CHAPTER 3.

LOWER BOUNDS FOR ESTIMATION

1. **Introduction and Examples**
2. **Cramér - Rao lower bounds for parametric models**
3. **Regular Estimates; Superefficiency; LAN and Le Cam's three lemmas**
4. **Hajek's convolution theorem and Local Asymptotic Minimax (LAM) theorem**

1. Introduction and Examples

One of the goals of statistical theory is to describe how well we can estimate parameters of interest in principle for any given model. Since we cannot estimate parameters perfectly, what is the best we *can* do?

A model $\mathbf{P}$ is simply a collection of probability distributions for the data we observe. Consider a parameter of interest $\nu = \nu(P)$ we want to estimate. Here are some frequent goals or questions:

Q1. Given a model $\mathbf{P}$ and a parameter ν of interest, how well can we estimate ν? [What is our "gold standard?"]

Q2. Can we compare absolute "in principle" standards for estimation of ν in a model $\mathbf{P}$ with estimation of ν in a submodel $\mathbf{P}_0 \subset \mathbf{P}$. [What is the effect of not knowing η on estimation of ν when $\theta = (\nu, \eta)$?]

Q3. For a fixed model $\mathbf{P}$ compare one or more estimators of ν to each other and to the best "in principle" bound.

The bounds we will discuss in this chapter provide some partial answers to these questions.

To indicate the scope of the questions we want to address, we begin with some examples of the models we would like to be able to handle. In all of the following examples we will suppose that we observe $X_1, \cdots, X_n$ iid as $X \sim P \in \mathbf{P}$ where $\mathbf{P}$ is the given model. Within each example the models increase in complexity from parametric to semiparametric to nonparametric. For further examples, see Bickel, Klaassen, Ritov, and Wellner (1993).

Example 1.1. Suppose that X is a rv on R^+; think of X as a survival time.

Example 1.1.A. Suppose that $X \sim \exp(\theta)$; i.e. $p_\theta(x) = \theta \exp(-\theta x) 1_{[0,\infty)}(x)$, $\theta > 0$. This is a simple parametric model $\mathbf{P} = \{P_\theta : \theta \in \Theta = R^+\}$.

Example 1.1.B. Suppose that $X \sim P$ on R^+ with density $p_G(x) = \int_0^\infty \lambda \exp(-\lambda x)\, dG(\lambda)$, the G mixture of exponentials. This might be viewed as a semiparametric model, $\mathbf{P} = \{P_G : G \in \mathbf{G}\}$ where $\mathbf{G} \equiv \{\text{all df's on } [0, \infty)\}$.

Example 1.1.C. Suppose that $X \sim P$ on R^+ with density $p = dP/d\lambda$ assumed to be nonincreasing; this is a family defined only by a shape restriction on the density; this is a nonparametric model.

Example 1.1.D. Suppose that $X \sim P$ on R^+ with P completely arbitrary; thus $\mathbf{P} = \mathbf{M}$, where $\mathbf{M}$ is the collection of all distributions on R^+. This is clearly a nonparametric model.

Example 1.2. Suppose that X is a rv on R; think of X as a measurement.

Example 1.2.A. Suppose that $X \sim N(\mu, \sigma)$, $\theta = (\mu, \sigma^2) \in \Theta \equiv R \times R^+$. Thus $\mathbf{P} = \{p_\theta = \phi((\cdot - \mu)/\sigma)/\sigma : \theta = (\mu, \sigma) \in \Theta\}$. This is perhaps *the* classic parametric model, the normal location - scale model. If we replace the standard normal density by some other density g_0 which is fixed and known (e.g. logistic, or Cauchy, or double exponential, or ...), the resulting model $\mathbf{P} = \{p_\theta = g_0((\cdot - \mu)/\sigma)/\sigma : \theta = (\mu, \sigma) \in \Theta\}$ is the g_0 – location - scale family.

Example 1.2.B. Suppose that $X \sim P_{\theta,G}$ on R with density $p_{\theta,G}(x) = g(x - \theta)$, with g symmetric about 0 (and, say, absolutely continuous). Thus

$$\mathbf{P} = \{P_{\theta,G} : \theta \in R, G \text{ with symmetric density } g\}.$$

This is a semiparametric model; the "one - sample symmetry model."

Example 1.2.C. Suppose that $X \sim P_{\theta,G}$ on R with density $p_{\theta,G}(x) = g(x - \theta)$, with g absolutely continuous but otherwise arbitrary. This is a nonparametric model with only a smoothness restriction g has a derivative g'.

Example 1.3. Suppose that $X = (Y, Z)$ is a random vector on $R^+ \times R^d$; think of Y as a survival time and Z as a vector of covariates.

Example 1.3.A. Suppose that $X = (Y, Z) \sim P_\theta$, with $(Y \mid Z = z) \sim \text{exponential}(e^{\theta^T z})$; i.e $\lambda(y|Z = z) = e^{\theta^T z}$ for $y \ge 0$. This is a parametric model with parameter space $\Theta = R^d$.

Example 1.3.B. Suppose that $X = (Y, Z) \sim P_\theta$, with $\lambda(y|Z = z) = e^{\theta^T z}\,\lambda(y)$ for $y \ge 0$ where $\theta \in R^d$ and λ is an arbitrary hazard function on R^+. This is a semiparametric model, the "Cox model" for survival analysis.

Example 1.3.C. Suppose that $X = (Y, Z) \sim P_\theta$, $\lambda(y|Z = z) = r_0(\theta^T z)\,\lambda(y)$ for $y \ge 0$. where $\theta \in R^d$ and $\lambda(y)$ is an arbitrary hazard function on R^+, but r_0 is some fixed and known nonnegative function. This is a variant of the "Cox model."

Example 1.3.D. Suppose that $X = (Y, Z) \sim P_\theta$, $\lambda(y|Z = z) = r(\theta^T z)\,\lambda(y)$ for $y \ge 0$. where $\theta \in R^d$, $\lambda(y)$ is an arbitrary hazard function on R^+, and r is an unknown nonnegative function.

This is a more involved variant of the "Cox model" with two infinite - dimensional "nuisance" parameters, λ and r.

Example 1.3.E. Suppose that $X = (Y, Z) \sim P$ on $R^+ \times R^d$ where P is completely arbitrary. This is a nonparametric model. (How do we define "effects of the covariates Z on Y" here?)

Example 1.4. Suppose that $X = (Y, Z)$ is a random vector on $R \times R^d$; think of Y as a measurement or response, and Z as a vector of covariates.

Example 1.4.A. Suppose that $X = (Y, Z) \sim P_{\theta,\sigma}$, $Y = \theta^T Z + \sigma\epsilon$ where $\theta \in R^d$, $\sigma > 0$, and $\epsilon \sim g_0$ is independent of Z. Here g_0 is a known density (such as the standard normal density ϕ), and $Z \sim H$ (supposed known for simplicity). This is a parametric model, the classical linear regression model (with $g_0 -$ errors).

Example 1.4.B. Suppose that $X = (Y, Z) \sim P_{\theta,G}$,

$$Y = \theta^T Z + \epsilon$$

where $\theta \in R^d$, and $\epsilon \sim g$ is independent of Z, but now g is an unknown density. This is a semiparametric model, the linear regression model with "arbitrary" error distribution.

Example 1.4.C. Suppose that $X = (Y, Z) \sim P_{\theta,\sigma,r}$,

$$Y = r(\theta^T Z) + \sigma\epsilon$$

where $\theta \in R^d$, $\sigma > 0$, $\epsilon \sim g_0$ is independent of Z, and r is an unknown function from R to R. This is again a semiparametric model, a model for "projection pursuit" regression with $g_0 -$ errors.

Example 1.4.D. Suppose that $X = (Y, Z) \sim P_{\sigma,r}$,

$$Y = r_1(Z_1) + \cdots + r_d(Z_d) + \sigma\epsilon$$

where $\sigma > 0$, $\epsilon \sim g_0$ is independent of Z, and $r = (r_1, \cdots, r_d)$ are unknown functions from R to R. This is again a semiparametric model, a model for "additive" regression with $g_0 -$ errors. Of course many other variants on models 4.A - 4.D are possible.

Example 1.4.E. Suppose that $X = (Y, Z) \sim P_{\sigma,r}$,

$$Y = r(Z) + \sigma\epsilon$$

where $\sigma > 0$, $\epsilon \sim g_0$ is independent of Z, and r is an unknown function from R^d to R. This is still a semiparametric model. However, because of dimensionality problems, this model becomes problematic for even moderately large d.

Example 1.4.F. Suppose that $X = (Y, Z) \sim P$, where P is an arbitrary probability distribution on $R \times R^d = R^{d+1}$.

Of course, not all problems are necessarily i.i.d. Here is one simple model which involves "pooling information" from three independent samples. There are many other related models -- as well as models in which the independence assumption is relaxed.

Example 1.5. Suppose that we observe data X as follows:

A. The first sample is a sample of iid pairs of size n_1 from a distribution P with cumulative distribution $H(x, y) \equiv P(X \le x, Y \le y)$ on R^2.

B. The second sample of size n_2 is a sample of iid X's from the marginal distribution P_X of P (with distribution function $F(x) = P(X \le x)$).

