Statistics 210A Lecture 23 Notes

Daniel Raban

November 16, 2021

1 Asymptotic Consistency of the MLE and Likelihood-Based Hypothesis Tests

1.1 Recap: Uniform convergence of random functions

Last time, we were interested in uniform convergence of the random functions given by the sample mean of $W_i(\theta; X_i) = \ell_1(\theta; X_i) - \ell_1(\theta_0; X_i)$. The nice thing about these is that

$$\mathbb{E}[W_i(\theta)] = D_{\mathrm{KL}}(\theta \,||\, \theta_0),$$

which is ≤ 0, with equality iff $P_\theta = P_{\theta_0}$. We saw that $\widetilde{\theta}_n \xrightarrow{p} \theta_0$ if the W_i are continuous and $\|\overline{W}_n - \mathbb{E}[\overline{W}_n]\|_\infty \xrightarrow{p} 0$ on compact Θ (otherwise, we need an extra argument).

We also proved the helpful fact

Proposition 1.1. *If $\|G_n - g\|_\infty \xrightarrow{p} 0$, $t_n \xrightarrow{p} t$, and G_n, g are continuous with compact domain, then*

$$G_n(t_n) \xrightarrow{p} g(t).$$

1.2 Asymptotic distribution of the MLE

Theorem 1.1. *Suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim} p_{\theta_0}$, where $\theta_0 \in \Theta^o \subseteq \mathbb{R}^d$. Assume that*

- *$\widehat{\theta}_n \xrightarrow{p} \theta_0$, where $\widehat{\theta}_n \in \operatorname{argmax}_{\theta \in \Theta} \ell_n(\theta; X)$*
- *In some neighborhood $B_\varepsilon(\theta_0) = \{\theta : \|\theta - \theta_0\| \leq \varepsilon\} \subseteq \Theta^o$,*
  - *(i) $\ell_1(\theta; X)$ has 2 continuous derivatives on $B_\varepsilon(\theta_0)$ for all x.*
  - *(ii) $\mathbb{E}_{\theta_0}[\sup_{\theta \in B_\varepsilon} \|\nabla^2 \ell_1(\theta; X_i)\|] < \infty$.*
  - *(iii) Fisher information condition:*

$$\mathbb{E}_{\theta_0}[\nabla \ell_1(\theta_0; X_i)] = 0, \qquad \mathrm{Var}_{\theta_0}(\nabla \ell_1(\theta)) = -\,\mathbb{E}_\theta[\nabla^2 \ell_1(\theta_0)] \succ 0.$$

Then

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \implies N_d(0, J_1(\theta_0)^{-1}),$$

i.e. the MLE is asymptotically efficent.

The conditions in this theorem can be relaxed somewhat.

Proof. Let A_n be the event $\{\|\widehat{\theta}_n - \theta_0\| \geq \varepsilon\}$. Then $\mathbb{P}_{\theta_0}(A_n) \to 0$ by assumption. All we care about is what happening on A_n^c. On A_n^c, $\widehat{\theta}_n \in B_\varepsilon(\theta_0)$, and

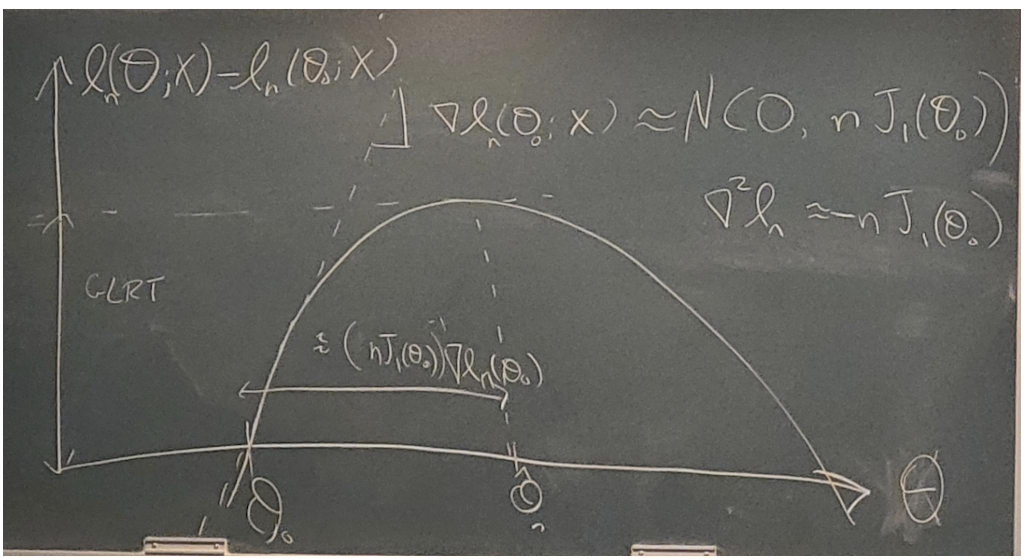
$$\begin{aligned} 0 &= \nabla \ell_n(\widehat{\theta}_n; X) \\ &= \nabla \ell_n(\theta_0; X) + \nabla^2 \ell_n(\widetilde{\theta}_0; X)(\widehat{\theta}_n - \theta_0) \end{aligned}$$

