36-700: Probability and Mathematical Statistics I **Fall 2016**

Lecture 12: September 26

Lecturer: Siva Balakrishnan

12.1 Review and Outline

Last class we discussed:

- The minimax risk of an estimator
- The Bayes risk of an estimator
- The Bayes rule with respect to a prior
- Connections between minimax and Bayes estimators

Today we are going to discuss in detail some properties of the MLE. Many of these results are in Chapter 9.4-9.8 of the Wasserman book. In order to get the main ideas across while still conveying some of the proof techniques we will be a bit non-rigorous in various places. In particular, many of the results we state hold under **regularity conditions**, i.e. typically they require that the space of possible parameters is compact, that the true parameter is in the interior of the space of possible parameters and that the likelihood function is smooth. Advanced mathematical statistics textbooks will be more rigorous about these conditions.

Finally, in this lecture we will also ignore computational aspects. The MLE in most cases is not calculated analytically. Rather, we use an algorithm like gradient ascent to try to maximize the likelihood. In some cases (for instance, when the likelihood is concave) we can guarantee that this algorithm will in fact find a parameter sufficiently close to the MLE. In general however computing the MLE can be intractable.

12.2 Consistency

The first important property of the MLE is that it is consistent under regularity conditions. As a reminder: consistency means that the MLE converges in probability to the true parameter.

In order to establish this we first need to recall the definition of the KL divergence. For two densities f, g we defined the KL divergence as:

$$D(f||g) = \int_x f(x) \log \left(\frac{f(x)}{g(x)} \right) dx.$$

For us, we will use this on parametric densities. In this case, two parameters θ_1, θ_2 will induce the KL divergence $D(f_{\theta_1}||f_{\theta_2})$. Two properties of the KL divergence that we will use are that the KL divergence is always positive and equal to 0 iff $f = g$ (almost everywhere).

We say a statistical model is **identifiable** if we have that if $\theta_1 \neq \theta_2$ then $D(f_{\theta_1}||f_{\theta_2}) > 0$. In words, this condition implies that different parameter values induce different distributions. We assume that the statistical model is identifiable in the sequel.

Let θ^* denote the true (unknown) value of the parameter. Then we can see that maximizing the log-likelihood is equivalent to maximizing the following function:

$$M_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log \frac{f_\theta(X_i)}{f_{\theta^*}(X_i)},$$

since the term in the denominator is just a constant. We can see that by the LLN:

$$M_n(\theta) \to \mathbb{E}_{f_{\theta^*}} \log \frac{f_\theta(X_i)}{f_{\theta^*}(X_i)} = -D(f_{\theta^*}||f_\theta).$$

Now, since the KL divergence is always positive, minimized when $\theta = \theta^*$, and uniquely minimized at this point (by identifiability) we can argue that in probability the maximum likelihood will be achieved at θ^*.

In order to be more rigorous we need that the convergence of $M_n(\theta)$ to $-D(f_{\theta^*}||f_\theta)$ hold for all possible values of the parameter θ. For a single θ this follows from the LLN; when we need the convergence in probability to hold over the entire parameter space we need to derive what are called *uniform* laws of large numbers (ULLNs). We have seen the techniques to do this: usually some combination of Hoeffding's inequality or Chebyshev's inequality and the union bound is used.

Let me first point out the main take away from this proof: **for large sample sizes maximizing the likelihood is roughly equivalent to minimizing the KL divergence to the *true model.***

The method-of-moments does not typically have such an interpretation so proving general results about its consistency is much more difficult.

12.3 Fisher Information

The MLE is often asymptotically Gaussian. In order to express its variance we need to first define what is called the Fisher information matrix.

We first define the **score function**:

$$s_\theta(X) = \frac{\partial \log f_\theta(X)}{\partial \theta}.$$

The score function at θ has mean zero when $X \sim f_\theta$, i.e.,

$$\mathbb{E}_\theta\left[s_\theta(X)\right] = 0.$$

One can see this by differentiating the expression that $1 = \int_\theta f_\theta(x)dx$, with respect to θ. This yields:

$$\begin{aligned} 0 &= \int_\theta \frac{\partial f_\theta(x)}{\partial \theta} dx \\ &= \int_\theta \frac{\partial \log f_\theta(x)}{\partial \theta} f_\theta(x)dx \\ &= \mathbb{E}_\theta\left[s_\theta(X)\right]. \end{aligned}$$

One can also interpret maximum likelihood as approximately trying to satisfy this condition, i.e., roughly the MLE $\hat{\theta}$ satisfies something like:

$$\frac{1}{n}\sum_{i=1}^n s_{\hat{\theta}}(X_i) = 0,$$

which by the LLN is like trying to solve the equation:

$$\mathbb{E}_{\theta^*} s_{\hat{\theta}}(X) = 0.$$

Under identifiability conditions this equation will be uniquely solved at $\hat{\theta} = \theta^*$.