C. The third sample of size n_3 is a sample of iid Y's from the marginal distribution P_Y of P (with distribution function $G(y) = P(Y \le y)$).

Question: How well can we estimate P (e.g. $\nu(P) \equiv P(X \le x_0, Y \le y_0)$) based on all the data?

Example 1.5.A. Suppose that P is bivariate normal with mean vector μ and covariance matrix Σ.

Example 1.5.B. Suppose that $P = P_{\theta,G,H}$ where $P_{\theta,G,H}$ has distribution function given by $F_{\theta,G,H}(x, y) = C_\theta(G(x), H(y))$ for some parametric family of distribution functions C_θ on the unit square $[0,1]^2$ with uniform marginals (such as the Morgenstern family $C_\theta(u, v) = uv(1 + \theta(1-u)(1-v))$ on the first day quiz). This is a semi-parametric model.

Example 1.5.C. Suppose that $P \in \mathbf{M}$, the collection of all distributions on R^2; this is the nonparametric version of the problem.

3.2

2. Cramér - Rao bounds for parametric models

We first discuss the elementary Cramér - Rao lower bound in the case of a one - dimensional parametric model: $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$ with $\Theta \subset R$; the reader may also wish to consult Lehmann, TPE, pages 115 ff.

Here are the assumptions we will need:

Assumptions:

A. $X \sim P_\theta$ on $(\mathbf{X}, \mathbf{A})$ with $\theta \in \Theta \subset R$.

B. $p_\theta \equiv dP_\theta / d\mu$ exists where μ is $\sigma -$ finite.

C. $T(X) \equiv T$ estimates $q(\theta)$ has $E_\theta |T(X)| < \infty$; set $b(\theta) \equiv E_\theta T - q(\theta) \equiv$ bias of T.

D. $q'(\theta) \equiv \dot{q}(\theta)$ exists.

Theorem 2.1. (Cramér - Rao inequality). Suppose that:

(1) Θ is an open subset of the real line.

(2) A. There exists a set B with $\mu(B) = 0$ such that: for $x \in B^c$

$$\frac{\partial}{\partial \theta} p_\theta(x) \qquad \text{exists for all } \theta .$$

B. $A \equiv \{x : p_\theta(x) = 0\}$ does not depend on θ.

(3) $I(\theta) \equiv E_\theta[\dot{\mathbf{l}}_\theta(X)]^2 > 0$ where $\dot{\mathbf{l}}_\theta(x) \equiv \frac{\partial}{\partial \theta} \log p_\theta(x)$; here $I(\theta)$ is called the *Fisher information for* θ and $\dot{\mathbf{l}}_\theta$ is called the *score function* for θ.

(4) $\int p_\theta(x)\, d\mu(x)$ and $\int T(x) p_\theta(x)\, d\mu(x)$ can both be differentiated with respect to θ under the integral sign.

(5) $\int p_\theta(x)\, d\mu(x)$ can be differentiated twice under the integral sign.

If (1) - (4) hold, then

$$Var_\theta[T(X)] \geq \frac{[\dot{q}(\theta) + \dot{b}(\theta)]^2}{I(\theta)} \qquad \text{for all} \quad \theta \in \Theta$$

$$= \frac{[\dot{q}(\theta)]^2}{I(\theta)} \qquad \text{if } T \text{ is unbiased} .$$

Equality holds for all θ if and only if for some function $A(\theta)$ we have

$$\dot{\mathbf{l}}_\theta(x) \equiv A(\theta)\{T(x) - E_\theta(T(X))\} \qquad \text{a.e. } \mu .$$

If, in addition, (5) holds, then

$$I(\theta) \;=\; -\,E_\theta[\frac{\partial^2}{\partial\theta^2}\log p_\theta(X)] \;=\; -\,E_\theta[\ddot{\mathbf{l}}_\theta(X)]\;.$$

Proof. Now

$$q(\theta) \;+\; b(\theta) \;=\; \int_{\mathbf{X}} T(x)\,p_\theta(x)\,d\mu(x) \;=\; \int_{\mathbf{X}\,\cap\,A^c\cap B^c} T(x)\,p_\theta(x)\,d\mu(x)\;;$$

hence

$$\begin{aligned}\dot q(\theta) \;+\; \dot b(\theta) \;&=\; \int_{\mathbf{X}\,\cap\,A^c\cap B^c} T(x)\,\frac{\partial}{\partial\theta}\,p_\theta(x)\,d\mu(x)\;;\\ &=\; \int_{\mathbf{X}\,\cap\,A^c\cap B^c} T(x)\,\dot{\mathbf{l}}_\theta(x)\,p_\theta(x)\,d\mu(x)\;;\\ &=\; E_\theta\{T(X)\dot{\mathbf{l}}_\theta(X)\}\\ &=\; Cov_\theta[T(X)\,,\dot{\mathbf{l}}_\theta(X)]\end{aligned}$$

since $\int p_\theta(x)\,d\mu(x) \;=\; 1$ implies, by arguing as above, that

$$0 \;=\; \int \frac{\partial}{\partial\theta}\,p_\theta\,d\mu \;=\; E_\theta[\dot{\mathbf{l}}(X)]\;.$$

Thus, by the Cauchy - Schwarz inequality,

$$[\dot q(\theta) \;+\; \dot b(\theta)]^2 \;=\; |Cov_\theta[T(X)\,,\dot{\mathbf{l}}_\theta(X)]|^2 \;\le\; Var_\theta[T(X)]\,I(\theta)\;.$$

The only inequality was Cauchy - Schwarz, and it is an equality for a fixed θ if and only if

$$\dot{\mathbf{l}}_\theta(x) \;=\; A(\theta)\{T(x) \;-\; q(\theta) \;-\; b(\theta)\} \qquad \text{a.s. } P_\theta$$

for some constant $A(\theta)$. By (2.B) this implies that this holds a.e. μ. Under further regularity conditions this holds if and only if P_θ is an exponential family; see e.g. Lehmann, page 123.

Finally if (5) holds, since

$$0 \;=\; \int \dot{\mathbf{l}}_\theta(x)\,p_\theta(x)\,d\mu(x)$$

differentiation once more (which is possible by (5)) yields

$$\begin{aligned}0 \;&=\; \int \ddot{\mathbf{l}}_\theta(x)\,p_\theta(x)\,d\mu(x) \;+\; \int \dot{\mathbf{l}}_\theta^2(x)\,p_\theta(x)\,d\mu(x)\\ &=\; \int \ddot{\mathbf{l}}_\theta\,p_\theta\,d\mu \;+\; I(\theta)\;. \qquad \square\end{aligned}$$

Example 2.2. (Poisson(θ); an exponential family) Suppose that $X_1\,,\cdots,X_n$ are iid $Poisson(\theta)$ with $\Theta \;=\; (0\,,\infty)$; i.e. $p_\theta(x) \;=\; e^{-n\theta}\theta^{\sum_i x_i}/\prod_{i=1}^n x_i!$ with respect to counting measure μ . Then

$$\log p_\theta(x) \; = \; -\, n\,\theta \; + \; (\sum_{i=1}^{n} X_i)\log\theta \; - \; \sum_{i=1}^{n} log(x_i!)$$

and

$$\dot{\mathbf{l}}_\theta(x) \; = \; -\, n \; + \; (\sum_{i=1}^{n} x_i)\frac{1}{\theta} = \frac{n}{\theta}(\overline{x}_n \; - \; \theta) \, .$$

Note that (1), (2), and (3) are trivial; and (4), (4′) hold if $E_\theta|T(X)| \; < \; \infty$ for all θ, since an absolutely convergent power series can be differentiated term by term. Thus the Cramér - Rao inequality holds for all T having $E_\theta|T(X)| \; < \; \infty$ for all θ. However only $q(\theta) \; = \; \theta$ (or a linear function of this) has a Minimum Variance Bound Unbiased (MVBU) estimator, and the MVBU estimator of θ is $\overline{X}$ which has variance $|\dot{q}(\theta)/A(\theta)| \; = \; \theta/n$. Thus $I_n(\theta) \; = \; n/\theta$. The bound for estimating $q(\theta) \; = \; \theta^2$ is $[\dot{q}(\theta)]^2/I(\theta) \; = \; (2\theta)^2/(n/\theta) \; = \; 4\theta^3/n$; but this bound cannot be achieved (for $n < \infty$). In fact we know that $\sum_{i=1}^{n} X_i$ is a complete sufficient statistic. It is easy to check that $T^* \; = \; \overline{X}^2 \; - \; n^{-1}\overline{X} \; = \; (n^{-1}\sum_{i=1}^{n} X_i)^2 \; - \; n^{-2}\sum_{i=1}^{n} X_i$ is unbiased; hence it is a UMVUE of θ^2. Also, its variance is $(4\theta^3)/n \; + \; (2\theta^2)/n^2 \; > \;$ the C-R bound.

