

557: MATHEMATICAL STATISTICS II

LARGE SAMPLE AND ASYMPTOTIC RESULTS

We now assess the properties of statistical procedures when the sample size n becomes **large** (*large sample theory*), or in the limit as n **tends to infinity** (*asymptotic theory*).

1 Asymptotic Assessment of Point Estimators

Consider a potentially infinite sequence of random variables $X_1, X_2, \dots, X_n, \dots$, and a corresponding sequence of estimators $\{T_n, n \geq 1\}$ of parameter $\tau(\theta)$, where, for each n , $T_n \equiv T_n(X_1, \dots, X_n)$.

- **Consistency and Asymptotic Unbiasedness:** The sequence $\{T_n, n \geq 1\}$ is **consistent** for $\tau(\theta)$ if

$$T_n \longrightarrow \tau(\theta) \quad \forall \theta$$

in probability (*weak consistency*), almost surely (*strong consistency*), or in r th mean for some r (for $r = 2$, *mean-square consistency*). The sequence of estimators is **asymptotically unbiased** for $\tau(\theta)$ if

$$\lim_{n \rightarrow \infty} E_{f_{T_n|\theta}}[T_n] = \tau(\theta).$$

It can be shown that $X_n \xrightarrow{r=2} X \implies X_n \xrightarrow{p} X$, so that mean-square consistency implies weak consistency. Then

$$E_{f_{T_n|\theta}}[(T_n - \tau(\theta))^2] = \text{Var}_{f_{T_n|\theta}}[T_n] + \left(E_{f_{T_n|\theta}}[T_n - \tau(\theta)]\right)^2$$

so mean-square consistency follows if T_n is asymptotically unbiased and has variance converging to zero. The *asymptotic variance* of T_n is σ^2 if, for some sequence of constants $\{k_n\}$,

$$\lim_{n \rightarrow \infty} k_n \text{Var}_{f_{T_n|\theta}}[T_n] = \sigma^2 < \infty$$

In most applications, $k_n = n$ suffices.

- **Efficiency:** A sequence of asymptotically unbiased estimators $T_n = T_n(\underline{X})$ of $\tau(\theta)$ is **efficient** if the variance of $\sqrt{n}(T_n - \tau(\theta))$ converges to the lower bound on variance dictated by the Cramér-Rao result, that is

$$\lim_{n \rightarrow \infty} n E_{f_{T_n|\theta}}[(T_n - \tau(\theta))^2] = (\dot{\tau}(\theta))^2 \mathcal{I}(\theta)^{-1}$$

Note: For finite n , an unbiased estimator T is sometimes termed efficient if its variance attains the Cramér-Rao lower bound; the efficiency, $e_T(\theta)$, of an unbiased estimator of θ is defined by

$$e_T(\theta) = \frac{\mathcal{I}(\theta)^{-1}}{\text{Var}_{f_{T|\theta}}[T]}.$$

These definitions can be extended to the multivariate case.

- **Asymptotic Relative Efficiency:** Consider two estimators $\tau(\theta)$, $T_{1n} = T_{1n}(\underline{X})$ and $T_{2n} = T_{2n}(\underline{X})$. The **Asymptotic Relative Efficiency** (ARE) of T_{1n} with respect to T_{2n} is defined as the ratio of their rescaled asymptotic mean-square errors (AMSE):

$$\text{ARE}_\theta(T_{1n}, T_{2n}) = \frac{\text{AMSE}_\theta(T_{2n})}{\text{AMSE}_\theta(T_{1n})} = \frac{\lim_{n \rightarrow \infty} n E_{f_{T_{2n}|\theta}}[(T_{2n} - \tau(\theta))^2]}{\lim_{n \rightarrow \infty} n E_{f_{T_{1n}|\theta}}[(T_{1n} - \tau(\theta))^2]}.$$

For two asymptotically unbiased estimators, the ARE is the ratio of the asymptotic variances.

2 Asymptotic Behaviour Of The Maximum Likelihood Estimator

Consider a random sample $x_1, \dots, x_n$ from a probability model indexed by parameter $\underline{\theta} \in \Theta \subseteq \mathbb{R}^d$, with density denoted $f_{X|\underline{\theta}}$ with support $\mathbb{X}$. Denote the true value of $\underline{\theta}$ by $\underline{\theta}_0$. Denote by $L(\underline{\theta}|\underline{x})$ and $l(\underline{\theta}|\underline{x})$ the likelihood and log likelihood respectively, and denote by

$$i_j(\underline{\theta}) = \frac{\partial l(\underline{\theta}|\underline{x})}{\partial \theta_j} \quad \ddot{l}_{jk}(\underline{\theta}|\underline{x}) = \frac{\partial^2 l(\underline{\theta}|\underline{x})}{\partial \theta_j \partial \theta_k} \quad \ddot{\ddot{l}}_{jkl}(\underline{\theta}|\underline{x}) = \frac{\partial^3 l(\underline{\theta}|\underline{x})}{\partial \theta_j \partial \theta_k \partial \theta_l}$$

the partial derivatives up to order three of $l(\underline{\theta}|\underline{x}) = \log f_{X|\underline{\theta}}(x|\underline{\theta})$. Note that

$$l_n(\underline{\theta}) = l(\underline{\theta}|\underline{x}) = \sum_{i=1}^n l(\underline{\theta}|x_i) \quad \dot{l}_n(\underline{\theta}) = \dot{l}(\underline{\theta}|\underline{x}) = \sum_{i=1}^n \dot{l}(\underline{\theta}|x_i) = \sum_{i=1}^n \frac{\dot{f}_{X|\underline{\theta}}(x_i|\underline{\theta})}{f_{X|\underline{\theta}}(x_i|\underline{\theta})} \quad (1)$$

with similar results for the other derivatives. Under mild regularity conditions, we prove that a solution to the equation found by equating (1) to zero provides an estimate for which the corresponding estimator that is weakly consistent for $\underline{\theta}_0$.

Regularity Conditions

A1. **Identifiability** : $f_{X|\underline{\theta}_1}(x|\underline{\theta}_1) = f_{X|\underline{\theta}_2}(x|\underline{\theta}_2) \forall x \in \mathbb{X} \iff \underline{\theta}_1 = \underline{\theta}_2$

A2. $\mathbb{X}$ **does not depend on** $\underline{\theta}$.

A3. Θ contains an **open neighbourhood**, $\Theta_0 \subset \mathbb{R}$, of θ_0 on which

(i) $l(\underline{\theta}|\underline{x})$ is **twice continuously differentiable with respect to** θ for all $x \in \mathbb{X}$.

(ii) Third derivatives of $l(\underline{\theta}|\underline{x})$ **exist** and are **absolutely bounded**, that is for $\theta \in \Theta_0$

$$\left| \ddot{\ddot{l}}(\underline{\theta}|\underline{x}) \right| \leq M(x) \quad \text{where} \quad \mathbb{E}_{f_{X|\underline{\theta}}} [M(X)|\underline{\theta}_0] < m < \infty$$

A4.

$$\mathbb{E}_{f_{X|\underline{\theta}}} [\dot{l}(\underline{\theta}_0|X)] = 0 \quad \mathbb{E}_{f_{X|\underline{\theta}}} [(\dot{l}(\underline{\theta}_0|X))^2] < \infty.$$

