Asymptotic Properties and Computation of Maximum Likelihood Estimators

BS2 Statistical Inference, Lecture 8 Michaelmas Term 2004

Steffen Lauritzen, University of Oxford; November 8, 2004

Asymptotics of MLE in general case

First, use Taylor's theorem on the likelihood equation

$$0 = S(\hat{\theta}) = S(\theta) - j(\theta)(\hat{\theta} - \theta) + j'(\theta^*)(\hat{\theta} - \theta)^2/2,$$

rearrange the equation to

$$(\hat{\theta} - \theta)\{j(\theta) - j'(\theta^*)(\hat{\theta} - \theta)/2\} = S(\theta),$$

solve for $(\hat{\theta} - \theta)$ and multiply with $\sqrt{n}$ to yield

$$\sqrt{n}(\hat{\theta} - \theta) = \frac{S(\theta)/\sqrt{n}}{j(\theta)/n - j'(\theta^*)(\hat{\theta} - \theta)/2n}.$$

By the Law of Large Numbers, $j(\theta)/n \xrightarrow{P} i(\theta)$.

By the Central Limit Theorem

$$S(\theta)/\sqrt{n} = \sqrt{n}\left\{\frac{1}{n}\sum S_i(\theta)\right\} \xrightarrow{D} \mathcal{N}\{0, i(\theta)\}.$$

Assume that for an individual observation,
$|l'''(x;t)| < M(x)$ *for* $|t-\theta| < \delta$, *where* $\mathbf{E}_\theta\{M(X)\} < \infty$
and that $\hat{\theta}$ *is consistent*, i.e. that $\hat{\theta} \xrightarrow{P} \theta$.

It then holds for sufficiently large n that

$$|j'(\theta^*)(\hat{\theta}-\theta)/2n| \leq \frac{1}{2n}\sum M(X_i)|\hat{\theta}-\theta| \xrightarrow{P} 0.$$

Slutsky's theorems now yields

$$\hat{\theta}_n \overset{\text{a}}{\sim} \mathcal{N}\{\theta, i(\theta)^{-1}i(\theta)i(\theta)^{-1}/n\} = \mathcal{N}\{\theta, i(\theta)^{-1}/n\}.$$

Cramér's conditions

The precise conditions for the above argument to hold are

1. Θ is an open subset of the real line;
2. $A = \{x \mid f(x;\theta) > 0\}$ does not depend on θ:
3. The log-likelihood function is three times continuously differentiable so that for some $\delta > 0$ and all t with $|\theta - t| < \delta$, we have $l^{(i)}(x;t) < M_i(x)$, where $\mathbf{E}_\theta\{M_i(X)\} < \infty$;
4. $i(\theta) = -\mathbf{E}\{l''(\theta)\}$ is positive.

The first two conditions ensure that there are no problems with defining derivatives.

The boundedness of the derivatives implies that integration and differentiation can be interchanged and that the remainder term in the Taylor expansion of the likelihood equation is negligble.

The assumption of positive information ensures we can divide when solving the equation for $\hat{\theta}$.

Consistency of the MLE

Cramér's conditions imply by themselves that the MLE is consistent, more precisely *that there is at least one consistent, asymptotically normal, and efficient root $\hat{\theta}$ to the likelihood equation.*

The heuristics of the argument is as follows. Fix θ at the true value and divide the equation for $\hat{\theta}$ everywhere by n to obtain

$$0 = S(\theta)/n - j(\theta)/n(\hat{\theta} - \theta) + \frac{1}{2n} j'(\theta^*)(\hat{\theta} - \theta)^2.$$

Let $x = \hat{\theta} - \theta$. The equation then becomes

$$0 = a_n + b_n x + c_n x^2$$

where $a_n \xrightarrow{P} 0$, $b_n \xrightarrow{P} b < 0$ and $|c_n| < d_n \xrightarrow{P} d$ is bounded.

With a little care and precision it can now be shown that for any $\epsilon > 0$, n can be chosen large enough for this equation to have a root in the region $|x| < \epsilon$.

If $\hat{\theta}_n$ denotes this root, we have thus established that $\hat{\theta}_n \xrightarrow{P} \theta$.

For details of this argument, see the very clear exposition in the original proof on pp. 497 ff. of

H. Cramér (1946). *Mathematical Methods in Statistics.* Princeton University Press, NJ.

Digression on convexity

A real valued function $g(x)$ is said to be *convex* if

$$g\{(x_1 + x_2)/2\} \leq \{g(x_1) + g(x_2)\}/2 \text{for all } x_1, x_2.$$

It is *stricly convex* if the inequality is strict unless $x_1 = x_2$.

A function $g(x)$ is *concave* if $-g(x)$ is convex. *A function which is both convex and concave is affine,* i.e. has the form $g(x) = a + bx$

If g is twice differentiable, g is convex if and only if $g''(x) \geq 0$ for all x.

The function $g(x) = x^2$ is strictly convex, whereas $g(x) = \sqrt{x}$ and $g(x) = \log x$ are strictly concave.

Jensen's inequality

One of the most used inequalities in probability theory is due to J.L.J Jensen, a Danish mathematician who worked in the Copenhagen Telephone Company around 1900.

If $g(x)$ convex it holds for any probability distribution that

$$\mathbf{E}\{g(X)\} \geq g\{\mathbf{E}(X)\}.$$

If g is strictly convex, equality holds if and only if X is constant.