Example 2.3. (Location with known "shape" g) Suppose that $X_1, \cdots, X_n$ are iid with density

$$p_\theta(x) \; = \; g(x \, - \, \theta)$$

where g is a known density; e.g. $N(0,1)$, or Cauchy, or logistic, or double exponential, or extreme value, or ... . Then, assuming g' exists a.e. (Lebesgue) and the other regularity conditions hold,

$$\dot{\mathbf{l}}_\theta(x) \; = \; \frac{\partial}{\partial\theta}\log g(x \, - \, \theta) \; = \; - \, \frac{g'(x \, - \, \theta)}{g(x \, - \, \theta)} \; \equiv \; - \, \frac{g'}{g}(x \, - \, \theta)$$

so that

$$I(\theta) \; = \; E_\theta(\dot{\mathbf{l}}_\theta^2(X)) \; = \; \int [\frac{g'}{g}(x \, - \, \theta)]^2 \, g(x \, - \, \theta)\, dx \; = \; \int [\frac{g'}{g}(y)]^2 \, g(y)\, dy$$

$$= \; \int \frac{[g'(y)]^2}{g(y)} \, dy \; \equiv \; I_g$$

and

$$I_n(\theta) \; = \; nI(\theta) \; = \; n\, I_g \, .$$

Thus for any unbiased estimator $\hat{\theta}_n$ of θ we have

$$Var_\theta(\hat{\theta}_n) \; \geq \; \frac{1}{n\, I_g} \, , \quad \text{or} \quad Var_\theta(\sqrt{n}(\hat{\theta}_n \; - \; \theta)) \; \geq \; \frac{1}{I_g} \, .$$

Example 2.4. (Scale with known shape) Suppose that $X_1, \cdots, X_n$ are iid with density

$$p_\theta(x) = \frac{1}{\theta} g(\frac{x}{\theta})$$

where g is known (e.g. exponential(1), Cauchy, gamma(5,1), ...). Then, assuming g' exists a.e. (Lebesgue) and the other regularity conditions hold,

$$\begin{aligned}\dot{\mathbf{l}}_\theta(x) &= \frac{\partial}{\partial\theta}\log\{\frac{1}{\theta}g(\frac{x}{\theta})\} = -\frac{1}{\theta} + \frac{g'(x/\theta)}{g(x/\theta)}(\frac{-x}{\theta^2}) \\ &= \frac{1}{\theta}\{-1 - \frac{x}{\theta}\frac{g'}{g}(\frac{x}{\theta})\}\end{aligned}$$

so that

$$\begin{aligned}I(\theta) &= E_\theta(\dot{\mathbf{l}}_\theta^2(X)) = \frac{1}{\theta^2}\int[-1 - \frac{x}{\theta}\frac{g'}{g}(\frac{x}{\theta})]^2\frac{1}{\theta}g(\frac{x}{\theta})\,dx \\ &= \frac{1}{\theta^2}\int[-1 - y\frac{g'}{g}(y)]^2\,g(y)\,dy \equiv \frac{1}{\theta^2}I_g(scale)\,.\end{aligned}$$

and

$$I_n(\theta) = nI(\theta) = \frac{n}{\theta^2}I_g(scale)\,.$$

Thus for any unbiased estimator $\hat{\theta}_n$ of θ we have

$$Var_\theta(\hat{\theta}_n) \ge \frac{\theta^2}{n\,I_g(scale)}, \qquad \text{or} \qquad Var_\theta(\sqrt{n}(\hat{\theta}_n - \theta)) \ge \frac{\theta^2}{I_g(scale)}\,.$$

Example 2.5. (Elementary mixture model) Suppose that f_0, f_1 are two known $\mu-$ densities, and that

$$\begin{aligned}p_\theta(x) &= \theta f_0(x) + (1-\theta)f_1(x)\,, \qquad \theta \in (0,1) \equiv \Theta \\ &= \theta f_0(x) + \bar{\theta}f_1(x)\,.\end{aligned}$$

Then $\log p_\theta(x) = \log[\theta f_0(x) + (1-\theta)f_1(x)]$ and hence

$$\dot{\mathbf{l}}_\theta(x) = \frac{f_0(x) - f_1(x)}{\theta f_0(x) + \bar{\theta}f_1(x)}$$

so that

$$I(\theta) = \int_{-\infty}^{\infty}\frac{(f_0(x) - f_1(x))^2}{\theta f_0(x) + \bar{\theta}f_1(x)}\,d\mu(x)\,.$$

Note that if $f_0 \ne f_1$ (on a set of positive μ measure), then the information $I(\theta)$ is finite and positive for all $\theta \in (0,1)$, while $I(\theta)$ converges to

$$\int_{-\infty}^{\infty}\frac{(f_0(x) - f_1(x))^2}{f_1(x)}\,d\mu(x) \qquad \text{as} \quad \theta \to 0\,;$$

similarly, $I(\theta)$ converges to

$$\int_{-\infty}^{\infty} \frac{(f_0(x) - f_1(x))^2}{f_0(x)}\, d\mu(x) \qquad \text{as} \quad \theta \to 1 \, .$$

These limiting values may be infinite.

The multiparameter Cramér - Rao inequality

Now we extend theorem 2.1 to the case in which the model is a $k-$dimensional parametric family: $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$ with $\Theta \subset R^k$.

Assumptions:

A. $X \sim P_\theta$ on $(\mathbf{X}, \mathbf{A})$ with $\theta \in \Theta \subset R^k$.

B. $p_\theta \equiv dP_\theta/d\mu$ exists where μ is $\sigma-$finite.

C. $T(X) \equiv T$ estimates $q(\theta)$ where $q : \Theta \to R$, and $E_\theta|T(X)| < \infty$; set $b(\theta) \equiv E_\theta T - q(\theta) \equiv$ bias of T .

D. $\dot{q}(\theta) = \nabla q(\theta)$ ($1 \times k$) exists.

Theorem 2.6. (Multiparameter Cramér - Rao inequality). Suppose that:

(1) Θ is an open subset of R^k .

(2) A. There exists a set $B \subset \mathbf{X}$ with $\mu(B) = 0$ such that if $x \notin B$, then $\frac{\partial}{\partial\theta_i} p_\theta(x)$ exists for $i = 1, \cdots, k$.

B. $A = \{x : p_\theta(x) = 0\}$ does not depend on θ .

(3) The $k \times k$ matrix $I(\theta) = (I_{ij}(\theta)) = (E_\theta[\dot{\mathbf{l}}_{\theta_i}(X)\dot{\mathbf{l}}_{\theta_j}(X)])$ is positive definite where $\dot{\mathbf{l}}_{\theta_i}(x) \equiv \frac{\partial}{\partial\theta_i} \log p_\theta(x)$.

(4) $\int p_\theta\, d\mu$ and $\int T\, p_\theta\, d\mu$ can be differentiated under the integral sign.

(5) $\int p_\theta\, d\mu$ can be differentiated twice under the integral sign.

If (1) - (4) hold, then

$$\begin{aligned} Var_\theta[T(X)] &\geq \alpha\, I^{-1}(\theta)\alpha^T \\ &= \dot{q}(\theta) I^{-1}(\theta) \dot{q}^T(\theta) \qquad \text{if } T \text{ is unbiased} \end{aligned}$$

where

$$\alpha \equiv (\alpha_1, \cdots, \alpha_k) \equiv \nabla(q(\theta) + b(\theta)) = \nabla E_\theta(T(X)) \, .$$

If, in addition, (5) holds, then

$$I(\theta) = -E_\theta \ddot{\mathbf{l}}_{\theta\theta}(X) = -\left(E_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j} \mathbf{l}(X)\right]\right) .$$

Proof. The following proof is for the case of an unbiased estimator. Since

$$q(\theta) \;=\; E_\theta T(X) \;=\; \int T(x) p_\theta(x) d\mu(x) \;,$$

differentiating w.r.t. each θ_j gives:

$$\begin{aligned}
\dot q(\theta) \;&=\; \Big(\int_{\mathbf{X} \cap B^c \cap A^c} T(x) \, \frac{\frac{\partial}{\partial \theta_j} \, p_\theta(x)}{p_\theta(x)} \, p_\theta(x) \, d\mu(x) \Big) \\
&=\; E_\theta T(X) \underline{\dot{\mathbf{l}}}_\theta(X) \\
&=\; E_\theta \{ (T(X) - E_\theta T(X)) \underline{\dot{\mathbf{l}}}_\theta(X) \} \;=\; Cov_\theta(T(X) \, , \dot{\mathbf{l}}_\theta(X)) \; .
\end{aligned}$$

By taking the transpose of this, and multiplying by $I^{-1}(\theta)\dot q(\theta)$ we find that

$$\dot q^T(\theta) I^{-1}(\theta) \dot q(\theta) \;=\; Cov_\theta(T(X) \, , \underline{\dot{\mathbf{l}}}^T(X) I^{-1}(\theta) \dot q(\theta)) \; .$$

Hence by the Cauchy-Schwarz inequality

$$\begin{aligned}
|\dot q^T(\theta) I^{-1}(\theta) \dot q(\theta)| \;&=\; |Cov_\theta(T(X) \, , \underline{\dot{\mathbf{l}}}^T(X) I^{-1}(\theta) \dot q(\theta))| \\
&\le\; \{ Var_\theta(T(X)) \, \dot q^T(\theta) I^{-1}(\theta) I(\theta) I^{-1}(\theta) \dot q(\theta) \}^{1/2} \; ,
\end{aligned}$$

and hence

$$Var_\theta(T(X)) \;\ge\; \dot q^T(\theta) I^{-1}(\theta) \dot q(\theta) \; .$$

Note that the function (for sample size $n = 1$) involved here, namely

$$\tilde{\mathbf{l}}_\theta(X_1) \;=\; \dot q^T(\theta) I^{-1}(\theta) \underline{\dot{\mathbf{l}}}_\theta(X_1) \;\equiv\; \text{the efficient influence function} \; ;$$

that is, if T_n is an asymptotically efficient estimator of $q(\theta) \;=\; \nu(P_\theta)$, then T_n is asymptotically linear with influence function exactly $\tilde{\mathbf{l}}_\nu$:

$$\begin{aligned}
\sqrt{n}(T_n \;-\; q(\theta)) \;&=\; \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\mathbf{l}}_\nu(X_i) \;+\; o_p(1) \\
&\to_d \; N(0 \, , \dot q^T(\theta) I^{-1}(\theta) \dot q(\theta)) \; .
\end{aligned}$$

This proof should be compared with the proof given in Lehmann TPE, pages 123 - 128. □

Rather than dwell further on the proof of theorem 2.6, we will concentrate on its interpretation in terms of elementary geometric ideas.

