May 6, 2003

3.8 Efficiency of maximum likelihood estimators. Let $K >> M$ for $m \times m$ matrices K, M mean that $K - M$ is nonnegative definite. Let T_n be a sequence of estimators such that the distribution of $\sqrt{n}(T_n - \theta)$ under $\Pr_\theta$ is asymptotically $N(0, v(\theta))$. By Theorem 3.7.11, under its assumptions, $v(\theta) >> I(\theta)^{-1}$ for Lebesgue almost all θ. Thus, the sequence $\{T_n\}$ will be called "efficient" if for all θ, under $\Pr_\theta$, $\sqrt{n}(T_n - \theta)$ is asymptotically $N(0, v(\theta))$ with $I(\theta)^{-1} >> v(\theta)$. In practice, efficient estimators will have $v(\theta) = I(\theta)^{-1}$ for all θ. The definition allows for superefficiency for some set of θ which, under the conditions of Sec. 3.7, will have Lebesgue measure 0. The efficiency of maximum likelihood estimators with $v(\theta) \equiv I(\theta)^{-1}$ will be proved under the following assumptions.

(EML-1) $\{P_\theta, \ \theta \in \Theta\}$ is an equivalent family of laws on a sample space $(X, \mathcal{B})$ having densities $f(\theta, \cdot) > 0$ with respect to a σ-finite measure μ, where Θ is an open subset of a Euclidean space $\mathbb{R}^m$. The observations $X_1, X_2, \ldots,$ are i.i.d. (P_{θ_0}) for some $\theta_0 \in \Theta$.

Let $L(\theta, x) := \log f(\theta, x)$ and $\psi(\theta, x) := \nabla_\theta L(\theta, x)$ where ∇_θ denotes gradient with respect to θ.

(EML-2) For each $x \in X$, $f(\cdot, x)$ is C^1 with respect to θ, and the Fisher information matrix $I(\cdot)$ exists on Θ and is continuous and non-singular at θ_0.

If $E_\theta(\nabla_\theta L(\theta, x) = 0$, which will be proved in Theorem 3.8.1 to follow from the given assumptions, then $I(\theta)$ is the covariance matrix C of $\psi(\theta, x)$.

(EML-3) $\{T_n\}$ is a sequence of maximum likelihood estimators and is consistent, in other words $T_n \to \theta$ in $\Pr_\theta$-probability as $n \to \infty$ for all θ.

Conditions for consistency of M-estimators were given in Sections 3.3 and 3.5.

Conditions (AN-4) and (AN-5)(ii) in Section 3.6 will be assumed, locally uniformly in θ_0. Specifically, recall that for $\delta > 0$ small enough, depending on θ,

$$u(\theta, x, \delta) \ := \ \sup\{|\psi(\eta, x) - \psi(\theta, x)| : \ |\eta - \theta| \le \delta\}.$$

(EML-4) (i) For each $\theta, \phi \in \Theta$, $\lambda_\phi(\theta) \ := \ E_\phi \psi(\theta, x)$ exists in $\mathbb{R}^m$. Let $\lambda(\cdot) \ := \ \lambda_{\theta_0}(\cdot)$.

(ii) For some numbers $b > 0$ and $\gamma > 0$, and some neighborhood U of θ_0, for all $\phi, \theta \in U$, $|\eta - \phi| < \gamma$ implies $\eta \in \Theta$, and $\max(E_\theta u(\phi, x, \delta), E_\theta[u(\phi, x, \delta)^2]) \ \le \ b\delta$ for any δ such that $0 \le \delta \le \gamma/2$.

(EML-5) For some neighborhood V of θ_0, $\sup_{\theta \in V} E_\theta |\psi(\theta, x)|^2 < \infty$.

As in Theorem 3.6.15, let A be the Fréchet derivative of $\lambda(\cdot)$ at θ_0 if it exists.

3.8.1 Theorem. Assume (EML-1) through (EML-5). Then $\lambda(\theta_0) = 0$ and A exists with $A = -I(\theta_0)$. Also, the distribution of $\sqrt{n}(T_n - \theta_0)$ converges to $N(0, I(\theta_0)^{-1})$ as $n \to \infty$.

Proof. Take $b, \gamma > 0$ such that (EML-4)(ii) holds and such that $|\phi - \theta_0| < \gamma$ implies $\phi \in U \cap V$ for V in (EML-5). For $\theta \ne \theta_0$ with $|\theta - \theta_0| < \gamma/2$ and $0 \le t \le 1$, let $\theta_t \ := \ \theta_0 + t(\theta - \theta_0)$. Then for each x, by (EML-2),

$$L(\theta, x) - L(\theta_0, x) \ = \ \int_0^1 \psi(\theta_t, x) dt \cdot (\theta - \theta_0).$$