2.1 Asymptotic behaviour of the Likelihood Ratio at $\underline{\theta}_0$

To find the maximum likelihood estimate, we solve the system of **likelihood equations**

$$\dot{l}_n(\underline{\theta}) = \dot{l}(\underline{\theta}|\underline{x}) = 0 \quad (\text{LE})$$

that is, a system of d equations based on the first partial derivative vector $\dot{l}$. First, note that if $\underline{\theta} \neq \underline{\theta}_0$,

$$T_n(\underline{x}, \underline{\theta}_0, \underline{\theta}) = \frac{1}{n} \frac{l_n(\underline{\theta})}{l_n(\underline{\theta}_0)} = \frac{1}{n} \sum_{i=1}^n \log \left\{ \frac{f_{X|\underline{\theta}}(x|\underline{\theta})}{f_{X|\underline{\theta}}(x|\underline{\theta}_0)} \right\}$$

then, as $n \rightarrow \infty$, by the weak law of large numbers (WLLN), say,

$$T_n(\underline{X}, \underline{\theta}_0, \underline{\theta}) \xrightarrow{p} \mathbb{E}_{f_{X|\underline{\theta}}} \left[\log \frac{f_{X|\underline{\theta}}(X|\underline{\theta})}{f_{X|\underline{\theta}}(X|\underline{\theta}_0)} \right] = \int \log \left\{ \frac{f_{X|\underline{\theta}}(x|\underline{\theta})}{f_{X|\underline{\theta}}(x|\underline{\theta}_0)} \right\} f_{X|\underline{\theta}}(x|\underline{\theta}_0) dx = -K(\underline{\theta}_0, \underline{\theta}) < 0$$

where $K(\underline{\theta}_0, \underline{\theta})$ is the Kullback-Leibler divergence between the pdfs with parameters $\underline{\theta}_0$ and $\underline{\theta}$. Hence, by A1, $T_n(\underline{X}, \underline{\theta}_0, \underline{\theta})$ converges to something negative. Thus, for all $\underline{\theta} \neq \underline{\theta}_0$,

$$\Pr[L(\underline{\theta}_0|\underline{X}) > L(\underline{\theta}|\underline{X})|\underline{\theta}_0] \rightarrow 1 \quad (2)$$

as $n \rightarrow \infty$; with probability converging to 1, the likelihood at $\underline{\theta}_0$ is greater than the likelihood elsewhere in Θ .

2.2 Consistency and Asymptotic Normality: Univariate Case

In the case $d = 1$, it is now straightforward to show that a solution - not necessarily the maximum likelihood solution - to the equation (LE) is weakly consistent for θ_0 , under additional regularity conditions: provided that the log-likelihood is suitably **differentiable** with respect to θ on Θ .

- **Consistency:** Let $a > 0$ and consider the set

$$B_a \equiv \{x : L(\theta_0 - a|x) < L(\theta_0|x) \text{ and } L(\theta_0 + a|x) < L(\theta_0|x)\} \subset \Theta$$

By equation (2), $\Pr(B_a) \rightarrow 1$ as $n \rightarrow \infty$. Therefore, with probability tending to one,

$$L(\theta_0 - a|x) < L(\theta_0|x) > L(\theta_0 + a|x).$$

As the log-likelihood is differentiable in a neighbourhood of θ_0 , $L(\theta|x)$ has a local maximum, $\tilde{\theta}_n(a)$, in the set $(\theta_0 - a, \theta_0 + a)$, at which $\dot{l}_n(\tilde{\theta}_n(a)) = 0$. Hence, for a arbitrarily small, as $n \rightarrow \infty$,

$$\Pr[|\tilde{\theta}_n(a) - \theta_0| < a|\theta_0] \rightarrow 1$$

so therefore the sequence of estimators $\{\tilde{\theta}_n(a), n \geq 1\}$ converges in probability to θ_0 . To obtain the required result independent of a , let $\tilde{\theta}_n$ be the root of the likelihood equations closest to θ_0 .

- **Asymptotic Normality:** Consider a Taylor expansion of $\dot{l}_n(\theta)$ around θ_0

$$\dot{l}_n(\theta) = \dot{l}_n(\theta_0) + (\theta - \theta_0)\ddot{l}_n(\theta_0) + (\theta - \theta_0)^2 \ddot{l}_n(\theta^*)/2$$

where θ^* lies between θ_0 and θ . Evaluating this at $\theta = \tilde{\theta}_n$, a root of the likelihood equation

$$0 = \dot{l}_n(\tilde{\theta}_n) = \dot{l}_n(\theta_0) + (\tilde{\theta}_n - \theta_0)\ddot{l}_n(\theta_0) + (\tilde{\theta}_n - \theta_0)^2 \ddot{l}_n(\theta_n^*)/2$$

so that on rearrangement

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

Now, in terms of $X_1, \dots, X_n$ as $n \rightarrow \infty$, by the CLT, if $V(\theta_0) = \text{Var}_{f_{X|\theta}}[S(X; \theta_0)] = \mathcal{I}(\theta_0)$,

$$\frac{1}{\sqrt{n}} \dot{l}_n(\theta_0) = \sqrt{n} \frac{1}{n} \left\{ \sum_{i=1}^n \frac{\dot{f}_{X|\theta}(X_i|\theta_0)}{f_{X|\theta}(X_i|\theta_0)} \right\} = \sqrt{n} S(\underline{X}; \theta_0) \xrightarrow{d} Z \sim \text{Normal}(0, V(\theta_0))$$

Similarly, by the Weak Law of Large Numbers, as $n \rightarrow \infty$,

$$-\frac{1}{n} \ddot{l}_n(\theta_0) = \frac{1}{n} \sum_{i=1}^n \Psi(\theta_0; X_i) \xrightarrow{p} \mathcal{I}(\theta_0).$$

Finally, with probability tending to 1,

$$\left| \frac{1}{n} \ddot{l}_n(\theta_n^*) \right| = \left| \frac{1}{n} \sum_{i=1}^n \ddot{l}_n(\theta_n^*; X_i) \right| < \frac{1}{n} \sum_{i=1}^n M(X_i) \xrightarrow{p} \mathbb{E}_{f_{X|\theta}}[M(X)|\theta_0] = O(1).$$

Hence, as $\tilde{\theta}_n \rightarrow \theta_0$, $(\tilde{\theta}_n - \theta_0) \xrightarrow{p} 0$, and, using Slutsky's Theorem

$$\frac{1}{n}(\tilde{\theta}_n - \theta_0) \ddot{l}_n(\theta_n^*) \xrightarrow{p} 0. \quad \therefore \quad \sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{d} \text{Normal}(0, \mathcal{I}(\theta_0)^{-1})$$

2.3 Extension to the Multivariate Case

With extensions to the regularity conditions, we can provide a similar result in the multivariate case.

A3. Θ contains an **open neighbourhood**, $\Theta_0 \subset \mathbb{R}^d$, of $\underline{\theta}_0$ on which

- (i) $l(\underline{\theta}|x)$ is **twice continuously differentiable with respect to $\underline{\theta}$** for all $x \in \mathbb{X}$.
- (ii) Third derivatives of $l(\underline{\theta}|x)$ **exist and are absolutely bounded**, that is

$$\left| \ddot{l}_{jkl}(\underline{\theta}|x) \right| \leq M_{jkl}(x) \quad \underline{\theta} \in \Theta_0$$

for all j, k, l , for some function $M_{jkl}(x)$ where

$$E_{f_{X|\underline{\theta}_0}} [M_{jkl}(X)|\underline{\theta}_0] < m_{jkl} < \infty$$

A4. (i) $E_{f_{X|\underline{\theta}_0}} [\dot{l}_j(\underline{\theta}_0|X)] = 0$ for $j = 1, \dots, d$.

(ii) $E_{f_{X|\underline{\theta}_0}} [(\dot{l}_j(\underline{\theta}_0|X))^2] < \infty$ for $j = 1, \dots, d$.

(iii) The $k \times k$ Fisher information matrix $\mathcal{I}(\underline{\theta}_0)$ is **positive definite**.

Theorem Suppose that conditions A1 to A4 hold. Then, as $n \rightarrow \infty$, with probability converging to 1, there exist solutions $\tilde{\underline{\theta}}_n$ of the likelihood equations (LE) such that $\tilde{\underline{\theta}}_n \xrightarrow{p} \underline{\theta}_0$. In addition

$$\sqrt{n}(\tilde{\underline{\theta}}_n - \underline{\theta}_0) \xrightarrow{d} \text{Normal}(\underline{0}, \mathcal{I}(\underline{\theta}_0)^{-1})$$

Proof: See Appendix A.

