Chapter 8

Local Asymptotic Normality

8.1 LAN and Gaussian shift families

AN::efficiency.LAN

In Chapter 3, pointwise Taylor series expansion gave quadratic approximations to to criterion functions $G_n(\theta) = n^{-1}\sum_{i\le n} g(X_i,\theta)$ for independent, identically distributed X_i's. For observations from a density $f(x,\theta)$, the choice $g(x,\theta) = \log f(x,\theta)$ then gives a quadratic approximation for the logarithm of the likelihood ratio process in $n^{-1/2}$ neighborhoods of some particular θ_0. As you see in the next few chapters, this quadratic approximation captures the most important features of the model, features that will lead us to a rigorous treatment of the concept of efficiency introduced in Chapter 1.

The quadratic approximation of log liklihood ratios is so important it is given a name, ***Local Asymptotic Normality***, or ***LAN*** for short. It turns out that the details leading to LAN are unimportant for the efficiency theorems. It might be that independence assumptions are needed to establish the approximation in particular cases, or it might be that the approximation is established by probabilistic arguments involving dependent random variables; but for the purposes of the efficiency arguments, it is only the LAN approximation itself that matters.

Recall the convention for probability measures $\mathbb{P}$ and $\mathbb{Q}$ on the same sigma-field: the likelihood ratio equals $d\mathbb{Q}/d\mathbb{P} = d\widetilde{\mathbb{Q}}/d\mathbb{P}$, the density with respect to $\mathbb{P}$ of the part $\widetilde{\mathbb{Q}}$ of $\mathbb{Q}$ that is dominated by $\mathbb{P}$.

LAN.defn <1>

Definition. *For each n let $\{\mathbb{Q}_{t,n} : t \in_n\}$ be a family of probabilty measures indexed by a subset T_n of $\mathbb{R}^k$. Suppose each point of $\mathbb{R}^k$ belongs to T_n for all n large enough. Let Γ be a positive definite matrix not depending on t. Say that the LAN condition holds at 0 with asymptotic variance matrix Γ if, for each fixed t in $\mathbb{R}^k$, the likelihood ratio has the representation*

$$L_n(t) := \frac{d\mathbb{Q}_{t,n}}{d\mathbb{Q}_{0,n}} = (1+\epsilon_n(t))\exp\left(t'Z_n - \tfrac{1}{2}t'\Gamma t\right),$$

where $\epsilon_n(t) = o_p(1)$ and $Z_n \rightsquigarrow N(0,\Gamma)$ under $\mathbb{Q}_{0,n}$. Say that the standardized LAN condition holds if Γ equals I_k, the identity matrix.

Remark. Many authors replace the $1 + \epsilon_n(t)$ factor by a $o_p(1)$ in the exponent, which leaves one to ponder whether a $o_p(1)$ can take the value $-\infty$ to accomodate the low probability cases where $L_n(t)$ is zero.

Typically the $\mathbb{Q}$'s will come from a local reparametrization of a model $\{\mathbb{P}_{\theta,n} : \theta \in \Theta\}$ around some point θ_0 in the interior of the parameter space: $\mathbb{Q}_{t,n} := \mathbb{P}_{\theta_0+t/\sqrt{n}}$ or, more generally, $\mathbb{Q}_{t,n} := \mathbb{P}_{\theta_0+A_nt}$ for some sequence of deterministic matrices $\{A_n\}$ that converges to the zero matrix as $n \to \infty$.

Remark. The cautious wording in Definition <1> regarding the covering property of the T_n sets is merely to handle cases such as: $T_n = \{t \in \mathbb{R}^k : \theta_0 + A_nt \in \Theta\}$ for a fixed proper subset Θ of $\mathbb{R}^k$. If θ_0 lies in the interior of Θ, the T_n sets expand to cover the whole of $\mathbb{R}^k$; each t in $\mathbb{R}^k$ is eventually a member of T_n. Of course the limit assertions only make sense when interpreted as statements that apply to all n greater than some $n_0(t)$, but it would be too tedious if I were to spell out such a minor subtlety repeatedly.

The dual role for Γ in the LAN definition—as a limiting variance matrix for Z_n and as the matrix in the quadratic form—is no accident. As you saw in Chapter 5, we need it to get the contiguity, $\mathbb{Q}_{t,n} \lhd \mathbb{Q}_{0,n}$. Without contiguity many asymptotic arguments would fail for subtle reasons involving sets of small measure.

A simple reparametrization, with $\Gamma^{-1/2}\delta$ replacing t and $\Gamma^{-1/2}Z_n$ replacing Z_n, would reduce the LAN property to the simpler standardized form where Γ equals the identity matrix. It is notationally simpler to work with the standardized case. The translation to the case of general Γ should never present more than a notational problem.

