb{P}_n f(\cdot,\theta) - g(\theta)) \le 0 \qquad \text{a.s.}$$

By Theorem 4.2 applied to $\{-h(x,\theta) : \theta \in \Theta\}$,

$$\text{limsup}_{n\to\infty} \sup_{\theta\in\Theta} (g(\theta)) - \mathbb{P}_n f(\cdot,\theta)) \le 0 \qquad \text{a.s.}$$

The conclusion of Theorem 4.1 follows since

$$\begin{aligned}
0 &\le \sup_{\theta\in\Theta} |\mathbb{P}_n f(\cdot,\theta) - g(\theta)| \\
&= \sup_{\theta\in\Theta} (\mathbb{P}_n f(\cdot,\theta) - g(\theta)) \vee \sup_{\theta\in\Theta} (g(\theta) - \mathbb{P}_n f(\cdot,\theta))\,.
\end{aligned}$$

$\square$

For our application of Theorem 4.2 to consistency of maximum likelihood, the following Lemma will be useful.

**Lemma 4.1** If the conditions of Theorem 4.2 hold, then $g(\theta)$ is upper-semicontinuous: i.e.

$$\text{limsup}_{\theta' \to \theta} g(\theta') \le g(\theta).$$

**Proof.** Since $f(x, \theta)$ is upper semicontinuous,

$$\text{limsup}_{\theta' \to \theta} f(x, \theta') \le f(x, \theta) \qquad \text{for all } x\,;$$

i.e.

$$\text{liminf}_{\theta' \to \theta} \left\{ f(x, \theta) - f(x, \theta') \right\} \ge 0 \qquad \text{for all } x\,.$$

Hence it follows by Fatou's lemma that

$$\begin{aligned} 0 &\le E\text{liminf}_{\theta' \to \theta} \left\{ f(X, \theta) - f(X, \theta') \right\} \\ &\le \text{liminf}_{\theta' \to \theta} E \left\{ f(X, \theta) - f(X, \theta') \right\} \\ &= Ef(X, \theta) - \text{limsup}_{\theta' \to \theta} Ef(X, \theta')\,; \end{aligned}$$

i.e.

$$\text{limsup}_{\theta' \to \theta} Ef(X, \theta') \le Ef(X, \theta) = g(\theta)\,.$$

□

Now we are prepared to tackle consistency of maximum likelihood estimates.

**Theorem 4.3** (Wald, 1949). Suppose that $X, X_1, \ldots, X_n$ are i.i.d. $P_{\theta_0}$, $\theta_0 \in \Theta$ with density $p(x, \theta_0)$ with respect to the dominating measure $\nu$, and that:
(a) $\Theta$ is compact.
(b) $p(x, \cdot)$ is upper semi-continuous in $\theta$ for all $x$.
(c) There exists a function $F(x)$ such that $EF(X) < \infty$ and

$$f(x, \theta) \equiv \log p(x, \theta) - \log p(x, \theta_0) \le F(x)$$

for all $x \in \mathcal{X}$, $\theta \in \Theta$.
(d) For all $\theta$ and all sufficiently small $\rho > 0$

$$\sup_{|\theta' - \theta| < \rho} p(x, \theta')$$

is measurable in $x$.
(e) $p(x, \theta) = p(x, \theta_0)$ a.e. $\nu$ implies that $\theta = \theta_0$.
Then for any sequence of maximum likelihood estimates $\widehat{\theta}_n$ of $\theta_0$,

$$\widehat{\theta}_n \to_{a.s.} \theta_0\,.$$

**Proof.** Let $\rho > 0$. The functions $\{f(x, \theta) : \theta \in \Theta\}$ satisfy the conditions of theorem 4.2. But we will apply Theorem 4.2 with $\Theta$ replaced by the subset

$$S \equiv \{\theta : |\theta - \theta_0| \ge \rho\} \subset \Theta\,.$$

Then $S$ is compact, and by Theorem 4.2

$$P_{\theta_0}\left(\text{limsup}_{n\to\infty}\sup_{\theta\in S}\mathbb{P}_n f(\cdot,\theta)\le\sup_{\theta\in S}g(\theta)\right)=1$$

where

$$g(\theta)=E_{\theta_0}f(X,\theta)=E_{\theta_0}\left\{\log\frac{p(X,\theta)}{p(X,\theta_0)}\right\}=-K(P_{\theta_0},P_\theta)<0\qquad\text{for}\ \ \theta\in S.$$

