LECTURE 6: LARGE-SAMPLE PROPERTIES OF MLE

Michal Kolesár*

October 9, 2024

REFERENCE CB, Chapter 10.1; HMC, Chapter 6.1–6.2, 6.4, or H22, Chapter 10.12–10.13

Last time we discussed the maximum likelihood estimator (MLE) as one general way of coming up with an estimator. But is it any good? Last time we saw that one way to check its optimality under mean squared error (MSE) was to compute the information (Cramér-Rao) bound. Even though the bound is not possible to attain outside simple cases, by comparing the MSE of $\hat{\theta}_{ML}$ to it, we can check "near optimality" of MLEs in finite samples. But we have to do this for each model—there is no result saying that the MLE is "close" to the bound in general.

In this lecture, we will consider large-sample arguments for using the MLE that *are* generic, in that they apply to all "standard" models. We'll show, in particular that under regularity conditions, MLEs are consistent and asymptotically normal. What's more, it'll turn out to be minimax in large samples. In my view, this result is justification for using the MLE.

1. CONSISTENCY AND ASYMPTOTIC NORMALITY

Let $X = (X_1, \ldots, X_n)$ be a random sample, with joint probability density function (PDF) $f_n(x \mid \theta) = \prod_{i=1}^n f(x_i \mid \theta)$, and $\theta \in \Theta$. The log-likelihood is $\ell_n(\theta \mid x) = \sum_{i=1}^n \log f(x_i \mid \theta)$. The MLE is, by definition, $\hat{\theta}_{ML} = \arg\max_{\theta\in\Theta} \ell_n(\theta \mid x)$. We let $\mathcal{I}(\theta) = \operatorname{var}(\partial \log f(X_i \mid \theta)/\partial\theta)$ denote the information based on a single observation. The first information equality is $E[\mathcal{S}_n(\theta \mid X)] = 0$, where $\mathcal{S}_n(\theta \mid X) = \partial\ell_n(\theta \mid X)/\partial\theta$. Since $\mathcal{S}_n(\hat{\theta}_{ML} \mid X)$ if the likelihood is maximized in the interior of the parameter space, as we remarked before, MLE can be viewed as a method of moments estimator based on the first information equality. So we can expect that the MLE is consistent. Indeed, the theorem below gives the consistency result for the MLE. In what follows, it will be convenient to distinguish between an arbitrary element θ of the parameter space, and the parameter that actually generated the data. To that end, we'll denote by θ_0 the "true" value of θ.

*Email: mkolesar@princeton.edu.

Remark 1. The regularity conditions for the theorems below are from Newey and McFadden (1994, Theorem 2.5, Theorem 3.3), those given in the textbooks are stronger than necessary.

Theorem 2 (Consistency, CB Theorem 10.1.6, HMC, Theorem 6.1.2). Suppose that $X_1, \ldots, X_n$ is a random sample with density $f(x_i \mid \theta_0)$. Suppose that

1. *θ is identifiable, i.e. for any $\theta \neq \theta_0$, $f(x_i \mid \theta) \neq f(x_i \mid \theta_0)$;*
2. *$\log f(X_i \mid \theta)$ is continuous at each θ with probability one, and $E_{\theta_0}[\sup_{\theta \in \Theta} |\log f(X_i \mid \theta)|] < \infty$; and*
3. *$\theta_0 \in \Theta$, which is compact.*

Then $\hat{\theta}_{ML} \xrightarrow{p} \theta_0$.

Proof. A rigorous proof of MLE consistency will be given in later courses, we give a sketch. Let $Q_n(\theta) = n^{-1}\ell_n(\theta) - n^{-1}\ell_n(\theta_0) = n^{-1}\sum_i \log(f(x_i \mid \theta)/f(x_i \mid \theta_0))$. By the law of large numbers, $\hat{Q}_n(\theta)$ converges in probability to $Q(\theta) = E_{\theta_0}[\log(f(x_i \mid \theta)/f(x_i \mid \theta_0))]$. By the identifiability assumption, $f(x_i \mid \theta)/f(x_i \mid \theta_0)$ has a non-degenerate distribution, so that by a strict version of Jensen's inequality, for $\theta \neq \theta_0$,

$$Q(\theta) < \log(E_{\theta_0}[f(x_i \mid \theta)/f(x_i \mid \theta_0)]) = \log \int_{\mathcal{X}} \frac{f(x_i \mid \theta)}{f(x_i \mid \theta_0)} f(x_i \mid \theta_0)\, \mathrm{d}x_i = \log 1 = 0.$$

