

Maximum likelihood: Asymptotic normality

Patrick Breheny

Intro

- Today, we continue with our goal of deriving the asymptotic properties of maximum likelihood estimators
- Previously, we established conditions under which the MLE was consistent: $\|\hat{\theta} - \theta^*\| \xrightarrow{P} 0$
- Today, we will see that under those same conditions, $\sqrt{n}(\hat{\theta} - \theta^*)$ converges in distribution to a multivariate normal
- After establishing this, we will consider how these results change if we remove the log-concavity assumption and allow for the possibility of multiple maxima

Vector-valued functions

- The main idea behind the proof is to take a Taylor expansion not of the likelihood, but rather the score
- However, this requires us to extend our earlier Taylor expansion results a bit
- Up until this point, we have always expanded scalar-valued functions, but the score function is vector-valued, which impacts the Lagrange form of the expansion

Taylor expansion of vector-valued functions

- In particular:
 - Taylor's theorem still applies to each component of the vector-valued function
 - However, the intermediate point may differ from one component to the next
- **Theorem:** Suppose $\mathbf{f}: \mathbb{R}^d \rightarrow \mathbb{R}^k$ is differentiable on $N_r(\mathbf{x}_0)$. Then for any $\mathbf{x} \in N_r(\mathbf{x}_0)$,

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{x}_0) + \mathbf{D}(\mathbf{x}, \mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$$

where row j of $\mathbf{D}(\mathbf{x}, \mathbf{x}_0)$ is $\nabla f_j(\tilde{\mathbf{x}}_j)^\top$, with each $\tilde{\mathbf{x}}_j$ on the line segment between $\mathbf{x}$ and $\mathbf{x}_0$

Asymptotic linearity of the score

- Applying this expansion to the score vector, we obtain the following corollary, which will come up many times in this course
- **Theorem:** Suppose regularity conditions (A)-(C) from the previous lecture are met. Then for any consistent estimator $\hat{\theta}$, we have

$$\frac{1}{\sqrt{n}}\mathbf{u}(\hat{\theta}) = \frac{1}{\sqrt{n}}\mathbf{u}(\theta^*) - \{\mathcal{J}(\theta^*) + o_p(1)\}\sqrt{n}(\hat{\theta} - \theta^*);$$

if $\hat{\theta}$ is $\sqrt{n}$ -consistent, then

$$\frac{1}{\sqrt{n}}\mathbf{u}(\hat{\theta}) = \frac{1}{\sqrt{n}}\mathbf{u}(\theta^*) - \mathcal{J}(\theta^*)\sqrt{n}(\hat{\theta} - \theta^*) + o_p(1)$$

Corollary for two estimators

Similarly, for any two consistent estimators $\hat{\theta}_1$ and $\hat{\theta}_2$, we have

$$\frac{1}{\sqrt{n}}\mathbf{u}(\hat{\theta}_1) = \frac{1}{\sqrt{n}}\mathbf{u}(\hat{\theta}_2) - \{\mathcal{J}(\theta^*) + o_p(1)\}\sqrt{n}(\hat{\theta}_1 - \hat{\theta}_2);$$

if both estimators are $\sqrt{n}$ -consistent,

$$\frac{1}{\sqrt{n}}\mathbf{u}(\hat{\theta}_1) = \frac{1}{\sqrt{n}}\mathbf{u}(\hat{\theta}_2) - \mathcal{J}(\theta^*)\sqrt{n}(\hat{\theta}_1 - \hat{\theta}_2) + o_p(1)$$

Slutsky with asymptotically invertible matrices

- There is one more preliminary result that we need in order to prove asymptotic normality, an application of Slutsky's theorem for matrices that converge to a positive definite target
- **Theorem:** Suppose $\mathbf{y}_n \xrightarrow{d} \mathbf{y}$, where $\mathbf{y}$ is a $d \times 1$ random vector, $\mathbf{A}_n \xrightarrow{P} \mathbf{A}$, where $\mathbf{A}$ is a positive definite matrix, and that $\mathbf{y}_n = \mathbf{A}_n \mathbf{x}_n$. Then

$$\mathbf{x}_n \xrightarrow{d} \mathbf{A}^{-1} \mathbf{y}$$

- Note that this would be trivial if $\mathbf{A}_n$ were invertible for all n , but we don't know that — only that it converges to a positive definite matrix

