18.650: Fundamentals of Statistics	October 2, 2023

Lecture 9— Properties of the MLE

Prof. Philippe Rigollet — *Scribe: Anya Katsevich*

1 Overview.

The MLE $\hat{\theta}^{\text{MLE}}$ satisfies three important properties:

1. Consistency: $\hat{\theta}^{\text{MLE}} \xrightarrow{\mathbb{P}} \theta^*$.
2. Asymptotic normality: $\sqrt{n}(\hat{\theta}^{\text{MLE}} - \theta^*) \rightsquigarrow \mathcal{N}(0, \sigma^2_{\text{MLE}})$.
3. Asymptotic efficiency: consider any other estimator $\hat{\theta}$ which is also asymptotically normal, i.e. such that $\sqrt{n}(\hat{\theta} - \theta^*) \rightsquigarrow \mathcal{N}(0, \sigma^2)$. Then $\sigma^2 \geq \sigma^2_{\text{MLE}}$.

In this lecture we go over consistency and asymptotic normality. The key concept involved in the latter property is the *Fisher information*.

Remarks.

1. **Caution:** the MLE is not always asymptotically normal! As a rule of thumb, if the log likelihood is differentiable, then the MLE is asymptotically normal. In this case, you can find the MLE by setting $\nabla \ell_n(\theta)$ to zero. If the log likelihood is not differentiable then the MLE is not guaranteed to be asymptotically normal.

   As an example, the log likelihood for the model $\{\text{Unif}[0, \theta] \mid \theta > 0\}$ is not differentiable. The MLE in this case is $\hat{\theta}^{\text{MLE}} = \max_i X_i$, and one can show that $\sqrt{n}(\max_i X_i - \theta) \rightsquigarrow 0$. In fact, there can be no sequence $a_n \to \infty$ such that $a_n(\max_i X_i - \theta)$ converges to a normal distribution because $\max_i X_i - \theta$ is always negative, but a normal distribution takes both positive and negative values.

2. Different MLEs for different statistical models can be studied on a case by case basis to determine whether they are asymptotically normal, and to compute their asymptotic variance. E.g. the Bernoulli MLE is $\hat{\theta}^{\text{MLE}} = \bar{X}_n$, and you can directly invoke the CLT to get the asymptotic variance. But we will see in the next section that there is a general formula for the asymptotic variance of the MLE.

2 Consistency and asymptotic normality

2.1 Proof of Consistency

Let $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{\theta^*}$, and let X denote a generic random variable with distribution $\mathbb{P}_{\theta^*}$. By the LLN and the calculations from Lecture 8 (see equation (3) on page 5), it holds

$$\tilde{\ell}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log f_\theta(X_i) \overset{n\to\infty}{\to} \mathbb{E}_{\theta^*}[\log f_\theta(X)] = -D_{\mathrm{KL}}(\mathbb{P}_{\theta^*} \,||\, \mathbb{P}_\theta) + \text{const.} \tag{1}$$

where the constant term depends on θ^* but not on the variable θ. (We have defined $\tilde{\ell}_n = \frac{1}{n}\ell_n$, i.e. the $1/n$-normalized log likelihood.) Since $\tilde{\ell}_n$ converges to the negative of the KL divergence (neglecting the constant term), we deduce that the *maximizer* of $\tilde{\ell}_n$ converges to the *maximizer* of $-D_{\mathrm{KL}}(\mathbb{P}_{\theta^*} \,||\, \mathbb{P}_\theta)$:

$$\begin{aligned}\hat{\theta}^{\mathrm{MLE}} &= \operatorname{argmax}_\theta \tilde{\ell}_n(\theta) \overset{n\to\infty}{\to} \operatorname{argmax}_\theta \left[-D_{\mathrm{KL}}(\mathbb{P}_{\theta^*} \,||\, \mathbb{P}_\theta)\right] \\ &= \operatorname{argmin}_\theta D_{\mathrm{KL}}(\mathbb{P}_{\theta^*} \,||\, \mathbb{P}_\theta) = \theta^*.\end{aligned}$$

Here we used that the θ which minimizes the KL divergence between $\mathbb{P}_{\theta^*}$ and $\mathbb{P}_\theta$ is the point θ^* itself.

2.2 Asymptotic normality and asymptotic variance

Definition 2.1: Fisher information

Consider a statistical model with pdf f_θ. If θ is a one-dimensional parameter, the Fisher information is defined as

$$I(\theta) = \mathbb{E}_\theta\left[-\frac{d^2}{d\theta^2}\log f_\theta(X)\right] = \mathbb{V}_\theta\left[\frac{d}{d\theta}\log f_\theta(X)\right].$$

If θ is a multi-dimensional parameter, then the Fisher information *matrix* is defined as

$$I(\theta) = \mathbb{E}_\theta[-\nabla_\theta^2 \log f_\theta(X)] = \mathbb{V}_\theta[\nabla \log f_\theta(X)].$$

The Fisher information (matrix) is important because the asymptotic variance of the MLE is given by its inverse, as the following theorem shows:

Theorem 2.2: Asymptotic variance of MLE

Suppose the ground truth parameter is θ^*. For a sufficiently regular model (i.e. a well-behaved density f_θ), the MLE $\hat{\theta}^{\mathrm{MLE}}$ has the following limit:

$$\sqrt{n}\left(\hat{\theta}^{\mathrm{MLE}} - \theta^*\right) \rightsquigarrow \mathcal{N}(0, I(\theta^*)^{-1}).$$

Remarks.