So $Q_n(\theta) \xrightarrow{p} Q(\theta)$, and the probability limit $Q(\theta)$ is maximized at the true parameter value θ_0. That is $\theta_0 = \operatorname{argmax}_\theta \operatorname{plim} Q_n(\theta)$. This doesn't quite imply that $\hat{\theta}_{ML} \xrightarrow{p} \theta_0$, i.e. that $\theta_0 = \operatorname{argmax}_\theta \operatorname{plim} Q_n(\theta) = \operatorname{plim} \operatorname{argmax}_\theta Q_n(\theta) = \operatorname{plim} \hat{\theta}_{ML}$. To be able to swap the argmax and the plim, we would need to strengthen the convergence to $\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \xrightarrow{p} 0$. □

Once we know that the estimator is consistent, we can think about its asymptotic distribution:

Theorem 3 (Asymptotic normality, CB, Theorem 10.1.12, HMC, Theorem 6.2.2). Consider the setup and assumptions from the previous theorem. In addition, assume that

1. *θ_0 is in the interior of Θ;*
2. *$f(x \mid \theta)$ is twice times continuously differentiable with respect to θ, and regularity conditions similar to Condition 2 of Theorem 2 hold for the score and the second derivative of the log-likelihood;*
3. *$\mathcal{I}(\theta_0)$ is non-singular.*

Then

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \Rightarrow \mathcal{N}(0, \mathcal{I}^{-1}(\theta_0)).$$

Proof. We'll sketch the proof for θ scalar. Let $\mathcal{S}_n(\theta) = \sum_{i=1}^n \partial \log f(X_i \mid \theta)/\partial\theta$. By definition of $\hat{\theta}_{ML}$, $\mathcal{S}_n(\hat{\theta}_{ML}) = 0$ (at least in large samples—why?). By the mean value theorem,

$$0 = \frac{1}{n}\mathcal{S}_n(\hat{\theta}_{ML}) = \frac{1}{n}\mathcal{S}_n(\theta_0) + \frac{1}{n}\frac{\partial \mathcal{S}_n(\tilde{\theta}_n)}{\partial\theta}(\hat{\theta}_{ML} - \theta_0),$$

where $\tilde{\theta}_n$ sits between θ_0 and $\hat{\theta}_{ML}$. Now, since $\hat{\theta}_{ML} \xrightarrow{p} \theta_0$, it means that $\tilde{\theta}_n \xrightarrow{p} \theta_0$. Furthermore, $\frac{1}{n}\partial \mathcal{S}_n(\theta)/\partial\theta \xrightarrow{p} E_{\theta_0}[\partial^2 \log f(X_i \mid \theta)/\partial\theta]$ at each θ by the law of large numbers. This should imply that $\frac{1}{n}\partial \mathcal{S}_n(\tilde{\theta}_n)/\partial\theta \xrightarrow{p} E_{\theta_0}[\partial^2 \log f(X_i \mid \theta_0)/\partial\theta] = -\mathcal{I}(\theta_0)$. This doesn't follow from the continuous mapping theorem since we have a sequence of functions $\frac{1}{n}\partial \mathcal{S}_n(\theta)/\partial\theta$ instead of just one function. Proving it requires the concept of asymptotic equicontinuity which we do not cover in this class. Suppose we nonetheless believe this result. Then in large samples, $\frac{1}{n}\partial \mathcal{S}_n(\tilde{\theta})/\partial\theta \neq 0$, so that rearranging the preceding display yields

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) = -\left(\frac{1}{n}\frac{\mathcal{S}_n(\tilde{\theta}_n)}{\partial\theta}\right)^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathcal{S}(\theta_0 \mid X_i)$$

Now, by the first information equality, $E[\mathcal{S}(\theta_0 \mid X_i)] = 0$. Thus, by the central limit theorem, $\frac{1}{\sqrt{n}}\sum_{i=1}^{n} \mathcal{S}(\theta_0 \mid X_i) \Rightarrow \mathcal{N}(0, \mathcal{I}(\theta_0))$. Since $\frac{1}{n}\mathcal{S}_n(\tilde{\theta}_n)/\partial\theta \xrightarrow{p} -\mathcal{I}(\theta_0)$, by Slutsky's theorem,

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \Rightarrow \mathcal{N}(0, \mathcal{I}^{-1}(\theta_0)). \qquad \square$$

MLEs are not necessarily unbiased in finite samples, but the result implies that the asymptotic bias is zero. Furthermore, we have shown that MLEs have an asymptotic variance equal to the information bound for unbiased estimators. So, heuristically, MLE is "asymptotically efficient": we don't expect to find other asymptotically unbiased estimator can have smaller asymptotic variance. We'll make this claim a bit more formal in Section 2 below.