Asymptotic normality of the MLE

- Our main result for today is proving the following central limit theorem-like result for the MLE of any smooth log-concave model (which is pretty simple given all the earlier results)
- **Theorem (Asymptotic normality of the MLE):** Suppose assumptions (A)-(D) from the previous lecture are met. Then the maximum likelihood estimator $\hat{\theta}$ satisfies

$$\sqrt{n}(\hat{\theta} - \theta^*) \xrightarrow{d} N(\mathbf{0}, \mathcal{J}(\theta^*)^{-1}).$$

- Note that (A)-(D) therefore ensure not just consistency, but $\sqrt{n}$ -consistency
- Note also another interpretation of the information: as information increases, the variance of the MLE $\hat{\theta}$ decreases

Non-standard problems

- Unlike the consistency proof, we do need differentiability requirements for asymptotic normality to hold
- For example, we remarked previously that for $X_i \stackrel{\text{iid}}{\sim} \text{Unif}(0, \theta)$, the MLE is consistent despite the likelihood not being continuous or differentiable at θ^*
- However, today's theorem does not hold for the uniform distribution:
 - Converges much faster: $\hat{\theta} - \theta^*$ is $O_p(1/n)$, not $O_p(1/\sqrt{n})$
 - $\mathcal{J}(\theta)$ is not even defined in the uniform case
 - Asymptotic distribution is not normal:

$$n(\theta^* - \hat{\theta}) \xrightarrow{d} \text{Exp}(1/\theta^*)$$

Influence function

- Since the MLE is $\sqrt{n}$ -consistent, we have

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \frac{1}{\sqrt{n}} \mathcal{J}^{-1}(\boldsymbol{\theta}^*) \mathbf{u}(\boldsymbol{\theta}^*) + o_p(1)$$

or in other words,

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + \frac{1}{n} \sum_i \mathcal{J}^{-1}(\boldsymbol{\theta}^*) \mathbf{u}_i(\boldsymbol{\theta}^*) + o_p(1/\sqrt{n})$$

- In statistics, this relationship is known as the *influence function* (formal definition on next slide); essentially, the derivative of the estimator with respect to the data

A connection with nonparametric statistics

- Influence functions play an important role in semiparametric and nonparametric statistics
- Suppose we are interested in estimating some function T of a distribution F ; the influence function is defined as

$$\text{IF}(x) = \lim_{\epsilon \rightarrow 0} \left[\frac{T\{(1 - \epsilon)F + \epsilon\delta_x\} - T(F)}{\epsilon} \right],$$

where δ_x is a distribution with all of its mass at x

- Then given some assumptions regarding the smoothness of T , the *von Mises expansion* essentially extends all of this Taylor expansion reasoning to the empirical CDF $\hat{F}$:

$$T(\hat{F}) = T(F) + \frac{1}{n} \sum \text{IF}_F(X_i) + o_p(1/\sqrt{n})$$

Asymptotically linear estimators

- Such estimators are said to be *asymptotically linear*
 - All asymptotically linear estimators are therefore asymptotically normal by the CLT
 - And the influence function determines the asymptotic variance
- Influence functions therefore provide a unifying thread connecting parametric likelihood theory with semiparametric and nonparametric statistics
- The influence function is potentially much more complicated in such problems, but for the parametric MLE, it has a simple form:

$$\text{IF}(x) = \mathcal{J}^{-1}(\theta^*) \mathbf{u}(\theta^*|x)$$

Multiple roots

- Finally, let's consider what happens if we drop assumption (D), that our likelihood is log-concave
- In this case, there are potentially many solutions to the likelihood equations

$$\mathbf{u}(\boldsymbol{\theta}) = \mathbf{0},$$

even if the MLE is unique

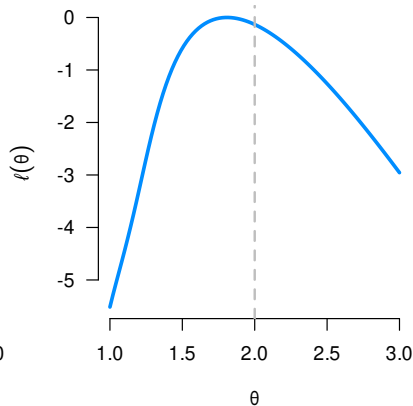
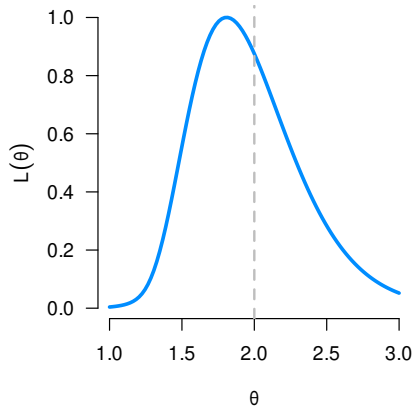
- Furthermore, as our counterexample at the beginning of the last lecture shows, if the likelihood has multiple modes there is no guarantee that the MLE is even consistent