2.4 Observed Information

The asymptotic distribution of the maximum likelihood estimator is given by previous results as

$$\sqrt{n}(\tilde{\underline{\theta}}_n - \underline{\theta}_0) \xrightarrow{p} Z \sim \text{Normal}(\underline{0}, \mathcal{I}(\underline{\theta}_0)^{-1})$$

For practical purposes, it is often necessary to produce an estimator of $\mathcal{I}(\underline{\theta}_0)$. The *observed information*, $\hat{I}_n = \hat{I}_n(\underline{X})$, can be used to approximate $\mathcal{I}(\underline{\theta}_0)$, where any of the definitions

$$\hat{I}_n(\tilde{\underline{\theta}}_n) = \begin{cases} \mathcal{I}(\tilde{\underline{\theta}}_n) \\ \frac{1}{n} \sum_{i=1}^n \underline{S}(X_i, \tilde{\underline{\theta}}_n) \underline{S}(X_i, \tilde{\underline{\theta}}_n)^\top \\ -\frac{1}{n} \sum_{i=1}^n \Psi(\tilde{\underline{\theta}}_n, X_i) \end{cases}$$

may be used, where $\tilde{\underline{\theta}}_n$ is a consistent estimator of $\underline{\theta}_0$. If, as in some hypothesis testing examples, $\underline{\theta}_0$ is known (or hypothesized) it replaces $\tilde{\underline{\theta}}_n$ in the calculation.

2.5 Quadratic Approximation

A second-order Taylor *approximation* to the log-likelihood around the MLE, $\hat{\theta}_n$, yields a quadratic approximation to the log-likelihood: As $\dot{l}_n(\hat{\theta}_n) = 0$,

$$l_n(\theta) \simeq l_n(\hat{\theta}_n) + (\theta - \hat{\theta}_n)^\top \dot{l}_n(\hat{\theta}_n) + \frac{1}{2}(\theta - \hat{\theta}_n)^\top \ddot{l}_n(\hat{\theta}_n)(\theta - \hat{\theta}_n) = l_n(\hat{\theta}_n) + \frac{1}{2}(\theta - \hat{\theta}_n)^\top \ddot{l}_n(\hat{\theta}_n)(\theta - \hat{\theta}_n)$$

In the one-dimensional case, we have

$$l_n(\theta) \simeq l_n(\hat{\theta}) + \frac{\ddot{l}_n(\hat{\theta})}{2}(\theta - \hat{\theta})^2$$

Example: Poisson(λ). Let $s_n = \sum_{i=1}^n x_i$. Then

$$l_n(\lambda) = \text{constant} + s_n \log \lambda - n\lambda \quad \dot{l}_n(\lambda) = s_n/\lambda - n \quad \ddot{l}_n(\lambda) = -s_n/\lambda^2 \quad (3)$$

and as the MLE is $\hat{\lambda}_n = \bar{x}$, we have the likelihood approximation

$$l_n(\lambda) = l_n(\hat{\lambda}_n) - \frac{1}{2} \frac{s_n}{\hat{\lambda}_n^2} (\lambda - \hat{\lambda}_n)^2 = l_n(\bar{x}) - \frac{n(\lambda - \bar{x})^2}{2\bar{x}}$$

Example: Normal($0, \sigma^2$) $\equiv$ Normal($0, \theta$). Let $q_n = \sum_{i=1}^n x_i^2$, we have

$$l_n(\theta) = \text{constant} - \frac{n}{2} \log \theta - \frac{q_n}{2\theta} \quad \dot{l}_n(\theta) = -\frac{n}{2\theta} + \frac{q_n}{2\theta^2} \quad \ddot{l}_n(\theta) = \frac{n}{2\theta^2} - \frac{q_n}{\theta^3}$$

The MLE is $\hat{\theta}_n = q_n/n$, and thus

$$l_n(\hat{\theta}_n) = \frac{n}{2\hat{\theta}_n^2} - \frac{q_n}{\hat{\theta}_n^3} = -\frac{n^3}{2q_n^2}$$

so we have the likelihood approximation

$$l_n(\theta) = \dot{l}_n(\hat{\theta}_n) - \frac{1}{2} \frac{n^3}{2q_n^2} (\theta - \hat{\theta}_n)^2 = l_n(q_n/n) - \frac{n^3(\theta - q_n/n)^2}{4q_n^2}.$$

3 Hypothesis Testing

3.1 Asymptotic Behaviour of the Likelihood Ratio Test Statistic under the Null

Recall that the **Likelihood Ratio Test (LRT)** statistic for testing H_0 against H_1

$$\begin{aligned} H_0 &: \theta \in \Theta_0 \\ H_1 &: \theta \in \Theta_1 \end{aligned}$$

is based on the statistic

$$\lambda_{\tilde{X}}(\underline{x}) = \frac{\sup_{\theta \in \Theta_0} f_{\tilde{X}|\theta}(\underline{x}|\theta)}{\sup_{\theta \in \Theta} f_{\tilde{X}|\theta}(\underline{x}|\theta)} = \frac{L(\hat{\theta}_{n0} | \underline{x})}{L(\hat{\theta}_n | \underline{x})}$$

say, where $\Theta \equiv \Theta_0 \cup \Theta_1$; H_0 is **rejected** if $\lambda_{\tilde{X}}(\underline{x})$ is **small enough**, that is, $\lambda_{\tilde{X}}(\underline{x}) \leq k$ for some k to be defined. **Note that Θ is not necessarily the entire parameter space, just the union of Θ_0 and Θ_1 .**

Theorem Consider using the likelihood ratio statistic $\lambda_{\underline{X}}(\underline{X})$ for testing

$$\begin{aligned} H_0 &: \theta = \theta_0 \\ H_1 &: \theta \neq \theta_0 \end{aligned} \quad (4)$$

Then under assumptions A1-A4, the maximum likelihood estimator for θ under H_1 , $\hat{\theta}_n$, is consistent and asymptotically normally distributed, and as $n \rightarrow \infty$ under H_0 ,

$$-2 \log \lambda_{\underline{X}}(\underline{X}) = -2(l_n(\theta_0) - l_n(\hat{\theta}_n)) \xrightarrow{d} Q \sim \chi_1^2$$

Proof A second-order Taylor expansion to the log-likelihood around the MLE, $\hat{\theta}_n$

$$l_n(\theta) = l_n(\hat{\theta}_n) + (\theta - \hat{\theta}_n) \dot{l}_n(\hat{\theta}_n) + \frac{1}{2}(\theta - \hat{\theta}_n)^2 \ddot{l}_n(\hat{\theta}_n) + \frac{1}{6}(\theta - \hat{\theta}_n)^3 \dddot{l}_n(\theta_n^*)$$

where θ_n^* lies between θ and $\hat{\theta}_n$. Now as $\hat{\theta}_n$ is the maximum likelihood estimate $\dot{l}_n(\hat{\theta}_n) = 0$, and therefore on rearrangement, evaluating at $\theta = \theta_0$,

$$-2(l_n(\theta_0) - l_n(\hat{\theta}_n)) = -(\theta_0 - \hat{\theta}_n)^2 \ddot{l}_n(\hat{\theta}_n) - \frac{1}{3}(\theta_0 - \hat{\theta}_n)^3 \dddot{l}_n(\theta_n^*).$$

In this expression, consider the asymptotic behaviour in terms of the random variables. We have

$$-(\theta_0 - \hat{\theta}_n)^2 \ddot{l}_n(\hat{\theta}_n) = (\sqrt{n}(\theta_0 - \hat{\theta}_n))^2 \left(-\frac{1}{n} \ddot{l}_n(\hat{\theta}_n) \right)$$

where

$$\sqrt{n}(\theta_0 - \hat{\theta}_n) \xrightarrow{d} Z \sim \text{Normal}(0, \mathcal{I}(\theta_0)^{-1}) \quad -\frac{1}{n} \ddot{l}_n(\hat{\theta}_n) \xrightarrow{p} \mathcal{I}(\theta_0)$$

so that, by Slutsky's Theorem

$$-(\theta_0 - \hat{\theta}_n)^2 \ddot{l}_n(\hat{\theta}_n) \xrightarrow{d} \chi_1^2$$

