Asymptotic Efficiency of the Maximum Likelihood Estimator

Guy Lebanon

January 29, 2009

In this note we provide a short proof based on [1] and [2] for the asymptotic normality and efficiency of the multivariate maximum likelihood estimator (mle). Asymptotic efficiency refers to the situation when the asymptotic variance equals the inverse Fisher information which is the best possible variance (Cramer-Rao lower bound). It is assumed that the reader is familiar with the notes on *Relative Efficiency, Efficiency, and the Fisher Information* and *Consistency of the Maximum Likelihood Estimator*.

We assume that (i) $X_1, X_2, \ldots$ are sampled iid from p_{θ_0}, which is assumed to be continuous and twice differentiable in θ. We also assume that (ii) the parameter space $\Theta \subset \mathbb{R}^k$ is a convex open set. For a function $g : \mathbb{R}^k \to \mathbb{R}$ we denote by $\nabla g(z)$ its gradient $k \times 1$ vector and by $\nabla^2 g(z)$ the $k \times k$ matrix containing its second order derivatives i.e. $[\nabla^2 g(z)]_{ij} = \frac{\partial^2}{\partial z_i \partial z_j} g(z)$.

The Fisher information is defined as the $k \times k$ matrix $J(\theta) = \mathsf{E}_{p_\theta}\{\nabla_\theta \log p_\theta(X)(\nabla_\theta \log p_\theta(X))^\top\}$. It is well known that $J(\theta) = -\mathsf{E}_{p_\theta}\{\nabla^2_\theta \log p_\theta(X)\}$ and $\mathsf{E}_{p_\theta}\{\nabla_\theta \log p_\theta(X)\} = 0$ (see the note *Relative Efficiency, Efficiency, and the Fisher Information* for the one dimensional case).

The following theorem establishes the asymptotic normality of the mle and its efficiency by demonstrating that the asymptotic variance is the inverse Fisher information.

Proposition 1 (Cramer)**.** *In addition to the assumptions above, we assume that (iii) there exists a function $K(x)$ such that $\mathsf{E}_{p_{\theta_0}} K(X) < \infty$ and each component of $\nabla_\theta \log p_\theta(x)$ is bounded in absolute value by $K(x)$ uniformly in some neighborhood of θ_0, (iv) $J(\theta_0)$ is a positive definite matrix, and (v) identifiability i.e. $p_\theta \equiv p_{\theta_0} \Leftrightarrow \theta = \theta_0$. Then there exists a strongly consistent sequence $\hat{\theta}_n$ of likelihood (local) maximizers for which the following convergence in distribution holds*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N(0, J(\theta_0)^{-1}).$$

Proof. Strong consistency follows from Proposition 3 in the note *Consistency of the Maximum Likelihood Estimator* applied to a compact neighborhood $A = \{\theta \in \Theta : \|\theta - \theta_0\| \leq \alpha\}$ (Chapters 17-18 in [1] demonstrate why the conditions in that proposition hold). We thus focus below on proving asymptotic normality based on [2].

We denote $H_n(\theta) = -\frac{1}{n}\sum_{i=1}^n \log p_\theta(X_i)$ and $H(\theta) = -\mathsf{E}_{p_{\theta_0}} \log p_\theta(X)$. The mean value theorem implies that there exists $\lambda \in (0,1)$ with $\theta' = \theta_0 + \lambda(\hat{\theta}_n - \theta_0)$ such that

$$\nabla_\theta H_n(\hat{\theta}_n) = \nabla_\theta H_n(\theta_0) + \nabla^2_\theta H_n(\theta')(\hat{\theta}_n - \theta_0).$$

Since $\hat{\theta}_n$ is the mle, we have $\nabla H_n(\hat{\theta}_n) = 0$ and

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\sqrt{n}(\nabla^2 H_n(\theta'))^{-1} \nabla H_n(\theta_0). \tag{1}$$

Since we already established consistency, $\hat{\theta}_n \xrightarrow{\mathrm{p}} \theta_0$ which implies that $\theta' \xrightarrow{\mathrm{p}} \theta_0$. Furthermore,

$$(\nabla^2 H_n(\theta'))^{-1} \xrightarrow{\mathrm{p}} (\nabla^2 H_n(\theta_0))^{-1} \xrightarrow{\mathrm{p}} J(\theta_0)^{-1} \tag{2}$$

where we used the fact that if $X_n \xrightarrow{\mathrm{p}} X$ then $g(X_n) \xrightarrow{\mathrm{p}} g(X)$ for continuous g and the law of large numbers.