Local log-concavity

- However, as you probably noticed, when proving consistency we only used assumption (D) at the very last step
- If we remove assumption (D), every step of the proof remains, except for the fact that at the end, all we can say is that there is a local maximum (i.e., a solution to the likelihood equations, not *the* solution to the likelihood equations) inside Θ^* that is consistent and asymptotically normal
- In other words, the likelihood isn't log-concave everywhere, but if the other conditions are met, and in particular if $\mathcal{J}(\theta^*)$ is positive definite, then there is a neighborhood Θ^* inside of which the likelihood is log-concave, and our theorems hold in a local sense

Revisiting our inconsistent MLE

The MLE isn't consistent but there is local solution which is:



Restating our earlier theorems

More general restatements of our earlier theorems:

- **Theorem (Consistency of the MLE):** Suppose assumptions (A)-(C) are met. Then with probability tending to 1, there exists a consistent sequence of solutions $\hat{\theta}_n$ to the likelihood equations:

$$\hat{\theta}_n \xrightarrow{P} \theta^*.$$

- **Theorem (Asymptotic normality of the MLE):** Suppose assumptions (A)-(C) from the previous lecture are met. Then with probability tending to 1, there exists a consistent sequence of solutions $\hat{\theta}_n$ to the likelihood equations satisfying

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} N(\mathbf{0}, \mathcal{J}(\theta^*)^{-1}).$$

Useful?

- Now, is this a useful generalization?
- Not necessarily — in theory, you can just choose the correct root, but that's an extremely strong assumption
 - Whatever algorithm we're using to maximize the likelihood is probably only going to return a single solution — we have no guarantees about its properties
 - Even if we were able to find all solutions of the likelihood equations, we have no way of knowing which one to choose

Useful? (cont'd)

- But also ...maybe?
- Suppose we have an estimator $\tilde{\theta}$, not the MLE, that we knew to be consistent
- We could, for example, pick the solution to the likelihood equations closest to $\tilde{\theta}$
- More ambitiously, we could take a Taylor expansion of the likelihood equations about the point $\tilde{\theta}$, then estimate θ via:

$$\hat{\theta} = \tilde{\theta} + \mathcal{I}_n(\tilde{\theta})^{-1} \mathbf{u}(\tilde{\theta})$$

- You can iterate this process if desired, repeating the above calculation until convergence (this is Newton's method), or just stop after one application (the “one-step estimator”)

One-step estimator theorem

- We'll skip the proof of this, but if $\tilde{\theta}$ is not only consistent but $\sqrt{n}$ -consistent, then our results hold not just for some mysterious, unknown root of the likelihood equations, but for the unique root defined on the previous slide
- **Theorem:** Suppose conditions (A)-(C) from the previous lecture are met, and that $\tilde{\theta}$ is a $\sqrt{n}$ -consistent estimator of θ . Define $\hat{\theta}_n = \tilde{\theta}_n + \mathcal{I}_n^{-1}(\tilde{\theta}_n)\mathbf{u}(\tilde{\theta}_n)$. Then

$$\sqrt{n}(\hat{\theta}_n - \theta^*) \xrightarrow{d} N(\mathbf{0}, \mathcal{J}(\theta^*)^{-1}).$$

- One can also use $\mathcal{J}(\tilde{\theta})$ to construct $\hat{\theta}$ and the theorem still holds

Cauchy example

- For example, suppose $X_i \stackrel{\text{iid}}{\sim} \text{Cauchy}(\theta)$; as we have already seen, this likelihood has multiple local maxima and it is unclear whether any given solution to the likelihood equations is consistent and asymptotically normal
- However, it can be shown that the sample median, $\tilde{\theta}$, is not the MLE but is a $\sqrt{n}$ -consistent estimator of θ
- Thus, the procedure on the previous slide can be used to obtain the likelihood root with known consistency and asymptotic normality properties

A word of caution

- The Cauchy distribution is a nice success story of maximum likelihood in the presence of multiple roots, but is arguably more of the exception than the rule
- In practice, it's inherently risky to go around constructing inference based on maximum likelihood in the presence of a likelihood with multiple maxima
- In other words, don't misinterpret the theory to mean that *any* root of the likelihood equations will be asymptotically normal