Furthermore by the Lemma, $g(\theta)$ is upper semicontinuous and hence achieves its supremum on the compact set $S$. Let $\delta=\sup_{\theta\in S}g(\theta)$. Then by Lemma 4.1.1 and the identifiability assumption (e), it follows that $\delta<0$ and we have

$$P_{\theta_0}\left(\text{limsup}_{n\to\infty}\sup_{\theta\in S}\mathbb{P}_n f(\cdot,\theta)\le\delta\right)=1\,.$$

Thus with probability 1 there exists an $N$ such that for all $n>N$

$$\sup_{\theta\in S}\mathbb{P}_n f(\cdot,\theta)\le\delta/2<0\,.$$

But

$$\mathbb{P}_n f(\cdot,\widehat{\theta}_n)=\sup_{\theta\in\Theta}\mathbb{P}_n f(\cdot,\theta)=\sup_{\theta\in\Theta}\frac{1}{n}\left\{l_n(\theta)-l_n(\theta_0)\right\}\ge 0\,.$$

Hence $\widehat{\theta}_n\notin S$ for $n>N$; that is, $|\widehat{\theta}_n-\theta_0|<\rho$ with probability 1. Since $\rho$ was arbitrary, $\widehat{\theta}_n$ is a.s. consistent. $\square$

**Remark 4.1** Theorem 4.3 is due to Wald (1949). The present version is an adaptation of Chapters 16 and 17 of Ferguson (1996). For further Glivenko - Cantelli theorems, see chapter 2.4 of Van der Vaart and Wellner (1996).

## 5 The EM algorithm

In statistical applications it is a fairly common occurrence that the observations involve "missing data" or "incomplete data", and this results in complicated likelihood functions for which there is no explicit formula for the MLE. Our goal in this section is to introduce one quite general scheme for maximizing likelihoods, the *EM - algorithm*, which is useful for dealing with missing or incomplete data.

Suppose that we observe $Y \sim Q_\theta$ on $(\mathcal{Y}, \mathcal{B})$ for some $\theta \in \Theta$; we assume that $Q_\theta$ has density $q_\theta$ with respect to a dominating measure $\nu$. This is the "observed" or "incomplete data".

On the other hand there is often an $X \sim P_\theta$ on $(\mathcal{X}, \mathcal{A})$ which has a simpler likelihood or for which the MLE's can be calculated explicitly, and satisfying $Y = T(X)$. Here we will assume that $P_\theta$ has density $p_\theta$ with respect to $\mu$, and we refer to $X$ as the "unobserved" or "complete data". Since $Y = T(X)$, it follows that

$$Q_\theta(B) \;=\; Q_\theta(Y \in B) = P_\theta(T(X) \in B) = P_\theta(X \in T^{-1}(B)),$$

or $Q_\theta = P_\theta \circ T^{-1}$. We want to compute

$$\hat{\theta}^Y = \text{argmax}_\theta \log q_\theta(Y),$$

but this is often difficult because $q_\theta$ is complicated. We don't get to observe $X$, but if we did, then often computation of

$$\hat{\theta}^X = \text{argmax}_\theta \log p_\theta(X)$$

is much easier.

How to proceed? Choose an initial estimator $\theta^{(0)}$ of $\theta$, and hence an initial estimator $P_{\theta^{(0)}}$ of $P_\theta$. Then we would proceed via an

**E-step:** compute, for $\theta \in \Theta$,

$$\phi_0(\theta) = \phi_0(\theta, Y) = E_{P_{\theta^{(0)}}}\{\log p_\theta(X) | T(X) = Y\}.$$

This is the "estimated log-likelihood based on our current guess $\theta^{(0)}$ of $\theta$ and our observations $Y$. Then carry out an

**M-step:** maximize $\phi_0(\theta) = \phi_0(\theta, Y)$ as a function of $\theta$ to find

$$\theta^{(1)} = \text{argmax}\phi_0(\theta).$$

Now iterate the above two steps:

The following examples illustrate this general scheme.

