Section 8: Asymptotic Properties of the MLE

In this part of the course, we will consider the asymptotic properties of the maximum likelihood estimator. In particular, we will study issues of consistency, asymptotic normality, and efficiency. Many of the proofs will be rigorous, to display more generally useful techniques also for later chapters.

We suppose that $\boldsymbol{X}_n = (X_1, \ldots, X_n)$, where the X_i's are i.i.d. with common density $p(x;\theta_0) \in \mathcal{P} = \{p(x;\theta) : \theta \in \Theta\}$. We assume that θ_0 is *identified* in the sense that if $\theta \neq \theta_0$ and $\theta \in \Theta$, then $p(x;\theta) \neq p(x;\theta_0)$ with respect to the dominating measure μ.

For fixed $\theta \in \Theta$, the joint density of $\boldsymbol{X}_n$ is equal to the product of the individual densities, i.e.,

$$p(\boldsymbol{x}_n; \theta) = \prod_{i=1}^{n} p(x_i; \theta)$$

As usual, when we think of $p(\boldsymbol{x}_n; \theta)$ as a function of θ with $\boldsymbol{x}_n$ held fixed, we refer to the resulting function as the likelihood function, $L(\theta; \boldsymbol{x}_n)$. The *maximum likelihood estimate* for observed $\boldsymbol{x}_n$ is the value $\theta \in \Theta$ which maximizes $L(\theta; \boldsymbol{x}_n)$, $\hat{\theta}(\boldsymbol{x}_n)$. Prior to observation, $\boldsymbol{x}_n$ is unknown, so we consider the *maximum likelihood estimator*, MLE, to be the value $\theta \in \Theta$ which maximizes $L(\theta; \boldsymbol{X}_n)$, $\hat{\theta}(\boldsymbol{X}_n)$. Equivalently, the MLE can be taken to be the maximum of the standardized log-likelihood,

$$\frac{l(\theta; \boldsymbol{X}_n)}{n} = \frac{\log L(\theta; \boldsymbol{X}_n)}{n} = \frac{1}{n} \sum_{i=1}^{n} \log p(X_i; \theta) = \frac{1}{n} \sum_{i=1}^{n} l(\theta; X_i)$$

We will show that the MLE is often

1. consistent, $\hat{\theta}(\boldsymbol{X}_n) \xrightarrow{P} \theta_0$
2. asymptotically normal, $\sqrt{n}(\hat{\theta}(\boldsymbol{X}_n) - \theta_0) \overset{D(\theta_0)}{\rightarrow}$ Normal R.V.
3. asymptotically efficient, i.e., if we want to estimate θ_0 by any other estimator within a "reasonable class," the MLE is the most precise.

To show 1-3, we will have to provide some regularity conditions on the probability model *and (for 3)* on the class of estimators that will be considered.

Section 8.1 Consistency

We first want to show that if we have a sample of i.i.d. data from a common distribution which belongs to a probability model, then under some regularity conditions on the form of the density, the sequence of estimators, $\{\hat{\theta}(\boldsymbol{X}_n)\}$, will converge in probability to θ_0.

So far, we have not discussed the issue of whether a maximum likelihood estimator exists or, if one does, whether it is unique. We will get to this, but first we start with a heuristic proof of consistency.

Heuristic Proof

The MLE is the value $\theta \in \Theta$ that maximizes $Q(\theta; \boldsymbol{X}_n) := \frac{1}{n}\sum_{i=1}^{n} l(\theta; X_i)$. By the WLLN, we know that

$$\begin{aligned} Q(\theta; \boldsymbol{X}_n) = \frac{1}{n}\sum_{i=1}^{n} l(\theta; X_i) \xrightarrow{P} Q_0(\theta) \quad &:= \quad E_{\theta_0}[l(\theta; X)] \\ &= \quad E_{\theta_0}[\log p(X; \theta)] \\ &= \quad \int \{\log p(x; \theta)\} p(x; \theta_0) d\mu(x) \end{aligned}$$

We expect that, on average, the log-likelihood will be close to the expected log-likelihood. Therefore, we expect that the maximum likelihood estimator will be close to the maximum of the expected log-likelihood. We will show that the expected log-likelihood, $Q_0(\theta)$ is maximized at θ_0 (i.e., the truth).