By Theorem 3.3.15 (about Kullback-Leibler divergence),

$$0 \;\geq\; \int [L(\theta,x) - L(\theta_0,x)] f(\theta_0,x)d\mu(x) \;=$$

$$\int \int_0^1 \psi(\theta_t,x)dt\ f(\theta_0,x)d\mu(x)\cdot(\theta-\theta_0) \;=\; \int_0^1 \lambda(\theta_t)dt\cdot(\theta-\theta_0)$$

where the interchange of integrals is justified since by (EML-4)(ii) for $\phi = \theta_0$, $|\psi(\theta_t,x)| \leq |\psi(\theta_0,x)| + u(\theta_0,x,\gamma/2)$, an integrable function for P_{θ_0}. Since $\lambda(\cdot)$ is continuous on U, also by (EML-4)(ii),

$$\int_0^1 \lambda(\theta_{tu})dt \to \lambda(\theta_0) \quad \text{as} \quad u \downarrow 0,$$

and $0 \geq \int_0^1 \lambda(\theta_{tu})dt\cdot(\theta_u - \theta_0)/u = \int_0^1 \lambda(\theta_{tu})dt\cdot(\theta-\theta_0)$. So $\lambda(\theta_0)\cdot(\theta-\theta_0) \leq 0$ for any θ in a neighborhood of θ_0, which implies $\lambda(\theta_0) = 0$. Also, by the same argument applied to ϕ such that $|\phi - \theta_0| \leq \gamma/2$ in place of θ_0, $\int \psi(\phi,x)f(\phi,x)d\mu(x) = 0$ for all $\phi \in U$. For (column) vectors $\eta,\zeta \in \mathbb{R}^m$, $\eta'\zeta = \eta\cdot\zeta \in \mathbb{R}$ and $\eta\zeta' = \eta\otimes\zeta$ is the $m\times m$ matrix $\{\eta_i\zeta_j\}_{i,j=1}^m$. Next, for $|\theta-\theta_0| \leq \gamma/2$,

$$\lambda(\theta) - \lambda(\theta_0) \;=\; \lambda(\theta) \;=\; \int \psi(\theta,x)f(\theta_0,x)d\mu(x)$$

$$= \; -\int \psi(\theta,x)[f(\theta,x) - f(\theta_0,x)]d\mu(x)$$

$$= \; -\int \psi(\theta,x)\left[\int_0^1 \psi(\theta_t,x)f(\theta_t,x)dt\cdot(\theta-\theta_0)\right] d\mu(x)$$

$$= \; -\int \psi(\theta,x)\left[\int_0^1 \psi(\theta_t,x)f(\theta_t,x)dt\right]' d\mu(x)(\theta-\theta_0).$$

Now,

$$\int \psi(\theta,x)\left[\int_0^1 \psi(\theta_t,x)f(\theta_t,x)dt\right]' d\mu(x)$$

$$= \; \int\int_0^1 \psi(\theta_t,x)\psi(\theta_t,x)'f(\theta_t,x)dt\ d\mu(x) + r(\theta) \;=\; \int_0^1 I(\theta_t)dt + r(\theta)$$

where, interchanging integrals by the Tonelli-Fubini theorem,

$$|r(\theta)| \;\leq\; \int_0^1 \int u(\theta_t,x,|\theta-\theta_0|)|\psi(\theta_t,x)|f(\theta_t,x)d\mu(x)dt \;\leq\; O(|\theta-\theta_0|^{1/2})$$

by the Cauchy-Bunyakovsky-Schwarz inequality applied to the two functions $g_t(\theta,x) := u(\theta_t,x,|\theta-\theta_0|)f(\theta_t,x)^{1/2}$ using (EML-4)(ii) for $\phi = \theta_t$ and $h_t(\theta,x) := |\psi(\theta_t,x)|f(\theta_t,x)^{1/2}$ using (EML-5). Since $I(\cdot)$ is continuous at θ_0, it follows that $A = -I(\theta_0)$ as stated. Recall that the covariance C of $\psi(\theta_0,x)$ is $I(\theta_0)$.

Next, we need to check the hypotheses of Section 3.6. (AN-1) follows from (EML-1) and (EML-3). In (AN-2), measurability of $\psi(\theta, \cdot)$ follows from that of $f(\theta, \cdot)$ as a density, and the fact that the components of the gradient of the measurable function $L(\cdot, \cdot)$ with respect to θ, which exist by (EML-2), are measurable as limits of a sequence of measurable functions along sequences $\phi_k = \theta + (1/k)e_i$ as $k \to \infty$ where e_i is one of the m standard unit vectors. Separability of ψ follows from its continuity with respect to θ, (EML-2), since $f(\cdot, \cdot) > 0$ (EML-1). In (AN-3), existence of $\lambda(\theta)$ is assumed in (EML-4)(i), and $\lambda(\theta_0) = 0$ has been proved. (AN-4)(i) follows from $A = -I(\theta_0)$, as proved, and the fact that $I(\theta_0)$ is non-singular (EML-2). (AN-4)(ii) follows from (EML-4)(ii) and (AN-5) from (EML-5). So we have all the hypotheses (AN-1) through (AN-5). By Theorem 3.6.15, recalling that in this section $\psi(\theta, x) = \nabla_\theta L(\theta, x)$ with covariance $I(\theta_0)$ at $\theta = \theta_0$, the distribution of $\sqrt{n}(T_n - \theta_0)$ converges to $N(0, I(\theta_0)^{-1})$, proving the theorem. □