for some $\widetilde{\theta}_n$. Now

$$\begin{aligned} \sqrt{n}(\widehat{\theta}_n - \theta_0) &= \underbrace{\left(\frac{1}{n}\nabla^2 \ell_n(\widetilde{\theta}_n)\right)^{-1}}_{\xrightarrow{p} J_1(\theta_0)^{-1}} \underbrace{\frac{1}{\sqrt{n}}\nabla \ell_n(\theta_0)}_{\implies N_d(0, J_1(\theta))}, \\ &\implies N_d(0, J_1(\theta_0)^{-1}). \end{aligned}$$

□

The proof basically says that the second derivative of the likelihood is approximately non-random and equals the Fisher information.



If the fisher information is very large, the second derivative of the likelihood function is huge at θ_0. This makes the likelihood more strongly peaked, so the MLE won't be so far from θ_0.

1.3 Likelihood-based hypothesis tests

We can develop likelihood-based tests based on measuring different aspects of the above MLE picture. Let $X_1, \ldots, X_n \overset{\text{iid}}{\sim} p_\theta(x)$, where $p_\theta(x)$ is "smooth" in θ. Assume that

$$\mathbb{E}_\theta[\nabla \ell_1(\theta; X_i)] = 0, \qquad \mathrm{Var}_\theta(\nabla \ell_1(\theta; X_i)) = -\mathbb{E}_\theta[\nabla^2 \ell_1(\theta; X_i)] = J_1(\theta) \succ 0,$$

and $\widehat{\theta}_{\mathrm{MLE}} \xrightarrow{p} \theta_0$. Then if $\theta = \theta_0$,

$$\frac{1}{\sqrt{n}} \nabla \ell_n(\theta_0) \implies N_d(0, J_1(\theta_0)),$$

$$-\frac{1}{n} \nabla^2 \ell_n(\theta_0) \xrightarrow{p} J_1(\theta_0),$$

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \implies N_d(0, J_1(\theta)^{-1}).$$

1.3.1 Wald-type confidence regions

Assume we have an estimator $\widehat{J}_n \succ 0$ such that $\frac{1}{n}\widehat{J}_n \xrightarrow{p} J_1(\theta_0) \succ 0$. Then

$$(J_1(\theta_0))^{1/2} \sqrt{n}(\widehat{\theta}_n - \theta_0) \implies N_d(0, I_d),$$

and by Slutsky's theorem,

$$\widehat{J}_n^{1/2}(\widehat{\theta}_n - \theta_0) \implies N_d(0, I_d).$$

To get a test statistic, we can do the simplest (but not always the best) thing and take the 2-norm:

$$\|\widehat{J}_n^{1/2}(\widehat{\theta}_n - \theta_0)\|^2 \implies \chi^2_d.$$

Here,

$$\mathbb{P}(\|\widehat{J}_n^{1/2}(\widehat{\theta}_n - \theta_0)\|^2 > \chi^2_d(\alpha)) \to \alpha,$$

where $\chi^2_d(\alpha)$ is the upper-α quantile.