Lemma 8.1: If θ_0 is identified and $E_{\theta_0}[|\log p(X;\theta)|] < \infty$ for all $\theta \in \Theta$, $Q_0(\theta)$ is uniquely maximized at $\theta = \theta_0$.

Proof: By Jensen's inequality, we know that for any strictly convex function $g(\cdot)$, $E[g(Y)] > g(E[Y])$. Take $g(y) = -\log(y)$. So, for $\theta \neq \theta_0$,

$$E_{\theta_0}[-\log(\frac{p(X;\theta)}{p(X;\theta_0)})] > -\log(E_{\theta_0}[\frac{p(X;\theta)}{p(X;\theta_0)}])$$

Note that

$$E_{\theta_0}[\frac{p(X;\theta)}{p(X;\theta_0)}] = \int \frac{p(x;\theta)}{p(x;\theta_0)} p(x;\theta_0) d\mu(x) = \int p(x;\theta) = 1$$

So, $E_{\theta_0}[-\log(\frac{p(X;\theta)}{p(X;\theta_0)})] > 0$ or

$$Q_0(\theta_0) = E_{\theta_0}[\log p(X;\theta_0)] > E_{\theta_0}[\log p(X;\theta)] = Q_0(\theta)$$

This inequality holds for all $\theta \neq \theta_0$.

Under technical conditions for the limit of the maximum to be the maximum of the limit, $\hat{\theta}(\boldsymbol{X}_n)$ should converge in probability to θ_0. Sufficient conditions for the maximum of the limit to be the limit of the maximum are that the convergence is uniform and the parameter space is compact.

The discussion so far only allows for a compact parameter space. In theory compactness requires that one know bounds on the true parameter value, although this constraint is often ignored in practice. It is possible to drop this assumption if the function $Q(\theta; \boldsymbol{X}_n)$ cannot rise too much as θ becomes unbounded. We will discuss this later.

Definition (Uniform Convergence in Probability): $Q(\theta; \boldsymbol{X}_n)$ converges uniformly in probability to $Q_0(\theta)$ if

$$\sup_{\theta \in \Theta} |Q(\theta; \boldsymbol{X}_n) - Q_0(\theta)| \overset{P(\theta_0)}{\to} 0$$

More precisely, we have that for all $\epsilon > 0$,

$$P_{\theta_0}[\sup_{\theta \in \Theta} |Q(\theta; \boldsymbol{X}_n) - Q_0(\theta)| > \epsilon] \to 0$$

Why isn't pointwise convergence enough? Uniform convergence guarantees that for almost all realizations, the paths in θ are in the ϵ-sleeve. This ensures that the maximum is close to θ_0. For pointwise convergence, we know that at each θ, most of the realizations are in the ϵ-sleeve, but there is no guarantee that for another value of θ the same set of realizations are in the sleeve. Thus, the maximum need not be near θ_0.

Theorem 8.2: Suppose that $Q(\theta; \boldsymbol{X}_n)$ is continuous in θ and there exists a function $Q_0(\theta)$ such that

1. $Q_0(\theta)$ is uniquely maximized at θ_0
2. Θ is compact
3. $Q_0(\theta)$ is continuous in θ
4. $Q(\theta; \boldsymbol{X}_n)$ converges uniformly in probability to $Q_0(\theta)$.

then $\hat{\theta}(\boldsymbol{X}_n)$ defined as the value of $\theta \in \Theta$ which for each $\boldsymbol{X}_n = \boldsymbol{x}_n$ maximizes the objective function $Q(\theta; \boldsymbol{X}_n)$ satisfies $\hat{\theta}(\boldsymbol{X}_n) \xrightarrow{P} \theta_0$.

Proof: For a positive ϵ, define the ϵ-neighborhood about θ_0 to be

$$\Theta(\epsilon) = \{\theta : \|\theta - \theta_0\| < \epsilon\}$$

We want to show that

$$P_{\theta_0}[\hat{\theta}(\boldsymbol{X}_n) \in \Theta(\epsilon)] \to 1$$

as $n \to \infty$. Since $\Theta(\epsilon)$ is an open set, we know that $\Theta \cap \Theta(\epsilon)^C$ is a compact set (Assumption 2). Since $Q_0(\theta)$ is a continuous function (Assumption 3), then $\sup_{\theta \in \Theta \cap \Theta(\epsilon)^C}\{Q_0(\theta)\}$ is a achieved for a θ in the compact set. Denote this value by θ^*. Since θ_0 is the unique max, let $Q_0(\theta_0) - Q_0(\theta^*) = \delta > 0$.