Finally,

$$(\theta_0 - \hat{\theta}_n)^3 \dddot{l}_n(\theta_n^*) = (\theta_0 - \hat{\theta}_n) \left((\theta_0 - \hat{\theta}_n)^2 \dddot{l}_n(\theta_n^*) \right) \xrightarrow{p} 0$$

so

$$-2(l_n(\theta_0) - l_n(\hat{\theta}_n)) \stackrel{d}{=} Q + o_P(1) \xrightarrow{d} Q \sim \chi_1^2$$

3.2 Asymptotic Behaviour of the Likelihood Ratio Test Statistic under the Alternative

In the test of the hypotheses in (4) using the likelihood ratio test, suppose that, in fact $\theta = \theta_T$. Then, under conditions A1-A4,

$$-\frac{2}{n}(l_n(\theta_0) - l_n(\hat{\theta}_n)) = \frac{2}{n}(l_n(\theta_T) - l_n(\theta_0)) - \frac{2}{n}(l_n(\theta_T) - l_n(\hat{\theta}_n)) = \frac{2}{n} \sum_{i=1}^n \log \frac{f_{X|\theta}(X_i|\theta_T)}{f_{X|\theta}(X_i|\theta_0)} - \frac{2}{n}(l_n(\theta_T) - l_n(\hat{\theta}_n))$$

But, as $n \rightarrow \infty$

$$\frac{2}{n} \sum_{i=1}^n \log \frac{f_{X|\theta}(X_i|\theta_T)}{f_{X|\theta}(X_i|\theta_0)} \xrightarrow{p} 2K(\theta_T, \theta_0) \quad \frac{2}{n}(l_n(\theta_T) - l_n(\hat{\theta}_n)) \xrightarrow{p} 0$$

as $\hat{\theta}_n \xrightarrow{p} \theta_T$, so

$$-\frac{2}{n}(l_n(\theta_0) - l_n(\hat{\theta}_n)) \xrightarrow{p} 2K(\theta_T, \theta_0)$$

d -dimensional Parameters

Consider using the likelihood ratio statistic for testing H_0 versus H_1 if $\underline{\theta}$ is d -dimensional parameter. Under assumptions A1-A4, as $n \rightarrow \infty$ under H_0 ,

$$-2 \log \lambda_{\underline{X}}(\underline{X}) = -2(l_n(\underline{\theta}_0) - l_n(\hat{\underline{\theta}}_n)) \xrightarrow{d} Q \sim \chi_d^2$$

This follows using identical methods to the $d = 1$ case. A second-order Taylor expansion to the log-likelihood around the MLE, $\hat{\underline{\theta}}_n$, yields

$$l_n(\underline{\theta}) = l_n(\hat{\underline{\theta}}_n) + (\underline{\theta} - \hat{\underline{\theta}}_n)^\top \dot{l}_n(\hat{\underline{\theta}}_n) + \frac{1}{2}(\underline{\theta} - \hat{\underline{\theta}}_n)^\top \ddot{l}_n(\hat{\underline{\theta}}_n)(\underline{\theta} - \hat{\underline{\theta}}_n) + o_P(1)$$

As $\hat{\underline{\theta}}_n$ is the maximum likelihood estimate

$$\dot{l}_n(\hat{\underline{\theta}}_n) = 0$$

and therefore on rearrangement, evaluating at $\underline{\theta} = \underline{\theta}_0$,

$$-2(l_n(\underline{\theta}_0) - l_n(\hat{\underline{\theta}}_n)) = -(\underline{\theta}_0 - \hat{\underline{\theta}}_n)^\top \ddot{l}_n(\hat{\underline{\theta}}_n)(\underline{\theta}_0 - \hat{\underline{\theta}}_n) + o_P(1)$$

But, by previous results

$$(\underline{\theta}_0 - \hat{\underline{\theta}}_n)^\top \ddot{l}_n(\hat{\underline{\theta}}_n)(\underline{\theta}_0 - \hat{\underline{\theta}}_n) = \sqrt{n}(\underline{\theta}_0 - \hat{\underline{\theta}}_n)^\top \left(-\frac{1}{n} \ddot{l}_n(\hat{\underline{\theta}}_n) \right) \sqrt{n}(\underline{\theta}_0 - \hat{\underline{\theta}}_n) \xrightarrow{d} \underline{Z}^\top \mathcal{I}(\underline{\theta}_0) \underline{Z}$$

as

$$\sqrt{n}(\hat{\underline{\theta}}_n - \underline{\theta}_0) \xrightarrow{d} \underline{Z} \sim \text{Normal}(\underline{0}, \mathcal{I}(\underline{\theta}_0)^{-1}) \quad -\frac{1}{n} \ddot{l}_n(\hat{\underline{\theta}}_n) \xrightarrow{p} \mathcal{I}(\underline{\theta}_0)$$

Hence

$$-2(l_n(\underline{\theta}_0) - l_n(\hat{\underline{\theta}}_n)) \xrightarrow{d} Q \sim \chi_d^2.$$

Theorem Consider testing the hypothesis

$$\begin{aligned} H_0 &: \underline{\theta} \in \Theta_0 \\ H_1 &: \underline{\theta} \in \Theta_1 \end{aligned}$$

using the likelihood ratio statistic

$$\lambda_{\underline{X}}(\underline{x}) = \frac{\sup_{\underline{\theta} \in \Theta_0} f_{\underline{X}|\underline{\theta}}(\underline{x}|\underline{\theta})}{\sup_{\underline{\theta} \in \Theta} f_{\underline{X}|\underline{\theta}}(\underline{x}|\underline{\theta})} = \frac{L(\hat{\underline{\theta}}_{n0} | \underline{x})}{L(\hat{\underline{\theta}}_n | \underline{x})}$$

say, where $\Theta \equiv \Theta_0 \cup \Theta_1$, and where H_0 specifies a model with k_1 free parameters (parameters not determined by the hypothesis), and H_1 specifies a model with k_2 free parameters, with $k_2 > k_1$. Then, under assumptions A1-A4, as $n \rightarrow \infty$, under H_0

$$-2 \log \lambda_{\underline{X}}(\underline{X}) = -2(l_n(\hat{\underline{\theta}}_{n0}) - l_n(\hat{\underline{\theta}}_n)) \xrightarrow{d} Q \sim \chi_{k_2 - k_1}^2$$

Note: Such hypotheses can often be specified in the form

$$\begin{aligned} H_0 &: \underline{\theta} = (\underline{\theta}_0, \underline{\theta}_1), \quad \underline{\theta}_1 \text{ unspecified} \\ H_1 &: \underline{\theta} \neq (\underline{\theta}_0, \underline{\theta}_1), \quad \underline{\theta}_1 \text{ unspecified} \end{aligned}$$

that is, H_0 places constraints on one component $\underline{\theta}$, but leaves the other unspecified.

3.3 Other Asymptotic Tests: The Wald and Score Statistics

The **Wald** and **Rao/Score** test statistics derived from a sample of size n , W_n and R_n , for testing (4) in the case of a vector-valued parameter are constructed as follows:

- **Wald Test** : The **Wald Statistic**, W_n , is defined by

$$W_n = n(\tilde{\theta}_n - \theta_0)^\top \hat{I}_n(\tilde{\theta}_n)(\tilde{\theta}_n - \theta_0) \quad (5)$$

where $\tilde{\theta}_n$ is a solution to the likelihood equations, $\hat{I}_n$ is the observed information.

- **Score Test** : Let

$$\underline{Z}_n \equiv \underline{Z}_n(\theta_0) = \frac{1}{\sqrt{n}} \dot{l}_n(\theta_0).$$

Then the (Rao) Score Test Statistic, R_n , is defined by

$$R_n = \underline{Z}_n^\top \mathcal{I}(\theta_0)^{-1} \underline{Z}_n \quad (6)$$

where $\mathcal{I}(\theta_0)$ can be replaced by the observed information $\hat{I}_n(\theta_0)$ if necessary.

