Statistics 210A Lecture 22 Notes

Daniel Raban

November 9, 2021

1 Asymptotic Consistency of the Maximum Likelihood Estimator

1.1 Recap: Maximum likelihood estimation

Last time, we introduced maximum likelihood estimation. If our model is $\mathcal{P}$ with densities $p_\theta(x)$ with respect to μ and if our sample is $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p_{\theta_0}$, then the **maximum likelihood estimator (MLE)** is

$$\begin{aligned}
\widehat{\theta}_n &= \arg\max_{\theta \in \Theta} p_\theta(X) \\
&= \arg\max_{\theta \in \Theta} \ell_n(\theta; X) \\
&= \arg\max_{\theta \in \Theta} \sum_{i=1}^n \underbrace{\ell_1(\theta; X_i) - \ell_1(\theta_0, X_i)}_{W_i(\theta)} \\
&= \arg\max_{\theta \in \Theta} \overline{W}_n(\theta),
\end{aligned}$$

where

$$\overline{W}_n(\theta) = \frac{1}{n} \sum_{i=1}^n W_i(\theta).$$

We are interested in how $\overline{W}_n$ converges to its expectation.

Last time, we made a quadratic expansion of the log-likelihood (or a linear expansion of the score),

$$0 = \nabla \ell_n(\widehat{\theta}_n) = \nabla \ell_n(\theta_0) + \nabla^2 \ell(\widetilde{\theta}_n)(\widehat{\theta}_n - \theta_0),$$

where $\widetilde{\theta}_n$ is some value given by the mean value theorem. Then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \left(-\frac{1}{n} \nabla^2 \ell(\widehat{\theta}_n)\right)^{-1} \underbrace{\left(\frac{1}{\sqrt{n}} \nabla \ell_n(\theta_0)\right)}_{\Longrightarrow N_d(0, J_1(\theta_0))}.$$

We want to say that the first term converges in probability to $J_1(\theta_0)^{-1}$. We need a few ingredients:

- $\widehat{\theta}_n \xrightarrow{p} \theta_0$.
- $J_1(\theta_0) \succ 0$.
- We need to deal with a random function at the random value $\widehat{\theta}_n$.

1.2 Pointwise convergence of likelihood ratio averages

We can say $\overline{W}_n(\theta)$ is a sample mean of iid $W_1(\theta), \ldots, W_n(\theta)$. Recall the **KL-Divergence**

$$D_{\mathrm{KL}}^{(1)}(\theta_0 \,||\, \theta) = \mathbb{E}_{\theta_0}\left[\log \frac{p_{\theta_0}(X_1)}{p_\theta(X_1)}\right].$$

Then by Jensen's inequality,

$$\begin{aligned} -D_{\mathrm{KL}}^{(1)}(\theta_0 \,||\, \theta) &\leq \log \mathbb{E}_{\theta_0}\left[\frac{p_{\theta_0}(X_1)}{p_\theta(X_1)}\right] \\ &\leq \log 1 \\ &= 0. \end{aligned}$$

Since log is strictly concave, this is a strict inequality unless $p_{\theta_0} = p_\theta$.

Now let's calculate the expectation of the Ws:

$$\begin{aligned} \mathbb{E}_{\theta_0}[\overline{W}_n(\theta)] &= \mathbb{E}_{\theta_0}[W_i(\theta)] \\ &= \mathbb{E}_{\theta_0}[\ell_1(\theta; X_1) - \ell_1(\theta_0; X_i)] \\ &= -D_{\mathrm{KL}}(\theta_0 \,||\, \theta) \\ &< 0, \end{aligned}$$

unless $p_{\theta_0} = p_0$. Then

$$\overline{W}_n(\theta) \xrightarrow{p} -D_{\mathrm{KL}}(\theta_0 \,||\, \theta) < 0$$

unless $p_{\theta_0} = p_0$. We need a way to make this convergence uniform.

1.3 Uniform convergence of random functions

Definition 1.1. For a compact K, let $C(K)$ be the set of all continuous functions $f : K \to \mathbb{R}$.

Definition 1.2. For any $f \in C(K)$, the L^∞ **norm is**

$$\|f\|_\infty = \sup_{t \in K} |f(t)|.$$

Definition 1.3. We say that $f_n \to f$ in this norm (f_n **converges uniformly** to f) if $\|f_n - f\|_\infty \to 0$.