Now for any θ, we distinguish between two cases.

Case 1: $\theta \in \Theta \cap \Theta(\epsilon)^C$.

Let A_n be the event that $\sup_{\theta \in \Theta \cap \Theta(\epsilon)^C} |Q(\theta; \boldsymbol{X}_n) - Q_0(\theta)| < \delta/2$. Then,

$$\begin{aligned} A_n \Rightarrow Q(\theta; \boldsymbol{X}_n) &< Q_0(\theta) + \delta/2 \\ &\leq Q_0(\theta^*) + \delta/2 \\ &= Q_0(\theta_0) - \delta + \delta/2 \\ &= Q_0(\theta_0) - \delta/2 \end{aligned}$$

Case 2: $\theta \in \Theta(\epsilon)$.

Let B_n be the event that $\sup_{\theta \in \Theta(\epsilon)} |Q(\theta; \boldsymbol{X}_n) - Q_0(\theta)| < \delta/2$. Then,

$$\begin{aligned} B_n \Rightarrow\ & Q(\theta; \boldsymbol{X}_n) > Q_0(\theta) - \delta/2 \text{ for all } \theta \\ \Rightarrow & Q(\theta_0; \boldsymbol{X}_n) > Q_0(\theta_0) - \delta/2 \end{aligned}$$

By comparing the last expressions for each of cases 1,2, we conclude that if both A_n and B_n hold then $\hat{\theta} \in \Theta(\epsilon)$. But by uniform convergence, $\text{pr}(A_n \cap B_n) \to 1$, so $\text{pr}(\hat{\theta} \in \Theta(\epsilon)) \to 1$.

A key element of the above proof is that $Q(\theta; \boldsymbol{X}_n)$ converges uniformly in probability to $Q_0(\theta)$. This is often difficult to prove. A useful condition is given by the following lemma:

Lemma 8.3: If $X_1, \ldots, X_n$ are i.i.d. $p(x; \theta_0) \in \{p(x; \theta) : \theta \in \Theta\}$, Θ is compact, $\log p(x; \theta)$ is continuous in θ for all $\theta \in \Theta$ and all $x \in \mathcal{X}$, and if there exists a function $d(x)$ such that $|\log p(x; \theta)| \leq d(x)$ for all $\theta \in \Theta$ and $x \in \mathcal{X}$, and $E_{\theta_0}[d(X)] < \infty$, then

i. $Q_0(\theta) = E_{\theta_0}[\log p(X; \theta)]$ is continuous in θ

ii. $\sup_{\theta \in \Theta} |Q(\theta; \boldsymbol{X}_n) - Q_0(\theta)| \xrightarrow{P} 0$

Example: Suicide seasonality and von Mises' distribution (in class)

Proof: We first prove the continuity of $Q_0(\theta)$. For any $\theta \in \Theta$, choose a sequence $\theta_k \in \Theta$ which converges to θ. By the continuity of $\log p(x;\theta)$, we know that $\log p(x;\theta_k) \to \log p(x;\theta)$. Since $|\log p(x;\theta_k)| \leq d(x)$, the dominated convergence theorem tells us that $Q_0(\theta_k) = E_{\theta_0}[\log p(X;\theta_k)] \to E_{\theta_0}[\log p(X;\theta)] = Q_0(\theta)$. This implies that $Q_0(\theta)$ is continuous.