In the one parameter case, the statistics can be expressed as

$$W_n = -(\tilde{\theta}_n - \theta_0)^2 \ddot{l}_n(\tilde{\theta}_n) \quad R_n = -\left\{ \dot{l}_n(\theta_0) \right\}^2 \left\{ \ddot{l}_n(\theta_0) \right\}^{-1}$$

3.4 Asymptotic Behaviour of the Wald and Score Statistics

(a) Under the **null hypothesis**

- **Wald Test** : For the Wald test, as

$$\underline{D}_n = \sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{d} \underline{Z} \sim \text{Normal}(\underline{0}, \mathcal{I}(\theta_0)^{-1})$$

it follows that

$$W_n = n(\tilde{\theta}_n - \theta_0)^\top \hat{I}_n(\tilde{\theta}_n)(\tilde{\theta}_n - \theta_0) = \underline{D}_n^\top \hat{I}_n(\tilde{\theta}_n) \underline{D}_n \xrightarrow{d} \underline{Z}^\top \mathcal{I}(\theta_0) \underline{Z} \sim \chi_d^2$$

- **Score Test** : For the Score test,

$$\underline{Z}_n = \frac{1}{\sqrt{n}} \dot{l}_n(\theta_0) = \sqrt{n} \left(\frac{1}{n} \dot{l}_n(\theta_0) \right) \xrightarrow{d} \underline{Z} \sim \text{Normal}(\underline{0}, \mathcal{I}(\theta_0))$$

and hence for the Score Test Statistic,

$$R_n = \underline{Z}_n^\top \mathcal{I}(\theta_0)^{-1} \underline{Z}_n \xrightarrow{d} \underline{Z}^\top \mathcal{I}(\theta_0) \underline{Z} \sim \chi_d^2$$

(b) If the null hypothesis is **not true**, let θ_T denote the true value of the parameter.

- **Wald Test** : For the Wald test,

$$\frac{1}{n} W_n = (\tilde{\theta}_n - \theta_0)^\top \hat{I}_n(\tilde{\theta}_n)(\tilde{\theta}_n - \theta_0) \xrightarrow{p} (\theta_T - \theta_0)^\top \mathcal{I}(\theta_T)(\theta_T - \theta_0) > 0$$

- **Score Test** : For the Score test, from above

$$\frac{1}{\sqrt{n}} \underline{Z}_n = \frac{1}{n} \dot{l}_n(\theta_0) \xrightarrow{p} E_{f_{X|\theta}}[\dot{l}_n(\theta_0)] = \underline{\mu}(\theta_T, \theta_0)$$

say, and hence for the Score Test Statistic,

$$\frac{1}{n} R_n = \frac{1}{n} \underline{Z}_n^\top \mathcal{I}(\theta_0)^{-1} \underline{Z}_n \xrightarrow{p} \underline{\mu}(\theta_T, \theta_0)^\top \mathcal{I}(\theta_0)^{-1} \underline{\mu}(\theta_T, \theta_0) > 0$$

3.5 Composite Hypotheses

Consider testing the hypothesis

$$\begin{aligned} H_0 &: \underline{\theta} = (\underline{\theta}_0, \underline{\theta}_1), & \underline{\theta}_1 \text{ unspecified} \\ H_1 &: \underline{\theta} \neq (\underline{\theta}_0, \underline{\theta}_1), & \underline{\theta}_1 \text{ unspecified} \end{aligned}$$

where $\underline{\theta}_0$ is $k_1 \times 1$, and $\underline{\theta}_1$ is $(d - k_1) \times 1$, that is, H_0 places constraints on one component $\underline{\theta}$, but leaves the other unspecified.

Let $\tilde{\underline{\theta}}_{n0} = (\underline{\theta}_0, \tilde{\underline{\theta}}_{n01})^\top$ and $\tilde{\underline{\theta}}_{n1} = (\tilde{\underline{\theta}}_{n10}, \tilde{\underline{\theta}}_{n11})^\top$ denote consistent estimators (possibly MLEs) under H_0 and H_1 respectively. Suppose that

$$\mathcal{I}(\underline{\theta}) = \begin{bmatrix} \mathcal{I}_{00}(\underline{\theta}) & \mathcal{I}_{01}(\underline{\theta}) \\ \mathcal{I}_{10}(\underline{\theta}) & \mathcal{I}_{11}(\underline{\theta}) \end{bmatrix}$$

denotes the Fisher information for $\underline{\theta}$, with blocks $\mathcal{I}_{00}(\underline{\theta})$ ($k_1 \times k_1$), $\mathcal{I}_{01}(\underline{\theta})$ ($k_1 \times (d - k_1)$) etc. Let

$$\mathcal{I}(\underline{\theta})^{-1} = \begin{bmatrix} \tilde{\mathcal{I}}_{00}(\underline{\theta}) & \tilde{\mathcal{I}}_{01}(\underline{\theta}) \\ \tilde{\mathcal{I}}_{10}(\underline{\theta}) & \tilde{\mathcal{I}}_{11}(\underline{\theta}) \end{bmatrix}$$

Then for the hypotheses above the Wald and Score tests are constructed as follows:

- **Wald Test :** The **Wald Statistic**, W_n , is defined by

$$W_n = n(\tilde{\underline{\theta}}_{n10} - \underline{\theta}_0)^\top \hat{I}_{n00.1}(\tilde{\underline{\theta}}_{n0})(\tilde{\underline{\theta}}_{n10} - \underline{\theta}_0) \quad (7)$$

where $\tilde{\underline{\theta}}_n$ is a solution to the likelihood equations, $\hat{I}_n$ is the observed information, and $\hat{I}_{n00.1}$ is the upper $k_1 \times k_1$ block of the inverse if $\hat{I}_n$. It can be shown that if

$$\hat{I}_n(\underline{\theta}) = \begin{bmatrix} \hat{I}_{n00}(\underline{\theta}) & \hat{I}_{n01}(\underline{\theta}) \\ \hat{I}_{n10}(\underline{\theta}) & \hat{I}_{n11}(\underline{\theta}) \end{bmatrix}$$

then

$$\hat{I}_{n00.1}(\tilde{\underline{\theta}}_{n0}) = \left(\hat{I}_{n00}(\tilde{\underline{\theta}}_{n0}) - \hat{I}_{n01}(\tilde{\underline{\theta}}_{n0}) \hat{I}_{n11}^{-1}(\tilde{\underline{\theta}}_{n0}) \hat{I}_{n10}(\tilde{\underline{\theta}}_{n0}) \right)^{-1}$$

- **Score Test :** Let

$$\underline{Z}_n \equiv \underline{Z}_n(\underline{\theta}_0) = \frac{1}{\sqrt{n}} \dot{l}_n(\underline{\theta}_0).$$

Then the (Rao) Score Test Statistic, R_n , is defined by

$$R_n = \underline{Z}_n^\top(\tilde{\underline{\theta}}_{n0}) \mathcal{I}(\tilde{\underline{\theta}}_{n0})^{-1} \underline{Z}_n(\tilde{\underline{\theta}}_{n0}) \quad (8)$$

where $\mathcal{I}(\underline{\theta}_0)$ can be replaced by the observed information $\hat{I}_n(\underline{\theta}_0)$ if necessary.

