

Asymptotics for ML-Estimates

Model: $X = (X_1, \dots, X_N)^T \in \mathbb{R}^N$, $\mathcal{L}(X) \in \{P_{\theta, N}; \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}$, where $P_{\theta, N}$ has density $p_\theta(x)$, $x \in \mathbb{R}^N$.

Assumption: (A1) $p_\theta(x)$ continuously differentiable w.r.t. θ for almost all x .

Likelihood and log likelihood: $L(\theta|X) = p_\theta(X)$, $\ell(\theta|X) = \log L(\theta|X)$

Notation: $p'_\theta(x) = \frac{\partial}{\partial \theta} p_\theta(x)$, $p''_\theta(x) = \frac{\partial^2}{\partial \theta^2} p_\theta(x), \dots$

Definition: score function $\psi(\theta|X) = \frac{\partial}{\partial \theta} \ell(\theta|X) = \frac{p'_\theta(x)}{p_\theta(x)}$

Fisher information $I(P_{\theta, N}) = \mathcal{E}_\theta \psi^2(\theta|X)$

Lemma 1 *If (A1) holds, we have*

a) $\mathcal{E}_\theta \psi(\theta|X) = 0$

b) $I(P_{\theta, N}) = -\mathcal{E} \frac{\partial^2}{\partial \theta^2} \ell(\theta|X)$, if $p''_\theta(x)$ exists for almost all x .

Idea of proof: By (A1), differentiation and integration may be exchanged. Use $\int p_\theta(x) dx = 1$ independently of θ .

If $\theta \in \Theta \subseteq \mathbb{R}^d$, then

$$\psi(\theta|X) = \text{grad}_\theta \ell(\theta|X) = \left(\frac{\partial}{\partial \theta_1} \ell(\theta|X), \dots, \frac{\partial}{\partial \theta_d} \ell(\theta|X) \right)^T$$

$$I(P_{\theta, N}) = -\mathcal{E}_\theta \psi(\theta|X) \psi^T(\theta|X) = -\mathcal{E}_\theta \text{Hess}_\theta \ell(\theta|X) = -\mathcal{E}_\theta \left(\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ell(\theta|X) \right)_{i,j=1, \dots, d}.$$

The results and proofs of this section can be generalized to $d > 1$ straightforwardly.

Theorem 1 *(Cramér-Rao or information inequality).*

Let (A1) be satisfied. Then, for any estimate T_N of θ with bias $b_T(\theta)$:

$$\text{mse}_\theta(T_N) \geq \frac{(1 + b'_T(\theta))^2}{I(P_{\theta, N})}.$$

Idea of proof: Differentiate $b_T(\theta)$, write $p'_\theta(x) = \psi(\theta(x))p_\theta(x)$, apply Cauchy-Schwarz inequality.

Corollary 0.1 *(Cramér-Rao for i.i.d. data)*

Let X_1, \dots, X_N be i.i.d. with $\mathcal{L}(X_j) = P_\theta$, density $f_\theta(x)$, $x \in \mathbb{R}$.

Under the assumptions of Theorem 1:

$$\text{mse}_\theta(T_N) \geq \frac{(1 + b'_T(\theta))^2}{N I(P_\theta)}.$$

Idea of proof: In this case, $\psi(\theta|X_1, \dots, X_N) = \sum_{j=1}^N \psi(\theta|X_j)$, $\psi(\theta(x)) = \frac{\partial}{\partial \theta} \log f_\theta(x)$. Use independence and Lemma 1 a) to show $I(P_{\theta, N}) = N I(P_\theta)$.

Definition 2 Let $Q_N : \mathbb{R}^N \times \Theta \rightarrow \mathbb{R}$ be given. $\hat{\theta}_N$ is a M-estimate of θ if

$$\hat{\theta}_N = \arg \max_{\theta \in \Theta} Q_N(X, \theta).$$

Theorem 3 (*Consistency of M-estimates*).

Let Θ be compact and

C1) $Q_N(x, \theta)$ continuous in θ

C2) For some continuous function $q : \Theta \rightarrow \mathbb{R}$

$$\frac{1}{N} Q_N(X, \theta) \xrightarrow{p} q(\theta) \quad \text{uniformly in } \theta \in \Theta$$

C3) $q(\theta)$ has a unique global maximum at $\theta_0 \in \Theta$.

Then,

$$\hat{\theta}_N = \arg \max_{\theta} Q_N(X, \theta) \xrightarrow{p} \theta_0 = \arg \max_{\theta} q(\theta).$$

Idea of proof: For $\delta = q(\theta_0) - \max_{|\theta - \theta_0| > \varepsilon} q(\theta)$ show, using $Q_N(X, \theta_0) \leq Q_N(X, \hat{\theta}_N)$, that $|\hat{\theta}_N - \theta_0| > \varepsilon$ implies $\sup_{\theta} |\frac{1}{N} Q_N(X, \theta) - q(\theta)| > \frac{\delta}{2}$. Apply C2).