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

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

Therefore

$$
\begin{aligned}
q_\theta(\underline{y}) &= \frac{n!}{y_1!y_2!y_3!y_4!}\left(\frac{1}{2}+\frac{\theta}{4}\right)^{y_1}\left(\frac{1-\theta}{4}\right)^{y_2}\left(\frac{1-\theta}{4}\right)^{y_3}\left(\frac{\theta}{4}\right)^{y_4}, \\
l(\theta|\underline{Y}) &= Y_1\log(1/2+\theta/4)+(Y_2+Y_3)\log(1-\theta)+Y_4\log\theta+\text{constant}, \\
\dot{\mathbf{l}}_\theta(\underline{Y}) &= Y_1\frac{1/4}{1/2+\theta/4}-(Y_2+Y_3)\frac{1}{1-\theta}+Y_4\frac{1}{\theta}, \\
\ddot{\mathbf{l}}(\underline{Y}) &= -Y_1\frac{(1/4)^2}{(1/2+\theta/4)^2}-(Y_2+Y_3)\frac{1}{(1-\theta)^2}-Y_4\frac{1}{\theta^2},
\end{aligned}
$$

and

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

If $\overline{\theta}_n$ is a preliminary estimator of $\theta$, then a one-step estimator of $\theta$ is given by

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

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

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

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

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

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

so that

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

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

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

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

(4) $$E(\underline{X}|\underline{Y})=\left(Y_1\frac{1/2}{1/2+\theta/4},Y_1\frac{\theta/4}{1/2+\theta/4},Y_2,Y_3,Y_4\right),$$

so we set

(5) $$\underline{\widehat{X}}^{(p)}\equiv\left(Y_1\frac{1/2}{1/2+\theta^{(p)}/4},Y_1\frac{\theta^{(p)}/4}{1/2+\theta^{(p)}/4},Y_2,Y_3,Y_4\right)$$

where $\theta^{(p)}$ is the estimator of $\theta$ at the $p-$th step of the algorithm.

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

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

yields

$$\widehat{\theta}_n = \frac{X_2 + X_5}{X_2 + X_5 + X_3 + X_4}$$

as an estimator of $\theta$ based on the full data. Since we can only observe $\underline{Y}$, we take

$$\widehat{\theta}^{(p+1)} = \frac{\widehat{X}_2^{(p)} + X_5}{\widehat{X}_2^{(p)} + X_5 + X_3 + X_4}\,, \tag{6}$$

and alternative between (5) and (6) with a "reasonable" guess $\widehat{\theta}_n^0$ of $\theta$; say $\widehat{\theta}^{(0)} = 1/2$. This yields the following table:

Table 4.1: Iterates of E and M steps in the Multinomial example

| $p$ | $\widehat{\theta}^{(p)}$ | $\widehat{\theta}^{(p)} - \widehat{\theta}_n$ | $\frac{\widehat{\theta}^{(p+1)} - \widehat{\theta}_n}{\widehat{\theta}^{(p)} - \widehat{\theta}_n}$ |
|---|---|---|---|
| 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 $\widehat{\theta}_n = .62682149....$

**Example 5.2** (Exponential mixture model). Suppose that $Y \sim Q_\theta$ on $R^+$ where $Q_\theta$ has density

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

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

$$\begin{aligned}
\dot{\mathbf{l}}_p(Y) &= \frac{\lambda e^{-\lambda Y} - \mu e^{-\mu Y}}{q_\theta(Y)}\,, \\
\dot{\mathbf{l}}_\lambda(Y) &= \frac{p e^{-\lambda Y}(1-\lambda Y)}{q_\theta(Y)}\,, \qquad \text{and} \\
\dot{\mathbf{l}}_\mu(Y) &= \frac{(1-p) e^{-\mu Y}(1-\mu Y)}{q_\theta(Y)}\,.
\end{aligned}$$

It is easily seen that $E_\theta \dot{\mathbf{l}}_p^2(Y) < \infty$, $E_\theta \dot{\mathbf{l}}_\lambda^2(Y) < \infty$, and $E_\theta \dot{\mathbf{l}}_\mu^2(Y) < \infty$, and, moreover, that that the information matrix 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

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

Thus the "incomplete data" $Y$ is just the first coordinate of $X$. Furthermore, maximum likelihood estimation of $\theta$ in the complete data problem is easy, as can be seen by the following calculations. The log of the density $p_\theta(x)$ is

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

so that

$$\begin{aligned}
\dot{\mathbf{l}}_p(X) &= \frac{\Delta}{p} - \frac{1-\Delta}{1-p}, \qquad \widehat{p} = \frac{1}{n}\sum_{i=1}^n \Delta_i\,, \\
\dot{\mathbf{l}}_\lambda(X) &= \Delta\left(\frac{1}{\lambda} - Y\right), \qquad \frac{1}{\widehat{\lambda}} = \frac{\sum_{i=1}^n \Delta_i Y_i}{\sum_{i=1}^n \Delta_i}\,, \\
\dot{\mathbf{l}}_\mu(X) &= (1-\Delta)\left(\frac{1}{\mu} - Y\right), \qquad \frac{1}{\widehat{\mu}} = \frac{\sum_{i=1}^n (1-\Delta_i) Y_i}{\sum_{i=1}^n (1-\Delta_i)}\,.
\end{aligned}$$