In the standardized case, the limiting distribution of the likelihood ratio $L_n(t)$ under $\mathbb{Q}_{0,n}$ is just the distribution of

$$L(t) := d\mathbb{Q}_t/d\mathbb{Q}_0 = \exp(t'z - \frac{1}{2}|t|^2) \qquad \text{under } \mathbb{Q}_0,$$

for the ***Gaussian shift family*** $\mathcal{Q} = \{\mathbb{Q}_t : t \in \mathbb{R}^k\}$ with $\mathbb{Q}_t = N(0, I_k)$. Moreover, for each finite subset F of $\mathbb{R}^k$, the random vector $\{L_n(t) : t \in F\}$ converges in distribution (under $\mathbb{Q}_{0,n}$ to $\{L(t) : t \in F\}$ under $\mathbb{Q}_0$.

In some asymptotic sense, the model $\mathcal{Q}_n = \{\mathbb{Q}_{t,n} : t \in T_n\}\}$ measures behaves like a gaussian shift family. It was one of Lucien Le Cam's great insights that the solutions of certain asymptotic problems related to the local behavior of $\mathcal{P}_n := \{\mathbb{P}_{\theta,n} : \theta \in \Theta\}$ near a fixed θ_0 can be reduced to the solutions of the corresponding Gaussian shift problems. Chapter 10 will explore this idea in more detail, showing that the results for Gaussian

Just Le Cam? Wald? Hájek?

shifts (to be established in Chapter 9) provide lower asymptotic bounds for efficiency quantities calculated for $\mathcal{P}_n$.

8.2 Bahadur's rescue of efficiency

LAN::bahadur

Under a regularity assumption weaker than LAN (but in the same spirit) Bahadur (1964) was one of the first to rescue Fisher's concept of efficiency. Essentially he considered the behavior of an estimator T_n under two alternatives $\{\mathbb{P}_n\}$ or $\{\mathbb{Q}_n\}$, for which

nearLAN<2>
$$\frac{d\mathbb{Q}_n}{d\mathbb{P}_n} \rightsquigarrow \exp\left(tZ - \tfrac{1}{2}t^2\sigma^{-2}\right) \qquad \text{with } Z \sim N(0,\sigma^{-2}) \text{ under } \mathbb{P},$$

for some constant $t > 0$. The constant σ^{-2} corresponds to the Fisher information function evaluated at the value θ_0 that defined $\mathbb{P}_n$. The only other vestige of the underlying parameter is an assumption about the asymptotic behavior of some estimator T_n under each sequence of alternatives. Specifically, suppose there is a number θ_0 and a constant $\tau > 0$ for which

<3>
$$\sqrt{n}\,(T_n - \theta_0) \rightsquigarrow N(0,\tau^2) \qquad \text{under } \mathbb{P}_n.$$

<4>
$$\liminf_n \mathbb{Q}_n\{\sqrt{n}\,(T_n - \theta_n) < 0\} \le \tfrac{1}{2} \qquad \text{where } \theta_n := \theta_0 + t/\sqrt{n}.$$

The second assumption is weaker than an assumption of a $N(0,\tau^2)$ limiting distribution$\sqrt{n}(T_n - \theta_n)$ under $\mathbb{Q}_n$.

Assuming <2>, <3>, and <4>, Bahadur (1964) was able to rule out the possibility that $\tau^2 < \sigma^2$, the inequality corresponding to superefficiency of T_n at θ_0.

The proof that $\tau^2 \ge \sigma^2$ will follow as a simple consequence of the next Lemma, which captures the essence of Bahadur's main argument.

Bahadur.lemma <5>
Lemma. *Suppose $\mathbb{P}_n$ and $\mathbb{Q}_n$ are probability measures with $\mathbb{Q}_n$ contiguous to $\mathbb{P}_n$. Suppose $d\mathbb{Q}_n/d\mathbb{P}_n$, as random variables on $(\mathcal{X}_n,\mathcal{A}_n,\mathbb{P}_n)$, converge in distribution to a random variable L on $(\mathcal{X},\mathcal{A},\mathbb{P})$. Then for each sequence of measurable functions ψ_n with $0 \le \psi_n \le 1$, and each positive constant C,*

$$\liminf_n \left(\mathbb{P}_n\psi_n + C\mathbb{Q}_n\bar{\psi}_n\right) \ge \|\mathbb{P} \wedge (C\mathbb{Q})\|_1,$$

where $\mathbb{Q}$ is the probability measure on $(\mathcal{X},\mathcal{A})$ defined by $d\mathbb{Q}/d\mathbb{P} = L$.