Theorem 4 (*Asymptotic normality of M-estimates*)

Let $\hat{\theta}_N, \theta_0$ be as in Theorem 3, assume $\hat{\theta}_N \xrightarrow{p} \theta_0$ and

B1) $Q_N(x, \theta)$ is twice continuously differentiable w.r.t. θ for almost all x

B2) $\frac{1}{N} Q_N''(X, \theta_N^*) \xrightarrow{p} a(\theta_0) \neq 0$ if $\theta_N^* \xrightarrow{p} \theta$

B3) $\frac{1}{\sqrt{N}} Q_N'(X, \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, b(\theta_0))$ with $0 < b(\theta_0) = \lim_{N \rightarrow \infty} \frac{1}{N} \mathcal{E}[Q_N'(\theta_0)]^2 < \infty$.

Then,

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{b(\theta_0)}{a^2(\theta_0)}\right).$$

Idea of proof: **Delta-method** or linearization around θ_0

$$0 = Q_N'(X, \hat{\theta}_N) = Q_N'(X, \theta_0) + (\hat{\theta}_N - \theta_0) Q_N''(X, \theta_N^*)$$

Solve for $\hat{\theta}_N - \theta_0$, apply B2), B3) and Slutsky's Lemma.

We now apply the previous theorems to the special case of ML-estimates based on i.i.d. data X_1, \dots, X_N with density $f_{\theta_0}(x)$. Here,

$$Q_N(x, \theta) = \ell(\theta|x) = \sum_{j=1}^N \log f_{\theta}(x_j), \quad x = (x_1, \dots, x_N)^T \in \mathbb{R}^N$$

By LLN,

$$\frac{1}{N} Q_N(X, \theta) \xrightarrow{p} \mathcal{E}_{\theta_0} \log f_{\theta}(X_1) =: q(\theta)$$

if the right-hand side exists. To get the crucial assumption C2), we need:

Theorem 5 (*Uniform law of large numbers*):

Let X_1, \dots, X_N be i.i.d. real random variables, $\Theta \subseteq \mathbb{R}$ compact, $g : \mathbb{R} \times \Theta \rightarrow \mathbb{R}$ measurable such that

U1) $\mathcal{E}|g(X_1, \theta)| < \infty$ for all $\theta \in \Theta$

U2) Lipschitz continuity in θ : $|g(x, \theta) - g(x, \eta)| \leq L(x)|\theta - \eta|$ for all $\theta, \eta \in \Theta$, $x \in \mathbb{R}$

U3) $\mathcal{E} L(X_1) < \infty$

Then,

$$\sup_{\theta \in \Theta} \left| \frac{1}{N} \sum_{j=1}^N g(X_j, \theta) - \mathcal{E}g(X_1, \theta) \right| \xrightarrow{p} 0 \quad (N \rightarrow \infty).$$

Idea of proof: By compactness, we have finitely many $\theta_1, \dots, \theta_K$ such that Θ is covered by the intervals $(\theta_k - \delta, \theta_k + \delta)$, $k = 1, \dots, K$. By LLN,

$$\frac{1}{N} \sum_{j=1}^N g(X_j, \theta) \xrightarrow{p} \mathcal{E}g(X_1, \theta),$$

and, therefore,

$$\sup_{1 \leq k \leq K} \left| \frac{1}{N} \sum_{j=1}^N g(X_j, \theta_k) - \mathcal{E}g(X_1, \theta_k) \right| \xrightarrow{p} 0.$$

The supremum overall θ is bounded from above by the sum of this term and of

$$\sup_{1 \leq k \leq K} \sup_{|\theta - \theta_k| < \delta} \left| \frac{1}{N} \sum_{j=1}^N \{g(X_j, \theta) - g(X_j, \theta_k)\} - \mathcal{E}\{g(X_1, \theta) - g(X_1, \theta_k)\} \right|.$$

By assumption U2), this is bounded by

$$\delta \left\{ \frac{1}{N} \sum_{j=1}^N L(X_j) - \mathcal{E} L(X_1) + \text{const.} \right\},$$

$\frac{1}{N} \sum_{j=1}^N L(X_j) \xrightarrow{p} \mathcal{E} L(X_1)$ by LLN and δ can be chosen arbitrarily small.

Theorem 6 (*Consistency of ML-estimates*)

Let X_1, \dots, X_N be i.i.d. with density $f_{\theta_0}(x)$ for some $\theta_0 \in \Theta \subseteq \mathbb{R}$, Θ compact. Assume

M1) support of $f_{\theta} = \overline{\{x; f_{\theta}(x) > 0\}}$ independent of θ

M2) $f_{\theta}(x)$ continuously differentiable w.r.t. θ with $|f'_{\theta}(x)| \leq L(x)f_{\theta}(x)$ for all θ, x

M3) $\mathcal{E}_{\theta_0} L(X_1) < \infty$.