> Regarding Theorem 2.2: note that in the one-dimensional case, $I(\theta^*)$ is just a scalar, so $I(\theta^*)^{-1} = 1/I(\theta^*)^{-1}$. In the multidimensional case, $I(\theta^*)^{-1}$ is the matrix inverse of $I(\theta^*)$.
>
> Regarding Definition 2.1 of the Fisher information:
>
> - The fact that the second derivative of $\log f_\theta$ equals the variance of the first derivative of $\log f_\theta$ is a property which needs to be proved (it's not immediately obvious just by looking at the formulas).
> - Note that we take the derivative of $f_\theta(X)$ with respect to θ, *not* with respect to X. The notation $\mathbb{E}_\theta$ and $\mathbb{V}_\theta$ indicates that the random variable X has pdf f_θ.
> - In the multi-dimensional case, let's check that both formulas for $I(\theta)$ really do give matrices. The quantity $\nabla_\theta^2 \log f_\theta(X)$ is indeed a matrix (the Hessian), so its expectation is also a matrix. Meanwhile, $\nabla_\theta \log f_\theta(X)$ is a vector, and the variance of a random vector is actually a whole covariance matrix.

2.3 Proof of asymptotic normality and asymptotic variance formula

For brevity, in this section, we write $\hat{\theta}$ to denote $\hat{\theta}^{\mathrm{MLE}}$. Our goal is to prove

$$\sqrt{n}(\hat{\theta} - \theta^*) \rightsquigarrow \mathcal{N}_k(0, I(\theta^*)^{-1}).$$

We introduce the following two functions:

$$\begin{aligned}\tilde{\ell}_n(\theta) &:= \frac{1}{n}\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^n \log f_\theta(X_i), \\ \ell(\theta) &= \mathbb{E}_{\theta^*}\left[\log f_\theta(X)\right].\end{aligned}$$

The first function is just the normalized sample log likelihood. The second function is the expectation of the first one (recall that $X_i \overset{\text{i.i.d.}}{\sim} X \sim \mathbb{P}_{\theta^*}$). The second function is known as the *population* log likelihood. Now, note that

$$\nabla\tilde{\ell}_n(\hat{\theta}) = 0, \qquad \nabla\ell(\theta^*) = 0. \tag{2}$$

This is because the log likelihood is maximized at $\hat{\theta}$ (this is the definition of $\hat{\theta}$) and because the population log likelihood $\ell(\theta)$ is the negative of the KL divergence, which is minimized at θ^*.

Next, let us Taylor expand $\nabla\tilde{\ell}_n(\hat{\theta})$ around the point θ^*:

$$\nabla\tilde{\ell}_n(\hat{\theta}) \approx \nabla\tilde{\ell}_n(\theta^*) + \nabla^2\tilde{\ell}_n(\theta^*)(\hat{\theta} - \theta^*).$$

Rearranging terms, this implies

$$\begin{aligned} \hat{\theta} - \theta^* &\approx \nabla^2\tilde{\ell}_n(\theta^*)^{-1}\left(\nabla\tilde{\ell}_n(\hat{\theta}) - \nabla\tilde{\ell}_n(\theta^*)\right) \\ &= -\nabla^2\tilde{\ell}_n(\theta^*)^{-1}\nabla\tilde{\ell}_n(\theta^*). \end{aligned} \tag{3}$$

In the second line we dropped $\nabla\tilde{\ell}_n(\hat{\theta})$, since it equals zero. Now, recall that

$$\nabla\tilde{\ell}_n(\theta^*) = \frac{1}{n}\sum_{i=1}^{n} \left(\nabla_\theta \log f_\theta(X_i)\right)\big|_{\theta=\theta^*}.$$

This is just an average of i.i.d. random vectors whose mean and covariance matrix are given by

$$\begin{aligned} &\mathbb{E}\left[\nabla_\theta \log f_\theta(X_i)\big|_{\theta=\theta^*}\right] = \nabla_\theta \mathbb{E}[\log f_\theta(X_i)]\big|_{\theta=\theta^*} = \nabla\ell(\theta)\big|_{\theta=\theta^*} = 0, \\ &\mathbb{V}\left[\nabla_\theta \log f_\theta(X_i)\big|_{\theta=\theta^*}\right] = I(\theta^*) \end{aligned}$$

The second line is just by the definition of Fisher information matrix. Therefore,

$$\sqrt{n}\nabla\tilde{\ell}_n(\theta^*) \rightsquigarrow \mathcal{N}(0, I(\theta^*)) \tag{4}$$

by the CLT. Finally, note that

$$\nabla^2\tilde{\ell}_n(\theta^*) \xrightarrow{\mathbb{P}} \nabla^2\ell(\theta^*) = \nabla_\theta^2 \mathbb{E}[\log f_\theta(X)]\big|_{\theta=\theta^*} = I(\theta^*) \tag{5}$$

by the LLN. Combining (4) and (5) in (3) and applying Slutsky, we infer that

$$\begin{aligned} \sqrt{n}\left(\hat{\theta} - \theta^*\right) &\approx -\nabla^2\tilde{\ell}_n(\theta^*)^{-1}\left[\sqrt{n}\nabla\tilde{\ell}_n(\theta^*)\right] \\ &\rightsquigarrow I(\theta^*)^{-1}\mathcal{N}(0, I(\theta^*)) = \mathcal{N}(0, I(\theta^*)^{-1}). \end{aligned}$$

To get the final equality, we used that if $Y \sim \mathcal{N}(0, \Sigma)$ then $AY \sim \mathcal{N}(0, A\Sigma A^T)$.