Properties of MLEs - proofs of results in learning material 8

Properties of the expected log-likelihood

The key to proving and understanding the large sample (asymptotic) properties of maximum likelihood estimators lies in obtaining some results for the expectation of the log-likelihood. In this section, some simple properties of the expected log-likelihood are derived.

Let $x_1, x_2, \ldots, x_n$ be independent observations from a p.d.f. $f(\mathbf{x}, \theta)$ where θ is an unknown parameter with true value θ_T. Treating θ as unknown, the likelihood and log-likelihood for θ are:

$$L(\theta) \propto \prod_{i=1}^{n} f(x_i, \theta)$$

$$\ell(\theta) = \log_e \left(\prod_{i=1}^{n} f(x_i, \theta) \right) = \sum_{i=1}^{n} \log_e [f(x_i, \theta)] = \sum_{i=1}^{n} \ell_i(\theta)$$

where ℓ_i is the log-likelihood given only the single observation x_i. Treating ℓ as a function of random variables $X_1, X_2, \ldots, X_n$ means that ℓ is itself a random variable (and the ℓ_i are independent random variables). Hence we can consider expectations of ℓ and its derivatives.

Result 1:

$$\mathbb{E}_T \left(\left. \frac{\partial \ell}{\partial \theta} \right|_{\theta_T} \right) = 0$$

Where the subscript (T) on the expectation is to emphasize that the expectation is w.r.t. $f(x, \theta_T)$. The proof goes as follows (where it is to be taken that all differentials are evaluated at θ_T, and there is sufficient regularity that the order of differentiation and integration can be exchanged)

$$\begin{aligned} \mathbb{E}_T \left(\frac{\partial \ell_i}{\partial \theta} \right) = \mathbb{E}_T \left(\frac{\partial}{\partial \theta} \log[f(x, \theta)] \right) &= \int \frac{1}{f(x, \theta_T)} \frac{\partial f}{\partial \theta} f(x, \theta_T) dx \\ &= \int \frac{\partial f}{\partial \theta} dx = \frac{\partial}{\partial \theta} \int f dx \\ &= \frac{\partial 1}{\partial \theta} = 0. \end{aligned}$$

That the same holds for ℓ follows immediately.

Result 1 has the following obvious consequence in Result 2, since $\left(\mathbb{E}_T \left(\left. \frac{\partial \ell}{\partial \theta} \right|_{\theta_T} \right) \right)^2 = 0$:

Result 2:

$$\mathrm{Var} \left(\left. \frac{\partial \ell}{\partial \theta} \right|_{\theta_T} \right) = \mathbb{E}_T \left[\left(\left. \frac{\partial \ell}{\partial \theta} \right|_{\theta_T} \right)^2 \right]$$

(Remember, $\mathrm{Var}(X) = \mathbb{E}(X^2) - (\mathbb{E}(X))^2$)

Fishers Information I_θ is defined by:

$$I_\theta \equiv \mathbb{E}_T \left[\left(\left. \frac{\partial \ell}{\partial \theta} \right|_{\theta_T} \right)^2 \right]$$

It can be shown that:

Result 3:

$$I_\theta \equiv \mathbb{E}_T\left[\left(\left.\frac{\partial \ell}{\partial \theta}\right|_{\theta_T}\right)^2\right] = -\mathbb{E}_T\left[\left.\frac{\partial^2 \ell}{\partial \theta^2}\right|_{\theta_T}\right]$$

I_θ is referred to as the **information (i.e. Fishers information)** about θ contained in the data. This terminology refers to the fact that if the data tie down θ very closely (and accurately) then the log-likelihood will be sharply peaked in the vicinity θ_T (i.e. high I_θ), whereas data containing little information about θ will lead to an almost flat likelihood and low I_θ.

The proof of result 3 is as follows. For a single observation, result 1 says that

$$\int \frac{\partial \log(f)}{\partial \theta} f dx = 0$$

Differentiating again w.r.t. θ yields

$$\int \frac{\partial^2 \log(f)}{\partial \theta^2} f + \frac{\partial \log(f)}{\partial \theta}\frac{\partial f}{\partial \theta} dx$$

but

$$\frac{\partial \log(f)}{\partial \theta} = \frac{1}{f}\frac{\partial f}{\partial \theta}$$

and hence

$$\frac{\partial \log(f)}{\partial \theta} f = \frac{\partial f}{\partial \theta}$$

and so

$$\int \frac{\partial^2 \log(f)}{\partial \theta^2} f dx = -\int \left[\frac{\partial \log(f)}{\partial \theta}\right]^2 f dx$$

which is

$$\mathbb{E}_T\left[\left.\frac{\partial^2 \ell_i}{\partial \theta^2}\right|_{\theta_T}\right] = -\mathbb{E}_T\left[\left(\left.\frac{\partial \ell_i}{\partial \theta}\right|_{\theta_T}\right)^2\right]$$

The result follows from this (given the independence of the ℓ_i).