The proof can be found in most probability books. For $g(x) = x^2$ we get the well-known

$$\mathbf{E}(X^2) \geq \{\mathbf{E}(X)\}^2$$

with equality if and only if X is constant.

Wald's proof of consistency

Cramér's argument yields a consistent root of the likelihood equation, but there may be many roots, so is this root related to the maximum?

Wald (1949) showed consistency of the MLE along the following lines:

Let θ_0 be the true value of θ and define

$$\lambda(\theta) = \mathbf{E}_{\theta_0}\{\log f(X;\theta)\} = \mathbf{E}_{\theta_0}\{l(\theta)\}.$$

The function $\lambda(\theta)$ is the expected value of the log-likelihood function, and we first show that this function has its unique maximum at $\theta = \theta_0$.

Since the function $\log x$ is concave, Jensen's inequality yields

$$\begin{aligned}
\lambda(\theta) - \lambda(\theta_0) &= \mathbf{E}_{\theta_0}\left\{\log \frac{f(X;\theta)}{f(X;\theta_0)}\right\} \\
&\leq \log\left[\mathbf{E}_{\theta_0}\left\{\frac{f(X;\theta)}{f(X;\theta_0)}\right\}\right] \\
&= \log\left\{\int \frac{f(x;\theta)}{f(x;\theta_0)} f(x;\theta_0)\, dx\right\} = 0,
\end{aligned}$$

where equality holds if and only if everywhere $f(x;\theta) = f(x;\theta_0)$.

So if $f(x;\theta) = f(x;\theta_0) \implies \theta = \theta_0$ we have

$$\lambda(\theta_0) \geq \lambda(\theta)$$

with equality only when $\theta = \theta_0$.

The law of large numbers now implies that,

$$\bar{l}_n(\theta) = \frac{1}{n} \sum \log f(X_i; \theta) \xrightarrow{P} \lambda(\theta).$$

So the issue is whether it holds that

$$g_n(\theta) \to g(\theta) \implies \arg\max_\theta g_n(\theta) \to \arg\max_\theta g(\theta)??$$

Unfortunately this is not true in general, but verifiable conditions can be given for this to hold. Note that Wald's consistency proof has a very different flavour from Cramér's, as smoothness conditions do not play an essential role.

Observed Fisher information

As argued, we may in wide generality assume that the MLE satisfies

$$\hat{\theta}_n \overset{\text{a}}{\sim} \mathcal{N}\{\theta, i_n(\theta)^{-1}\}$$

where $i_n(\theta) = ni(\theta)$ is the information in the full sample. But how can we use this to judge the uncertainty of $\hat{\theta}$ when θ is unknown?

On possibility is to use $i_n(\hat{\theta})$ instead of $i_n(\theta)$ but an alternative would be to go directly into the Taylor approximation and use

$$j(\hat{\theta}) = -\sum l''(X_i; \hat{\theta})$$

and it turns out that this in many ways is preferable.

If a maximum of the likelihood function has been found, we must have $j(\hat{\theta})$ is positive. The quantity $j(\hat{\theta})$ is known as the *observed Fisher information*

In any case, it is an easy consequence of the consistency of $\hat{\theta}$ and Slutsky's theorem that, under Cramér's conditions, any of

$$\sqrt{ni(\theta)}(\hat{\theta}-\theta), \quad \sqrt{ni(\hat{\theta})}(\hat{\theta}-\theta), \quad \sqrt{j(\hat{\theta})}(\theta_n-\theta)$$

converge in distribution to $\mathcal{N}(0,1)$ (Problem sheet 4).

This fact holds also in the multivariate case with square roots of matrices properly interpreted.

Computation of the MLE

Generally the solution to the likelihood equation must be calculated by iterative methods.

One of the most common methods is the *Newton–Raphson method* and is based on successive approximations to the solution, using Taylor's theorem to approximate the equation.

Thus, we take an initial value θ_0 and write

$$0 = S(\theta_0) - j(\theta_0)(\theta - \theta_0)$$

ignoring the remainder term. Solving this equation for θ then yields a new value θ_1

$$\theta_1 = \theta_0 + j(\theta_0)^{-1} S(\theta_0)$$

and we keep repeating this procedure as long as $|S(\theta_j)| > \epsilon$. Clearly, $\hat{\theta}$ is a fixed point of this algorithm as $S(\hat{\theta}) = 0$.

Formally the iteration becomes

- Choose an initial value θ and calculate $S(\theta)$ and $j(\theta)$;
- `While` $|S(\theta)| > \epsilon$ `Repeat`
    1. $\theta \leftarrow \theta + j(\theta)^{-1} S(\theta)$
    2. Calculate $S(\theta)$ and $j(\theta)$ `go to 1`
- Return θ;

Other criteria for terminating the iteration can be used. To get a criterion which is insensitive to scaling of the variables, one can instead use the criterion $j(\theta)S(\theta)^2 > \epsilon$.

Properties of the Newton–Raphson method

If θ_0 is chosen sufficiently near $\hat{\theta}$ convergence is very fast.

It can be computationally expensive to evaluate $j(\theta)$ a large number of times. This is sometimes remedied by only changing j every 10 iterations or similar.

Another problem with the Newton–Raphson method is its lack of stability. When the initial value θ_0 is far from θ it might wildly oscillate and not converge at all.

This is sometimes remedied by making smaller steps as

$$\theta \leftarrow \theta + \gamma j(\theta)^{-1} S(\theta)$$

where $0 < \gamma < 1$ is a constant.