M4) $\mathcal{E}_{\theta_0} \log f_{\theta}(X_1)$ exists for all θ and has a unique global maximum in θ_0 .

Idea of proof: Check assumptions of Theorem 3, where, in particular, M2) guarantees Lipschitz continuity of $\log f_{\theta}(x)$ such that Theorem 5 may be applied.

Theorem 7 (*Asymptotic normality of ML-estimates*)

Let the assumptions of Theorem 6 be satisfied and

M5) $f_\theta(x)$ is twice continuously differentiable w.r.t. θ

M6) $\psi'(\theta|x) = \frac{\partial^2}{\partial \theta^2} \log f_\theta(x)$ is Lipschitz continuous in θ :

$$|\psi(\theta|x) - \psi(\eta|x)| \leq H(x)|\theta - \eta| \quad \text{for all } \theta, \eta \in \Theta, x \in \mathbb{R}$$

with $\mathcal{E} H(X_1) < \infty$.

M7) $0 < I(P_{\theta_0}) < \infty$.

Then,

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \frac{1}{I(P_{\theta_0})})$$

Idea of proof: Check assumptions of Theorem 4. In particular, B3) follows from the CLT for i.i.d. data applied to $Q_N(X, \theta_0) = \sum_{j=1}^N \psi(\theta_0|X_j)$, as $\mathcal{E} \psi(\theta|X_1) = 0$ by Lemma 1 a) and $\text{var} \psi(\theta_0|X_1) = I(P_{\theta_0})$.

B2) follows from

$$\frac{1}{N} Q_N''(X, \theta_N^*) = \frac{1}{N} \sum_{j=1}^N \{\psi'(\theta_N^*|X_j) - \psi'(\theta_0|X_j)\} + \frac{1}{N} \sum_{j=1}^N \psi'(\theta_0|X_j)$$

where the first term converges to 0 for $\theta_N^* \xrightarrow{p} \theta_0$ by M6) and the LLN, and the second term converges by the LLN to $\mathcal{E} \psi'(\theta_0|X_1) = -I(P_{\theta_0})$ by Lemma 1 b).

Remark 8 To apply the theorem to Θ which is not compact, one has to show as a first step that there is a compact subset $\Theta_0 \subseteq \Theta$ with $\theta_0 \in \Theta_0$ such that $\text{pr}(\hat{\theta}_N \in \Theta_0) \rightarrow 1$ for $N \rightarrow \infty$ fast enough. Then, $\hat{\theta}_N$ behaves asymptotically like $\hat{\theta}_N^0 = \arg \max_{\theta \in \Theta_0} Q_N(X, \theta)$, and the theorems may be applied to the latter.

Let $T_{1,n}, T_{2,n}$ denote two estimates of θ based on samples of size n , and let R denote the risk, e.g. mean-squared error.

Definition 9 a) If $R(T_{1,M}, \theta) = R(T_{2,N}, \theta)$, then $\frac{M}{N}$ is the relative efficiency of T_2 relative to T_1 (depending on R).

b) The asymptotic relative efficiency $ARE(T_2, T_1)$ is defined by

$$\frac{M}{N} \longrightarrow ARE(T_2, T_1) \quad \text{for } M, N \rightarrow \infty \text{ such that } R(T_{1,M}, \theta) = R(T_{2,N}, \theta).$$

In general, the efficiency depends on θ .

Example: For $R(T, \theta) = \text{mse}_\theta(T) = \mathcal{E}_\theta(T - \theta)^2$ and for two asymptotically unbiased and normal estimates, i.e.

$$\sqrt{n}(T_{1,n} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_1^2), \quad \sqrt{n}(T_{2,n} - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma_2^2)$$

we have

$$ARE(T_2, T_1) = \frac{\sigma_1^2}{\sigma_2^2}$$

as for large M, N

$$\frac{\sigma_1^2}{M} \sim \text{var}_\theta T_{1,M} \sim \text{mse}_\theta T_{1,M} \stackrel{!}{=} \text{mse}_\theta T_{2,N} \sim \text{var}_\theta T_{2,N} \sim \frac{\sigma_2^2}{N}$$

such that

$$\frac{M}{N} \sim \frac{\sigma_1^2}{\sigma_2^2}.$$

Corollary 0.2 Let $\hat{\theta}_N$ be the ML-estimate of θ_0 , and let T_N be any other unbiased estimate. Under the assumptions of Theorem 7

$$ARE(\hat{\theta}_N, T_N) \geq 1$$

i.e. $\hat{\theta}_N$ is asymptotically efficient.

Idea of proof: By Theorem 7, $\text{mse}(\hat{\theta}_N) \sim \frac{1}{NI(P_{\theta_0})}$,

By Corollary 0.1, $\text{mse}(T_N) \geq \frac{1}{NI(P_{\theta_0})}$.