Next, we work to establish the uniform convergence in probability. We need to show that for any $\epsilon, \eta > 0$ there exists $N(\epsilon,\eta)$ such that for all $n > N(\epsilon,\eta)$,

$$P[\sup_{\theta\in\Theta} |Q(\theta;\boldsymbol{X}_n) - Q_0(\theta)| > \epsilon] < \eta$$

Since $\log p(x;\theta)$ is continuous in $\theta \in \Theta$ and since Θ is compact, we know that $\log p(x;\theta)$ is uniformly continuous (see Rudin, page 90). Uniform continuity says that for all $\epsilon > 0$ there exists a $\delta(\epsilon) > 0$ such that $|\log p(x;\theta_1) - \log p(x;\theta_2)| < \epsilon$ for all $\theta_1, \theta_2 \in \Theta$ for which $\|\theta_1 - \theta_2\| < \delta(\epsilon)$. This is a property of a function defined on a set of points. In contrast, continuity is defined relative to a particular point. Continuity of a function at a point θ^* says that for all $\epsilon > 0$, there exists a $\delta(\epsilon, \theta^*)$ such that $|\log p(x;\theta) - \log p(x;\theta^*)| < \epsilon$ for all $\theta \in \Theta$ for which $\|\theta - \theta^*\| < \delta(\epsilon, \theta^*)$. For continuity, δ depends on ϵ *and* θ^*. For uniform continuity, δ depends only on ϵ. In general, uniform continuity is stronger than continuity; however, they are equivalent on compact sets.

Aside: Uniform Continuity vs. Continuity

Consider $f(x) = 1/x$ for $x \in (0, 1)$. This function is continuous for each $x \in (0, 1)$. However, it is not uniformly continuous. Suppose this function was uniformly continuous. Then, we know for any $\epsilon > 0$, we can find a $\delta(\epsilon) > 0$ such that $|1/x_1 - 1/x_2| < \epsilon$ for all x_1, x_2 such that $|x_1 - x_2| < \delta(\epsilon)$. Given an $\epsilon > 0$, consider the points x_1 and $x_2 = x_1 + \delta(\epsilon)/2$. Then, we know that

$$\left|\frac{1}{x_1} - \frac{1}{x_2}\right| = \left|\frac{1}{x_1} - \frac{1}{x_1 + \delta(\epsilon)/2}\right| = \frac{\delta(\epsilon)/2}{x_1(x_1 + \delta(\epsilon)/2)}$$

Take x_1 sufficiently small so that $x_1 < \min(\frac{\delta(\epsilon)}{2}, \frac{1}{2\epsilon})$. This implies that

$$\frac{\delta(\epsilon)/2}{x_1(x_1 + \delta(\epsilon)/2)} > \frac{\delta(\epsilon)/2}{x_1\delta(\epsilon)} = \frac{1}{2x_1} > \epsilon$$

This is a contradiction.

Uniform continuity also implies that

$$\Delta(x,\delta) = \sup_{\{(\theta_1,\theta_2):\|\theta_1-\theta_2\|<\delta\}} |\log p(x;\theta_1) - \log p(x;\theta_2)| \to 0 \qquad (1)$$

as $\delta \to 0$. By the assumption of Lemma 8.3, we know that $\Delta(x,\delta) \leq 2d(x)$ for all δ. By the dominated convergence theorem, we know that $E_{\theta_0}[\Delta(X,\delta)] \to 0$ as $\delta \to 0$.

Now, consider open balls of length δ about each $\theta \in \Theta$, i.e., $B(\theta,\delta) = \{\tilde{\theta} : \|\tilde{\theta} - \theta\| < \delta\}$. The union of these open balls contains Θ. This union is an open cover of Θ. Since Θ is a compact set, we know that there exists a finite subcover, which we denote by $\{B(\theta_j,\delta), j = 1,\ldots,J\}$.

Taking a $\theta \in \Theta$, by the triangle inequality, we know that

$$\begin{aligned} |Q(\theta; \boldsymbol{X}_n) - Q_0(\theta)| \quad \leq \quad & |Q(\theta; \boldsymbol{X}_n) - Q(\theta_j; \boldsymbol{X}_n)| + \qquad (2) \\ & |Q(\theta_j; \boldsymbol{X}_n) - Q_0(\theta_j)| + \qquad (3) \\ & |Q_0(\theta_j) - Q_0(\theta)| \qquad (4) \end{aligned}$$