In both cases, under H_0 , the statistics converge in distribution to a Chi-squared distribution,

$$W_n \xrightarrow{d} \chi_{k_1}^2 \quad R_n \xrightarrow{d} \chi_{k_1}^2$$

APPENDICES

NOT EXAMINABLE

A Proof of Consistency and Asymptotic Normality of the ML Estimator

Proof Let $a > 0$, and define Q_a such that $Q_a = \{\theta \in \Theta : \|\theta - \theta_0\| = a\}$. Consider a third order Taylor expansion of $l_n(\theta)$ of around θ_0 . Rearranging, and dividing by n , we have

$$\begin{aligned} \frac{1}{n}(l_n(\theta) - l_n(\theta_0)) &= \frac{1}{n} \sum_{j=1}^k A_j(\mathcal{X})(\theta_j - \theta_{0j}) + \frac{1}{2n} \sum_{j=1}^d \sum_{k=1}^d B_{jk}(\mathcal{X})(\theta_j - \theta_{0j})(\theta_k - \theta_{0k}) \\ &+ \frac{1}{6n} \sum_{j=1}^d \sum_{k=1}^d \sum_{l=1}^d (\theta_j - \theta_{0j})(\theta_k - \theta_{0k})(\theta_l - \theta_{0l}) \left\{ \sum_{i=1}^n \gamma_{jkl}(x_i) M_{jkl}(x_i) \right\} \quad (9) \\ &= s_1 + s_2 + s_3 \end{aligned}$$

say, where $0 \leq |\gamma_{jkl}(x)| \leq 1$, and, for $j, k = 1, \dots, d$,

$$A_j(\mathcal{X}) = \dot{l}_j(\theta_0|\mathcal{X}) \quad B_{jk}(\mathcal{X}) = \ddot{l}_{jk}(\theta_0|\mathcal{X})$$

Let S_1, S_2 and S_3 be the random variables corresponding to the quantities s_1, s_2 and s_3 . We aim to show that the supremum of $(l_n(\theta) - l_n(\theta_0))/n$ on Q_a is negative with probability tending to 1 if a is sufficiently small; to do this, we show that the supremum of S_2 is negative, while S_1 and S_3 are negligible compared to S_2 . Now, by the WLLN and assumption A3(i),

$$\frac{1}{n} A_j(\mathcal{X}) = \frac{1}{n} \dot{l}_j(\theta_0|\mathcal{X}) \xrightarrow{p} E_{f_{\mathcal{X}|\theta_0}} [\dot{l}_j(\theta_0|X)] = 0 \quad (10)$$

and by the WLLN

$$\frac{1}{n} B_{jk}(\mathcal{X}) = \frac{1}{n} \ddot{l}_{jk}(\theta_0|\mathcal{X}) \xrightarrow{p} E_{f_{\mathcal{X}|\theta_0}} [\ddot{l}_{jk}(\theta_0|X)] = -\mathcal{I}_{jk}(\theta_0) \quad (11)$$

On Q_a , we have

$$|S_1| \leq \frac{1}{n} a \sum_{j=1}^d |A_j(\mathcal{X})|$$

so that for any a , as $n \rightarrow \infty$, from equation (10), with probability tending to 1,

$$\frac{1}{n} |A_j(\mathcal{X})| < a^2 \quad \therefore \quad |S_1| < sa^3$$

Secondly,

$$\begin{aligned} 2S_2 &= \frac{1}{n} \sum_{j=1}^d \sum_{k=1}^d B_{jk}(\mathcal{X})(\theta_j - \theta_{0j})(\theta_k - \theta_{0k}) \\ &= \sum_{j=1}^d \sum_{k=1}^d \left(\frac{1}{n} B_{jk}(\mathcal{X}) - (-\mathcal{I}_{jk}(\theta_0)) \right) (\theta_j - \theta_{0j})(\theta_k - \theta_{0k}) + \sum_{j=1}^d \sum_{k=1}^d (-\mathcal{I}_{jk}(\theta_0)) (\theta_j - \theta_{0j})(\theta_k - \theta_{0k}) \end{aligned}$$

As before, as $n \rightarrow \infty$, from equation (11), with probability tending to 1,

$$\left| \sum_{j=1}^d \sum_{k=1}^d \left(\frac{1}{n} B_{jk}(\mathcal{X}) - (-\mathcal{I}_{jk}(\theta_0)) \right) (\theta_j - \theta_{0j})(\theta_k - \theta_{0k}) \right| < s^2 a^3 \quad (12)$$

whereas the second term is the constant quadratic form

$$\sum_{j=1}^d \sum_{k=1}^d (-\mathcal{I}_{jk}(\underline{\theta}_0))(\theta_j - \theta_{0j})(\theta_k - \theta_{0k}) = -(\underline{\theta} - \underline{\theta}_0)^\top \mathcal{I}(\underline{\theta}_0)(\underline{\theta} - \underline{\theta}_0).$$

Now, as $\mathcal{I}(\underline{\theta}_0)$ is positive definite, it has a singular value decomposition $\mathcal{I}(\underline{\theta}_0) = V^\top D V$, where D is the diagonal eigenvalue matrix $D = \text{diag}(\lambda_1, \dots, \lambda_d)$, with $0 < \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_d$, and V is the matrix of eigenvectors, with $V^\top V = I_d$. Thus

$$-(\underline{\theta} - \underline{\theta}_0)^\top \mathcal{I}(\underline{\theta}_0)(\underline{\theta} - \underline{\theta}_0) = -\sum_{j=1}^d \lambda_j \xi_j(\underline{\theta}_0, \underline{\theta})^2$$

where $\xi(\underline{\theta}_0, \underline{\theta}) = V(\underline{\theta} - \underline{\theta}_0)$, so that

$$\sum_{j=1}^d \xi_j(\underline{\theta}_0, \underline{\theta})^2 = \xi(\underline{\theta}_0, \underline{\theta})^\top \xi(\underline{\theta}_0, \underline{\theta}) = (\underline{\theta} - \underline{\theta}_0)^\top V^\top V (\underline{\theta} - \underline{\theta}_0) = (\underline{\theta} - \underline{\theta}_0)^\top (\underline{\theta} - \underline{\theta}_0) = \sum_{j=1}^d (\theta_j - \theta_{0j})^2$$

Now, on the surface of the hypersphere Q_a , $\|(\underline{\theta} - \underline{\theta}_0)\| = a$ so

$$\sum_{j=1}^d \xi_j(\underline{\theta}_0, \underline{\theta})^2 = a^2 \geq \lambda_1 \sum_{j=1}^d \xi_j(\underline{\theta}_0, \underline{\theta})^2 \geq \lambda_1 a^2 \quad \therefore \quad -(\underline{\theta} - \underline{\theta}_0)^\top \mathcal{I}(\underline{\theta}_0)(\underline{\theta} - \underline{\theta}_0) \leq -\lambda_1^2 a^2 \quad (13)$$

Hence, combining equations (12) and (13), with probability tending to 1, for a small enough, $S_2 < -ca^2$. Finally, for S_3 , with probability tending to 1,

$$\left| \frac{1}{n} \sum_{i=1}^n M_{jkl} \right| < 2m_{jkl} \quad \therefore \quad |S_3| < \frac{1}{6} s^3 a^3 \sum_{j=1}^d \sum_{k=1}^d \sum_{l=1}^d m_{jkl} = ba^3$$

say. Thus, combining results we have

$$\sup_{\underline{\theta} \in Q_a} (S_1 + S_2 + S_3) \leq \sup_{\underline{\theta} \in Q_a} S_2 + \sup_{\underline{\theta} \in Q_a} \|S_1 + S_3\| < -ca^2 + (b + s)a^2$$

which is **negative** if $a < c/(b + s)$. Thus, l has a local maximum inside Q_a , as for n large enough, with probability at least $1 - \epsilon$ that is, as $(l_n(\underline{\theta}) - l_n(\underline{\theta}_0))/n < 0$, or equivalently,

$$\Pr [l_n(\underline{\theta}) < l_n(\underline{\theta}_0) \text{ for all } \underline{\theta} \in Q_a \mid \underline{\theta}_0] \longrightarrow 1 \quad \text{as} \quad n \longrightarrow \infty.$$

Therefore, as the likelihood equations (LE) are satisfied at local maxima, it follows that (with probability converging to 1 as $n \longrightarrow \infty$) there **exists** a solution, $\tilde{\underline{\theta}}_n(a)$, inside Q_a , for any a small enough. Thus the result follows as

$$\lim_{n \rightarrow \infty} \Pr [\|\tilde{\underline{\theta}}_n(a) - \underline{\theta}_0\| < a] = 1 \quad \therefore \quad \tilde{\underline{\theta}}_n(a) \xrightarrow{p} \underline{\theta}_0.$$

