DA Freedman   Notes on the MLE   Fall 2003

The object here is to provide a sketch of the theory of the MLE. Rigorous presentations can be found in the references cited below.

Calculus. Let f be a smooth, scalar function of the $p \times 1$ vector x. We view f' as a $1 \times p$ vector of partial derivatives $\{\partial f / \partial x_i : i = 1, \ldots, p\}$. Likewise, f'' is a $p \times p$ matrix and f''' is a 3-D array of partials. As a matter of notation, f' is the derivative of f and g^T is the transpose of g. Moreover, $\|x\|$ is the Euclidean norm, $\|x\|^2 = \sum_{i=1}^p x_i^2$. Abbreviate

$$f'''_{ijk} = \frac{\partial^3 f}{\partial x_i \partial x_j \partial x_k}$$

(1) Lemma. Let

$$M = \max_{ijk} \max_{\|x\| \le \delta} |f'''_{ijk}|;$$

the indices i, j, and k need not be distinct. If $\|x\| \le \delta$, then

$$f(x) = f(0) + f'(0)x + \frac{1}{2} x^T f''(0) x + g(x)$$

where

$$|g(x)| \le \frac{1}{6} p^{3/2} M \|x\|^3.$$

Sketch of proof. We may assume that $f(0) = f'(0) = f''(0) = 0$. Fix x with $\|x\| \le \delta$; let $0 \le u \le 1$; view $\phi(u) = f(ux)$ as a scalar function of the scalar u; then $\phi(0) = \phi'(0) = \phi''(0) = 0$, so $\phi(u) = u^3 \phi'''(v)/3!$ by Taylor's theorem, with $0 \le v \le u$. Now $0 \le u^3 \le 1$, and

$$\phi'''(v) = \sum_{ijk} f'''_{ijk}(vx) x_i x_j x_k$$

so

$$|\phi(u)| \le \frac{1}{6} M \sum_{ijk} |x_i||x_j||x_k| = \frac{1}{6} M \Big(\sum_i |x_i| \Big)^3 \le \frac{1}{6} M p^{3/2} \|x\|^3$$

by the inequality of Cauchy-Schwarz.

Let g be a smooth $1 \times p$ function of x. We view g' as $p \times p$; and

$$g(x + \delta) = g(x) + \delta^T g'(x) + O(\|\delta\|^2) \quad \text{as } \delta \to 0.$$

(2) Lemma. Let $h = fg$, where f is scalar and g is $1 \times p$; both functions are smooth. Then h' is $p \times p$ and

$$h' = fg' + f'^T g.$$

Examples. Suppose a is a $1 \times p$ vector of reals and $f(x) = ax$. Then $f'(x) = a$ and $f''(x) = 0$. Suppose A is a $p \times p$ matrix of reals, perhaps asymmetric, and $g(x) = x^T A$. Then $g'(x) = A$ and $g''(x) = 0$. Let $h(x) = x^T Ax$. Then $h'(x) = x^T(A + A^T)$, $h''(x) = (A + A^T)$ and $h'''(x) = 0$.

Fisher information. Let $f_\theta(x)$ be a density, bounded, positive, vanishing rapidly as $|x| \to \infty$. There are problems at boundary points; we take θ and x to be Euclidean; $\theta, x \to f_\theta(x)$ is assumed smooth. Now $\int f_\theta(x)\,dx = 1$, so

$$\int \frac{\partial}{\partial\theta} f_\theta(x)\,dx = \int \frac{\partial^2}{\partial\theta^2} f_\theta(x)\,dx = 0. \tag{3}$$

The *Fisher Information Matrix* is

$$I(\theta) = -\int \Big(\frac{\partial^2}{\partial\theta^2} \log f_\theta(x)\Big) f_\theta(x)\,dx.$$

(4) Lemma.

$$\begin{aligned} I(\theta) &= \int \Big(\frac{\partial}{\partial\theta} \log f_\theta(x)\Big)^T \Big(\frac{\partial}{\partial\theta} \log f_\theta(x)\Big) f_\theta(x)\,dx \\ &= \int \Big(\frac{\partial}{\partial\theta} f_\theta(x)\Big)^T \Big(\frac{\partial}{\partial\theta} f_\theta(x)\Big) \frac{1}{f_\theta(x)}\,dx. \end{aligned}$$

Proof. To begin with,

$$\frac{\partial}{\partial\theta} \log f_\theta(x) = \frac{1}{f_\theta(x)} \frac{\partial}{\partial\theta} f_\theta(x). \tag{5}$$

Then by (2),

$$\frac{\partial^2}{\partial\theta^2} \log f_\theta(x) = \frac{1}{f_\theta(x)} \frac{\partial^2}{\partial\theta^2} f_\theta(x) - \frac{1}{f_\theta(x)^2} \Big(\frac{\partial}{\partial\theta} f_\theta(x)\Big)^T \Big(\frac{\partial}{\partial\theta} f_\theta(x)\Big);$$

and (3) completes the proof.