First, an easy example.

Example 2.7. (Weibull). If $(\mathbf{X} \, , \mathbf{A}) \;=\; (R^+ \, , \mathbf{B}^+)$, the nonnegative real numbers with its usual Borel $\sigma -$ field, then the *Weibull* family $\mathbf{P}$ is the parametric model with densities

$$p_\theta(x) \;=\; \frac{\beta}{\alpha} \, (\frac{x}{\alpha})^{\beta - 1} \, exp(\, - \, (\frac{x}{\alpha})^\beta) \, 1_{[0, \infty)}(x)$$

with respect to Lebesgue measure where $\theta \equiv (\alpha, \beta) \in (0, \infty) \times (0, \infty) \subset R^2$. For the Weibull family $\mathbf{P}$, $\log p_\theta(x)$ is differentiable at every $\theta \in \Theta$ and the scores are:

$$\dot{\mathbf{l}}_\alpha(x) = \frac{\beta}{\alpha}\{(\frac{x}{\alpha})^\beta - 1\}$$

$$\dot{\mathbf{l}}_\beta(x) = \frac{1}{\beta} - \frac{1}{\beta}\log\{(\frac{x}{\alpha})^\beta\}\{(\frac{x}{\alpha})^\beta - 1\}.$$

Thus $\dot{\mathbf{P}} \equiv [\dot{\mathbf{l}}_\theta]$ is the two-dimensional subspace of $L_2(P_\theta)$ spanned by $\dot{\mathbf{l}}_\alpha$ and $\dot{\mathbf{l}}_\beta$, and the Fisher information matrix is:

$$I(\theta) = E(\dot{\mathbf{l}}_\theta^T(X)\dot{\mathbf{l}}_\theta(X)) = \begin{pmatrix} \frac{\beta^2}{\alpha^2} & \frac{a}{\alpha} \\ \frac{a}{\alpha} & \frac{b^2}{\beta^2} \end{pmatrix}$$

where, with $Y \sim$ exponential(1), and $\gamma \equiv .577\cdots$ Euler's constant,

$$a = -E\{(Y-1)^2\log(Y)\} = -(1-\gamma),$$

and

$$b^2 = E\{(Y-1)\log(Y) - 1\}^2 = \frac{\pi^2}{6} + (1-\gamma)^2.$$

Note that $\det(I(\theta)) = (b^2 - a^2)\alpha^{-2} = (\pi^2/6)\alpha^{-2} > 0$, so $I(\theta)$ is nonsingular. Thus by theorem 2.6, if $q(\theta)$ is a real - valued function of θ (e.g. $q(\theta) \equiv E_\theta X = \alpha\Gamma(1 + 1/\beta)$), and T is any estimator of $q(\theta)$ based on $\underline{X} = (X_1, \cdots, X_n)$ (with X_i's iid P_θ) satisfying Assumption C and hypothesis (4), then

$$Var_\theta[T(\underline{X})] \ge \frac{(\alpha^T I^{-1}(\theta)\alpha)}{n}$$

$$= \frac{\dot{q}(\theta)I^{-1}(\theta)\dot{q}^T(\theta)}{n} \quad \text{if } T \text{ is unbiased.}$$

Some Geometry

The bounds given in theorems 1 and 2 lead us to the following definitions:

Suppose ν is a Euclidean parameter defined on a regular parametric model $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$. We can identify ν with the parametric function $q: \Theta \to R^m$ defined by

$$q(\theta) = \nu(P_\theta).$$

Fix $P = P_\theta$ and suppose q has a total differential matrix $\dot{q}_{m\times k}$ at θ. Define

(1) $\quad I^{-1}(P \mid \nu, \mathbf{P}) = \dot{q}(\theta)I^{-1}(\theta)\dot{q}^T(\theta)$, the *information bound* for ν and

(2) $\tilde{\mathbf{l}}(\,\bullet\,, P \mid \nu, \mathbf{P}) \;=\; \dot{q}(\theta) I^{-1}(\theta) \dot{\mathbf{l}}(\theta)$, the *efficient influence function* for ν .

As defined in (1) and (2), the information bound and influence function appear to depend on the parametrization $\theta \to P_\theta$ of $\mathbf{P}$. However, as our notation indicates:

Proposition 2.8. $I^{-1}(P \mid \nu, \mathbf{P})$ and $\tilde{\mathbf{l}}(\,\bullet\,, P \mid \nu, \mathbf{P})$ are invariant under smooth changes of parametrization.

Here is a formal calculation.

Suppose $\gamma \to \theta(\gamma)$ is a one-to-one continuously differentiable mapping of an open subset Γ of R^k onto Θ with nonsingular differential $\dot{\theta}$. We represent $\mathbf{P} \;=\; \{Q_\gamma : \gamma \in \Gamma\}$ where

$$Q_\gamma \;\equiv\; P_{\theta(\gamma)}.$$

Identify ν by

$$\nu(\gamma) \;\equiv\; \nu(Q_\gamma) \;\equiv\; q(\theta(\gamma)).$$

Then, by the chain rule, the Fisher information matrix for γ is

$$\dot{\theta}(\gamma) I(\theta(\gamma)) \dot{\theta}^T(\gamma)$$

while

$$\dot{\nu}(\gamma) \;=\; \dot{q}(\theta(\gamma)) \dot{\theta}^T(\gamma)\,.$$

Substituting back into (1) gives the same answer for $\gamma \to Q_\gamma$ as for $\theta \to P_\theta$. A similar calculation works for $\tilde{\mathbf{l}}$. $\square$

Now we specialize slightly: suppose that $\theta^T = (\nu^T, \eta^T)$ where $\nu \in N \subset R^m$, $\eta \in H \subset R^{k-m}$; here ν is the *parameter of interest* and η is a *nuisance parameter*. We can think of this as $q(\theta) \;=\; q(\nu\,, \eta) \;=\; \nu$ so that $\dot{q}(\theta) \;=\; (Ident\,, 0)$ ($m \times k$).

If $\theta_0 = (\nu_0, \eta_0) \in \Theta$ let $\mathbf{P}_1(\eta_0) \;\equiv\; \{P_\theta : \eta = \eta_0, \nu \in N\}$. This is the model when $\eta = \eta_0$ is known. We want to assess the cost of not knowing η by comparing the information bounds and efficient influence functions for ν at P_{θ_0} in $\mathbf{P}_1(\eta_0)$ and $\mathbf{P}$.

We let $<\,,\,>_0$ be the inner product in $L_2(P_{\theta_0})$, $||\cdot||_0$ the norm, and write E_0 for expectation under P_{θ_0} .