Proof Write L_n for the density of the part of $\mathbb{Q}_n$ that is absolutely contnuous with respect to $\mathbb{P}_n$. We are assumimg that $L_n \rightsquigarrow L$. Thus

$$\mathbb{P}_n\psi_n + C\mathbb{Q}_n\bar{\psi}_n \ge \inf_{0\le\psi\le1} \mathbb{P}_n\left(\psi + CL_n\bar{\psi}\right) = \mathbb{P}_n\left(\{CL_n \le 1\} + CL_n\{CL_n > 1\}\right).$$

That is, the infimum is achieved when $\psi := \{CL_n \le 1\}$. Rewrite the last expectation as $\mathbb{P}_n\,(1 \wedge (CL_n))$. The map $x \mapsto 1 \wedge (Cx)$ is bounded and continuous on $\mathbb{R}^+$. The lower bound converges to $\mathbb{P}\,(1 \wedge (CL)) = \|\mathbb{P} \wedge (C\mathbb{Q})\|_1$, as asserted.

□

PROOF (OF THEOREM <??>) Identify the limit distribution for L_n with the distribution of the density $d\mathbb{Q}/d\mathbb{P}$, where $\mathbb{P} := N(0,\sigma^2)$ and $\mathbb{Q} := N(t,\sigma^2)$. Invoke the Lemma with $\psi_n := \{\sqrt{n}(T_n - \theta_n) \ge 0\}$ and $C := \exp\left(-\sigma^{-2}t^2/2\right)$. The lim inf of $\mathbb{P}_n\psi_n + C\mathbb{Q}_n\bar{\psi}_n$ is less than $\mathbb{P}\{N(-t,\tau^2) \ge 0\} + \frac{1}{2}C$. To calculate the norm of $\mathbb{P} \wedge (C\mathbb{Q})$, note that the $N(0,\sigma^2)$ density is smaller than C times the $N(t,\sigma^2)$ density at those points x of the real line for which $-\frac{1}{2}x^2\sigma^{-2} \le -\frac{1}{2}\sigma^{-2}t^2 - \frac{1}{2}\sigma^{-2}(x-t)^2$, that is, when $x \ge t$. Thus

$$\|\mathbb{P} \wedge (C\mathbb{Q})\| = \mathbb{P}[t,\infty) + C\mathbb{Q}(-\infty,t] = \bar{\Phi}(t/\sigma) + \tfrac{1}{2}C.$$

In order that $\bar{\Phi}(t/\tau) + \frac{1}{2}C \ge \bar{\Phi}(t/\sigma) + \frac{1}{2}C$, we must have $\tau \ge \sigma$.

□

8.3 The negligible set of points of superefficiency

LAN::bahadur

Should any of this section be salvaged for variations on Bahadur, or should it wait until Chap 10?

If f is a real valued function, write $\bar{f}$ for $1 - f$.

Recall from Section **??** that the affinity between two finite measures λ and μ on a space $(\mathcal{X},\mathcal{A})$ is

$$\alpha_1(\lambda,\mu) = \|\lambda - \mu\|_1 = \inf_{0\le f\le 1} \lambda f + \mu\bar{f},$$

where the infimum runs over all measurable functions with $0 \le f \le 1$. When $\mathcal{X}$ equals $\mathbb{R}^k$, we get the same infimum if f is also restricted to be continuous (Problem [4]).

The particular case of two normal distributions will be important for LAN models. From Section **??**:

normal.affinity<6>

$$\|N(t_1,\sigma_2) - N(t_2,\sigma^2)\|_1 = 2\mathbb{P}\{\,|N(0,1)| \le |t_1 - t_2|/\sigma\}$$

extra.bahadur <7>

Lemma. *Suppose $\mathbb{P}_n$ and $\mathbb{Q}_n$ are probability measures with $\mathbb{Q}_n$ contiguous to $\mathbb{P}_n$. Suppose $d\mathbb{Q}_n/d\mathbb{P}_n$, as random variables on $(\mathcal{X}_n,\mathcal{A}_n,\mathbb{P}_n)$, converge in distribution to a random variable L on $(\mathcal{X},\mathcal{A},P)$. Then for each sequence of measurable functions ψ_n with $0 \le \psi_n \le 1$, and each positive constant C,*

$$\liminf_n \left(\mathbb{P}_n\psi_n + C\mathbb{Q}_n\bar{\psi}_n\right) \ge \|P \wedge (CQ)\|_1,$$

where Q is the probability measure on $(\mathcal{X},\mathcal{A})$ defined by $dQ/dP = L$.