The statistical model. Let X_i be measurable functions on $(\Omega, \mathcal{F})$ for $i = 1, \ldots, n$, with values in R^q. For $\theta \in R^p$, let P_θ be a probability on $(\Omega, \rhd)$. With respect to the probability P_θ, let X_i be independent random variables, having common probability density f_θ on R^q. In particular,

$$I(\theta) = -\mathrm{E}_\theta \Big\{ \frac{\partial^2}{\partial\theta^2} \log f_\theta(X_i) \Big\}.$$

We assume $I(\theta)$ is invertible. By (3) and (4),

$$(6) \qquad \mathrm{E}_\theta\left\{\frac{\partial}{\partial\theta}\log f_\theta(X_i)\right\} = 0, \qquad \mathrm{var}_\theta\left\{\frac{\partial}{\partial\theta}\log f_\theta(X_i)\right\} = I(\theta).$$

The *log likelihood function* is

$$L(\theta) = \sum_{i=1}^{n} \log f_\theta(X_i).$$

The X_i are often viewed as fixed, the variable is θ. Write θ_0 for the (unknown) true value of θ. The first derivative of the log likelihood function is

$$L'(\theta) = \sum_{i=1}^{n} \frac{\partial}{\partial\theta}\log f_\theta(X_i).$$

Of course, $L'(\theta)$ is random, because it depends on the X_i; this is suppressed in the notation. By (6),

$$(7) \qquad \mathrm{E}_\theta\{L'(\theta)\} = 0, \quad \mathrm{var}_\theta\{L'(\theta)\} = nI(\theta).$$

The MLE $\hat{\theta}$, by definition, maximizes the likelihood function. (Technically, there may be multiple maxima, but see below; with weaker conditions, there may be no maximum, but the theory can still be pushed through.) The main result to be discussed here says that asymptotically, the MLE is normal, with mean θ_0 and variance $I(\theta_0)^{-1}/n$. Asymptotic optimality is another idea, see the references below.

(8) Theorem. As $n \to \infty$, the P_{θ_0}-distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$ converges to normal, with mean 0 and variance $I(\theta_0)^{-1}$.

Sketch of proof. By entropy considerations, for large n, the MLE will almost surely be within a small neighborhood of the true parameter value θ_0. Indeed, if f and g are densities, then $\int f \log g < \int f \log f$ unless $g = f$. So $L(\theta)$ is much smaller than $L(\theta_0)$ unless $|\theta - \theta_0| \leq \delta$. Then the log likelihood function can be expanded in a Taylor series around θ_0:

$$L(\theta) = L(\theta_0) + L'(\theta_0)(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^T L''(\theta_0)(\theta - \theta_0) + R.$$

The lead term $L(\theta_0)$ is random; but since this term does not depend on θ, its behavior is immaterial. The first derivative $L'(\theta_0)$ is asymptotically normal with mean 0 and variance $nI(\theta_0)$ by the central limit theorem and (7). By the strong law, $L''(\theta_0) \approx -nI(\theta_0)$. The remainder term R has

$$|R| = O(n\|\theta - \theta_0\|^3)$$

by (1), and is negligible relative to the quadratic term. Thus, the MLE $\hat{\theta}$ essentially maximizes

$$\theta \to L'(\theta_0)(\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)^T L''(\theta_0)(\theta - \theta_0).$$

So

$$\hat{\theta} - \theta_0 \approx -L''(\theta_0)^{-1} L'(\theta_0)^T$$

and is asymptotically N(0, $I(\theta_0)^{-1}/n$), as required.

The maximum can be found by setting the derivative to 0. The “likelihood equation" $L'(\hat{\theta}) = 0$ (almost) boils down to

$$L'(\theta_0) + (\theta - \theta_0)^T L''(\theta_0) = 0.$$

The “observed information" $L''(\hat{\theta})/n$ can be used to approximate Fisher information. There is similar theory for integer-valued random variables, for random variables with fairly general range spaces, for Θ a half-space, an open subset of R^p, etc., etc.

