1 Estimating Equations and Maximum Likelihood asymptotics

For an array $A \in \mathbb{R}^{d_1 \times \cdots \times d_p}$, let

$$|A| = \sum_{1 \leq i_1 \leq d_1, \ldots, j=1 \ldots p} |A_{i_1, \ldots, i_p}|,$$  \hspace{1cm} (1)

and for $x \in \mathbb{R}^{d_p}$ define $A[x] \in \mathbb{R}^{d_1 \times \cdots \times d_p-1}$ by specifying the array entries by

$$A[x]_{i_1, \ldots, i_p-1} = \sum_{1 \leq i_p \leq d_p} A_{i_1, \ldots, i_p-1, i_p} x_{i_p}.$$

The multiplication so defined reduces to the usual scalar and matrix-vector product when $p = 1$ and $p = 2$ respectively; we may write $Ax$ for $A[x]$. Note that with norm (1),

$$|A[x]| = \sum_{i_1, \ldots, i_{p-1}} |\sum_{i_p} A_{i_1, \ldots, i_p} x_{i_p}| \leq \sum_{i_1, \ldots, i_p} |A_{i_1, \ldots, i_p}| |x|_{i_p} \leq \sum_{i_1, \ldots, i_p} |A_{i_1, \ldots, i_p}| |x| = |A||x|.$$

For $x_j \in \mathbb{R}^{d_j}, j = p, \ldots, p-k+1$, let

$$A[x_p, \ldots, x_{p-k+1}] = (((Ax_p) x_{p-1}) \cdots) x_{p-k+1} \in \mathbb{R}^{d_1 \times \cdots \times d_{p-k}},$$

we have therefore, for instance

$$(A[x_p, \ldots, x_{p-k+1}])[x_{p-k}] = A[x_p, \ldots, x_{p-k}].$$

We adopt the usual convention that an empty product of the $d_j$’s is 1. For example, if $A \in \mathbb{R}^{d_1 \times d_2}$, and $x_j \in \mathbb{R}^{d_j}, j = 2, 1$ then $A[x_2, x_1] = x_1^T Ax_2 \in \mathbb{R}^1$. If $x_p = \ldots = x_{p-k+1} = x$, we let

$$A[x^{\otimes k}] = A[x, \ldots, x].$$

Let the parameter space $\Theta$ be an open subset of $\mathbb{R}^d$, and for $n \geq 1$ let $X_n \in \mathbb{R}^n$ be a random vector and $U_n(\theta) = U_n(X_n, \theta) \in \mathbb{R}^d$ with

$$U_n(X_n, \theta) = (U_{n,j}(X_n, \theta))_{1 \leq j \leq d}$$

where

$$U_{n,j} : \mathbb{R}^n \times \Theta \rightarrow \mathbb{R}.$$
We will assume that for the sequence of random vectors $X_n \in \mathbb{R}^n$, for all $n$ sufficiently large, there exists a solution $\hat{\theta}_n = \hat{\theta}_n(X_n)$ to the estimating equation

$$U_n(X_n, \theta) = 0. \quad (2)$$

With $\Theta_0 \subset \Theta$ open and $\theta_0 \in \Theta_0$ we assume that $U_n(\theta) = U_n(X_n, \theta)$ is twice differentiable for $\theta \in \Theta_0$, and define the arrays

$$U'_n(\theta) = \left( \frac{\partial U_{n,j}(\theta)}{\partial \theta_a} \right)_{1 \leq j, a \leq d} \quad \text{and} \quad U''_n(\theta) = \left( \frac{\partial^2 U_{n,j}(\theta)}{\partial \theta_a \partial \theta_b} \right)_{1 \leq j, a, b \leq d}.$$

**Theorem 1.1** Suppose there exists a sequence of real numbers $a_n$ such that

$$a_n U'_n(\theta_0) \to_p 0, \quad (3)$$

and a matrix $\Gamma$ such that

$$a_n U''_n(\theta_0) \to_p \Gamma, \quad (4)$$

with

$$\inf_{|h|=1} h^\top \Gamma h = \beta > 0. \quad (5)$$

Further, for any $\eta \in (0,1)$, suppose there exists a $K$ such that for all $n$ sufficiently large,

$$P(|a_n U''_n(\theta)| \leq K, \theta \in \Theta_0) \geq 1 - \eta. \quad (6)$$

Then for any given $\epsilon > 0$ and $\eta \in (0,1)$, for all $n$ sufficiently large, with probability at least $1 - \eta$ there exists $\hat{\theta}_n \in \Theta$ satisfying $U_n(\hat{\theta}_n) = 0$ and $|\hat{\theta}_n - \theta_0| \leq \epsilon$. Thus, there exists a consistent sequence of roots $\hat{\theta}_n$ to the estimating equation (2).