Proof Write L_n for the density $d\widetilde{\mathbb{Q}}_n/d\mathbb{P}_n$, where $\widetilde{\mathbb{Q}}_n$ denotes the part of $\mathbb{Q}_n$ that is absolutely contnuous with respect to $\mathbb{P}_n$. We are assumimg that $L_n \rightsquigarrow L$. Then

$$\begin{aligned}\mathbb{P}_n\psi + C\mathbb{Q}_n\bar{\psi} &\geq \mathbb{P}_n\left(\psi_n + CL_N\bar{\psi}\right)\\ &\geq \mathbb{P}_n\left(\{CL_n \leq 1\} + CL_N\{CL_n > 1\}\right))\end{aligned}$$

That is, the minimum is achieved when $\psi_n = \{CL_n \leq 1\}$. Rewrite the last expectation as $\mathbb{P}_n 1 \wedge (CL_N)$. The map $x \mapsto 1 \wedge (Cx)$ is bounded and continuous on $\mathbb{R}_+$. The lower bound converges to

$$P\left(1 \wedge (CL)\right) = P \min\left(\frac{dP}{dP}, \frac{d(CQ)}{dP}\right) = \|P \wedge (CQ)\|_1,$$

as asserted. □

shrink <8> **Corollary.** *Suppose $\{Y_n\}$ is sequence of random vectors for which $Y_n \rightsquigarrow \lambda$ under $\mathbb{P}_n$ and $Y_n \rightsquigarrow \mu$ under $\mathbb{Q}_n$, where λ and μ are probability measures on $\mathbb{R}^k$. Then for each positive constant C,*

$$\|\lambda \wedge (C\mu)\|_1 \geq \|P \wedge (CQ)\|_1$$

In particular, $\|\lambda - \mu\|_1 \leq \|P - Q\|_1$.

Proof For each continuous g with $0 \leq g \leq 1$ invoke the Lemma with $\psi_n = g(Y_n)$ to get

$$\lambda g(y) + C\mu\bar{g}(y) \geq \|P \wedge (CQ)\|_1$$

Take the infimum over all such g. □

One-dimensional? Suppose $(T_n - \theta)/\delta_n \rightsquigarrow N(0, \sigma_\theta^2)$ under $\mathbb{P}_{n,\theta}$. Assume LAN. Give Bahadur method for median unbiased, after first noting the proof via Corollary <8> and equality <6>.

<9> **Corollary.** *Suppose $\{T_n\}$ is sequence of random vectors, and λ_0 is a probability distribution on $\mathbb{R}^k$, for which $\sqrt{n}(T_n - \theta_0) \rightsquigarrow \lambda_0$ under $\mathbb{P}_n$ and $\sqrt{n}(T_n - \theta_0) - t \rightsquigarrow \lambda_0$ under $\mathbb{Q}_n$. Then, for each continuous g with $0 \leq g \leq 1$,*

$$\lambda_0 g(x) + C\lambda_0 g(x - t) \geq \|P \wedge (CQ)\|_1$$

In particular, if P is the $N(0,1)$ distribution and Q is the $N(t,1)$ distribution, then

$$\lambda_0(-\infty, z) + \lambda_0[z + t, \infty) \geq \mathbb{P}\{\,|N(0,1)| \geq t/2\} \qquad \textit{for every real } z$$

8.4 A classical sufficient condition for LAN

LAN::classicalLAN

Let $\{P_\theta : \theta \in \Theta\}$ be a family of probability measures on a space $(\mathcal{X}, \mathcal{A})$, indexed by a subset Θ of $\mathbb{R}^k$, with corresponding densities $\{f_\theta(x)\}$ with respect to a measure λ. Suppose observations $\{x_i\}$ are drawn independently from the distribution P_{θ_0}, where θ_0 is an interior point of Θ.

Under the classical regularity conditions—twice continuous differentiability of $\log f(x, \theta)$ with respect to θ, with a dominated second derivative—the log of the likelihood ratio

$$L_n(\theta) = \prod_{i \le n} \frac{f(x_i, \theta)}{f(x_i, \theta_0)}$$

has a local quadratic approximation in $1/\sqrt{n}$ neighborhoods of θ_0. Formally, the approximation results from the usual pointwise Taylor expansion of the log density $g(x, \theta) = \log f(x, \theta)$, following a familiar style of argument. For example, in one dimension,

$$\begin{aligned} \log L_n(\theta_0 + t/\sqrt{n}) &= \sum\nolimits_{i \le n} \big(g(x_i, \theta_0 + t/\sqrt{n}) - g(x_i, \theta_0) \big) \\ &= \frac{t}{\sqrt{n}} \sum\nolimits_{i \le n} g^\bullet(x_i, \theta_0) + \frac{t^2}{2n} \sum_{i \le n} g^{\bullet\bullet}(x_i, \theta_0) + \ldots \\ &\approx t Z_n - \frac{t^2}{2} \Gamma, \end{aligned}$$

where $\Gamma = -P_{\theta_0} g^{\bullet\bullet}(x, \theta_0)$ and

$$Z_n = \sum\nolimits_{i \le n} g^\bullet(x_i, \theta_0)/\sqrt{n} \rightsquigarrow N\big(0, \operatorname{var}_{\theta_0} g^\bullet(x, \theta_0)\big).$$

The limiting variance for Z_n and the coefficient Γ from the quadratic term both equal the information function evaluated at θ_0.