Suppose the model is regular and write $\dot{\mathbf{l}}$ for the score function at θ_0 and $\tilde{\mathbf{l}} \;=\; I^{-1}(\theta_0)\dot{\mathbf{l}}$ for the efficient influence function of the parameter θ at P_{θ_0} in $\mathbf{P}$. Decompose

$$\dot{\mathbf{l}} \;=\; \begin{pmatrix} \dot{\mathbf{l}}_1 \\ \dot{\mathbf{l}}_2 \end{pmatrix}\,, \qquad \tilde{\mathbf{l}} \;=\; \begin{pmatrix} \tilde{\mathbf{l}}_1 \\ \tilde{\mathbf{l}}_2 \end{pmatrix}$$

with $\tilde{\mathbf{l}}_1$, $\dot{\mathbf{l}}_1$ m-vectors, $\tilde{\mathbf{l}}_2$, $\dot{\mathbf{l}}_2$ $(k-m)$-vectors. Write $I(\theta_0)$ in block matrix form, suppressing dependence on θ_0, as

$$I \;=\; [\,I_{ij}\,]_{i,j=1,2} \;=\; \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix}$$

with I_{11} $m \times m$, I_{12} $m \times (k-m)$, I_{21} $(k-m) \times m$, I_{22} $(k-m) \times (k-m)$, and similarly decompose $I^{-1}(\theta_0)$ into I^{ij}, $i,j = 1,2$. By well-known block matrix forms of matrix inverses we have

(3) $$I^{-1}(\theta_0) \;=\; [I^{ij}]_{i,j\,=\,1,2} \;=\; \begin{pmatrix} I_{11\cdot 2}^{-1} & -\,I_{11\cdot 2}^{-1} I_{12} I_{22}^{-1} \\ -\,I_{22\cdot 1}^{-1} I_{21} I_{11}^{-1} & I_{22\cdot 1}^{-1} \end{pmatrix}$$

where

(4) $$\begin{aligned} I_{11\cdot 2} &\equiv I_{11} \;-\; I_{12} I_{22}^{-1} I_{21} \\ I_{22\cdot 1} &\equiv I_{22} \;-\; I_{21} I_{11}^{-1} I_{12} \end{aligned} \;.$$

By (1) and (2) on page 12, the information bound for estimating ν in $\mathbf{P}$ is $I^{11} = I_{11\cdot 2}^{-1}$ and the efficient influence function for ν in $\mathbf{P}$ is

(5) $$\begin{aligned} \tilde{\mathbf{l}}_1 &= I^{11}\dot{\mathbf{l}}_1 \;+\; I^{12}\dot{\mathbf{l}}_2 \\ &= I_{11\cdot 2}^{-1}(\dot{\mathbf{l}}_1 \;-\; I_{12} I_{22}^{-1}\dot{\mathbf{l}}_2) \qquad \text{by (1)} \\ &\equiv I_{11\cdot 2}^{-1}\mathbf{l}_1^{*} \,. \end{aligned}$$

Since

$$\begin{aligned} I_{11\cdot 2} &= E_0(\,\dot{\mathbf{l}}_1 \;-\; I_{12}I_{22}^{-1}\dot{\mathbf{l}}_2)(\dot{\mathbf{l}}_1 \;-\; I_{12}I_{22}^{-1}\dot{\mathbf{l}}_2)^T \\ &= E\,\mathbf{l}_1^{*}\mathbf{l}_1^{*T} \,, \end{aligned}$$

we see that (5) has the same form as $\tilde{\mathbf{l}} = I^{-1}(\theta_0)\dot{\mathbf{l}}$ with $\tilde{\mathbf{l}}$ replaced by $\tilde{\mathbf{l}}_1$, $I(\theta_0) = E_0(\,\dot{\mathbf{l}}\dot{\mathbf{l}}^T)$ replaced by $I_{11\cdot 2} = E_0(\,\mathbf{l}_1^{*}\mathbf{l}_1^{*T})$, and $\dot{\mathbf{l}}$ replaced by

(6) $$\mathbf{l}_1^{*} \;\equiv\; \dot{\mathbf{l}}_1 \;-\; I_{12}I_{22}^{-1}\dot{\mathbf{l}}_2 \,.$$

We therefore call $\mathbf{l}_1^{*}$ the *efficient score function* for ν in $\mathbf{P}$, and call $I_{11\cdot 2}$ the *information for* ν in $\mathbf{P}$.

If on the other hand $\eta = \eta_0$ is treated as known, the information bound (for ν in $\mathbf{P}_1(\eta_0)$) is I_{11}^{-1} and the corresponding efficient influence curve (for ν in $\mathbf{P}_1(\eta_0)$) is just

(7) $$I_{11}^{-1}\,\dot{\mathbf{l}}_1 \,.$$

From the block matrix formulas relating $[\,I_{ij}\,]$ and $[\,I^{ij}\,]$, we can derive some important relations between these quantities. First note from (3) and (4) that

(8) $$(I^{11})^{-1} \;=\; I_{11\cdot 2} \;=\; I_{11} \;-\; I_{12}I_{22}^{-1}I_{21} \,,$$

so not knowing η decreases the information for ν by $I_{12}I_{22}^{-1}I_{21}$. Similarly,

$$I_{11}^{-1} = I^{11} - I^{12}(I^{22})^{-1}I^{21}$$

or

$$(9) \qquad I^{11} = I_{11\cdot 2}^{-1} = I_{11}^{-1} + I^{12}(I^{22})^{-1}I^{21},$$

so not knowing η increases the information bound (inverse information) by $I^{12}(I^{22})^{-1}I^{21}$. Moreover, from (8),

$$(10) \qquad I_{11\cdot 2} = I_{11} \quad \text{and} \quad I_{11\cdot 2}^{-1} = I_{11}^{-1}$$

if and only if

$$(11) \qquad I_{12} = 0\,.$$

In this case it also follows from (5), (6) and (10) that

$$(12) \qquad \tilde{\mathbf{l}}_1 = I_{11}^{-1}\dot{\mathbf{l}}_1 \quad \text{and} \quad \mathbf{l}_1^* = \dot{\mathbf{l}}_1\,.$$

Definition 2.9. $\{\hat{\nu}_n\}$ is an *adaptive estimate* of ν in the presence of η if $\hat{\nu}_n$ is regular on $\mathbf{P}$ and efficient for each of the models $\mathbf{P}_1(\eta)$, for all η.

If an adaptive estimate exists, we can do as well not knowing η as knowing it. By (11) and (12), a necessary condition for the existence of adaptive estimates in regular parametric models is

$$(13) \qquad I_{12}(\theta) = 0 \quad \text{for all} \quad \theta\,.$$

Adaptation is very much a feature of the parametrization, as the following examples show.

Example 1.3.2.A, continued.

Suppose that $P_\theta \equiv N(\nu, \eta)$, $\nu \in R$, $\eta > 0$. Thus $I_{12}(\theta) = 0$ for all θ. In this, the normal location-scale model, we can estimate the mean equally well whether we know or do not know the variance. □

Example 2.10. Reparametrization of normal location - scale.

Now suppose that $P_\theta = N(\nu, \eta - \nu^2)$, $\eta > \nu^2$. Then easy calculation shows that

$$I_{12}(\theta) = -\frac{\nu}{(\eta - \nu^2)^2}$$

by the classical formula

$$I(\theta) = -\left[E_\theta\left(\frac{\partial^2 \mathbf{l}(\theta)}{\partial\theta_i \partial\theta_j}\right)\right].$$

□

We can think of $\mathbf{l}_1^*$ as the $\dot{\mathbf{l}}_1$ corresponding to the reparametrization $(\nu, \eta) \to (\nu\,, \eta\, +\, I_{22}^{-1}(\theta_0) I_{21}(\theta_0)(\nu - \nu_0))$. With this reparametrization, adaptation at θ_0 becomes possible since $\dot{\mathbf{l}}_2$ is unchanged and condition (3) is satisfied. If we can paste together these local reparametrizations and find $(\nu, \eta) \to (\nu, \gamma(\nu, \eta))$ such that

$$\gamma(\nu, \eta) \;-\; \gamma(\nu_0, \eta_0) \;=\; \eta \;-\; \eta_0 \;+\; I_{22}^{-1}(\theta_0) I_{21}(\theta_0)(\nu - \nu_0) \;+\; o(|\theta - \theta_0|)$$

for every $\theta_0 = (\nu_0, \eta_0)$, then under this reparametrization the necessary condition for adaptation holds. For instance in example 6 we can take $\gamma(\nu, \eta) = \eta - \nu^2$. These remarks have little practical significance since the initial parametrization is usually natural and the reparametrization is not.

The efficient influence function $\tilde{\mathbf{l}}_1$ and efficient score function $\mathbf{l}_1^*$ can be interpreted geometrically in the Hilbert space $L_2(P_{\theta_0})$; see sections A.1 and A.2 for elementary Hilbert space theory. First suppose $m = 1$. Let $[\dot{\mathbf{l}}_2]$ be the linear span of the components of $\dot{\mathbf{l}}_2$ in $L_2(P_{\theta_0})$. Then by example A.2.1, $I_{12} I_{22}^{-1} \dot{\mathbf{l}}_2$ is the projection of $\dot{\mathbf{l}}_1$ on $[\dot{\mathbf{l}}_2]$, and by (6) the efficient score function $\mathbf{l}_1^*$ is the projection of $\dot{\mathbf{l}}_1$ on the orthocomplement of $[\dot{\mathbf{l}}_2]$.

We can also relate the efficient influence functions $\tilde{\mathbf{l}}_1$ and $I_{11}^{-1} \dot{\mathbf{l}}_1$ for ν in $\mathbf{P}$ and $\mathbf{P}_1(\eta_0)$. In particular, $I_{11}^{-1} \dot{\mathbf{l}}_1$ is the projection of $\tilde{\mathbf{l}}_1$ on $[\dot{\mathbf{l}}_1]$. We need only check that $\tilde{\mathbf{l}}_1 - I_{11}^{-1} \dot{\mathbf{l}}_1 = (I^{11} - I_{11}^{-1}) \dot{\mathbf{l}}_1 + I^{12} \dot{\mathbf{l}}_2$ is orthogonal to $\dot{\mathbf{l}}_1$, and this follows from $I^{11} I_{11} + I^{12} I_{21} = 1$.

If $m > 1$ these relationships continue to hold if projection is interpreted componentwise. The following basic proposition can be viewed as providing the rationale for two different approaches to computing information bounds in semiparametric models which will be presented in chapter 3 of Bickel, Klaassen, Ritov, and Wellner (1993).