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

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

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

$$\begin{aligned}
\widehat{p}^{(m+1)} &= \frac{1}{n}\sum_{i=1}^n \widehat{\Delta}_i^{(m)}, \qquad \widehat{\Delta}(Y_i, \widehat{\theta}^{(m)})\,, \\
\frac{1}{\widehat{\lambda}^{(m+1)}} &= \frac{\sum_{i=1}^n Y_i \widehat{\Delta}_i^{(m)}}{\sum_{i=1}^n \widehat{\Delta}_i^{(m)}}\,, \\
\frac{1}{\widehat{\mu}^{(m+1)}} &= \frac{\sum_{i=1}^n Y_i (1-\widehat{\Delta}_i^{(m)})}{\sum_{i=1}^n (1-\widehat{\Delta}_i^{(m)})}\,.
\end{aligned}$$

## 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 $X_1, \ldots, X_n$ be i.i.d. $\mathbb{X}-$ valued random variables with common probability distribution (measure) $P$ on $(\mathbb{X}, \mathcal{A})$. For an arbitrary probability distribution (measure) $Q$ on $(\mathbb{X}, \mathcal{A})$, let $Q(\{X_i\}) \equiv q_i$. If there are no ties in the $X_i$'s, then

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

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

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

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

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

with equality if and only if $\sum q_i = 1$ and $q_1 = \cdots = q_n = 1/n$. Thus the MLE of $P$ is $\mathbb{P}_n$ with $\mathbb{P}_n(\{X_i\}) = 1/n$, or

$$\mathbb{P}_n(A) = \frac{1}{n}\#\{i \le n : \ X_i \in A\} = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(A)\,.$$

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

**Theorem 6.1** The nonparametric maximum likelihood estimator of a completely arbitrary distribution $P$ on any measurable space based on i.i.d. data is the empirical measure $\mathbb{P}_n = n^{-1}\sum_1^n \delta_{X_i}$.

In particular, if $\mathbb{X} = R$, the empirical distribution function $\mathbb{F}_n(x) \equiv n^{-1}\sum_1^n 1_{[X_i \le x]}$ is the MLE of $F$. Recall from Chapter 2 that Donsker's theorem yields

$$\sqrt{n}(\mathbb{F}_n - F) \stackrel{d}{=} \mathbb{U}_n(F) \Rightarrow \mathbb{U}(F)$$

where $\mathbb{U}_n$ is the empirical process of i.i.d. $U[0,1]$ random variables and $\mathbb{U}$ is a Brownian bridge process.

**Example 6.2** (Censored data). Suppose that $X_1, \ldots, X_n$ are i.i.d. $F$ and $Y_1, \ldots, Y_n$ are i.i.d. $G$. Think of the $X$'s as survival times and the $Y$'s as censoring times. Suppose that we can observe only the i.i.d. pairs $(Z_1, \Delta_1), \ldots, (Z_n, \Delta_n)$ where $Z_i \equiv X_i \wedge Y_i$ and $\Delta_i \equiv 1\{X_i \le Y_i\}$. Thus the joint distribution of $(Z, \Delta)$ 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

$$1 - H(z) = P(Z > z) = P(X > z, Y > z) = (1 - F(z))(1 - G(z))\,.$$

Now suppose for simplicity that $Z_{n:1} \le \cdots \le Z_{n:n}$ are all distinct. Let $\Delta_{n:1}, \ldots, \Delta_{n:n}$ be the corresponding $\Delta$'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 $q_i \equiv G\{Z_{n:i}\} = G(Z_{n:i}) - G(Z_{n:i}-) = \Delta G(Z_{n:i})$, $i = 1, \ldots, n$, $p_{n+1} = 1 - F(Z_{n:n}) = 1 - \sum_{j=1}^{n} p_j$, and $q_{n+1} = 1 - G(Z_{n:n}) = 1 - \sum_{j=1}^{n} q_j$, then a nonparametric likelihood for the censored data problem is

$$\prod_{i=1}^{n} p_i^{\Delta_{n:i}} \left( \sum_{j=i}^{n+1} q_j \right)^{\Delta_{n:i}} q_i^{1-\Delta_{n:i}} \left( \sum_{j=i+1}^{n+1} p_j \right)^{1-\Delta_{n:i}} = \prod_{i=1}^{n} p_i^{\Delta_{n:i}} \left( \sum_{j=i+1}^{n+1} p_j \right)^{1-\Delta_{n:i}} \times B \tag{2}$$

where $B$ depends only on the $q_i$'s, and hence only on $G$. Thus we can find the nonparametric maximum likelihood estimator of $F$ by maximizing the first term over the $p_i$'s. This is easy once the log-likelihood is re-written in terms of $\lambda_i \equiv p_i / \sum_{j=i}^{n+1} p_j$, $i = 1, \ldots, n+1$.