Example 1. Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}(1/\theta, 1)$. The MLE is $\hat{\theta} = 1/\overline{X}_n$ (by the invariance property of MLE). The mean of $\hat{\theta}_n$ doesn't exist (and hence the bias is not defined for any n, no matter how large n is), yet we have (either by delta method, or by applying Theorem 3)

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \Rightarrow \mathcal{N}(0, \theta_0^4). \qquad \boxtimes$$

Example 2. Let $X_1, \ldots, X_n$ be a random sample from a distribution with PDF $f(x \mid \theta) = \theta \exp(-\theta x)$. This distribution is called exponential, with $\theta \in \mathbb{R}_+$, and the support is also $\mathbb{R}_+$. The cdf is $1 - e^{-\theta x}$. We have

$$\mathcal{S}(\theta \mid x_i) = 1/\theta - x_i,$$

and $\partial\mathcal{S}/\partial\theta = -1/\theta^2$, so that Fisher information is $\mathcal{I}(\theta) = 1/\theta^2$. From the previous display and the second-order condition, it follows that $\hat{\theta}_{ML} = 1/\overline{X}_n$. Its asymptotic distribution is given by $\sqrt{n}(\hat{\theta}_{ML} - \theta_0) \Rightarrow \mathcal{N}(0, \theta_0^2)$. $\boxtimes$

Example 3. For asymptotic normality of MLE, we need the likelihood to be smooth, which in Theorem 3 is formalized by requiring f to be twice continuously differentiable. Although this can be slightly relaxed, a certain amount of smoothness of $\theta \mapsto f(\cdot, \theta)$ is essential: otherwise MLE may be neither asymptotically normal, nor asymptotically efficient. In contrast, we do not need smoothness for consistency.

For example, consider a random sample $X_1, \ldots, X_n$ from $U[0, \theta]$. Then $\hat{\theta}_{ML} = X_{(n)}$. The conditions of Theorem 2 hold, so $\hat{\theta}_{ML} \xrightarrow{p} \theta_0$. On the other hand, $\sqrt{n}(\hat{\theta}_{ML} - \theta_0)$ is

always non-positive, so it certainly cannot be asymptotically mean-zero normal. In fact, for any $x \leq 0$, we have

$$P_\theta\left(n(X_{(n)} - \theta) \leq x\right) = \prod_{i=1}^{n} P_\theta\left(n(X_i - \theta) \leq x\right)$$
$$= P_\theta\left(X_i \leq \theta + x/n\right)^n = \left(\frac{\theta + x/n}{\theta}\right)^n \to e^{x/\theta},$$

where last equality holds for n large enough, since for any fixed $x \leq 0$, $\theta + x/n \in [0, \theta]$ for n large enough, and $P_\theta(X_i \leq y) = y/\theta$ for any $y \in [0, \theta]$. The previous display implies that for any $x \geq 0$,

$$P_{\theta_0}(-n(X_{(n)} - \theta_0) \leq x) = 1 - P_{\theta_0}(-n(X_{(n)} - \theta_0) \geq x)$$
$$= 1 - P_{\theta_0}(n(X_{(n)} - \theta_0) \leq -x) \to 1 - e^{-x/\theta_0}.$$

Comparing this with the cdf of an exponential distribution from the previous example, we see that $n(X_{(n)} - \theta_0) \Rightarrow -\text{expo}(1/\theta_0)$. So the rate of convergence is faster than $\sqrt{n}$, and the limiting distribution is not normal, but exponential. Is MLE asymptotically efficient in the sense that the MSE of the limit distribution is as small as possible? Let $Z \sim \text{expo}(1/\theta_0)$. The asymptotic MSE is $E[Z^2] = 2\theta_0^2$. On the other hand, consider the estimator $\hat{\theta} = (n+1)X_{(n)}/n$. Since $X_{(n)} \xrightarrow{p} \theta_0$, it follows by Slutsky's theorem that

$$-n(\hat{\theta} - \theta_0) = -n(X_{(n)} - \theta_0) - X_{(n)} \Rightarrow Z - \theta_0.$$