The methods from Chapter 3 can be used to establish such a quadratic approximation rigorously, with even a uniform $o_p(1)$ bound on the remainder for t in bounded neighborhoods of the origin, which is more than we need for LAN.

For simplicity of notation, again suppose $\theta_0 = 0$. Write N_0 for the set $\{x : f_0(x) = 0\}$, and $\alpha(\theta)$ for $P_\theta N_0$, the total mass of the part of P_θ that is not absolutely continuous with respect to P_0. The $\mathcal{F}_n$-measurable set $F_n := \bigcup_{i \le n} \{x_i \in N_0\}$ has zero P_0^n probability, but

$$P_\theta^n F_0^c = \prod\nolimits_{i \le n} P_\theta N_0^c = \big(1 - \alpha(\theta)\big)^n.$$

If $\alpha(\theta)$ were not of order $o(\theta^2)$ we could find a sequence $\{\theta_n\}$ of order $O(n^{-1/2})$ and an $\epsilon > 0$ for which $\alpha(\theta_n) \geq \epsilon/n$ infinitely often. We would then have a sequence for which $\liminf_n P^n_{\theta_n} F_n \geq 1 - e^{-\epsilon} > 0$ but $P^n_0 F_n \equiv 0$, ruling out contiguity. Thus a necessary condition for contiguity, $P^n_{\theta_n} \lhd P^n_0$ whenever $\theta_n = O(n^{-1/2})$ is

ctg.nec<10>
$$P_\theta\{x : f_0(x) = 0\} = o(\theta^2) \qquad \text{as } \theta \to 0.$$

Assumption <10> takes care of one difficulty in the the case when the sets $\{f_\theta > 0\}$ are not the same as $\{f_0 > 0\}$. Another, more subtle, problem arises with the definition of $\log f_\theta$. If $f_0(x) > 0$ then, by continuity, we know that $f_\theta(x) > 0$ for $|\theta| \leq \delta(x)$, but there is no guarantee of a fixed neighborhood U of 0 on which $f_\theta(x) > 0$ for all x. We might have $P_0 \log f_\theta(x) = -\infty$ for all $\theta \neq 0$, which would cast doubt on some of the calculations sketched at the start of this Section. The function $\ell_\theta(x) := \log f_\theta(\theta)$ might only be finite on an interval of θ values that depend on x. It still makes sense to work with the pointwise derivative $\dot\ell_0(x)$, but we might encounter the value $-\infty$ with positive P_0 probability when studying $\ell_\theta(x)$ for a fixed $\theta \neq 0$.

In view of these worrisome details, it is better to impose the regularity conditions directly on $f_\theta(x)$, and not on $\log f_\theta(x)$. It also simplifies matters greatly if we take densities with respect to P_{θ_0} and not with respect to an arbitrary dominating λ.

LANclassical<11> **Theorem.** *Let $p_\theta(x) = dP_\theta/dP_{\theta_0}$, the density with respect to P_{θ_0} of the part of P_θ that is dominated by P_{θ_0}. Suppose the map $\theta \mapsto p_\theta$ is twice differentiable in a neighborhood $\mathcal{U}$ of θ_0, which is an interior point of Θ. Let $\mathbb{P}_n = P^n_{\theta_0}$ and $\mathbb{Q}_n = P^n_{\theta_n}$, for a sequence $\theta_n := \theta_0 + t_n/\sqrt{n}$ with $\{t_n\}$ bounded. Suppose also that p is twice differentiable with:*

(i) $\theta \mapsto p^{\bullet\bullet}_\theta(x)$ is continuous at θ_0;

(ii) there exists an $M(x)$ in $\mathcal{L}^1(P_{\theta_0})$ for which $\sup_{\theta\in\mathcal{U}} |p^{\bullet\bullet}_\theta(x)| \leq M(x)$;

(iii) $|p^\bullet_\theta| \in \mathcal{L}^2(P_{\theta_0})$ for each $\theta \in \mathcal{U}$ and $P_{\theta_0}|p^\bullet_\theta|^2 \to P^x_{\theta_0}|p^\bullet_{\theta_0}|^2 < \infty$ as $\theta \to \theta_0$;

(iv) $P^\perp_\theta \mathcal{X} = o(|\theta - \theta_0|^2)$ as $\theta \to \theta_0$.

Then

(a) $P_{\theta_0} p^\bullet_{\theta_0} = 0$ and $P_{\theta_0} p^{\bullet\bullet}_{\theta_0} = 0$.

(b)
$$\frac{d\mathbb{Q}_n}{d\mathbb{P}_n} = (1 + o_p(1;\mathbb{P}_n)) \exp\left(t_n' Z_n - \tfrac{1}{2} t_n' \mathbb{I}_0 t_n\right),$$

where $\mathbb{I}_0 := P_{\theta_0}(p^\bullet_{\theta_0} p^\bullet_{\theta_0}{}')$ *and* $Z_n := \sum_{i\le n} p^\bullet_0(x_i)/\sqrt{n} \rightsquigarrow N(0,\mathbb{I}_0)$ *under* $\mathbb{P}_n$.