The Fisher information is just the variance of the score function, i.e.:

$$I(\theta) = \mathrm{Var}_\theta\left(s_\theta(X)\right) = \mathbb{E}\left(s_\theta(X)^2\right).$$

Under some mild regularity conditions the Fisher information can also be written as:

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

Let us prove this: differentiating the score equation we obtain that,

$$\begin{aligned} \frac{\partial}{\partial \theta}\mathbb{E}_\theta\left[\frac{\partial}{\partial \theta}\log f_\theta(X)\right] &= 0, \\ \int_X \left[\frac{\partial}{\partial \theta}\left(f_\theta(X)\frac{\partial}{\partial \theta}\log f_\theta(X)\right)\right] dX &= 0, \end{aligned}$$

which after some algebra yields the expression that:

$$\mathbb{E}_\theta \left[\frac{\partial^2}{\partial \theta^2} \log f_\theta(X) \right] + \mathbb{E}_\theta \left[\frac{\partial}{\partial \theta} \log f_\theta(X) \right]^2 = 0.$$

This gives us the desired result.

So effectively, the Fisher information measures the variance of the score function, but also the Hessian or the curvature of the log-likelihood.

One can intuitively imagine that the curvature of the log-likelihood is related to how well we can estimate the unknown parameter. Roughly, if the log-likelihood is very flat then even if our estimate $\hat{\theta}$ is very close in likelihood it need not be the case that $\hat{\theta}$ is close to θ^*. We will try to further formalize this intuition in the next section.

12.4 Asymptotic Normality of the MLE

Another important property of the MLE is that under regularity conditions:

$$\sqrt{nI(\theta^*)}(\hat{\theta} - \theta^*)$$

converges in distribution to a standard normal. We have previously seen how one can use this type of convergence in distribution to construct (asymptotic) confidence intervals. A simple argument will also show that this convergence implies that the MSE behaves like:

$$\mathbb{E}(\hat{\theta} - \theta^*)^2 \to \frac{1}{nI(\theta^*)},$$

so we can see that the Fisher information plays a crucial role in determining the quality of the MLE.

We once again have the problem that the confidence intervals above are not constructive since the true parameter θ^* is not known. Under regularity conditions we also have that:

$$\sqrt{nI(\hat{\theta})}(\hat{\theta} - \theta^*) \to N(0,1).$$

There are also some cases in which you cannot analytically compute the Fisher information. In these cases often the *observed information* is useful:

$$I_n(\hat{\theta}) = -\frac{1}{n} \sum_{i=1}^n \left[\frac{\partial^2}{\partial \theta^2} \log f_\theta(X_i)|_{\theta=\hat{\theta}} \right],$$

Under regularity conditions it will also be the case that:

$$\sqrt{nI_n(\hat{\theta})}(\hat{\theta} - \theta^*) \to N(0,1).$$

Let us try to roughly argue that the asymptotic normality makes sense. Lets define the log-likelihood function as

$$\ell(\theta) = \log L(\theta|X_1, \ldots, X_n).$$

Since $\hat{\theta}$ maximizes the log-likelihood we have that:

$$0 = \ell'(\hat{\theta}) \approx \ell'(\theta^*) + \ell''(\theta^*)(\hat{\theta} - \theta^*),$$

by a Taylor expansion. Re-arranging this we get that:

$$\sqrt{n}(\hat{\theta} - \theta^*) \approx \frac{\frac{1}{\sqrt{n}}\ell'(\theta^*)}{-\frac{1}{n}\ell''(\theta^*)}.$$

We can analyze the numerator and denominator separately: first the denominator,

$$-\frac{1}{n}\ell''(\theta^*) = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2}{\partial\theta^2} \log f_\theta(X_i)|_{\theta=\theta^*} \to I(\theta^*).$$

Now the numerator:

$$\begin{aligned}\frac{1}{\sqrt{n}}\ell'(\theta^*) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \frac{\partial}{\partial\theta} \log f_\theta(X_i)|_{\theta=\theta^*} \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_{\theta^*}(X_i).\end{aligned}$$

This is a sum of i.i.d. random variables with mean zero, and variance $I(\theta^*)$ so by the CLT we can conclude that:

$$\frac{1}{\sqrt{n}}\ell'(\theta^*) \to N(0, I(\theta^*))$$

in distribution. Putting these two pieces together we obtain the desired result.

At a high-level from a statistical point of view the MLE is a great estimator when appropriate regularity conditions hold: it is equivariant, it is consistent, and asymptotically normal with a variance that we can compute. Often it will also turn out that the MLE is "optimal" in the sense that amongst all unbiased estimators with an asymptotic Gaussian limit it has the lowest variance. We will delve into this in a future lecture.

Much of modern statistical theory tries to understand cases when: (1) the regularity conditions fail, (2) the asymptotic theory above is not valid (a common example of this is high-dimensional statistics, where the number of parameters greatly exceeds the number of data points) (3) the MLE is intractable to compute (in for instance graphical models and latent variable models).