To test $H_0 : \theta = \theta_0$, we reject if $\|\widehat{J}_n^{1/2}(\widehat{\theta}_n - \theta_0)\|_2^2 > \chi^2_d(\alpha)$. Equivalently, we can say we reject θ_0 iff $\widehat{J}_n^{1/2}(\widehat{\theta}_n - \theta_0) \notin B_{\chi^2_d(\alpha)}(0)$. So we can reject θ_0 if and only if $\theta_0 \notin \widehat{\theta} + \widehat{J}_n^{1/2} B_{\chi^2_d(\alpha)}(0)$. This gives a *confidence ellipsoid.*

Here are some options for $\widehat{J}_n$:

1.

$$\begin{aligned} \widehat{J}_n &= n J_1(\widehat{\theta}_n) \\ &= n \,\mathrm{Var}_\theta(\nabla \ell_n(\theta; X))\big|_{\theta = \widehat{\theta}_n} \\ &= n \,\mathrm{Var}_{\widehat{\theta}_n}(\nabla \ell_n(\widehat{\theta}_n; X)) \end{aligned}$$

2. Observed Fisher information:

$$\widehat{J}_n = -\nabla^2 \ell_n(\widehat{\theta}_n; X)$$

The observed Fisher information is generally preferred and is used in practice. We can get a *Wald interval* for θ_j by

$$\theta_n \approx N_d(\theta_0, J_n(\theta_0)^{-1}),$$

which tells us that

$$\widehat{\theta}_{n,j} \approx N(\theta_{0,j}, (J_n(\theta_0)^{-1})_{j,j}).$$

So the **univariate Wald interval** for θ_j is

$$\begin{aligned} C_j &= \widehat{\theta}_{n,j} \pm \widehat{\text{s.e.}}(\widehat{\theta}_{n,j}) z_{\alpha/2} \\ &= \widehat{\theta}_{n,j} \pm \sqrt{(\widehat{J}_n^{-1})_{j,j}} z_{\alpha/2} \end{aligned}$$

1.3.2 The score test

Here is a test which only assumes normality of the Fisher information. Test $J_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$. Then

$$\frac{1}{\sqrt{n}} \nabla \ell_n(\theta_0; X) \overset{H_0}{\Longrightarrow} N_d(0, J_1(\theta_0)),$$

and the **score statistic** looks like

$$J_n(\theta_0)^{-1/2} \nabla \ell_n(\theta_0; X) \overset{H_0}{\Longrightarrow} N_d(0, I_d).$$

So we reject H_0 if $\|J_n(\theta_0)^{-1/2} \nabla \ell_n(\theta_0; X)\|^2 > \chi_d^2(\alpha)$.

If $d = 1$, this looks like

$$\frac{\dot{\ell}_n(\theta_0)}{\sqrt{J_n(\theta_0)}} \Longrightarrow N(0,1).$$

This is actually invariant of parameterization. For simplicity of notation, assume $d = 1$ for now. Let $\theta = g(\zeta)$ with $\dot{\zeta}) > 0$ be a reparameterization, and denote $q_\zeta(x) = p_{g(\zeta)}(x)$. Then the score is

$$\begin{aligned} \dot{\ell}^{(\zeta)}(\zeta, x) &= \frac{d}{d\zeta} \log p_{g(\zeta)}(x) \\ &= \dot{\ell}(g(\zeta)) \dot{g}(\zeta) \end{aligned}$$

by the chain rule. The Fisher information is

$$J^{(\zeta)}(\zeta) = J^{(\theta)}(g(\zeta) \dot{g}(\zeta)^2.$$

So the score statistic is unchanged by the parameterization.

Example 1.1. Let $X_1, \dots, X_n \overset{\text{iid}}{\sim} e^{\eta^\top T(x) - A(\eta)} h(x)$ be an s-parameter exponential family. THen

$$\nabla \ell_n(\eta) = \left(\sum_{i=1}^n T(X_i) \right) - n\mu(\zeta), \qquad \text{where } \mu(\eta) = \mathbb{E}_\eta[T(X_i)].$$

Then

$$\left\| J_n(\eta_0)^{-1/2} \left(\sum_i T(X_i) - n\mu(\eta) \right) \right\|_2^2 \implies \chi_d^2$$

gives us our test. In particular, if $d = 1$, we get

$$\frac{\sum_i T(X_i) - n\mu(\eta_0)}{\sqrt{n \operatorname{Var}_{\eta_0}(T(X_1))}} \overset{H_0}{\implies} N(0,1),$$

so this is a Z-test.

The test statistic for the score test is

$$\|(J_1(\theta_0))^{-1/2} \tfrac{1}{\sqrt{n}} \nabla \ell_n(\theta_0)\|^2,$$

while the test statistic for the Wald test is

$$\|\widehat{J}_1^{1/2} \sqrt{n}(\widehat{\theta}_n - \theta_0)\|^2,$$

where $\sqrt{n}(\widehat{\theta}_n - \theta_0) \approx J_1(\theta_0^{-1}) \frac{1}{n} \nabla \ell_n(\theta_0)$. So these are asymptotically the same test.