The main Taylor expansion ideas in the proof are captured by the following Lemma. It is worthwhile separating these arguments from the rest of the proof because the same Lemma will also be needed when establishing LAN under a DQM assumption in Section 6.

epsin<12> **Lemma.** *Suppose* $L_n = \prod_i (1+\epsilon_{i,n})$ *where* $\{\epsilon_{i,n} : 1 \le i \le k_n\}$ *are random variables for which*

(i) $\max_i |\epsilon_{i,n}| = o_p(1)$ *as* $n \to \infty$

(ii) $\sum_i \epsilon^2_{i,n} = O_p(1)$ *as* $n \to \infty$

Then $L_n = (1+o_p(1)) \exp(\sum_i \epsilon_{i,n} - \frac{1}{2}\sum_i \epsilon^2_{i,n})$.

(iii) If $\epsilon_{i,n} = U_{i,n} + V_{i,n}$ *with*

(a) $\sum_i U^2_{i,n} = O_p(1)$ *and* $\max_i |U_{i,n}| = o_p(1)$

(b) $\sum_i V_{i,n} = o_p(1)$ *and* $\sum_i V^2_{i,n} = o_p(1)$

then assumptions (i) and (ii) hold and

$$\sum\nolimits_i \epsilon_{i,n} - \tfrac{1}{2}\sum\nolimits_i \epsilon^2_{i,n} = \sum\nolimits_i U_{i,n} - \tfrac{1}{2}\sum\nolimits_i U^2_{i,n} + o_p(1)$$

Remark. In (iii) the extra $o_p(1)$ in the exponent can be absorbed into the $1+o_p(1)$ factor.

PROOF For the first assertion use the Taylor approximation

$$|\log(1+z) - z + \tfrac{1}{2}z^2| \le |z|^3 \qquad \text{for } |z| \le 1/2..$$

on the set $A_n = \{\max_i |\epsilon_{i,n}| \le 1/2\}$ to get

$$\begin{aligned} |\log(L_n) - \sum\nolimits_i \epsilon_{i,n} + \tfrac{1}{2}\sum\nolimits_i \epsilon^2_{i,n}| &\le \sum\nolimits_i |\epsilon_{i,n}|^3 \\ &\le \max_i |\epsilon_{i,n}| \sum\nolimits_i \epsilon^2_{i,n} = o_p(1), \end{aligned}$$

that is,

$$L_n A_n = A_n \exp\left(\delta_n Z_n - \tfrac{1}{2}\delta_n^2 \mathbb{I}_0 + o_p(1)\right).$$

The $1+o_p(1)$ factor in the statement of the Theorem absorbs the $o_p(1)$ in the exponent, as well as allowing for arbitrarily bad behavior of L_n on A_n^c.

□

Proof (of Theorem <11>) For simplicity of notation assume $\theta_0 = 0$ and write P for P_{θ_0}. Also, intepret all $o_p(\cdot)$ and $O_p(\cdot)$ as $o_p(\cdot;\mathbb{P}_n)$ and $O_p(\cdot;\mathbb{P}_n)$.

□

8.5 Differentiation of unit vectors

LAN::unit.vector

For a differentiable map $\theta \mapsto \xi_\theta$, the Cauchy-Schwarz inequality implies that $\langle \xi(\theta_0), r(\theta)\rangle = o(|\theta - \theta_0|)$. It would usually be a blunder to assume naively that the bound must therefore be of order $O(|\theta - \theta_0|^2)$; typically, higher-order differentiability assumptions are needed to derive approximations with smaller errors. However, if $\|\xi(\theta)\|$ is constant—that is, if the function is constrained to take values lying on the surface of a sphere—then the naive assumption turns out to be no blunder. Indeed, in that case, $\langle \xi(\theta_0), r(\theta)\rangle$ can be written as a quadratic in $\theta - \theta_0$ plus an error of order $o(|\theta - \theta_0|^2)$.

UNITvector<13>

Lemma. *Let $\theta \mapsto \xi_\theta$ be a map from $\mathbb{R}$ into $\mathcal{L}^2(\lambda)$ that is Hellinger differentiable at some θ_0, that is, $\xi_\theta = \xi_0 + (\theta - \theta_0)\Delta + r_\theta$, with $\Delta \in \mathcal{L}^2(\lambda)$ and $\|r_\theta\| = o(|\theta - \theta_0|)$ near θ_0. Then*

(i) $\langle \xi_0, \Delta\rangle = 0$

(ii) $2\langle \xi_0, r_\theta\rangle = -\theta^2\|\Delta\|^2 + o(|\theta - \theta_0|^2)$.