Note that $\partial U_{n,j}(\theta)/\partial \theta_a$ is not necessarily equal to $\partial U_{n,a}(\theta)/\partial \theta_j$, so that the limiting matrix $\Gamma$ is not necessarily symmetric. Condition (5) is equivalent to the condition that $\Gamma + \Gamma^\top$ is positive definite.

\[2\]
Proof: By replacing $U_n$ by $a_n U_n$, we may assume the conditions of Theorem 1.1 hold for $U_n$ with $a_n = 1$, and without loss of generality, by replacing $\theta$ by $\theta - \theta_0$, we take $\theta_0 = 0$. By Taylor's theorem, for $\theta \in \Theta_0$,

$$U_n(\theta) = U_n(0) + U'_n(0)[\theta] + \frac{1}{2}U''_n(\theta_0^*)[\theta^2]$$

where $\theta_0^*$ lies on the line connecting $\theta$ and 0. Evaluation at $\theta$ yields

$$U_n(\theta)[\theta] = U_n(0)[\theta] + U'_n(0)[\theta^2] + \frac{1}{2}U''_n(\theta_0^*)[\theta^3].$$

For the given $\eta \in (0, 1)$, choose $K$ such that (6) holds with $\eta$ replaced by $\eta/2$, and for the given $\epsilon > 0$, take $0 < \delta < \epsilon$ such that

$$B_\delta \subset \Theta_0 \quad \text{and} \quad \beta > C\delta^2 \quad \text{where} \quad C = 2 + \frac{1}{2}K,$$

where

$$B_\delta = \{ \theta : |\theta| \leq \delta \}.$$

Now choose $n_0$ such that for $n \geq n_0$, with probability exceeding $1 - \eta/2$,

$$|U_n(0)| < \delta^2 \quad \text{and} \quad |U'_n(0) - \Gamma| < \delta.$$

Hence, for $n \geq n_0$ and $\theta \in B_\delta$, with probability at least $1 - \eta$, from (7), (8) and (6),

$$|U_n(\theta) - \Gamma\theta| \leq |U_n(\theta) - U'_n(0)[\theta]| + |(U'_n(0) - \Gamma)\theta|$$

$$= |U_n(0) + \frac{1}{2}U''_n(\theta_0^*)[\theta^2]| + |(U'_n(0) - \Gamma)\theta|$$

$$\leq \delta^2 + \frac{1}{2}K|\theta|^2 + \delta|\theta| \leq C\delta^2,$$

so

$$|U_n(\theta)[\theta] - \theta^T\Gamma\theta| \leq C\delta^3.$$

Hence, if $|\theta| = \delta$,

$$U_n(\theta)[\theta] \geq \theta^T\Gamma\theta - C\delta^3 \geq \beta|\theta| - C\delta^3 = \delta(\beta - C\delta^2) > 0.$$

Applying the argument in Lemma 2 of [1], assume for the sake of contradiction that $U_n(\theta)$ does not have a root in $B_\delta$. Then for $\theta \in B_\delta$, the function $f(\theta) = -\delta U_n(\theta)/|U_n(\theta)|$ continuously maps $B_\delta$ to itself. By the Brouwer fixed point theorem, there exists $\vartheta \in B_\delta$, with $f(\vartheta) = \vartheta$. Since $|f(\theta)| = \delta$ for all $\theta \in B_\delta$, we have $|f(\vartheta)| = |\vartheta| = \delta$, which gives the contradiction $\delta^2 = |\vartheta|^2 = \vartheta^T\vartheta = f(\vartheta)[\vartheta] < 0$. \qed
**Theorem 1.2** Suppose that the solution $\hat{\theta}_n$ to (2) is consistent for $\theta_0$, that is, $\hat{\theta}_n \to_p \theta_0$. Let $a_n$ be as in (4) and (6), and assume that the matrix $\Gamma$ is non-singular. Further, let $b_n$ be such that for some random variable $Y$,

$$b_n U_n(\theta_0) \to_d Y.$$  

Then

$$c_n(\hat{\theta}_n - \theta_0) \to_d -\Gamma^{-1}Y \quad \text{where} \quad c_n = \frac{b_n}{a_n}.$$  