Proposition 2.11.

A. The efficient score function $\mathbf{l}_1^*(\,\bullet\,, P_{\theta_0} \mid \nu, \mathbf{P})$ is the projection of the score function $\dot{\mathbf{l}}_1$ on the orthocomplement of $[\,\dot{\mathbf{l}}_2\,]$ in $L_2(P_{\theta_0})$.

B. The efficient influence function $\tilde{\mathbf{l}}(\,\bullet\,, P_{\theta_0} \mid \nu\,, \mathbf{P}_1(\eta_0))$ is the projection of the efficient influence function $\tilde{\mathbf{l}}_1(\,\bullet\,, P_{\theta_0} \mid \nu\,, \mathbf{P})$ on $[\,\dot{\mathbf{l}}_1\,]$ in $L_2(P_{\theta_0})$.

See figures 1 and 2.

Here is another relationship between the influence and score functions of $\mathbf{P}_1(\eta_0)$ and its companion $\mathbf{P}_2(\nu_0) \equiv \{P_{(\nu_0\,, \eta)} : \eta \in H\}$. We use the subscript 2 for score and influence function in the companion model. The efficient influence function $\tilde{\mathbf{l}}_1$ can be written as

(14) $$\tilde{\mathbf{l}}_1 \;=\; I_{11}^{-1} \dot{\mathbf{l}}_1 \;-\; I_{11}^{-1} I_{12} \tilde{\mathbf{l}}_2 \,.$$

This relationship was implicit in section 4 of Begun, Hall, Huang, and Wellner (1983); it will appear again in the context of semiparametric models (with ν infinite - dimensional and η finite - dimensional) in section 5.4. Note that (14) provides an immediate proof, by orthogonality of $\tilde{\mathbf{l}}_2$ to $[\dot{\mathbf{l}}_1]$, of the formula

$$I_{11\cdot 2}^{-1} = I_{11}^{-1} + I_{11}^{-1} I_{12} I_{22\cdot 1}^{-1} I_{21} I_{11}^{-1} , \tag{15}$$

which is another way of writing (9).

Proof of (14). From (3),

$$\begin{aligned}
&\tilde{\mathbf{l}}_1 + I_{11}^{-1} I_{12} \tilde{\mathbf{l}}_2 \\
&\quad = I^{11}\dot{\mathbf{l}}_1 + I^{12}\dot{\mathbf{l}}_2 + I_{11}^{-1} I_{12} (I^{21}\dot{\mathbf{l}}_1 + I^{22}\dot{\mathbf{l}}_2) \\
&\quad = I_{11}^{-1}\Big\{ (I_{11}I^{11} + I_{12}I^{21})\dot{\mathbf{l}}_1 + (I_{11}I^{12} + I_{12}I^{22})\,\dot{\mathbf{l}}_2 \Big\} \\
&\quad = I_{11}^{-1}\,\dot{\mathbf{l}}_1 .
\end{aligned}$$

□

The following table summarizes the efficient score functions, efficient influence functions, information, and inverse information for the two models $\mathbf{P}$ and $\mathbf{P}_1(\eta_0)$.

		Model	
name	**notation**	$\mathbf{P}$	$\mathbf{P}_1(\eta_0)$
efficient score	$\mathbf{l}_1^*(\cdot\,, P \mid \nu\,, \cdot)$	$\mathbf{l}_1^* = \dot{\mathbf{l}}_1 - I_{12} I_{22}^{-1} \dot{\mathbf{l}}_2$	$\dot{\mathbf{l}}_1$
information	$I(P \mid \nu\,, \cdot)$	$E\,\mathbf{l}_1^*\mathbf{l}_1^{*T} = I_{11} - I_{12} I_{22}^{-1} I_{21}$ $\equiv I_{11\cdot 2}$	I_{11}
efficient influence function	$\tilde{\mathbf{l}}_1(\cdot\,, P \mid \nu\,, \cdot)$	$\tilde{\mathbf{l}}_1 = I^{11}\dot{\mathbf{l}}_1 + I^{12}\,\dot{\mathbf{l}}_2$ $= I_{11\cdot 2}^{-1}\mathbf{l}_1^*$ $= I_{11}^{-1}\dot{\mathbf{l}}_1 - I_{11}^{-1} I_{12}\,\tilde{\mathbf{l}}_2$	$I_{11}^{-1}\dot{\mathbf{l}}_1$
information bound	$I^{-1}(P \mid \nu\,, \cdot)$	$I^{11} = I_{11\cdot 2}^{-1}$ $= I_{11}^{-1} + I_{11}^{-1} I_{12}\; I_{22\cdot 1}^{-1} I_{21} I_{11}^{-1}$	I_{11}^{-1}

Table 1.

Proposition 2.11 can be put in a broader context.

Proposition 2.12. Let $m = 1$ and suppose that T_n is an asymptotically linear estimator of ν with influence function ψ. Then:

A. T_n is Gaussian regular if and only if

(16) $$\psi - \tilde{\mathbf{l}}_1 \perp \dot{\mathbf{P}} = [\,\dot{\mathbf{l}}_1, \dot{\mathbf{l}}_2]\,,$$

or, equivalently, if and only if both

(17) $$< \psi, \dot{\mathbf{l}}_1 >_0 = 1$$

and

(18) $$\psi \perp [\dot{\mathbf{l}}_2]\,.$$

B. If T_n is regular, then $\psi \in \dot{\mathbf{P}} = [\dot{\mathbf{l}}_1, \dot{\mathbf{l}}_2]$ if and only if $\psi = \tilde{\mathbf{l}}_1$.

We note in passing that (17) and (18) are asymptotic versions of the equations leading to the Cramér-Rao information bound. Consider the problem of minimizing $\Sigma(P_{\theta_0}, T) = E_0\,\psi^2$ subject to (17) and (18). For simplicity take $k = 2$. If we write

$$\psi = c\dot{\mathbf{l}}_1 + d\dot{\mathbf{l}}_2 + \Delta\,,$$

where $\Delta \perp [\dot{\mathbf{l}}_1, \dot{\mathbf{l}}_2]$, then (18) holds if and only if

$$\psi = c(\,\dot{\mathbf{l}}_1 - I_{12}I_{22}^{-1}\dot{\mathbf{l}}_2) + \Delta = c\mathbf{l}_1^* + \Delta\,,$$

while (17) forces

$$c = ||\dot{\mathbf{l}}_1 - I_{12}I_{22}^{-1}\dot{\mathbf{l}}_2||_0^{-2}\,.$$

Finally

$$||\psi||_0^2 = ||\mathbf{l}_1^*||_0^{-2} + ||\Delta||_0^2\,.$$

Therefore, the minimizing $\Delta = 0$ and as expected the minimizing ψ is the efficient influence function. This argument makes clear the characterizing features of the efficient influence function implied in proposition 1.B:

(i) $\tilde{\mathbf{l}}_1$ and all other influence functions are orthogonal to $[\,\dot{\mathbf{l}}_2\,]$.

(ii) $\tilde{\mathbf{l}}_1$ is the unique influence function belonging to $[\,\dot{\mathbf{l}}_1, \dot{\mathbf{l}}_2\,]$.

(iii) $\tilde{\mathbf{l}}_1$ can be obtained by projecting any influence function ψ corresponding to a regular estimate for ν into $[\,\dot{\mathbf{l}}_1, \dot{\mathbf{l}}_2\,]$.

Here is a slight generalization of proposition 2 to a general function $\nu(P_\theta) = q(\theta)$.

Proposition 2.13. Suppose that T_n is an asymptotically linear estimator at θ_0 of $\nu(P_\theta) = q(\theta)$ with influence function ψ where $q : \Theta \to R^m$. Then:

A. T_n is (Gaussian) regular at θ_0 if and only if $q(\theta)$ is differentiable at θ_0 with derivative $\dot{q}(\theta_0)$ and, with $\tilde{\mathbf{l}} \equiv \tilde{\mathbf{l}}(\,\cdot\,, P_{\theta_0}\,|\,\nu, \mathbf{P})$,

(19) $$\psi - \tilde{\mathbf{l}} \perp \dot{\mathbf{P}} = [\dot{\mathbf{l}}_1, \dot{\mathbf{l}}_2]$$

where (19) is equivalent to

(20) $$E_0\,\psi\,\dot{\mathbf{l}}^T \;=\; \dot{q}(\theta_0)\,.$$

B. If T_n is regular then $\psi \in \dot{\mathbf{P}}^m$ if and only if

(21) $$\psi \;=\; \tilde{\mathbf{l}} \;=\; \dot{q}(\theta_0)\,I^{-1}(\theta_0)\,\dot{\mathbf{l}}(\theta_0)\,.$$