It can be interesting to investigate the possibility that the assumption in (EML-2) that $f(\cdot, x)$ be C^1 in θ might be weakened. Huber (1967) proposed that the derivative of $L(\cdot, \cdot)$ with respect to θ need only exist "in measure," not necessarily at all x or θ, one possible interpretation of which is: for each θ, there is a vector-valued function $\psi(\theta, x)$ such that for each $\phi \in \mathbb{R}^m$,

$$\lim_{t \to 0} [L(\theta + t\phi, x) - L(\theta, x)]/t \;=\; \phi \cdot \psi(\theta, x), \tag{3.8.2}$$

where the convergence is in probability with respect to x, and $\theta + t\phi \in \Theta$ for t small enough. Consider the following

Example. Let $X = \Theta$ be the open interval $(0, 1) \subset \mathbb{R}$ and let

$$f(\theta, x) \;:=\; (1 + \theta)^{-1} \left[1 + 1_{(0,\theta]}(x)\right]$$

with respect to Lebesgue measure. Since $2\theta + (1 - \theta) \;\equiv\; 1 + \theta$ this does give probability densities. We have

$$L(\theta, x) \;=\; -\log(1 + \theta) + (\log 2) 1_{(0,\theta]}(x),$$

and $\partial L(\theta, x)/\partial\theta = -1/(1 + \theta)$ for $x \neq \theta$, so this is the derivative in probability $\psi(\theta, x)$ by the definition (3.8.2). Strangely, it does not depend on x. Thus $\lambda(\theta) = -1/(1 + \theta)$ also.

For n i.i.d. observations $X_1, \ldots, X_n$, $1/n$ times the log likelihood is

$$L_n\left(\theta, \{X_j\}_{j=1}^n\right) \;:=\; \frac{1}{n} \sum_{j=1}^n L(\theta, X_j) \;=\; -\log(1 + \theta) + (\log 2) F_n(\theta),$$

where F_n is the empirical distribution function based on $X_1, \ldots, X_n$. Let the true parameter $\theta_0 = \phi$ for some $\phi \in (0, 1)$. We know by the Glivenko-Cantelli theorem (RAP, Theorem 11.4.2) that almost surely $F_n(t)$ converges to the true distribution function $F(t)$ uniformly in t. To find a maximum likelihood estimate of θ, we need approximately to maximize $\eta(\theta) \;:=\; -\log(1 + \theta) + (\log 2) F(\theta)$. The derivative of this with respect to θ is

$$\eta'(\theta) \;=\; -(1 + \theta)^{-1} + (\log 2)(1 + \phi)^{-1} \left[1 + 1_{(0,\phi]}(\theta)\right]$$

for $\theta \neq \phi$. The right term is piecewise constant in θ, and the derivative of $-(1+\theta)^{-1}$ is $(1+\theta)^{-2} > 0$. It follows that η equals the convex function $-\log(1+\theta)$ plus a piecewise linear function, so it is convex on each interval $[0, \phi]$ and $[\phi, 1]$. Clearly $\eta(0) = \eta(1) = 0$. We have $\eta(\phi) = -\log(1+\phi) + (\log 2)(2\phi)/(1+\phi)$. To show that this is strictly positive for $0 < \phi < 1$ we want to show that $(1+\phi)\log(1+\phi) < (2\log 2)\phi$. Both sides are 0 at 0 and equal $2\log 2$ at 1. The left side is strictly convex since its second derivative is $1/(1+\phi) > 0$, and the right side is linear, so it's true that $\eta(\phi) > 0$ for $0 < \phi < 1$. At $\theta = \phi$, the left and right derivatives of η satisfy $\eta'(\phi-) > 0$, $\eta'(\phi+) < 0$. Thus $\theta = \phi$ gives a local maximum of η. By the convexity on $[0, \phi]$ and $[\phi, 1]$ and since $\eta(0) = \eta(1) = 0$, $\theta = \phi$ gives the unique global maximum of $\eta(\cdot)$.

Since F_n is a right-continuous step function and increases at its jumps, maximum likelihood estimators will exist for all n and each equals one of $X_1, \ldots, X_n$. Almost surely the values of the log likelihood at the X_j are all different, so the MLE is unique. By the Glivenko-Cantelli theorem, the maximum likelihood estimators will be consistent (converge to ϕ). Thus (EML-3) holds. It is not hard to verify that (EML-1), (EML-4), and (EML-5) all hold and that the Fisher information $I(\theta)$ exists and is continuous and non-zero. Thus all of (EML-1) through (EML-5) hold except that in (EML-2), the C^1 condition has been weakened to differentiability in probability. The given proof of Theorem 3.8.1 doesn't work in this case, since in the first display, $L(\theta, x) - L(\theta_0, x)$, which does depend on x, can't be equal to an integral of ψ which doesn't depend on x.

NOTE

The proof on efficiency of the maximum likelihood estimator is from Huber (1967). Assumption (EML-2) is strengthened to make the proof work.

REFERENCE

Huber, P. J. (1967). See Sec. 3.3.