Theorem 1.1 (Law of large numbers for random functions)**.** *Assume K is comapct, and $W_1, W_2, \cdots \in C(K)$ are iid with $\mathbb{E}[\|W_i\|_\infty] < \infty$. Let $\mu(t) = \mathbb{E}[W_i(t)]$. Then $\mu(t) \in C(K)$, and*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} W_i - \mu \right\|_\infty \xrightarrow{p} 0.$$

*That is, $\frac{1}{n}\sum_{i=1}^{n} W_i \to \mu$ **uniformly in probability**.*

We won't prove this.

Theorem 1.2 (9.4 in Keener)**.** *Let $G_1, G_2, \ldots$ be random functions in $C(K)$ with K compact. Assume that $\|G_n - g\|_\infty \xrightarrow{p} 0$ for some fixed $g \in C(K)$. Then*

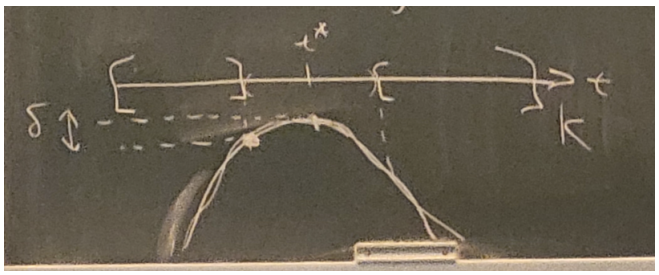
1. *If $t_n \xrightarrow{p} t^*$ with t_n random and $t^* \in K$ fixed, then $G_n(t_n) \xrightarrow{p} g(t^*)$.*
2. *If g is maximized at a unique value $t^* \in K$ and $G_n(t_n) = \max_t G_n(t)$, then $t_n \xrightarrow{p} t^*$.*
3. *If $K \subseteq \mathbb{R}$, $g(t) = 0$ has a unique solution t^*, and t_n solves $G_n(t_n) = 0$, then $t_n \xrightarrow{p} t^*$.*

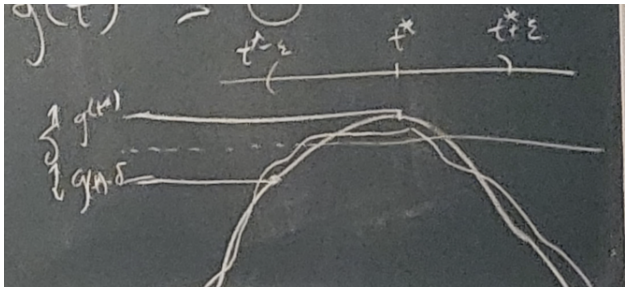
Proof.

1.
$$\begin{aligned} |G_n(t_n) - g(t^*)| &\le |G_n(t_n) - g(t-n)| + |g(t_n) - g(t^*)| \\ &\le \underbrace{\|G_n - g\|_\infty}_{\xrightarrow{p} 0} + \underbrace{|g(t_n) - g(t^*)|}_{\xrightarrow{p} 0}, \end{aligned}$$

   where the second term converges to 0 in probability by the continuous mapping theorem. So $G_n(t_n) \xrightarrow{p} g(t^*)$.

2. Fix $\varepsilon > 0$, and let $B_\varepsilon(t^*) = \{t : \|t - t^*\| < \varepsilon\}$. Let $K_\varepsilon = K \setminus B_\varepsilon(t^*)$; this intersection is also compact. Let
$$\delta = g(t^*) - \max_{t \in K_\varepsilon} g(t) > 0.$$





If $t_n \in K_\varepsilon$, then

$$G_n(t_n) \leq \max_{t \in K_\varepsilon} g(t) + \|G_n - g\|_\infty. = g(t^*) - \delta + \|G_n - g\|_\infty.$$

We also know that

$$G_n(t_n) \geq G_n(t^*) \geq g(t^*) - \|G_n - g\|_\infty.$$

Subtracting these inequalities gives

$$2\|G_n - g\|_\infty \geq \delta.$$

The probability of this is going to 0 by assumption, so $\mathbb{P}(t_n \in K_\varepsilon) \to 0$.