The proof of asymptotic normality proceeds in a similar fashion to the univariate case; by multivariate Taylor's Theorem in the $d \times 1$ system of equations

$$\frac{1}{\sqrt{n}} \dot{l}_n(\underline{\theta}_0) = -\frac{1}{\sqrt{n}} \ddot{l}_n(\underline{\theta}_0)(\tilde{\underline{\theta}}_n - \underline{\theta}_0) - \frac{1}{2\sqrt{n}} (\tilde{\underline{\theta}}_n - \underline{\theta}_0)^\top \ddot{l}_n(\underline{\theta}_n^*)(\tilde{\underline{\theta}}_n - \underline{\theta}_0)$$

The left hand side converges in probability (and in distribution) to $Z \sim \text{Normal}(0, \mathcal{I}(\theta_0))$, and for the right hand side,

$$-\frac{1}{n} \left[\ddot{l}_n(\theta_0) + \frac{1}{2}(\tilde{\theta}_n - \theta_0)^\top \ddot{l}_n(\theta_n^*) \right] \xrightarrow{p} \mathcal{I}(\theta_0)$$

by analogy with the univariate case. Hence by Slutsky's Theorem, for large n ,

$$-\frac{1}{\sqrt{n}} \left[\ddot{l}_n(\theta_0) + \frac{1}{2}(\tilde{\theta}_n - \theta_0)^\top \ddot{l}_n(\theta_n^*) \right] (\tilde{\theta}_n - \theta_0) = \mathcal{I}(\theta_0) \sqrt{n}(\tilde{\theta}_n - \theta_0) + o_P(1)$$

where $o_P(1)$ represents a term that converges in probability to zero. Hence

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{p} Z \sim \text{Normal}(0, \mathcal{I}(\theta_0)^{-1})$$

and the result is proved. ■

B Improving Inefficient Estimators: The One-Step Estimator

Objective : to produce a consistent estimator with asymptotic variance equal to the inverse Fisher information $\mathcal{I}(\theta_0)^{-1}$, as this is the best possible variance we can achieve for consistent estimators.

Suppose that $\hat{\theta}^{(0)}$ is a (consistent) estimator of θ with asymptotic variance $\Sigma^{(0)}$ where $\Sigma^{(0)} - \mathcal{I}(\theta_0)^{-1}$ is positive definite, so that

$$\Sigma^{(0)} \geq \mathcal{I}(\theta_0)^{-1} \quad \text{or} \quad x^\top (\Sigma^{(0)} - \mathcal{I}(\theta_0)^{-1}) x > 0 \quad \forall x \in \mathbb{R}^d$$

so that $\hat{\theta}^{(0)}$ is inefficient. This estimator can be improved by two iterative procedures that each define a sequence of estimators that are consistent, and will have increasingly better properties.

- **Newton's Method** For $k = 0, 1, \dots$, define

$$\hat{\theta}^{(k+1)} = \hat{\theta}^{(k)} - \left(\ddot{l}_n(\hat{\theta}^{(k)}) \right)^{-1} \dot{l}_n(\hat{\theta}^{(k)})$$

- **Method of Scoring** For $k = 0, 1, \dots$, define

$$\begin{aligned} \text{(i)} \quad \hat{\theta}^{(k+1)} &= \hat{\theta}^{(k)} + \left(\mathcal{I}(\hat{\theta}^{(k)}) \right)^{-1} \left(\frac{1}{n} \dot{l}_n(\hat{\theta}^{(k)}) \right) \\ \text{(ii)} \quad \hat{\theta}^{(k+1)} &= \hat{\theta}^{(k)} + \left(\hat{I}_n(\hat{\theta}^{(k)}) \right)^{-1} \left(\frac{1}{n} \dot{l}_n(\hat{\theta}^{(k)}) \right) \end{aligned}$$

if $\mathcal{I}(\cdot)$ is not known precisely. Note that

$$-\frac{1}{n} \ddot{l}_n(\hat{\theta}^{(k)}) \xrightarrow{p} \mathcal{I}(\theta)$$

The following theorem proves that only **one** iterative step is required to match the asymptotic efficiency of solutions to the likelihood equations, which have been shown to have asymptotic variance equal to the Cramér-Rao information bound. The method of proof is as follows

- 1 Find a consistent but possibly inefficient estimator $\tilde{\theta}_n$

2 Form the one-step Newton or Scoring Estimator using the formulae

$$\hat{\theta}^\dagger = \tilde{\theta}_n - \left(\ddot{l}_n(\tilde{\theta}_n) \right)^{-1} \dot{l}_n(\tilde{\theta}_n) \quad \hat{\theta}^\star = \tilde{\theta}_n + \left(\mathcal{I}(\tilde{\theta}_n) \right)^{-1} \frac{1}{n} \dot{l}_n(\tilde{\theta}_n)$$

3 Show that these estimators have the same asymptotic properties as solutions to the likelihood equations. That is, under the regularity conditions, if $\hat{\theta}_n$ satisfies

$$\dot{l}_n(\hat{\theta}_n) = 0 \quad (\text{LE})$$

then,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \mathcal{I}(\theta_0)^{-1})$$

and the following theorem shows that $\hat{\theta}^\dagger$ and $\hat{\theta}^\star$ also have these properties.

B.1 The Efficiency of One-Step Estimators

Theorem

Let $\tilde{\theta}_n, n = 1, 2, \dots$, be a consistent sequence of estimators of $\theta_0 \in \Theta$ such that

$$\sqrt{n}(\tilde{\theta}_n - \theta_0) \xrightarrow{d} N(0, \Sigma(\theta_0))$$

with $\Sigma(\theta_0)$ finite. Then, under the conditions A1-A4, the two estimators

$$\hat{\theta}^\dagger = \tilde{\theta}_n - \left(\ddot{l}_n(\tilde{\theta}_n) \right)^{-1} \dot{l}_n(\tilde{\theta}_n) \quad (14)$$

$$\hat{\theta}^\star = \tilde{\theta}_n + \left(\mathcal{I}(\tilde{\theta}_n) \right)^{-1} \frac{1}{n} \dot{l}_n(\tilde{\theta}_n) \quad (15)$$

are **asymptotically equivalent** to the MLE,

$$\sqrt{n}(\hat{\theta}^\dagger - \hat{\theta}_n) \xrightarrow{p} 0 \quad \sqrt{n}(\hat{\theta}^\star - \hat{\theta}_n) \xrightarrow{p} 0$$

and

$$\sqrt{n}(\hat{\theta}^\dagger - \theta_0) \xrightarrow{d} N(0, \mathcal{I}(\theta_0)^{-1}) \quad \sqrt{n}(\hat{\theta}^\star - \theta_0) \xrightarrow{d} N(0, \mathcal{I}(\theta_0)^{-1})$$

Proof See below.

Example: Cauchy(θ). We have for the derivatives of the likelihood

$$l_n(\theta) = -n \log \pi - \sum_{i=1}^n \log(1 + (x_i - \theta)^2) \quad \dot{l}_n(\theta) = 2 \sum_{i=1}^n \frac{(x_i - \theta)}{1 + (x_i - \theta)^2} \quad \ddot{l}_n(\theta) = 2 \sum_{i=1}^n \frac{(x_i - \theta)^2 - 1}{(1 + (x_i - \theta)^2)^2}$$

The MLE $\hat{\theta}_n$ can only be computed numerically. An alternative estimator is the sample median $\tilde{\theta}_n = \tilde{x}$, say, which is consistent for θ . Then the Newton one-step estimator is

$$\hat{\theta}^\dagger = \tilde{\theta}_n - \frac{\dot{l}_n(\tilde{\theta}_n)}{\ddot{l}_n(\tilde{\theta}_n)} = \tilde{x} - \frac{\sum_{i=1}^n \frac{(x_i - \tilde{x})}{1 + (x_i - \tilde{x})^2}}{\sum_{i=1}^n \frac{(x_i - \tilde{x})^2 - 1}{(1 + (x_i - \tilde{x})^2)^2}}.$$

n	Measure	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\hat{\theta}^\dagger$	n	Measure	$\hat{\theta}_n$	$\tilde{\theta}_n$	$\hat{\theta}^\dagger$
20	Variance	2.313	2.765	6.191	100	Variance	2.035	2.503	2.036
	Efficiency	1.000	0.837	0.374		Efficiency	1.000	0.813	1.000
40	Variance	2.164	2.647	2.169	10000	Variance	2.010	2.480	2.010
	Efficiency	1.000	0.818	0.998		Efficiency	1.000	0.811	1.000