Testing. Let Θ_0 be a p_0-dimensional subset of R^p. We wish to test the null hypothesis that $\theta_0 \in \Theta_0$. Let $\hat{\theta}_0$ be the MLE, where the maximization is restricted to Θ_0. For a simple hypothesis, Wald’s t-test compares the MLE to its SE and there is a version like Hotelling’s T^2 for composite hypotheses. The Neyman-Pearson (or Wilks) statistic is $2[L(\hat{\theta}) - L(\hat{\theta}_0)]$, which has under the null hypothesis an asymptotic $\chi^2_{p-p_0}$ distribution. Rao’s score test uses the statistic $L'(\hat{\theta}_0)I(\hat{\theta}_0)^{-1}L'(\hat{\theta}_0)^T/n$; again the asymptotic distribution is $\chi^2_{p-p_0}$. At interior points, these test statistics are asymptotically equivalent; at boundary points, Wald’s test and the Neyman-Pearson statistic get into trouble, while the score test often does fine. The leading special case for the null distribution of these tests has $n = 1,\ \ X \sim N(\theta_0, I)$, so $I(\theta_0)$ is the identity matrix, and $\Theta_0 = \{\theta : \theta_{p_0+1} = \cdots = \theta_p = 0\}$. The general case follows by change of variables and rotation.

Examples. Suppose the U_i are IID N(α, 1), the V_i are IID N(β, 1), the U’s and V’s are independent. Let $X_i = (U_i, V_i)$ and $\theta = (\alpha, \beta)$. Now

$$2L(\theta) = n \log \frac{1}{2\pi} - \sum_{i=1}^{n}(U_i - \bar{U})^2 - \sum_{i=1}^{n}(V_i - \bar{V})^2 - n(\bar{U} - \alpha)^2 - n(\bar{V} - \beta)^2.$$

The MLE is the sample mean. For testing the null hypothesis that $\beta = 0$, the Neyman-Pearson statistic and the Rao score statistic are both $n\bar{V}^2$. If you restrict β to be non-negative, $\hat{\beta}$ is 0 when $\bar{V} < 0$; the Neyman-Pearson statistic $n\hat{\beta}^2$ is not χ^2-like: the score statistic is still $n\bar{V}^2$, whose null distribution is χ^2_1.

Exercises. Suppose the X_i are IID Poisson, with mean λ. Write down $L,\ L',\ L'',\ I$. Find the MLE for λ. If $\lambda_1 > 0$, write down the Neyman-Pearson statistic and the score statistic for testing the null hypothesis that $\lambda = \lambda_1$. Verify the asymptotic distributions under the null. Which test is more powerful for $\lambda > \lambda_1$? For $\lambda < \lambda_1$?

Suppose the X_i are independent $\mathrm{N}(\theta_i, 1)$ for $i = 1, \ldots, p$; the θ_i are unrestricted real numbers. Find the MLE for θ. Find the Neyman-Pearson and Rao tests for the null hypothesis that

$$\theta_i = 0 \text{ for } i = p_0 + 1, \ldots, p$$

Let M be a non-random $n \times p$ matrix of full rank; suppose $Y = M\theta + \epsilon$, where the ϵ_i are IID $\mathrm{N}(0, \sigma^2)$. Write down $L,\ L',\ L'',\ I$. Find the MLE for θ and σ^2. Write down the Neyman-Pearson statistic and the score statistic for testing the null hypothesis that $\theta_1 = 0$. Derive the normal equations by differentiating $\theta \to \|Y - M\theta\|^2$ with respect to θ.

Let Φ be the standard normal distribution function, with $\Phi' = \phi$ being the density. According to the probit model, given $X_1, \ldots, X_n$, the variables $Y_1, \ldots, Y_n$ are independent 0–1 variables, each being 1 with probability $\Phi(X_i\beta)$. For $x > 0$, show that $1 - \Phi(x) < \phi(x)/x$. Conclude that Φ and $1 - \Phi$ are log concave. Conclude further that the log likelihood function for the probit model is concave. Hint: show first $1 - \Phi(x) < \int_x^\infty (z/x)\phi(z)\,dz$.

References

CR Rao (1973). *Linear Statistical Inference*. 2nd ed. Wiley.

Chapter 6 discusses likelihood techniques; pp.415ff cover Wald's t-test, the Neyman-Pearson likelihood ratio test, and Rao's score test.

Notation

$\mathcal{I} = I(\theta)$

$\ell = L$

$\mathbf{V} = L'(\theta)^T/\sqrt{n}$

$\mathbf{D} = \sqrt{n}(\hat{\theta} - \theta_0)$.

The subscript 0 means, substitute θ_0 for θ.

The superscript * means, substitute θ^* for θ, the former being the MLE over a restricted subset of parameter space; Rao's parameter θ is q-dimensional, and his restricted parameter space is s-dimensional.

EL Lehmann (1991). *Testing Statistical Hypotheses*. 2nd ed. Wadsworth & Brooks/Cole.

The χ^2 and likelihood ratio tests are discussed on pp.477ff.

EL Lehmann (1991). *Theory of Point Estimation*. Wadsworth & Brooks/ Cole.

The information inequality (aka the Cramèr-Rao inequality) is discussed on pp.115ff, and the theory of the MLE is developed in Chapter 6. For exponential families, the calculus is much more tractable; see pp.119, 417, 438.