3. The proof of this is analogous to the proof of the second statement. □

What if we don't need the exact maximizer or if there is no exact maximizer? We can modify part 2 of the theorem:

Theorem 1.3. *Let $G_1, G_2, \ldots$ be random functions in $C(K)$ with K compact. Assume that $\|G_n - g\|_\infty \xrightarrow{p} 0$ for some fixed $g \in C(K)$. Then if g is maximized at a unique value $t^* \in K$ and $G_n(t_n) = \max_t G_n(t) - \alpha_n$ with $\alpha_n \to 0$, then $t_n \xrightarrow{p} t^*$.*

Proof. We can repeat the same argument, except this time we get

$$F_n(t_n) \geq G_n(t^*) - \alpha_n \geq g(t^*) - \|G_n - g\|_\infty - \alpha_n.$$

This gives

$$2\|G_n - g\|_\infty \geq \delta - \alpha_n,$$

and the proof still works. □

1.4 Consistency results for the MLE

Theorem 1.4 (Consistency of the MLE for compact Θ)**.** *Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} p_{\theta_0}$, where $\mathcal{P}$ has continuous densities p_θ for $\theta \in \Theta$. Assume that*

- *Θ is compact,*
- *$\mathbb{E}_{\theta_0}[\|W_i\|_\infty] = \mathbb{E}_{\theta_0}[\|\ell_1(\theta; X_i) - \ell_1(\theta_0; X_i)\|_\infty] < \infty$,*
- *The model $\mathcal{P}$ is identifiable.*

Then $\widehat{\theta}_n \xrightarrow{p} \theta_0$ if $\widehat{\theta}_n \in \arg\max \ell_n(\theta; X)$.

So it doesn't matter which value we pick for the MLE; we still get consistency.

Proof. Since the densities are continuous, $W_i \in C(\Theta)$. They are iid with mean $\mu(\theta) = -D_{\mathrm{KL}}(\theta_0 \,||\, \theta)$, where $\mu(\theta_0) = 0$ and $\mu(\theta) < 0$ for all $\theta \neq \theta_0$. So θ_0 uniquely maximizes μ. By definition, $\widehat{\theta}_n$ maximizes $\overline{W}_n$, so $\|\overline{W}_n - \mu\|_\infty \xrightarrow{p} 0$ by the law of large numbers. Now apply the previous theorem. □

Here is a way (but not the only way) to restrict our attention to a compact set.

Theorem 1.5 (Keener 9.11 with slightly stronger assumptions)**.** *Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} p_{\theta_0}$, where the model $\mathcal{P}$ has continuous densities p_θ for $\theta \in \Theta \subseteq \mathbb{R}^d$. Assume*

- *The model is identifiable.*
- *For all compact $K \subseteq \mathbb{R}^d$, $\mathbb{E}[\sup_{\theta \in K} |W_i(\theta|] < \infty$.*
- *There exists an $r > 0$ such that*

$$\mathbb{E}\left[\sup_{\|\theta - \theta_0\| > r} W_i(\theta)\right] < 0.$$

Then $\widehat{\theta}_n \xrightarrow{p} \theta_0$ if $\widehat{\theta}_n \in \arg\max \ell_n(\theta; X)$.

Proof. Let $A = \{\theta : \|\theta - \theta_0\| > r\}$, and let $\alpha = \mathbb{E}[\sup_{\theta \in A} W_i(\theta)] < 0$. Then

$$\sup_{\theta \in A} \overline{W}_n(\theta) \leq \frac{1}{n} \sum_{i=1}^n \sup_{\theta \in A} W_i(\theta) \to \alpha < 0.$$

So

$$\mathbb{P}(\widehat{\theta}_n \in A) \leq \mathbb{P}(\overline{W}_n(\theta_0) \leq \sup_{\theta \in A} \overline{W}_n(\theta_)) \xrightarrow{0},$$

as $\alpha \xrightarrow{p} 0$ implies $\sup_{\theta \in A} \overline{W}_n(\theta) \xrightarrow{p} 0$. Now let

$$\widehat{\theta}_n^A = \widehat{\theta}_n \mathbb{1}_{\{\widehat{\theta}_n \in A^c\}} + \theta_0 \mathbb{1}\{\widehat{\theta}_n \in A\} \xrightarrow{p} \theta_0.$$

Then $\widehat{\theta}_n \xrightarrow{p} \theta_0$. □