Therefore, the asymptotic MSE is $E(Z - \theta_0)^2 = \text{var}(Z) = \theta_0^2$, half as big as that of MLE. It is possible to show that $\hat{\theta}$ is indeed asymptotically efficient under quadratic loss. ⊠

Example 4. Let us consider what might happen if the true parameter value were on the boundary of Θ. Let $X_1, \ldots, X_n$ be a random sample from distribution $\mathcal{N}(\theta, 1)$ with $\theta \geq 0$. As an exercise, check that $\hat{\theta}_{ML} = \max\{\overline{X}_n, 0\}$. Suppose that $\theta_0 = 0$. Then $\sqrt{n}(\hat{\theta}_{ML} - \theta_0)$ is always nonnegative. So it does not converge to mean zero normal distribution. Instead,

$$\sqrt{n}(\hat{\theta}_{ML} - \theta_0) = \max\{\sqrt{n}\overline{X}_n, 0\} \Rightarrow \max\{\mathcal{N}(0,1), 0\}. \qquad \boxtimes$$

Example 5. Finally, note that we implicitly assumed both theorems that Θ is fixed, i.e. independent of n. In particular, the number of parameters should not depend on n. Indeed, let

$$X_i = \begin{pmatrix} X_{1i} \\ X_{2i} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_i \\ \mu_i \end{pmatrix}, \begin{pmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{pmatrix}\right)$$

for $i = 1, \ldots, n$, and $X_1, \ldots, X_n$ be mutually independent. In the problem set, you'll show that if the sample size n increases to infinity, the MLE for σ^2 is inconsistent in this case, though a consistent estimator for σ^2 exists. ⊠

2. LARGE-SAMPLE OPTIMALITY OF MLE [[TODO: SMOOTH OUT WRITING]]

Loosely speaking, Theorem 3 suggests that MLE is asymptotically unbeatable, for two reasons, First, it achieves the Cramér-Rao bound in large samples: it seems like it shouldn't be possible to construct another estimator with a "better" asymptotic distribution, as Fisher conjectured in the 1920s.

The second reason is that the MLE estimator $\hat{\theta}_{ML}$ is a sufficient statistic, and in large samples it has the distribution

$$\hat{\theta}_{ML} \approx \mathcal{N}(\theta, \mathcal{I}(\theta_0)^{-1}/n). \tag{1}$$

Since $\hat{\theta}_{ML}$ is sufficient, we expect that the problem of estimating θ is the same as the finite-sample problem of estimating θ based on a single observation $X \sim \mathcal{N}(0, \Sigma)$, with Σ corresponding to the inverse of the information. Here we know that MLE is minimax, and also admissible if we are interested in a single θ_j.

So we expect that in large samples, we (i) can't find another estimator that's asymptotically unbiased, with a smaller asymptotic variance than $\hat{\theta}_{ML}$, and (ii) $\hat{\theta}_{ML}$ should be "asymptotically minimax".

Both properties turn out to more or less be true. However claim (i) is only true if you restrict attention to estimators that are "regular": this took a while to show, basically until Hájek (1970). However, this restriction is not particularly appealing, since many clever estimators, such as those based on model selection or pre-testing, are not regular.

Claim (ii) can be justified by a powerful *local asymptotic minimax theorem*, that considerably strengthens Theorem 3. In particular, this theorem shows that if we look at the *worst-case risk* for any bowl-shaped loss function (this includes squared loss, absolute value loss etc) over a shrinking neighborhood of the true θ_0, consider this risk as $n \to \infty$, one cannot beat the large-sample risk of MLE. Properly stating and understanding this result requires developing quite a bit of powerful machinery, called "limits of experiments", developed by Lucien Le Cam in the 1960s and 1970s. We'll leave that to a second-year course.

The takeaway for us for now is that, loosely speaking, the limiting distribution of the MLE is "asymptotically minimax": even though there exist "irregular" estimators that can beat this limiting distribution for some select values of θ_0, this comes at the expense of much worse performance for neighboring values of θ_0, so that, if we take the worst-case risk in the neighborhood of the true θ_0, they are dominated by MLE.