Proof. By asymptotic linearity of T_n and proposition 2.1.2,

(a) $$\mathbf{L}_{\theta_0}\begin{pmatrix}\sqrt{n}\,(T_n - q\,(\theta_0)) \\ L_n\,(\theta_0 + t_n\,/\,\sqrt{n}) - L_n\,(\theta_0)\end{pmatrix} \;\to\; N\left(\begin{pmatrix}0 \\ -\,\Sigma_{22}\,/\,2\end{pmatrix}, \Sigma\right)$$

where

(b) $$\Sigma \;=\; [\,\Sigma_{ij}\,],\;\; \Sigma_{11} \;=\; E_0\,\psi\psi^T,\;\; \Sigma_{12} \;=\; E_0\psi\,\dot{\mathbf{l}}^T\,t,\;\; \Sigma_{22} \;=\; t^T I(\theta_0)\,t\,,$$

$$t_n \;\to\; t\,.$$

Consequently, by Le Cam's third lemma (lemma A.9.3)

(c) $$\mathbf{L}_{\theta_0 + t_n/\sqrt{n}}\,(\sqrt{n}(T_n \;-\; q(\theta_0))) \;\to\; N(\Sigma_{12}\,,\,\Sigma_{11})\,.$$

Assume now that T_n is regular. Then

(d) $$\mathbf{L}_{\theta_0 + t_n/\sqrt{n}}\,(\sqrt{n}\,(T_n - q(\theta_0 + t_n\,/\,\sqrt{n}))) \;\to\; N(0, \Sigma_{11})$$

and from (c) and (d) we conclude

(e) $$\sqrt{n}\,(q(\theta_0 + t_n\,/\,\sqrt{n}) - q(\theta_0)) \;\to\; \Sigma_{12} \;=\; E_0\psi\,\dot{\mathbf{l}}^T\,t\,.$$

But this implies that q is differentiable at θ_0 with derivative $\dot{q}(\theta_0)$ satisfying (20) and hence (19).

On the other hand, if q is differentiable and (20) holds, then (e) is valid which together with (c) implies (d) and hence Gaussian regularity. The proof of A is complete.

As for B, note that A implies that $\dot{q}$ and hence $\tilde{\mathbf{l}}$ are well defined and that (19) holds. Since $\tilde{\mathbf{l}} \in \dot{\mathbf{P}}^m$, (19) yields $\psi \in \dot{\mathbf{P}}^m$ if and only if $\psi - \tilde{\mathbf{l}} \;=\; 0\,.$ □

Choosing $q(\theta) \;=\; q(\nu\,,\eta) \;=\; \nu$ in proposition 3 immediately yields a generalization of proposition 2 to $m \;>\; 1$. Now (20) becomes

(22) $$E_0\,\psi\,\dot{\mathbf{l}}_1^T \;=\; J_{m\times m}\;,$$

(23) $$E_0\,\psi\dot{\mathbf{l}}_2^T \;=\; 0$$

where J is the identity. In particular if $m \;=\; k$ we obtain that the influence function of any linear and Gaussian regular estimate of θ has

(24) $$E_0\,\psi\,\dot{\mathbf{l}}^T \;=\; J_{k\times k}\;.$$

3. Regular Estimates; Superefficiency; LAN & Le Cam's three lemmas

If $X_1, \cdots, X_n$ are iid p_θ, T_n is an unbiased estimator of $q(\theta)$, and the conditions of the Cramér - Rao inequality hold, then

$$(1) \qquad Var_\theta[T_n] \geq \frac{[\dot{q}(\theta)]^2}{n\, I(\theta)} .$$

If

$$(2) \qquad \sqrt{n}(T_n - q(\theta)) \to_d N(0, V(\theta)) ,$$

then it follows (from Fatou and Skorokhod) that

$$(3) \qquad V(\theta) \leq \liminf_{n \to \infty} \{n\, Var_\theta(T_n)\} .$$

If T_n is unbiased and

$$(3') \qquad V(\theta) = \lim_n \{n\, Var_\theta(T_n)\} ,$$

then (1) implies

$$(4) \qquad V(\theta) \geq \frac{[\dot{q}(\theta)]^2}{I(\theta)} .$$

Does the inequality in (4) holds under restrictions on p_θ only? The answer to this question is *no*, as is shown by the following example of Hodges.

Example 3.1. (Hodges). Let $X_1, \cdots, X_n$ be iid $N(\theta, 1)$ so $I(\theta) = 1$. Let $0 < a < 1$, and

$$(5) \qquad T_n \equiv \begin{cases} \overline{X}_n & \text{if } |\overline{X}_n| > n^{-1/4} \\ a\overline{X}_n & \text{if } |\overline{X}_n| \leq n^{-1/4} . \end{cases}$$

Then

$$(6) \qquad \sqrt{n}(T_n - \theta) \to_d N(0, V(\theta))$$

where

$$(7) \qquad V(\theta) = \begin{cases} 1 & \text{if } \theta \neq 0 \\ a^2 & \text{if } \theta = 0 \end{cases} .$$

Thus $V(\theta) \geq 1/I(\theta)$ fails at $\theta = 0$ if $|a| < 1$, and T_n is a *superefficient estimator* of θ at $\theta = 0$.

Proof of (6). Since $\sqrt{n}(\overline{X} - \theta) \stackrel{d}{=} Z \sim N(0, 1)$ for all $n \geq 1$,

$$\begin{aligned} \sqrt{n}(T_n - \theta) &= \sqrt{n}(\overline{X} - \theta) 1_{[|\overline{X}| > n^{-1/4}]} + \sqrt{n}(a\overline{X} - \theta) 1_{[|\overline{X}| \leq n^{-1/4}]} \\ &= \sqrt{n}(\overline{X} - \theta) 1_{[\sqrt{n}|\overline{X} - \theta + \theta| > n^{1/4}]} \end{aligned}$$

$$+ \ \{a\sqrt{n}(\bar{X} - \theta) + \sqrt{n}\theta(a-1)\}1_{[\sqrt{n}|\bar{X} - \theta + \theta| < n^{1/4}]}$$

$$\stackrel{d}{=} Z1_{[|Z + \sqrt{n}\theta| \geq n^{1/4}]} + \{aZ + \sqrt{n}\theta(a-1)\}1_{[|Z + \sqrt{n}\theta| \leq n^{1/4}]}$$

$$\rightarrow_{a.s.} \begin{cases} Z & \text{if } \theta \neq 0 \\ aZ & \text{if } \theta = 0 \end{cases} \sim N(0, V(\theta)) .$$

Note that $V(\theta)$ is a discontinuous function of θ .

If $\theta \equiv \theta_n = cn^{-1/2}$, then from (a), under P_{θ_n} we have

$$\sqrt{n}(T_n - \theta_n) \stackrel{d}{=} Z1_{[|Z + c| > n^{1/4}]} + \{aZ + c(a-1)\}1_{[|Z + c| \leq n^{1/4}]}$$

$$\rightarrow_{a.s.} aZ + c(a-1)$$

$$\sim N(c(a-1), a^2) .$$

Note that this limiting distribution depends on c , and hence Hodges superefficient estimator is not locally regular in the sense of the following definition:

Definition 3.2. $T = \{T_n\}$ is a *locally regular estimator* of θ at $\theta = \theta_0$ if, for every sequence $\{\theta_n\} \subset \Theta$ with $\sqrt{n}(\theta_n - \theta_0) \rightarrow t \in R^k$, under P_{θ_n}

$$\sqrt{n}(T_n - \theta_n) \rightarrow_d \mathbb{Z} \quad \text{as } n \rightarrow \infty$$

where the law of $\mathbb{Z}$ depends on θ_0 but not on t . Thus the limit distribution of $\sqrt{n}(T_n - \theta_n)$ does not depend on the direction of approach t of θ_n to θ_0 .

This will turn out to be a key hypothesis in the formulation of Hájek's convolution theorem in the next section.

Contiguity theory: Le Cam's three (or four!) lemmas and LAN

Consider a sequence of statistical problems (with only two sequences of probability measures) with:

measure spaces:	$(\mathbf{X}_n, \mathbf{A}_n, \mu_n)$
probability measures:	$P_n << \mu_n, \ Q_n << \mu_n$
densities:	$p_n \equiv \dfrac{dP_n}{d\mu_n}, \ q_n \equiv \dfrac{dQ_n}{d\mu_n}$
likelihood ratios:	$L_n \equiv \begin{cases} q_n/p_n & \text{if } p_n > 0 \\ 1 & \text{if } q_n = p_n = 0 \\ n & \text{if } q_n > 0 = p_n \end{cases}$.

Definition 3.3. The sequence $\{Q_n\}$ is *contiguous* to $\{P_n\}$ if for every sequence $B_n \in \mathbf{A}_n$ for which $P_n(B_n) \rightarrow 0$ it follows that $Q_n(B_n) \rightarrow 0$.

Thus contiguity of $\{Q_n\}$ to $\{P_n\}$ means that Q_n is "asymptotically absolutely continuous" with respect to P_n in the sense of domination of measures. We therefore denote contiguity of $\{Q_n\}$ to $\{P_n\}$ by $\{Q_n\} \lhd \{P_n\}$, a notation due to Witting and Nölle (1970). Two sequences are contiguous to each other if both $\{Q_n\} \lhd \{P_n\}$ and $\{P_n\} \lhd \{Q_n\}$ and we then write $\{P_n\} \lhd\rhd \{Q_n\}$.

Definition 3.4. The sequence $\{Q_n\}$ is *asymptotically orthogonal* to $\{P_n\}$ if there exist a sequence $B_n \in \mathbf{A}_n$ such that $Q_n(B_n) \to 1$ and $P_n(B_n) \to 0$.