Choose θ_j so that $\theta \in B(\theta_j; \delta)$. Since $\|\theta - \theta_j\| < \delta$, we know that (2) is equal to

$$|\frac{1}{n} \sum_{i=1}^{n} \{\log p(X_i; \theta) - \log p(X_i; \theta_j)\}|$$

and this is less than or equal

$$\frac{1}{n} \sum_{i=1}^{n} |\log p(X_i; \theta) - \log p(X_i; \theta_j)| \leq \frac{1}{n} \sum_{i=1}^{n} \Delta(X_i, \delta)$$

We also know that (4) is less than or equal to

$$\sup_{\{(\theta_1,\theta_2):\|\theta_1-\theta_2\|<\delta\}} |Q_0(\theta_1) - Q_0(\theta_2)|$$

Since $Q_0(\theta)$ is uniformly continuous, we know that this bound can be made arbitrarily small by choosing δ to be small. That is, this bound can be made less than $\epsilon/3$, for a $\delta < \delta_3$.

We also know that (3) is less than

$$\max_{j=1,\ldots,J} |Q(\theta_j; \boldsymbol{X}_n) - Q_0(\theta_j)|$$

Putting these results together, we have that for any $\delta < \delta_3$

$$\sup_{\theta\in\Theta} |Q(\theta; \boldsymbol{X}_n) - Q_0(\theta)| \leq \frac{1}{n}\sum_{i=1}^{n} \Delta(X_i, \delta) + \max_{j=1,\ldots,J} |Q(\theta_j; \boldsymbol{X}_n) - Q_0(\theta_j)| + \epsilon/3$$

So for any $\delta < \delta_3$, we know that $P_{\theta_0}[\sup_{\theta\in\Theta} |Q(\theta; \boldsymbol{X}_n) - Q_0(\theta)| > \epsilon]$

$$\begin{aligned}
&\leq P_{\theta_0}[\frac{1}{n}\sum_{i=1}^{n}\Delta(X_i,\delta) + \max_{j=1,\ldots,J}|Q(\theta_j;\boldsymbol{X}_n) - Q_0(\theta_j)| > 2\epsilon/3] \\
&\leq P_{\theta_0}[\frac{1}{n}\sum_{i=1}^{n}\Delta(X_i,\delta) > \epsilon/3] + \qquad (5) \\
&\quad P_{\theta_0}[\max_{j=1,\ldots,J}|Q(\theta_j;\boldsymbol{X}_n) - Q_0(\theta_j)| > \epsilon/3] \qquad (6)
\end{aligned}$$

Now, we can show that we can take n sufficiently large so that (5) and (6) can be made small.

Note that (5) is equal to

$$P_{\theta_0}[\frac{1}{n}\sum_{i=1}^{n}\{\Delta(X_i,\delta) - E_{\theta_0}[\Delta(X,\delta)]\} + E_{\theta_0}[\Delta(X,\delta)] > \epsilon/3]$$

We already have demonstrated that $E_{\theta_0}[\Delta(X,\delta)] \to 0$ as $\delta \to 0$. Choose δ small enough ($< \delta_1$)that $E_{\theta_0}[\Delta(X,\delta)] < \epsilon/6$. Call this number δ_1. Take $\delta < \min(\delta_1, \delta_3)$. Then (5) is less than

$$P_{\theta_0}[\frac{1}{n}\sum_{i=1}^{n}\{\Delta(X_i,\delta) - E_{\theta_0}[\Delta(X,\delta)]\} > \epsilon/6]$$

By the WLLN, we know that there exists $N_1(\epsilon,\eta)$ so that for all $n > N_1(\epsilon,\eta)$, the above term is less than $\eta/2$. So, (5) is less than $\eta/2$.

For the δ considered so far, find the finite subcover $\{B(\theta_j, \delta), j = 1, \ldots, J\}$. Now, (6) is equal to

$$P_{\theta_0}[\cup_{j=1}^{J}\{|Q(\theta_j; \boldsymbol{X}_n) - Q_0(\theta_j)| > \epsilon/3\}] \leq \sum_{j=1}^{J} P_{\theta_0}[|Q(\theta_j; \boldsymbol{X}_n) - Q_0(\theta_j)| > \epsilon/3]$$

By the WLLN, we know that for each θ_j and for any $\eta > 0$, we know there exists $N_{2j}(\epsilon, \eta)$ so that for all $n > N_{2j}(\epsilon, \eta)$

$$P_{\theta_0}[|Q(\theta_j; \boldsymbol{X}_n) - Q_0(\theta_j)| > \epsilon/3] \leq \eta/(2J)$$

Let $N_2(\epsilon, \eta) = \max_{j=1,\ldots,J}\{N_{2j}\}$. Then, for all $n > N_2(\epsilon, \eta)$, we know that

$$\sum_{j=1}^{J} P_{\theta_0}[|Q(\theta_j; \boldsymbol{X}_n) - Q_0(\theta_j)| > \epsilon/3] < \eta/2$$

This implies that (6) is less than $\eta/2$.