Stat 618 folks: Extend this lemma to cover the case where $\lambda = P_{\theta_0}$ and $P_\theta^\perp \mathcal{X} = o\left(|\theta - \theta_0|^2\right)$ near θ_0.

Proof Without loss of generality suppose $\theta_0 = 0$. Because both ξ_θ and ξ_0

have unit length,

$$\begin{aligned} 0 = \|\tau_n\|^2 - \|\tau_0\|^2 &= 2\alpha_n\langle\tau_0, W\rangle && \text{order } O(\alpha_n) \\ &\quad + 2\langle\tau_0, \rho_n\rangle && \text{order } o(\alpha_n) \\ &\quad + \alpha_n^2\|W\|^2 && \text{order } O(\alpha_n^2) \\ &\quad + 2\alpha_n\langle W, \rho_n\rangle + \|\rho_n\|^2 && \text{order } o(\alpha_n^2). \end{aligned}$$

On the right-hand side I have indicated the order at which the various contributions tend to zero. (The Cauchy-Schwarz inequality delivers the $o(|\theta|)$ and $o(|\theta|^2)$ terms.) The exact zero on the left-hand side leaves the leading $2\theta\langle\xi_0, \Delta\rangle$ unhappily exposed as the only $O(|\theta|)$ term. It must be of smaller order, which can happen only if $\langle\xi_0, \Delta\rangle = 0$, leaving

$$0 = 2\langle\xi_0, r_\theta\rangle + \theta^2\|\Delta\|^2 + o(|\theta|^2),$$

as asserted. □

Without the fixed length property, the inner product $\langle\xi_0, r_\theta\rangle$, which inherits $o(|\theta|)$ behaviour from $\|r_\theta\|$, might not decrease at the $O(|\theta|^2)$ rate.

8.6 LAN via DQM

LAN::DQMLAN

Le Cam (1970) showed that the LAN approximation holds under conditions much weaker than the classical smoothness and domination assumptions: Hellinger differentiability is almost enough. Only a few small, but very significant, details related to division by zero complicate the argument. To avoid these difficulties it is better to work with the DQM assumption, with densities p_θ with respect to P_{θ_0} for the part of P_θ that is dominated by P_{θ_0}.

To avoid notational fuss, let me again assume that $\theta_0 = 0$. Recall from Chapter 7 that DQM of $\theta \mapsto P_\theta$ at 0, with score function Δ, means

(i) $P_\theta^\perp(\mathcal{X}) = o(|\theta|^2)$ as $|\theta| \to 0$,

(ii) Δ is a vector of $\mathcal{L}^2(P_0)$ functions for which

$$\sqrt{p_\theta(x)} = 1 + \tfrac{1}{2}\theta'\Delta(x) + r_\theta(x) \qquad \text{with } P_0\left(r_\theta^2\right) = o(|\theta|^2) \text{ near } 0.$$

Remember also that $\mathbb{I}_0 = P_0(\Delta\Delta')$ is the Fisher information matrix at $\theta = 0$.

almost.LAN<14>

Theorem. *Suppose $\theta \mapsto P_\theta$ is DQM at 0, with score function Δ. Let $\mathbb{P}_n := P_0^n$ and $\mathbb{Q}_n := P_{\theta_n}^n$, with $\theta_n := t_n/\sqrt{n}$ for a bounded sequence $\{t_n\}$. Then, under $\{\mathbb{P}_n\}$,*

$$\frac{d\mathbb{Q}_n}{d\mathbb{P}_n} = (1 + o_p(1; \mathbb{P}_n)) \exp\left(\delta_n' Z_n - \tfrac{1}{t}'_n \mathbb{I}_0 t_n\right),$$

where

$$Z_n := n^{-1/2} \textstyle\sum_{i \le n} \Delta(x_i) \rightsquigarrow N(0, \mathbb{I}_0).$$

In particular, the reparametrized family $\mathbb{Q}_{t,n} = P^n_{t/\sqrt{n}}$ is LAN with asymptotic variance matrix $\mathbb{I}_0$.

PROOF

The asserted quadratic approximation follows. □

Problems

[1] Give hints for proof of LAN implies DQM, as on Le Cam (1986, page 584).

[2] Suppose $\{g_n\}$ is a sequence of vector-valued, measurable functions for which $g_n \to g_0$ a.e. $[\mu]$ and $\mu|g_n|^2 \to \mu|g_0|^2 < \infty$, for some measure μ.

(i) Use Fatou's lemma to show that

$$\liminf \mu\left(2|g_n|^2 + 2|g_0|^2 - |g_n - g_0|^2\right) \ge 4\mu|g_0|^2$$

(ii) Deduce that $\mu|G_n - g_0|^2 \to 0$.