Example 6 (Hodges estimator). The Hodges estimator was first described in Le Cam (1953), who says that Joseph Hodges discovered it two years earlier, in 1951. It punctured Fisher's original claims about asymptotic optimality of MLE. It is the simplest example of a model selection estimator, picking between two models: $X_i \sim \mathcal{N}(0,1)$ and $X_i \sim \mathcal{N}(\theta, 1)$: $\hat{\theta}_H = 0$ if $|\overline{X}_n| < n^{-1/4}$ (we pick the simpler models), and $\hat{\theta}_H = \overline{X}_n$ otherwise (we pick the more general model). The estimator buys us better performance at $\theta_0 = 0$

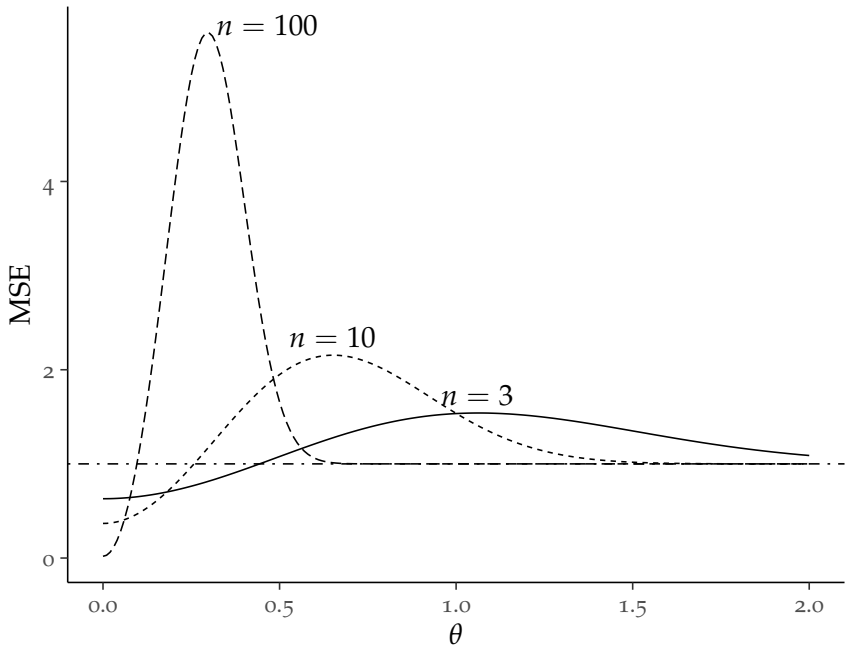

Figure 1: The normalized mean squared error $nE_\theta[(\hat{\theta}_H - \theta)^2]$ for the Hodges' estimator as a function of the sample size n and the true mean θ. The horizontal line denotes the normalized mean squared error for the maximum likelihood estimator $\overline{X}_n$.

at the expense of (much) worse performance near 0: see Figure 1. This feature is shared by basically all model selection procedures. ⊠

Remark 4 (Importance of the normal means model). Equation (1) suggests that the normal means model is important. Indeed, it turns out that in "regular models", in large samples, large-sample properties of tests and estimators are matched by tests and estimators based on the normal model. This is the sense in which the normal model is a "limiting experiment". For this reason, we'll spend some time thinking about optimality in this normal model—if we figure out what to do in this model, we'll know what to do in general in "nice" parametric models.

REFERENCES

Hájek, Jaroslav. 1970. "A Characterization of Limiting Distributions of Regular Estimates." *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 14 (4): 323–330. https://doi.org/10.1007/BF00533669.

Le Cam, Lucien M. 1953. "On Some Asymptotic Properties of Maximum Likelihood Estimates and Related Bayes Estimates." Edited by Jerzy Neyman, M Loève, and O Struve. *University of California publications in statistics,* no. 11 (January): 277–330.

Newey, Whitney K., and Daniel L. McFadden. 1994. "Large Sample Estimation and Hypothesis Testing." Chap. 36 in *Handbook of Econometrics,* edited by Robert F. Engle and Daniel L. McFadden, 4:2111–2245. New York, NY: Elsevier. https://doi.org/10.1016/S1573-4412(05)80005-4.