**Proof:** Without loss of generality replace $a_n U_n$ by $U_n$ and take $a_n = 1$, and set $\theta_0$ to 0. Since the limit in distribution does not depend on events of vanishingly small probability, we may assume $\hat{\theta}_n \in \Theta_0$ and $\Gamma_n^{-1}$ exists for all $n$ sufficiently large. For such $n$ setting $\theta = \hat{\theta}_n$ in (7), using $U_n(\hat{\theta}_n) = 0$,

$$U_n'(0)[\hat{\theta}_n] = -U_n(0) - \frac{1}{2} U_n''(\theta_0^*)[\hat{\theta}_n^2] = -U_n(0) - \epsilon_n[\hat{\theta}_n], \quad (10)$$

where

$$\epsilon_n = \frac{1}{2} U_n''(\theta_0^*)[\hat{\theta}_n].$$

Setting

$$\Gamma_n = U_n'(0) + \epsilon_n,$$

and multiplying by $c_n$, we rewrite (10) as

$$\Gamma_n c_n \hat{\theta}_n = -c_n U_n(0).$$

Using $\theta_0^* \in \Theta_0$, (6) and that $\hat{\theta}_n \to_p 0$, we have $|\epsilon_n| \to_p 0$, and consequently $\Gamma_n \to_p \Gamma$ and $\Gamma_n^{-1} \to_p \Gamma^{-1}$. Now by Slutsky’s Theorem (e.g. see [3]) and (9),

$$c_n \hat{\theta}_n = \Gamma_n^{-1} \left( \Gamma_n c_n \hat{\theta}_n \right) = -\Gamma_n^{-1} \left( c_n U_n(0) \right) \to_d -\Gamma^{-1}Y. \quad \blacksquare$$

**Example 1.1** Maximum Likelihood estimation for the i.i.d. sample $X = (X_1, \ldots, X_n)$ from a density $p(x; \theta_0), \theta_0 \in \Theta \subset \mathbb{R}^d$. The score function, by independence, is the sum of score contributions from the individual $X_i$

$$U_n(X_n; \theta) = \sum_{i=1}^n U(X_i; \theta) \quad \text{where} \quad U(x; \theta) = -\left( \frac{\partial \log p(x; \theta)}{\partial \theta_j} \right)_{1 \leq j \leq d} \in \mathbb{R}^d.$$
Further

\[ U'(x, \theta) = \left( \frac{\partial^2 \log p(x; \theta)}{\partial \theta_j \partial \theta_a} \right)_{1 \leq j, a \leq d} \quad \text{and} \quad U''(x, \theta) = \left( \frac{\partial^3 \log p(x; \theta)}{\partial \theta_j \partial \theta_a \partial \theta_b} \right)_{1 \leq j, a, b \leq d}. \]

Taking \( a_n = n^{-1} \) the weak law of large numbers implies Conditions 3 and (4) when, respectively

\[ E_{\theta} U(X, \theta_0) = 0, \quad \text{and} \quad E_{\theta} U'(X, \theta_0) = \Gamma, \]

for which it suffices that differentiation with respect to \( \theta \) is allowed under the integral in

\[ \int p(x, \theta) dx = 1, \]

as then the score has mean zero and \( \Gamma \) is the Fisher information matrix

\[ \Gamma = \text{Var}_{\theta_0}(U'(X, \theta_0)). \]

The matrix \( \Gamma \) will be positive definite, and hence satisfy (5), when \( U'(X, \theta_0) \) is a non-degenerate random vector.

With \( b_n = n^{-1/2} \), the classical CLT for i.i.d. variables gives

\[ b_n U_n(X_n, \theta_0) \rightarrow_d Y \quad \text{where} \quad Y \sim N(0, \Gamma), \]

so that \( b_n/a_n = \sqrt{n} \) and \( \Gamma^{-1} Y \sim N(0, \Gamma^{-1}) \). Hence, when the remainder term condition (6) is satisfied, Theorem 1.1 and 1.2 gives that there exists a consistent sequence of roots \( \hat{\theta}_n \) for \( \theta_0 \) such that

\[ \sqrt{n}(\hat{\theta}_n - \theta_0) \rightarrow_d N(0, \Gamma^{-1}). \]

References