[3] Let $W_1, W_2, \ldots$ be a sequence of independent, identically distributed random variables with $\mathbb{P}|W_i|^r < \infty$ for a constant $r \geq 1$. Prove that $\max_{i \leq n} |W_i| = o_p(n^{1/r})$. Hint: Show that

$$\mathbb{P}\{\max_{i \leq n} |W_i| > \epsilon n^{1/r}\} \leq \epsilon^{-r} \mathbb{P}|W_1|^r \{|Z_1| > \epsilon n^{1/r}\},$$

then invoke Dominated Convergence.

[4] Show that the affinity between two finite Borel measures λ and μ on a metric space $\mathcal{X}$ equals the infimum of $\lambda g + \mu \bar{g}$ taken over all continuous functions g for which $0 \leq g \leq 1$. Hint: Use the fact that the bounded continuous functions are dense in $\mathcal{L}^1(\lambda + \mu)$. Also, if $0 \leq f \leq 1$ show that $|f - g_0| \leq |f - g|$ where $g_0 = 1 \wedge g^+$.

[5] Let $\mathbb{P}_n$ and $\mathbb{Q}_n$ be as in Lemma <5>. Suppose $\{Y_n\}$ is sequence of random vectors for which $Y_n \rightsquigarrow \lambda$ under $\mathbb{P}_n$ and $Y_n \rightsquigarrow \mu$ under $\mathbb{Q}_n$, where λ and μ are probability measures on $\mathbb{R}^k$. For each positive constant C, show that $\|\lambda \wedge (C\mu)\|_1 \geq \|P \wedge (CQ)\|_1$. Deduce that $\|\lambda - \mu\|_1 \leq \|P - Q\|_1$. Hint: Invoke the Lemma with $\psi_n := g(Y_n)$, with g continuous and $0 \leq g \leq 1$, then appeal to Problem [4].

[6] Show that $\|N(t_1, \sigma^2) - N(t_2, \sigma^2)\|_1 = 2\mathbb{P}\{\, |N(0,1)| \leq |t_1 - t_2|/\sigma\}$.

8.7 Notes

LAN::Notes

These notes refer to material now in other chapters. They need to be updated.

The argument in Section 3 is an extension of the method of Bahadur (1964). He noted that there is an easy generalization to the case where the parameter is vector valued. Bahadur imposed classical regularity conditions to produce the required approximation for the likelihood ratio.

I borrowed the exposition for the last three Sections from Pollard (1997). The essential argument is fairly standard, but the interpretation of some of the details is novel. Compare with the treatments of Le Cam (1970, and 1986 Section 17.3), Ibragimov and Has'minskii (1981, page 114), Millar (1983, page 105), Le Cam and Yang (1990, page 101), or Strasser (1985, Chapter 12).

???

Hájek (1962) used Hellinger differentiability to establish limit behaviour of rank tests for shift families of densities. Most of results in Section **??**

are adapted from the Appendix to Hájek (1972), which in turn drew on Hájek and Šidák (1967, page 211) and earlier work of Hájek. For a proof of the multivariate version of Theorem <??> see Bickel, Klaassen, Ritov, and Wellner (1993, page 13). A reader who is puzzled about all the fuss over negligible sets, and behaviour at points where the densities vanish, might consult Le Cam (1986, pages 585–590) for a deeper discussion of the subtleties.

References

Bahadur, R. R. (1964). On Fisher's bound for asymptotic variances. *Annals of Mathematical Statistics 35*, 1545–1552.

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.

Hájek, J. (1962). Asymptotically most powerful rank-order tests. *Annals of Mathematical Statistics 33*, 1124–1147.

Hájek, J. (1972). Local asymptotic minimax and admissibility in estimation. In L. Le Cam, J. Neyman, and E. L. Scott (Eds.), *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Volume I, Berkeley, pp. 175–194. University of California Press.

Hájek, J. and Z. Šidák (1967). *Theory of Rank Tests*. Academic Press. Also published by Academia, the Publishing House of the Czechoslavak Academy of Sciences, Prague.

Ibragimov, I. A. and R. Z. Has'minskii (1981). *Statistical Estimation: Asymptotic Theory*. New York: Springer.

Le Cam, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimators. *Annals of Mathematical Statistics 41*, 802–828.

Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. New York: Springer-Verlag.

Le Cam, L. and G. L. Yang (1990). *Asymptotics in Statistics: Some Basic Concepts*. Springer-Verlag.

Millar, P. W. (1983). The minimax principle in asymptotic statistical theory. *Springer Lecture Notes in Mathematics 976*, 75–265.

Pollard, D. (1997). Another look at differentiability in quadratic mean. In D. Pollard, E. Torgersen, and G. L. Yang (Eds.), *A Festschrift for Lucien Le Cam*, pp. 305–314. New York: Springer-Verlag.

Strasser, H. (1985). *Mathematical Theory of Statistics: Statistical Experiments and Asymptotic Decision Theory*. Berlin: De Gruyter.