We will first take a different approach by using example 6.1 as follows: if $F$ has density $f$, then the hazard function $\lambda$ is given by $\lambda(t) = f(t)/(1 - F(t))$, and the cumulative hazard function $\Lambda$ 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 distribution function $F$, it turns out that the "right" way to define $\Lambda$, the cumulative hazard function corresponding to $F$, is:

$$\Lambda(t) = \int_{[0,t]} \frac{1}{1 - F(s-)} dF(s)\,. \tag{3}$$

Note that we can write $\Lambda$ as

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

Moreover, we can estimate both $H^{uc}$ and $H$ by their natural nonparametric estimators (from Example 6.1):

$$\mathbb{H}_n^{uc}(z) = \frac{1}{n} \sum_{i=1}^{n} 1_{[Z_i \le z, \Delta_i = 1]}, \qquad \mathbb{H}_n(z) = \frac{1}{n} \sum_{i=1}^{n} 1_{[Z_i \le z]}\,.$$

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

$$\widehat{\Lambda}_n(t) = \int_0^t \frac{1}{1 - \mathbb{H}_n(s-)} d\mathbb{H}_n^{uc}(s)\,. \tag{4}$$

It remains only to invert the relationship (3) 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

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

**Proposition 6.1** Suppose that $\Lambda$ is the cumulative hazard function corresponding to an arbitrary distribution function $F$ as defined by (3). Then

$$(5) \qquad 1 - F(t) = \exp(-\Lambda_c(t)) \prod_{s \le t} (1 - \Delta\Lambda(s)) \equiv \prod_{s \le t} (1 - d\Lambda(s))\,.$$

**Proof.** In the case of a continuous distribution function $F$, $\Lambda$ is also continuous, $\Lambda = \Lambda_c$, $\Delta\Lambda = 0$ identically, and we calculate $\Lambda(t) = -\log(1 - F(t))$ so that (5) holds.

In the case of a purely discrete distribution function $F$ the cumulative hazard function $\Lambda$ 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 $s_1, \ldots, s_k$ are the points of jump of $F$ which are less than or equal to $t$. Hence (5) also holds in this case. For a complete proof of the general case, which relies on rewriting (3) as

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

or equivalently

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

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. □

Now we return to the likelihood in (2) with the goal of maximizing it directly. We first use the discrete form of the identities above linking $F$ and $\Lambda$ to re-write (2) in terms of $\lambda_i \equiv p_i / \sum_{j=i}^{n+1} p_j$: note that

$$\prod_{i=1}^{n} p_i^{\Delta_{n:i}} \left( \sum_{j=i+1}^{n+1} p_j \right)^{1-\Delta_{n:i}} = \prod_{i=1}^{n} \left( \frac{p_i}{\sum_{j=i}^{n+1} p_j} \right)^{\Delta_{n:i}} \left( \frac{\sum_{j=i}^{n+1} p_j}{\sum_{j=i+1}^{n+1} p_j} \right)^{\Delta_{n:i}} \sum_{j=i+1}^{n+1} p_j$$

$$
\begin{aligned}
&= \prod_{i=1}^{n} \lambda_i^{\Delta_{n:i}} (1-\lambda_i)^{-\Delta_{n:i}} \left( \sum_{j=i+1}^{n+1} p_j \right) \\
&\qquad \text{since } \ 1-\lambda_i = 1 - \frac{p_i}{\sum_{j=i}^{n+1} p_j} = \frac{\sum_{j=i+1}^{n+1} p_j}{\sum_{j=i}^{n+1} p_j} \\
&= \prod_{i=1}^{n} \lambda_i^{\Delta_{n:i}} (1-\lambda_i)^{-\Delta_{n:i}} \cdot \prod_{i=1}^{n} \prod_{j=1}^{i} (1-\lambda_j) \\
&\qquad \text{since } \ \prod_{j=1}^{i} (1-\lambda_j) = \sum_{j=i+1}^{n+1} p_j \\
&= \prod_{i=1}^{n} \lambda_i^{\Delta_{n:i}} (1-\lambda_i)^{-\Delta_{n:i}} \prod_{j=1}^{n} (1-\lambda_j)^{n-j+1} \\
&= \prod_{i=1}^{n} \lambda_i^{\Delta_{n:i}} (1-\lambda_i)^{n-\Delta_{n:i}-i+1}.
\end{aligned}
$$