For the remaining term in (1) we have

$$-\sqrt{n}\,\nabla_\theta H_n(\theta_0) = \sqrt{n}\,\nabla_\theta \left(\frac{1}{n} \sum_{i=1}^{n} \log p_\theta(X_i) \right) \Bigg|_{\theta=\theta_0}.$$

The random vectors $\nabla \log p_\theta(X_i)|_{\theta=\theta_0}$, for $i = 1, \ldots, n$ are iid with mean $\mathsf{E}\{\nabla \log p_\theta(X)\} = 0$ and covariance $J(\theta_0)$ and so by the central limit theorem we have the following convergence in distribution

$$-\sqrt{n}\,\nabla_\theta H_n(\theta_0) \rightsquigarrow N(0, J(\theta_0)). \tag{3}$$

Using Slutsky's theorem we combine (1), (2) and (3) to obtain

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N(0, J^{-1}(\theta_0)).$$

□

Note that the proposition above refers to a sequence of likelihood maximizers $\hat{\theta}_n$. If there are multiple likelihood maximizers, it is not clear whether a particular sequence will achieve consistency. If, however, there is a single likelihood maximizer (as is often the case if the likelihood is concave) the maximum likelihood estimator is uniquely defined and is guaranteed to be consistent and asymptotically normal with the above variance.

Due to Proposition 1, it is often said that the mle $\hat{\theta}_n$ has $\sqrt{n}$-convergence rate to θ. An alternative way to write the convergence $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow N(0, J^{-1}(\theta_0))$ is as

$$\hat{\theta}_n \approx \theta_0 + \frac{1}{\sqrt{n}} N(0, J^{-1}(\theta_0))$$

which exposes the fact that for large n, $\hat{\theta}_n$ equals θ_0 plus a random remainder or noise term $\frac{1}{\sqrt{n}} N(0, J^{-1}(\theta_0))$ whose variance decreases as $(nJ(\theta))^{-1}$.

It is also possible to use the above asymptotic distribution for hypothesis testing or obtaining confidence intervals with respect to the mle. For example, using Slutsky's theorem and $\hat{\theta}_n \xrightarrow{\text{p}} \theta_0$ we have $J^{1/2}(\hat{\theta}_n) \xrightarrow{\text{p}} J^{1/2}(\theta_0)$. Additional applications of Slutsky's theorem provides

$$\begin{aligned} \sqrt{n}(\hat{\theta}_n - \theta_0)^\top J^{1/2}(\hat{\theta}_n) &\xrightarrow{\text{p}} \sqrt{n}(\hat{\theta}_n - \theta_0)^\top J^{1/2}(\theta_0) \rightsquigarrow N(0, I_k) \\ n^{-1}(\hat{\theta}_n - \theta_0)^\top J(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) &\rightsquigarrow \chi^2_k. \end{aligned}$$

The above statistics may be used to test hypothesis regarding θ_0. Similarly, inverting the statistics above with respect to $\hat{\theta}_n$ provide formulas for confidence intervals, for example using the first statistic we may obtain Wald's confidence for θ_0

$$\theta_0 \in (\hat{\theta}_n \pm r_{\alpha/2,k}\,(nJ(\hat{\theta}_n))^{-1/2}) \quad \text{with probability} \quad 1 - \alpha.$$

Note that the size of the confidence interval decreases as n increases (higher sample size) and as J increases (higher asymptotic accuracy of the mle).

References

[1] T. S. Ferguson. *A Course in Large Sample Theory.* Chapman & Hall, 1996.

[2] U. Grenander and M. Miller. *Pattern Theory: From Representation to Inference.* Oxford University Press, 2007.