Now notice that result 1 says that the expected log-likelihood has a turning point at θ_T, while since I_θ is non-negative, result 3 indicates that this turning point is a maximum. So the expected log-likelihood has a maximum at the true parameter value.

Note also that although the results presented here were derived assuming that the data were independent observations from the same distribution, this is in fact much more restrictive than is necessary, and the results hold more generally. Similarly the results generalize immediately to vector parameters.

In this case result 3 is:

Result 3 (vector parameter)

$$\boldsymbol{I}_\theta \equiv \mathbb{E}_T \begin{pmatrix} \left(\frac{\partial \ell}{\partial \theta_1}\right)^2 & \frac{\partial \ell}{\partial \theta_1}\frac{\partial \ell}{\partial \theta_2} & . \\ \frac{\partial \ell}{\partial \theta_2}\frac{\partial \ell}{\partial \theta_1} & \left(\frac{\partial \ell}{\partial \theta_2}\right)^2 & . \\ . & . & . \end{pmatrix} = -\mathbb{E}_T \begin{pmatrix} \frac{\partial^2 \ell}{\partial \theta_1^2} & \frac{\partial^2 \ell}{\partial \theta_1 \partial \theta_2} & . \\ \frac{\partial^2 \ell}{\partial \theta_2 \partial \theta_1} & \frac{\partial^2 \ell}{\partial \theta_1^2} & . \\ . & . & . \end{pmatrix}$$

Results 1 and 3 do not establish that this maximum is the global maximum of the expected log-likelihood but a more involved proof shows that this is the case. This will not be proved here.

Large sample distribution of $\hat{\theta}$

To obtain the large sample distribution of the MLE $\hat{\theta}$ we make a Taylor expansion of the derivative of the log-likelihood around the true parameter θ_T and evaluate this at $\hat{\theta}$.

$$\left.\frac{\partial \ell}{\partial \theta}\right|_{\hat{\theta}} \simeq \left.\frac{\partial \ell}{\partial \theta}\right|_{\theta_T} + \left(\hat{\theta} - \theta_T\right) \left.\frac{\partial^2 \ell}{\partial \theta^2}\right|_{\theta_T}$$

and from the definition of the MLE the left hand side must be zero, so we have that

$$\left(\hat{\theta} - \theta_T\right) \simeq \frac{\partial \ell / \partial \theta|_{\theta_T}}{-\,\partial^2 \ell / \partial \theta^2|_{\theta_T}}$$

with equality, in the large sample limit (by consistency of $\hat{\theta}$).

Now the top of this fraction has expected value zero and variance I_θ (from earlier), but it is also made up of a sum of independent and identically distributed random variables, $\partial \ell_i / \partial \theta$, so that by the central limit theorem as $n \to \infty$ its distribution will tend to $N(0, I_\theta)$. By the law of large numbers we also have that as $n \to \infty$, $-\partial^2 \ell / \partial \theta^2\big|_{\theta_0} \to I_\theta$ (in probability). So in the large sample limit $(\hat{\theta} - \theta_T)$ is distributed as an $N(0, I_\theta)$ random variable divided by I_θ. i.e. in the limit as $n \to \infty$

$$\left(\hat{\theta} - \theta_T\right) \sim N(0, I_\theta^{-1}).$$

The result generalizes to vector parameters:

$$\hat{\boldsymbol{\theta}} \sim N(\boldsymbol{\theta}_T, \boldsymbol{I_\theta}^{-1})$$

in the large sample limit. Again the result holds generally and not just for the somewhat restricted form of the likelihood which we have assumed here.

Usually, of course, $\boldsymbol{I_\theta}$ will not be known any more than $\boldsymbol{\theta}$ is and will have to be estimated by plugging $\hat{\boldsymbol{\theta}}$ into the expression for $\boldsymbol{I_\theta}$. In fact, often the **sample information matrix $\mathbf{K}(\mathbf{x})$**, which is just the negative of the hessian $(-\mathbf{H})$ of the log-likelihood evaluated at the MLE, is an adequate approximation to the information matrix $\boldsymbol{I_\theta}$ itself.

This provides:

$$\hat{\boldsymbol{\theta}}_{MLE} \sim N(\boldsymbol{\theta_T}, \mathbf{K}(\mathbf{x})^{-1})$$

.