Since each term of this likelihood has the same form as a Binomial likelihood, it is clear that it is maximized by

$$\hat{\lambda}_i = \frac{\Delta_{n:i}}{n-i+1}, \qquad i = 1, \ldots, n.$$

Note that this agrees exactly with our estimator $\hat{\Lambda}_n$ derived above (in the case of no ties): $\Delta\hat{\Lambda}_n(Z_{n:i}) = \Delta_{n:i}/(n-i+1)$.

The right side of (5) is called the *product integral*; see Gill and Johansen (1990) for a survey. It follows from proposition 6.1 that the nonparametric maximum likelihood estimator of $F$ in the case of censored data is the *product limit estimator* $\widehat{\mathbb{F}}_n$ given by

$$
\begin{aligned}
(6) \qquad 1 - \widehat{\mathbb{F}}_n(t) &= \prod_{s \le t} (1 - \Delta\widehat{\Lambda}_n(s)) \\
&= \prod_{i: Z_{n:i} \le t} \left( 1 - \frac{\Delta_{n:i}}{n-i+1} \right) \qquad \text{if there are no ties .}
\end{aligned}
$$

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

$$(7) \qquad \sqrt{n}(\widehat{\Lambda}_n - \Lambda) \Rightarrow \mathbb{B}(C) \qquad \text{in } \ D[0,\tau], \quad \tau < \tau_H;$$

and hence that

$$(8) \qquad \sqrt{n}(\widehat{\mathbb{F}}_n - F) \Rightarrow (1-F)\mathbb{B}(C) \stackrel{d}{=} \frac{1-F}{1-K}\mathbb{U}(K) \qquad \text{in } \ D[0,\tau]$$

for $\tau < \tau_H$ 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 and $F$ is continuous $K = F$ and the limit process in (8) becomes just $\mathbb{U}(F)$, the limit process of the usual empirical process. Martingale methods for proving the convergences (7) are due to Gill (1980), (1983); see Shorack and Wellner (1986), chapter 7.

**Example 6.3** (Cox's proportional hazards model and profile likelihood). Suppose that $T$ is a survival time and $Z$ is a covariate vector with values in $R^k$. Further, suppose that $(T|Z)$ has conditional hazard function

$$\lambda(t|z) = e^{\theta^T z}\lambda(t)\,.$$

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

$$\Lambda(t|z) = \exp(\theta^T z)\Lambda(t)\,,$$

and, assuming that $F$ is continuous, with $\overline{F} \equiv 1 - F$,

$$1 - F(t|z) = \overline{F}(t|z) = \overline{F}(t)^{\exp(\theta^T z)}\,,$$

or,

$$f(t|z) = \exp(\theta^T z)\overline{F}(t)^{\exp(\theta^T z)}\lambda(t)\,,$$

If we assume that $Z$ has density $h$, then

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

Hence

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

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

$$\begin{aligned} l(\theta, \lambda, h|\underline{X}) &= \sum_{i=1}^n \{\theta^T Z_{(i)} - \exp(\theta^T Z_{(i)})\sum_{j\le i}\lambda_j + \log\lambda_i + \log h_i\} \\ &= \sum_{i=1}^n \theta^T Z_{(i)} + \sum_{i=1}^n \left\{\log\lambda_i - \lambda_i\sum_{j\ge i}\exp(\theta^T Z_{(j)})\right\} + \sum_{i=1}^n \log h_i \end{aligned}$$

since

$$\begin{aligned} \sum_{i=1}^n\sum_{j\le i}\lambda_j e^{\theta^T Z_{(i)}} &= \sum_{j=1}^n\sum_{i=1}^n \lambda_j 1\{j \le i\}\exp(\theta^T Z_{(i)}) \\ &= \sum_{j=1}^n \lambda_j\left(\sum_{i\ge j}\exp(\theta^T Z_{(i)})\right)\,. \end{aligned}$$

Maximizing this with respect to $\lambda_i$ and $h_i$ (subject to the constraint $\sum_{i=1}^n h_i = 1$ and assuming that all the $Z_j$'s are distinct) yields

$$\widehat{\lambda}_i = \frac{1}{\sum_{j\ge i}\exp(\theta^T Z_{(j)})}\,, \qquad \widehat{h}_i = 1/n\,.$$

