Consistency of the Maximum Likelihood Estimator

Guy Lebanon

January 5, 2008

In this note we provide a short proof based on Chapters 16-17 of [1] for the consistency of the multivariate maximum likelihood estimator (mle). Consistency is a condition which ensures that for large datasets the mle will converge to the true parameter. We assume that at this point the reader is familiar with the note *Consistency of Estimators*.

We first introduce the uniform strong law of large numbers. We assume that $X_1, X_2, \ldots$ are iid samples from F and $U(x,\theta)$ is a function of x for all $\theta \in \Theta$. The strong law of large numbers state that $(1/n)\sum_{i=1}^n U(X_i,\theta) \to \mathsf{E}\,U(X,\theta) \stackrel{\text{def}}{=} \mu(\theta)$ almost surely (and therefore also in probability i.e. $\forall \epsilon > 0$, $P(|n^{-1}\sum_{i=1}^n U(X_i,\theta) - \mu(\theta)| > \epsilon) \to 0$). The uniform strong law of large numbers strengthen the convergence to be uniform over the space Θ i.e.

$$\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n U(X_i,\theta) - \mu(\theta)\right| \stackrel{\text{a.s.}}{\to} 0.$$

If the set Θ is finite, the above uniform convergence follows from the strong law of large numbers because the intersection of a finite numbers of sets of probability 1 has probability 1. We have the following result for a compact (potentially infinite) Θ, originally due to Le Cam (for a proof see Theorem 16(a) in [1]).

Proposition 1. *Let Θ be a compact parameter space and $U(x,\theta)$ an upper semi-continuous in θ for all x. If there exists a function $K(x)$ such that $\mathsf{E}K(X) < \infty$ and $|U(x,\theta)| \le K(x)$ for all x,θ, then $\frac{1}{n}\sum_{i=1}^n U(X_i,\theta) \stackrel{\text{a.s.}}{\to} \mu(\theta)$ uniformly i.e.*

$$P\left\{\lim_{n\to\infty}\sup_{\theta\in\Theta}\left|\frac{1}{n}\sum_{i=1}^n U(X_i,\theta) - \mu(\theta)\right| = 0\right\} = 1$$

Next, we need the following result of Shannon asserting the non-negativity of the KL divergence.

Proposition 2. *For any two densities or mass functions p, q,*

$$D(p||q) \stackrel{\text{def}}{=} \mathsf{E}_p \log\frac{p(X)}{q(X)} \ge 0$$

with equality iff $p \equiv q$.

Proof. Apply Jensen's inequality to obtain

$$-D(p||q) = \mathsf{E}_p \log\frac{q(X)}{p(X)} \le \log E_p \frac{q(X)}{p(X)} = \log\int q(x)\,dx = 0$$

with equality iff $p \equiv q$ (replace integral above with a sum if X is discrete). □

We now assume that $X_1, X_2, \ldots$ are sampled from p_{θ_0} and define the likelihood function as $L_n(\theta) = \prod_{i=1}^n p_\theta(x_i)$. A maximum likelihood estimator (mle) is defined as any function $\hat\theta_n = \hat\theta(x_1,\ldots,x_n)$ such that

$L_n(\hat\theta_n) = \sup_{\theta\in\Theta} L_n(\theta)$. The mle maximizes

$$\frac{1}{n}\log L_n(\theta) - \frac{1}{n}\log L_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\log\frac{p_\theta(X_j)}{p_{\theta_0}(X_j)} \overset{\text{a.s.}}{\to} -D(p_{\theta_0}||p_\theta) \le 0$$

which converges by the law of large numbers to an expression that is maximized at 0 iff $\theta_0 = \theta$ according to Proposition 2. This suffices to prove strong consistency of the mle if Θ is finite. However, in most cases Θ is infinite and we need to use Proposition 1 to extend the result. The proof below of the mle's strong consistency is due to Wald.

Proposition 3. *Let $X_1, X_2, \ldots$ be random vectors sampled iid from p_{θ_0} where the parameter space $\Theta \subset \mathbb{R}^k$ is compact and p is continuous in θ for all x. We assume further identifiability i.e. $p_\theta \equiv p_{\theta_0} \Leftrightarrow \theta = \theta_0$, and that there exists a function $K(x)$ with $\mathsf{E}_{\theta_0}|K(X)| < \infty$ and $\log p_\theta(x) - \log p_{\theta_0}(x) \le K(x)$ for all x, θ. Then for any sequence of mle $\hat\theta_n$ we have $\hat\theta_n \overset{\text{a.s.}}{\to} \theta_0$.*

Proof. The conditions of Proposition 1 are satisfied for $U(x,\theta) \overset{\text{def}}{=} \log p_\theta(x) - \log p_{\theta_0}(x)$ and $\mu(\theta) = \mathsf{E}\,U(X,\theta) = -D(p_{\theta_0}, p_\theta)$. Let $\rho > 0$ and define the compact set $S = \{\theta \in \Theta : \|\theta - \theta_0\| \ge \rho\}$. Since $\mu(\theta)$ is continuous it achieves its maximum on S denoted by $\delta = \sup_{\theta\in S}\mu(\theta)$. By Proposition 2, $\delta < 0$ and hence by Proposition 1 there exists N such that $\forall n > N$, $\sup_{\theta\in S} n^{-1}\sum_{i=1}^n U(x_i,\theta) < 0$ with probability 1. But since $n^{-1}\sum_{i=1}^n U(x_i,\theta)$ equals 0 for $\theta = \theta_0$ we have $n^{-1}\sum_{i=1}^n U(x_i,\hat\theta_n) \ge 0$ which shows that the mle is not in S. Since ρ was arbitrarily chosen, the proposition follows. □

We make the following comments.

- For the sake of simplicity we omit certain conditions in Proposition 3 concerning measurability.
- Proposition 3 also holds for upper semi-continuous p. This version, presented in [1], allows extending the mle consistency for families of non-continuous densities such as the uniform distribution.
- Similar consistency results apply to estimators maximizing the pseudo-likelihood or composite likelihood. Once identifiability is ensured, these extensions follow in a straightforward way by applying similar arguments to each conditional in the pseudo likelihood or each likelihood object in composite likelihood.
- The compactness of Θ may be seen as rather restrictive. It is possible, however, to extend the theorem to open sets of $\mathbb{R}^k$ assuming continuity and differentiability of p_θ in $\theta \in \Theta$. For a proof see Chapter 18 in [1].

References

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