Asymptotic Properties of Maximum Likelihood Estimators

BS2 Statistical Inference, Lecture 7 Michaelmas Term 2004

Steffen Lauritzen, University of Oxford; November 4, 2004

Convergence in probability and in distribution

A sequence of random variables $Y_1, Y_2, \ldots$ is said to *converge in probability* to a random variable Y if for all $\epsilon > 0$

$$\lim_{n \to \infty} P\{|Y_n - Y| > \epsilon\} = 0$$

and we then write $\operatorname{plim} Y_n = Y$ or $Y_n \xrightarrow{P} Y$. We mostly use the special case where $Y = c$ is constant.

$Y_1, Y_2, \ldots$ *converges in distribution* to Y if for all continuity points y of the distribution function $F(y) = P(Y \leq y)$ of Y it holds that

$$\lim_{n \to \infty} P(Y_n \leq y) = F(y).$$

We then write $\operatorname{dlim} Y_n = Y$ or $Y_n \xrightarrow{D} Y$.

Slutsky's theorems

Let $Y_1, Y_2 \ldots$ and $Z_1, Z_2, \ldots$ be sequences of random variables so that $Y_n \xrightarrow{D} Y$ and $Z_n \xrightarrow{P} c$ where c is a constant. It then holds that

- $Y_n + Z_n \xrightarrow{D} Y + c$;
- $Y_n Z_n \xrightarrow{D} cY$.

Indeed if $g(y, z)$ is continuous at all points (y, c), it holds that

$$g(Y_n, Z_n) \xrightarrow{D} g(Y, c).$$

Note that the two first statements are special cases of the latter. See Knight (1999) for proofs of these results.

The delta method

Suppose we have established convergence in distribution for a scaled and centered sequence of variables

$$Y_n = (X_n - b)/a_n \xrightarrow{D} Y \tag{1}$$

for a scaling sequence $a_n > 0$ which converges to 0 for $n \to \infty$.

If g is continuously differentiable at b with derivative g'

$$\{g(X_n) - g(b)\}/a_n \xrightarrow{D} g'(b)Y. \tag{2}$$

This result is mostly used when $Y \sim \mathcal{N}(0, 1)$ where we also write (1) as $X_n \overset{\text{a}}{\sim} \mathcal{N}(b, a_n^2)$ with the consequence (2) then as $g(X_n) \overset{\text{a}}{\sim} \mathcal{N}\{g(b), a_n^2 g'(b)^2\}$.

Proof of the delta method

This is an easy consequence of Taylor's theorem and Slutsky's theorems.

First realise that the convergence in distribution of $Y_n = (X_n - b)/a_n$ implies $X_n \xrightarrow{P} b$ because

$$\begin{aligned}
&\lim_{n\to\infty} P(|X_n - b| > \epsilon) \\
&\quad = \lim_{n\to\infty} P(|X_n - b|/a_n > \epsilon/a_n) \\
&\quad = \lim_{n\to\infty} \{1 - P(Y_n < \epsilon/a_n) + P(Y_n > -\epsilon/a_n)\} \\
&\quad = \lim_{n\to\infty} \{1 - F(\epsilon/a_n) + F(-\epsilon/a_n)\} = 1 - 1 + 0 = 0.
\end{aligned}$$

Taylor's theorem yields

$$g(X_n) - g(b) = g'(b_n^*)(X_n - b),$$

where b_n^* is between X_n and b. Now divide by a_n to get

$$\{g(X_n) - g(b)\}/a_n = g'(b_n^*)(X_n - b)/a_n.$$

Since $X_n \xrightarrow{P} b$ and b_n^* is between X_n and b, $b_n^* \xrightarrow{P} b$. As g is continuously differentiable, we conclude that $g'(b_n^*) \xrightarrow{P} g'(b)$ and Slutsky's theorem now yields the result.

Multivariate delta method

The delta method can be generalised immediately to the multivariate case. Suppose Y is multivariate in (1) so

$$X_n \overset{\text{a}}{\sim} \mathcal{N}_S(b, a_n^2 \Sigma),$$

and $g = (g_1, \ldots, g_R)$ are continuously differentiable at b with matrix of partial derivatives $g'(b)$ with

$$g'(b)_{rs} = \frac{\partial}{\partial b_s} g_r(b).$$

Then it holds that

$$g(X_n) \overset{\text{a}}{\sim} \mathcal{N}_R\{g(b), a_n^2 g'(b) \Sigma g'(b)^\top\}.$$

Asymptotics of MLE in canonical exponential families

Consider a sample $(X_1, \ldots, X_n)$ of size n from a d-dimensional canonical exponential family with individual densities

$$f(x;\theta) = b(x)e^{\theta^\top t(x) - c(\theta)}, \theta \in \Theta \subseteq \mathcal{R}^d.$$

We have seen that the MLE of θ is given as

$$\hat{\theta} = \hat{\theta}_n = \tau^{-1}(\bar{T}_n),$$

where τ is the mean value mapping and

$$\bar{T}_n = \frac{t(X_1) + \cdots + t(X_n)}{n}.$$

We will now show that *the MLE is asymptotically normally distributed, and asymptotically unbiased and efficient*, i.e.

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

The central limit theorem yields for $\eta = \tau(\theta)$ that

$$\bar{T}_n \overset{\text{a}}{\sim} \mathcal{N}_d\left\{\eta, \frac{1}{n} i(\theta)\right\}.$$

Using the inverse function theorem for $g = \tau^{-1}$ gives

$$g'(\eta) = \frac{d\theta}{d\eta} = \left\{\frac{d\eta}{d\theta}\right\}^{-1} = \{\tau'(\theta)\}^{-1} = i(\theta)^{-1}.$$

The delta method now yields

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

Asymptotics of MLE in general case

In the general case we give a heuristic argument for the fact that — under suitable regularity conditions — it holds that

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

where $i(\theta)$ is the Fisher information for an individual observation.

For simplicity we consider the one-dimensional case. As usual we let $l(\theta)$ be the log-likelihood function and $S(\theta)$ the score statistic. We also define

$$j(\theta) = -l''(\theta) = -S'(\theta)$$

so we have that

$$S(\theta) = \sum_i S(X_i; \theta), \quad j(\theta) = \sum_i j(X_i; \theta)$$

and

$$\mathbf{V}\{S(\theta)\} = \mathbf{E}\{j(\theta)\} = ni(\theta).$$

Now use Taylor's formula to write

$$0 = S(\hat{\theta}) = S(\theta) - j(\theta)(\hat{\theta} - \theta) + R(X, \theta, \hat{\theta})$$

and hope that the remainder term R is small and can be ignored.

Solve this equation to yield (ignoring the remainder term)

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

By the LLN the denominator converges in probability to $i(\theta)$.

By the central limit theorem the numerator converges in distribution to $\mathcal{N}(0, i(\theta))$.

Slutsky's theorems now yields

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

The issue is to find conditions which ensures the remainder term to be small.

Apart from the usual conditions which ensure interchange of differentiation and integration, so the Fisher information is well defined and equal to the variance of the score statistic, this typically involves the assumptions that

- the third derivative of $l(\theta)$ is uniformly bounded in a neighbourhood of θ;

- the estimator is consistent so that $\hat{\theta}_n \xrightarrow{P} \theta$.

In general, the last of these conditions is more tricky to establish than the first, but both are fulfilled in a large number of cases.