Thus the profile log-likelihood for $\theta$ is given by

$$(9) \qquad l^{prof}(\theta|\underline{X}) = \log\left\{\prod_{i=1}^{n}\frac{\exp(\theta^T Z_{(i)})}{\sum_{j\ge i}\exp(\theta^T Z_{(j)})}\frac{1}{(ne)^n}\right\}.$$

The first factor here is (the log of) *Cox's partial likelihood* for $\theta$; Cox (1972) derived this by other means. Maximizing it over $\theta$ yields Cox's partial likelihood estimator of $\theta$, which is in fact the maximum (nonparametric or semiparametric) profile likelihood estimator. Let

$$\widehat{\theta}_n \equiv \text{argmax}_\theta\, l^{prof}(\theta|\underline{X}).$$

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

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

where

$$\mathbb{H}_n(t) \equiv n^{-1}\sum_{i=1}^n 1_{[T_i\le t]}, \qquad \mathbb{Y}_n(t,\theta) \equiv n^{-1}\sum_{i=1}^n 1_{[T_i\ge t]}\exp(\theta Z_i).$$

This estimator was derived by Breslow (1972), (1974), and is now commonly called the *Breslow estimator* of $\Lambda$. It is also asymptotically efficient; see Begun, Hall, Huang, and Wellner (1983) and Bickel, Klaassen, Ritov, and Wellner (1993). Although our treatment here has not included right censoring, this can easily be incorporated in this model, and this was one of the key contributions of Cox (1972).

**Example 6.4** (Estimation of a concave distribution function and monotone decreasing density). Suppose that the model $\mathcal{P}$ is all probability distributions $P$ on $R^+ = [0,\infty)$ with corresponding distribution functions $F$ which are concave. It follows that the distribution function $F$ corresponding to $P \in \mathcal{P}$ has a density $f$ and that $f$ is nonincreasing. It was shown by Grenander (1956) that if $X_1, \ldots, X_n$ are i.i.d. $P \in \mathcal{P}$ with distribution function $F$, then the MLE of $F$ over $\mathcal{P}$ is the least concave majorant $\widehat{\mathbb{F}}_n$ of $\mathbb{F}_n$; and thus the MLE $\widehat{f}_n$ of $f$ is given by the slope of $\widehat{\mathbb{F}}_n$. See Barlow, Bartholomew, Bremner, and Brunk (1972) for this and related results. It was shown by Kiefer and Wolfowitz (1976) that

$$\sqrt{n}(\widehat{\mathbb{F}}_n - F) \Rightarrow \mathbb{U}(F),$$

and this phenomena of no improvement or reduction in asymptotic variance even though the model $\mathcal{P}$ is a proper subset of $\mathcal{M} \equiv \{\text{all } P \text{ on } R^+\}$ is explained by Millar (1979). Prakasa Rao (1969) showed that if $f(t) > 0$, then

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

where $\mathbb{Z}$ is the location of the maximum of the process $\{B(t) - t^2 : \ t \in R\}$ where $B$ is standard Brownian motion starting from 0; his proof has been greatly simplified and the limit distribution examined in detail by Groeneboom (1984), (1989). These results have been extended to estimation of a monotone density with right-censored data by Huang and Zhang (1994) and Huang and Wellner (1995).

**Example 6.5** (Interval censored or "current status" data). Suppose, as in example 6.2, that $X_1, \ldots, X_n$ are i.i.d. $F$ and $Y_1, \ldots, Y_n$ are i.i.d. $G$, but now suppose that we only observe $(Y_i, \Delta_i)$ where $\Delta_i = 1_{[X_i \le Y_i]}$. Again the goal is to estimate $F$, even though we never observe an $X$ directly, but only the indicators $\Delta$. If $G$ has density $g$, then the density $p(y, \delta)$ of the i.i.d. pairs $(Y, \delta)$ is

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

if we suppose that there are no ties in the $Y$'s, let $Y_{n:1} < Y_{n:2} < \cdots < Y_{n:n}$, and write $P_i \equiv F(Y_{n:i})$, $q_i \equiv G(\{Y_{n:i}\})$, then a nonparametric likelihood for the data is given by

$$L(\underline{P}, \underline{q} | \underline{Y}, \underline{\Delta}) = \prod_{i=1}^{n} P_i^{\Delta_{n:i}} (1 - P_i)^{1-\Delta_{n:i}} q_i\,,$$

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 $0 \le P_1 \le \ldots \le P_n \le 1$. This was solved by Ayer, Brunk, Ewing, Reid, and Silverman (1955) and also by van Eeden (1956), (1957). The following description of the solution is from Groeneboom and Wellner (1992).