The simulation below shows the variances estimated as the variance of $T_n = \sqrt{n}(\theta_n - \theta)$ for $\hat{\theta}_n$, $\tilde{\theta}_n$ and $\hat{\theta}^\dagger$, for 50000 replicate data sets of size $n = 20, 40, 100$ and 10000 . In this case

$$\mathcal{I}(\theta) = \int_{-\infty}^{\infty} (S(x, \theta))^2 f_{X|\theta}(x|\theta) dx = \int_{-\infty}^{\infty} \frac{4(x - \theta)^2}{\pi(1 + (x - \theta)^2)^3} dx = \int_{-\infty}^{\infty} \frac{4x^2}{\pi(1 + x^2)^3} dx = \frac{1}{2}$$

so the asymptotic variance of the efficient estimator is $\mathcal{I}(\theta)^{-1} = 2$.

B.2 The Efficiency of One-Step Estimators

Proof Suppose that $\hat{\theta}_n$ is a consistent sequence of estimators that satisfy (LE), with

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \mathcal{I}(\theta_0)^{-1})$$

Note: At no stage in the estimation will we actually have to find the numerical value of $\hat{\theta}_n$.

A first-order Taylor expansion of $\dot{l}_n$ about $\hat{\theta}_n$ yields the following equation:

$$\dot{l}_n(\tilde{\theta}_n) = \dot{l}_n(\hat{\theta}_n) + \ddot{l}_n(\theta_n^\dagger)(\tilde{\theta}_n - \hat{\theta}_n) = \ddot{l}_n(\theta_n^\dagger)(\tilde{\theta}_n - \hat{\theta}_n). \quad (16)$$

as $\hat{\theta}_n$ is a solution to equation (LE), for some $\theta_n^\dagger$ where

$$\|\theta_n^\dagger - \hat{\theta}_n\| \leq \|\tilde{\theta}_n - \hat{\theta}_n\|$$

In equation (16), the left hand side is a $d \times 1$ vector, and $\ddot{l}_n(\theta_n^\dagger)$ is a $d \times d$ matrix. By definition,

$$(\hat{\theta}^\dagger - \hat{\theta}_n) = (\tilde{\theta}_n - \hat{\theta}_n) - \left(\ddot{l}_n(\tilde{\theta}_n) \right)^{-1} \dot{l}_n(\tilde{\theta}_n)$$

so that, by equation (16),

$$\begin{aligned} \sqrt{n}(\hat{\theta}^\dagger - \hat{\theta}_n) &= \sqrt{n} \left[(\tilde{\theta}_n - \hat{\theta}_n) - \left(\ddot{l}_n(\tilde{\theta}_n) \right)^{-1} \dot{l}_n(\tilde{\theta}_n) \right] = \sqrt{n} \left[(\tilde{\theta}_n - \hat{\theta}_n) - \left(\ddot{l}_n(\tilde{\theta}_n) \right)^{-1} \ddot{l}_n(\theta_n^\dagger)(\tilde{\theta}_n - \hat{\theta}_n) \right] \\ &= \left[\mathbf{1}_d - \left(\ddot{l}_n(\tilde{\theta}_n) \right)^{-1} \ddot{l}_n(\theta_n^\dagger) \right] \sqrt{n}(\tilde{\theta}_n - \hat{\theta}_n) \end{aligned} \quad (17)$$

where $\mathbf{1}_d$ is the $d \times d$ identity matrix. Recall that both $\hat{\theta}_n$ and $\tilde{\theta}_n$ are consistent by assumption

$$\hat{\theta}_n \xrightarrow{p} \theta_0 \quad \tilde{\theta}_n \xrightarrow{p} \theta_0 \quad \therefore \quad \tilde{\theta}_n - \hat{\theta}_n \xrightarrow{p} 0.$$

Furthermore, by the Weak Law of Large Numbers, and the continuous mapping theorem

$$\frac{1}{n} \ddot{l}_n(\tilde{\theta}_n) \xrightarrow{p} -\mathcal{I}(\theta_0)$$

and

$$\widehat{\underline{\theta}}_n \xrightarrow{p} \underline{\theta}_0 \quad \text{and} \quad \widetilde{\underline{\theta}}_n - \widehat{\underline{\theta}}_n \xrightarrow{p} \underline{0} \quad \therefore \quad \underline{\theta}_n^\dagger \xrightarrow{p} \underline{\theta}_0$$

so therefore

$$\frac{1}{n} \ddot{l}_n(\underline{\theta}_n^\dagger) \xrightarrow{p} -\mathcal{I}(\underline{\theta}_0).$$

Therefore, in equation (17)

$$\left(\ddot{l}_n(\widetilde{\underline{\theta}}_n) \right)^{-1} \ddot{l}_n(\underline{\theta}_n^\dagger) \xrightarrow{p} \mathcal{I}(\underline{\theta}_0)^{-1} \mathcal{I}(\underline{\theta}_0) = \mathbf{1}_d$$

and so

$$\mathbf{1}_d - \left(\ddot{l}_n(\widetilde{\underline{\theta}}_n) \right)^{-1} \ddot{l}_n(\underline{\theta}_n^\dagger) \xrightarrow{p} \mathbf{1}_d - \mathbf{1}_d = \mathbf{0}_d \quad (18)$$

Also in equation (17),

$$\sqrt{n}(\widetilde{\underline{\theta}}_n - \widehat{\underline{\theta}}_n) = \sqrt{n}(\widetilde{\underline{\theta}}_n - \underline{\theta}_0) - \sqrt{n}(\widehat{\underline{\theta}}_n - \underline{\theta}_0)$$

and, by assumption

$$\left. \begin{array}{l} \sqrt{n}(\widetilde{\underline{\theta}}_n - \underline{\theta}_0) \xrightarrow{d} N(0, \Sigma(\underline{\theta}_0)) \\ \sqrt{n}(\widehat{\underline{\theta}}_n - \underline{\theta}_0) \xrightarrow{d} N(0, \mathcal{I}(\underline{\theta}_0)^{-1}) \end{array} \right\} \implies \sqrt{n}(\widetilde{\underline{\theta}}_n - \widehat{\underline{\theta}}_n) \xrightarrow{d} \underline{Z}_0 \sim N(0, \Sigma(\underline{\theta}_0) + \mathcal{I}(\underline{\theta}_0)^{-1})$$

Hence, from equations (17) and (18), and Slutsky's Theorem,

$$\sqrt{n}(\widehat{\underline{\theta}}^\dagger - \widehat{\underline{\theta}}_n) = \left[\mathbf{1}_d - \left(\ddot{l}_n(\widetilde{\underline{\theta}}_n) \right)^{-1} \ddot{l}_n(\underline{\theta}_n^\dagger) \right] \sqrt{n}(\widetilde{\underline{\theta}}_n - \widehat{\underline{\theta}}_n) \xrightarrow{p} \mathbf{0}_d \times \underline{Z}_0 = \underline{0}$$

Hence the two estimators are asymptotically equivalent, that is

$$\sqrt{n}(\widehat{\underline{\theta}}^\dagger - \widehat{\underline{\theta}}_n) \xrightarrow{p} \underline{0}.$$

But the asymptotic distribution of $\sqrt{n}(\widehat{\underline{\theta}}_n - \underline{\theta}_0)$ is known, and therefore

$$\sqrt{n}(\widehat{\underline{\theta}}^\dagger - \underline{\theta}_0) \xrightarrow{d} N(0, \mathcal{I}(\underline{\theta}_0)^{-1}).$$

and therefore the one-step estimator is efficient. ■