Maximum likelihood: Asymptotic normality

Patrick Breheny

October 19

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: $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\| \xrightarrow{\text{P}} 0$
- Today, we will see that under those same conditions, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ 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

Preliminary: Another Taylor series

- The main idea behind the proof is to take a Taylor series expansion not of the likelihood function, but rather the score function
- Since the score function is vector-valued, let us first derive an additional Taylor series expansion, this one pertaining to vector-valued functions
- **Theorem:** Suppose $\mathbf{f} : \mathbb{R}^d \to \mathbb{R}^k$ is twice differentiable on $N_r(\mathbf{x}_0)$, and that $\nabla^2 f$ is bounded on $N_r(\mathbf{x}_0)$. Then for any $\mathbf{x} \in N_r(\mathbf{x}_0)$,

$$\mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{x}_0) + \left[\nabla \mathbf{f}(\mathbf{x}_0) + O(\|\mathbf{x} - \mathbf{x}_0\|)\mathbf{1}_{d \times k}\right]^\top (\mathbf{x} - \mathbf{x}_0),$$

where $\mathbf{1}$ is a matrix of ones (i.e., every element equals one)

Application to the score

- Applying this expansion to the score vector, we obtain the following corollary, which is the main result driving the proof of asymptotic normality
- **Theorem:** Suppose regularity conditions (A)-(C) from the previous lecture are met. Then for any consistent estimator $\hat{\boldsymbol{\theta}}$, we have

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

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

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

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 $\hat{\boldsymbol{\theta}}$ satisfies

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \xrightarrow{\text{d}} \text{N}(\mathbf{0}, \boldsymbol{\mathscr{I}}(\boldsymbol{\theta}^*)^{-1}).$$

- We can now see another intuitive interpretation of the information: as information increases, the variance of the MLE $\hat{\boldsymbol{\theta}}$ decreases

Influence function

- Earlier, we saw that we can write

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \tfrac{1}{\sqrt{n}} \boldsymbol{\mathscr{I}}^{-1}(\boldsymbol{\theta}^*)\mathbf{u}(\boldsymbol{\theta}^*) + o_p(1)$$

- In other words,

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + \tfrac{1}{n}\sum_i \boldsymbol{\mathscr{I}}^{-1}(\boldsymbol{\theta}^*)\mathbf{u}_i(\boldsymbol{\theta}^*) + o_p(1/\sqrt{n})$$

- In statistics, the relationship between an estimate and the weight given to an individual observation is known as the *influence function* (formal definition on next slide)
- We can see here that for maximum likelihood estimators, the influence function has a very simple form (asymptotically): $\mathrm{IF}(x) = \boldsymbol{\mathscr{I}}^{-1}(\boldsymbol{\theta}^*)\mathbf{u}(\boldsymbol{\theta}^*|x)$

A connection with nonparametric statistics

- This forms an interesting theoretical bridge between maximum likelihood and nonparametric statistics
- Suppose we are interesting in estimating some function T of a distribution F; the influence function is defined as

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

  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 series reasoning to the empirical CDF $\hat{F}$:

$$T(\hat{F}) = T(F) + \frac{1}{n} \sum L_F(X_i) + o_p(1)$$

Non-standard problems

- Unlike the consistency proof, we do need differentiability requirements for asymptotic normality to hold
- For example, we remarked previously that for $X_i \overset{\text{iid}}{\sim} \text{Unif}(0, \theta)$, 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 $O_p(1/\sqrt{n})$
  - $\mathscr{I}(\theta)$ is not even defined in the uniform case
  - Asymptotic distribution is not normal: $n(\theta^* - \hat{\theta}) \xrightarrow{\text{d}} \text{Exp}(1/\theta)$

Alternative regularity conditions

- Thus, our conditions today are closer to being necessary conditions than those last time – there is less room to substitute weaker conditions and still obtain the result
- One exception worth noting is that we only really need a uniform bound on the second derivative for all the preceding proofs
- However, the proofs are simpler if we use the bound on the third derivatives that we already assumed in order to prove consistency

Local asymptotic normality

- A rather different approach to proving MLE asymptotics was pursued by Le Cam (1986), who abandoned the entire idea of $n \to \infty$ in favor of what he called local asymptotic normality (LAN)
- Instead of considering limits as $n \to \infty$, Le Cam showed that as the shape of the log-likelihood becomes more quadratic, the distribution of the MLE becomes more normal
- We won't go into any of the details here, but this is an interesting phenomenon to be aware of, since your sample size will never be infinite, but you can always plot the log-likelihood and assess how close to a quadratic it is

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

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

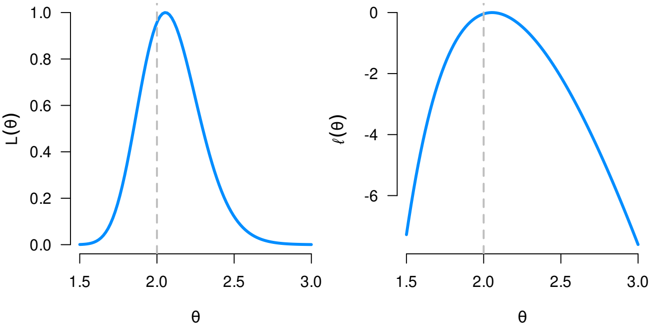
  even if the MLE is unique
- Furthermore, as our counterexample at the beginning of the last lecture shows, if the likelihood has multiple modes there is no guarantee that the MLE is even consistent

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 $\boldsymbol{\Theta}^*$ 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 $\mathscr{I}(\boldsymbol{\theta}^*)$ is positive definite, then there is a neighborhood $\boldsymbol{\Theta}^*$ 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

- With this in mind, we can offer 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 $\hat{\boldsymbol{\theta}}_n$ to the likelihood equations:

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

- **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 $\hat{\boldsymbol{\theta}}_n$ to the likelihood equations satisfying

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{0}, \boldsymbol{\mathscr{I}}(\boldsymbol{\theta}^*)^{-1}).$$

Useful?

- Now, is this a useful generalization?
- Not necessarily:
  - First of all, 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
  - Second of all, 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 $\tilde{\boldsymbol{\theta}}$, not the MLE, that we knew to be consistent
- We could, for example, pick the solution to the likelihood equations closest to $\tilde{\boldsymbol{\theta}}$
- More ambitiously, we could take a Taylor series expansion of the likelihood equations about the point $\tilde{\boldsymbol{\theta}}$, then estimate $\boldsymbol{\theta}$ via:

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

- 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 $\tilde{\boldsymbol{\theta}}$ 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 $\tilde{\boldsymbol{\theta}}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\theta}$. Define $\hat{\boldsymbol{\theta}}_n = \tilde{\boldsymbol{\theta}}_n + \boldsymbol{\mathcal{I}}(\tilde{\boldsymbol{\theta}}_n)^{-1}\mathbf{u}(\tilde{\boldsymbol{\theta}}_n)$. Then

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}^*) \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{0}, \boldsymbol{\mathscr{I}}(\boldsymbol{\theta}^*)^{-1}).$$

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

Cauchy example

- For example, suppose $X_i \overset{\text{iid}}{\sim} \text{Cauchy}(\theta)$; 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
- Every situation is different, of course, but my personal opinion is that it's inherently risky to go around constructing inference based entirely on maximum likelihood in the presence of a likelihood with multiple maxima, and runs the risk of producing completely misleading results