**(i)** Plot the points $\{(0,0), (i, \sum_{j \le i} \Delta_{n:j}),\ i = 1, \ldots, n\}$. This is called the *cumulative sum diagram*.

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

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

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

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

$$n^{1/3}(\widehat{\mathbb{F}}_n(t) - F(t)) \to_d \left(\frac{F(t)(1 - F(t)) f(t)}{2g(t)}\right)^{1/3} (2\mathbb{Z})$$

where $\mathbb{Z}$ is the location of the maximum of the process $\{B(t) - t^2 :\ t \in R\}$ as in example 6.4.

For further discussion of the definition of nonparametric maximum likelihood estimators see Kiefer and Wolfowitz (1956) and Scholz (1980). For applications to a mixture model, see Jewell (1982). Other applications of nonparametric maximum likelihod include the work of Vardi (1985) and Gill, Vardi, and Wellner (1988) on biased sampling models. 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. Some progress on the general theory of nonparametric maximum likelihood estimators has been made by Gill (1989), Gill and van der Vaart (1993), and van der Vaart (1995).

# 7 Limit theory for the statistical agnostic

In the preceding sections we studied the limit behavior of the MLE $\widehat{\theta}_n$ (or ELE $\widetilde{\theta}_n$) under the assumption that the model $\mathcal{P}$ is true; i.e. assuming that the data $X_1, \ldots, X_n$ were governed by a probability distribution $P_\theta \in \mathcal{P}$. Frequently however we are in the position of not being at all sure that the true $P$ is an element of the model $\mathcal{P}$, and it is natural to ask about the asymptotic behavior of $\widehat{\theta}_n$ (or $\widetilde{\theta}_n$ when, in fact, $P \notin \mathcal{P}$. 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), and Bickel, Klaassen, Ritov and Wellner (1993) appendix A.10 and sections 7.2 - 7.4.

**Heuristics for Maximum Likelihood**

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

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

is twice continuously differentiable in $\theta$ for $P-$ a.e. $x$. We *do not assume* that $P \in \mathcal{P} = \{P_\theta : dP_\theta/d\mu = p_\theta,\ \theta \in \Theta\}$. 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 $\mu$ which also dominates all $P_\theta$, $\theta \in \Theta$. Then the maximum likelihood estimator maximizes

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

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

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

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)

$$\begin{aligned}
\widehat{\theta}_n \quad &= \quad \mathrm{argmax}\{\frac{1}{n}\sum_{i=1}^{n} \rho(X_i;\theta)\} \\
&\to_p \quad \mathrm{argmax}\{E_P \log p(X_1) - K(P,P_{\theta_0})\} = \theta_0(P) = \mathrm{argmin}_{\theta\in\Theta} K(P,P_\theta)\,.
\end{aligned}$$

What about a central limit theorem? First note that since

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

and

$$\widehat{\theta}_n \quad \text{maximizes} \quad \frac{1}{n}\sum_{i=1}^{n} \rho(X_i;\theta) = \mathbb{P}_n \rho(x;\theta)\,,$$

we expect that

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

and

$$0 = \nabla_\theta \mathbb{P}_n \rho(X, \theta)|_{\theta=\widehat{\theta}_n} = \mathbb{P}_n \psi(X;\widehat{\theta}_n)\,.$$

Therefore, by Taylor expansion of $\mathbb{P}_n\psi(X;\theta)$ about $\theta_0$, it follows that

$$\begin{aligned} 0 &= \Psi_n(\widehat{\theta}_n) \equiv \mathbb{P}_n \psi(X;\widehat{\theta}_n) \\ &= \Psi_n(\theta_0) + \dot{\Psi}_n(\theta_n^*)(\widehat{\theta}_n - \theta_0) \end{aligned}$$

where

$$\begin{aligned} &(1) \qquad \sqrt{n}\Psi_n(\theta_0) &&= \sqrt{n}\mathbb{P}_n \psi(\cdot\,;\theta_0) \\ &(2) &&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \psi(X_i;\theta_0) \to_d \mathbb{Z} \sim N_d(0, K) \end{aligned}$$

with

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

We also have

$$\dot{\Psi}_n(\theta_0) = \mathbb{P}_n \dot{\psi}(X;\theta_0) \to_{a.s.} E_P \dot{\psi}(X;\theta_0)\,,$$

and hence we also expect to be able to show that

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

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

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

Note that if, in fact $P \in \mathcal{P}$, 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.