Combining the results for (5) and (6), we have demonstrated that there exists an $N(\epsilon, \eta) = \max(N_1(\epsilon, \eta), N_2(\epsilon, \eta))$ so that for all $n > N(\epsilon, \eta)$,

$$P_{\theta_0}[\sup_{\theta \in \Theta} |Q(\theta; \boldsymbol{X}_n) - Q_0(\theta)| > \epsilon] < \eta$$

Q.E.D.

Other conditions that might be useful to establish uniform convergence in probability are given in the lemmas below.

Lemma 8.4 may be useful when the data are not independent.

Lemma 8.4: If Θ is compact, $Q_0(\theta)$ is continuous in $\theta \in \Theta$, $Q(\theta; \boldsymbol{X}_n) \xrightarrow{P} Q_0(\theta)$ for all $\theta \in \Theta$, and there is an $\alpha > 0$ and $C(\boldsymbol{X}_n)$ which is bounded in probability such that for all $\tilde{\theta}, \theta \in \Theta$,

$$|Q(\tilde{\theta}, \boldsymbol{X}_n) - Q(\theta, \boldsymbol{X}_n)| \leq C(\boldsymbol{X}_n)\|\tilde{\theta} - \theta\|^{\alpha}$$

then,

$$\sup_{\theta \in \Theta} |Q(\theta; \boldsymbol{X}_n) - Q_0(\theta)| \xrightarrow{P} 0$$

Aside: $C(\boldsymbol{X}_n)$ is bounded in probability if for every $\epsilon > 0$ there exists $N(\epsilon)$ and $\eta(\epsilon) > 0$ such that

$$P_{\theta_0}[|C(\boldsymbol{X}_n)| > \eta(\epsilon)] < \epsilon$$

for all $n > N(\epsilon)$.

Θ is Not Compact

Suppose that we are not willing to assume that Θ is compact. One way around this is bound the objective function from above uniformly in parameters that are far away from the truth. For example, suppose that there is a compact set D such that

$$E_{\theta_0}[\sup_{\theta \in \Theta \cap D^C} \{\log p(X;\theta)\}] < Q_0(\theta_0) = E_{\theta_0}[\log p(X;\theta_0)]$$

By the law of large numbers, we know that with probability approaching one that

$$\begin{aligned} \sup_{\theta \in \Theta \cap D^C} \{Q(\theta; \boldsymbol{X}_n)\} &\leq \frac{1}{n} \sum_{i=1}^{n} \sup_{\theta \in \Theta \cap D^C} \log p(X_i;\theta) \\ &< \frac{1}{n} \sum_{i=1}^{n} \log p(X_i;\theta_0) = Q(\theta_0; \boldsymbol{X}_n) \end{aligned}$$

Therefore, with probability approaching one, we know that $\hat{\theta}(\boldsymbol{X}_n)$ is in the compact set D. Then, we can apply the previous theory to show consistency.

The following lemma can be used in cases where the objective function is concave.

Lemma 8.5: If there is a function $Q_0(\theta)$ such that

i. $Q_0(\theta)$ is uniquely maximized at θ_0

ii. θ_0 is an element of the interior of a convex set Θ (does not have to be bounded)

iii. $Q(\theta, \boldsymbol{x}_n)$ is concave in θ for each $\boldsymbol{x}_n$, and

iv. $Q(\theta; \boldsymbol{X}_n) \xrightarrow{P} Q_0(\theta)$ for all $\theta \in \Theta$

then $\hat{\theta}(\boldsymbol{X}_n)$ exists with probability approaching one and $\hat{\theta}(\boldsymbol{X}_n) \xrightarrow{P} \theta_0$.