Lemma 3.5. (Le Cam's first lemma.) Suppose that $\mathbf{L}(L_n \mid P_n) \to \mathbf{L}(L)$ and $E(L) = 1$. Then $\{Q_n\} \lhd \{P_n\}$.

Corollary 3.6. If $\mathbf{L}(\log L_n \mid P_n) \to \mathbf{L}(\log L) = N(-\frac{1}{2}\sigma^2, \sigma^2)$, then $\{Q_n\} \lhd\rhd \{P_n\}$.

Proof. Note that $\mathbf{L}(L) = \mathbf{L}(e^{\sigma Z - \sigma^2/2})$ where $\mathbf{L}(Z) = N(0,1)$ and hence $E(L) = 1$. □

Definition 3.7. A sequence of random variables $\{X_n\}$ (with X_n defined on $(\mathbf{X}_n, \mathbf{A}_n, P_n)$) is *uniformly integrable* if

$$\lim_{\lambda\to\infty} \overline{\lim_{n\to\infty}}\, P_n(\,|X_n| 1_{[|X_n| \ge \lambda]}) = 0 .$$

Proposition 3.8. $\{X_n\}$ is uniformly integrable if and only if both of the following hold:

(i) $\sup_{n\ge 1} P_n(|X_n|) < \infty$.

(ii) $B_n \in \mathbf{A}_n$ with $P_n(B_n) \to 0$ implies $P_n(|X_n| 1_{B_n}) \to 0$.

Proof. See e.g. Billingsley (1968) page 34 or Chow and Teicher (1978) pages 92 - 93. □

Lemma 3.9. (Hall and Loynes (1977).) $\{Q_n\} \lhd \{P_n\}$ if and only if $\{L_n\}$ is uniformly integrable with respect to $\{P_n\}$ and $Q_n(p_n = 0) \to 0$.

Now suppose that $\underline{X}_n = (X_1, \cdots, X_n) \in \mathbf{X}_n$ and that

$$p_n(\underline{x}_n) = \prod_{i=1}^n f_{ni}(x_i), \qquad P_n \equiv \prod_{i=1}^n P_{ni},$$

$$q_n(\underline{x}_n) = \prod_{i=1}^n g_{ni}(x_i), \qquad Q_n \equiv \prod_{i=1}^n Q_{ni},$$

so that

$$\text{(9)} \qquad \log L_n = \sum_{i=1}^n \log\left(\frac{g_{ni}}{f_{ni}}(X_i)\right) \quad \begin{cases} < \infty & \text{a.s. } P_n \\ > -\infty & \text{a.s. } Q_n . \end{cases}$$

Suppose that the summands in (1) satisfy the *uniform asymptotic negligibility (UAN)* condition

(10) $$\lim_{n \to \infty} \max_{1 \le i \le n} P_n\left(\left| \frac{g_{ni}}{f_{ni}}(X_i) - 1 \right| > \epsilon \right) = 0 \qquad \text{for all} \quad \epsilon > 0 .$$

To get random variables with finite variance (to which classical central limit theorems may be applied), let

(11) $$W_n \equiv 2 \sum_{i=1}^{n} \left\{ \frac{g_{ni}^{1/2}}{f_{ni}^{1/2}}(X_i) - 1 \right\} \equiv \sum_{i=1}^{n} T_{ni}$$

and note that

$$Var\left(\frac{g_{ni}^{1/2}}{f_{ni}^{1/2}}(X_i) \right) \le E\left(\frac{g_{ni}}{f_{ni}}(X_i) \right) = \int 1_{[f_{ni} > 0]}\, g_{ni}\, d\mu \le 1 .$$

The following lemma reduces the proof of asymptotic normality of $\log L_n$ to the problem of establishing asymptotic normality of W_n.

Lemma 3.10. (Le Cam's second lemma.) Suppose that the UAN condition (10) holds and $\mathbf{L}(W_n \,|\, P_n) \to N(-\frac{1}{4}\sigma^2 , \sigma^2)$. Then

(12) $$\log L_n - (W_n - \frac{1}{4}\sigma^2) = o_{P_n}(1)$$

(or, for every $\epsilon > 0$

$$P_n(|\log L_n - (W_n - \frac{1}{4}\sigma^2)| > \epsilon) \to 0 \qquad \text{as} \quad n \to \infty),$$

and hence

(13) $$\mathbf{L}(\log L_n \,|\, P_n) \to N(-\frac{1}{2}\sigma^2 , \sigma^2) .$$

The proof of lemma 3.10 involves a long truncation argument, and is therefore deferred to the end of the section.

Corollary 3.11. If f_n is a sequence of densities such that

$$|| \sqrt{n}\,(f_n^{1/2} - f^{1/2}) - \delta ||_2 \to 0 \qquad \text{as} \quad n \to \infty$$

where $|| \bullet ||_2$ is the $L_2(\mu)$ - metric and $\delta \in L_2(\mu)$, then with $p_n(\underline{x}) \equiv \prod_{i=1}^{n} f(x_i)$ and $q_n(\underline{x}) \equiv \prod_{i=1}^{n} f_n(x_i)$ it follows that

(14) $$\log L_n - \left(\frac{2}{\sqrt{n}} \sum_{i=1}^{n} \frac{\delta}{f^{1/2}}(X_i) - \frac{1}{2} ||2\delta||_2^2 \right) = o_{P_n}(1)$$

and hence

(15) $$\mathbf{L}(\log L_n \,|\, P_n) \to N(-\frac{1}{2}\sigma^2 , \sigma^2)$$

with $\sigma^2 = ||2\delta||_2^2 = 4||\delta||_2^2$.

Corollary 3.12. Suppose that $\mathbf{P} = \{P_\theta : \theta \in \Theta \subset R^k\}$ is a regular parametric model (dominated by μ) in the sense that

$$||\sqrt{p_{\theta+h}} - \sqrt{p_\theta} - \frac{1}{2}\dot{\mathbf{l}}_\theta h^T \sqrt{p_\theta}||_\mu = o(|h|) .$$

Then, with $\theta_n \equiv \theta + n^{-1/2}h$, $p_n(\underline{x}) \equiv \prod_{i=1}^n p_\theta(x_i)$, and $q_n(\underline{x}) \equiv \prod_{i=1}^n p_{\theta_n}(x_i)$, it follows that

$$\log L_n - (\frac{h}{\sqrt{n}} \sum_{i=1}^n \dot{\mathbf{l}}_\theta^T(X_i) - \frac{1}{2} hI(\theta)h^T) = o_{P_n}(1) ,$$

and hence

$$\mathbf{L}(\log L_n \,|\, P_n) \to N(-\frac{1}{2}\sigma^2, \sigma^2)$$

with $\sigma^2 = hI(\theta)h^T$.

Proof. This follows immediately from corollary 3.11 with the identification $2\delta/f^{1/2} = \dot{\mathbf{l}}_\theta$. □

Now suppose that under

(16) $\quad P_n : \quad X_{n1}, \cdots, X_{nn}$ are iid f ,

and under

(17) $\quad Q_n : \quad X_{n1}, \cdots, X_{nn}$ are independent with densities $f_{n1}, \cdots, f_{nn}$

with respect to μ . Assume that $a_{n1}, \cdots, a_{nn}$, $n \ge 1$ are constants which satisfy

(18) $$\max_{1 \le i \le n} \frac{a_{ni}^2}{a_n^T a_n} = \frac{\max_{1 \le i \le n} a_{ni}^2}{\sum_{i=1}^n a_{ni}^2} \to 0 \quad \text{as} \quad n \to \infty ,$$

and suppose there exists $\delta \in L_2(\mu)$ such that

(19) $$\sum_{i=1}^n ||(f_{ni}^{1/2} - f^{1/2}) - \frac{a_{ni}}{\sqrt{a_n^T a_n}} \delta||_2^2 \to 0 \quad \text{as} \quad n \to \infty .$$

Corollary 3.13.. Suppose that (16) - (19) hold. Then

(20) $$\log L_n - (Z_n - \frac{1}{2}||2\delta||_2^2) = o_{P_n}(1)$$

where

(21) $$Z_n \equiv 2\sum_{i=1}^n \frac{a_{ni}}{\sqrt{a_n^T a_n}} \frac{\delta}{f^{1/2}}(X_{ni})$$

and

(22) $\quad \mathbf{L}(Z_n) \to N(0, ||2\delta||_2^2) \quad$ as $\quad n \to \infty$.

Proof. See Shorack and Wellner (1986) page 154 and 163 - 165. Note that Corollary 3.11 is the special case of Corollary 3.13 with all $a_{ni} = 1$ and $f_{ni} = f_n$ for all $i = 1, \cdots, n$.

Lemma 3.14. (Le Cam's third lemma.) Suppose that a statistic T_n satisfies

(23)
$$\mathbf{L}(\,(T_n, \log L_n)^T \mid P_n\,) \to \mathbf{L}(\,(T, \log L)^T\,) \sim N_2\left(\begin{pmatrix}\mu \\ -\sigma^2/2\end{pmatrix}, \begin{pmatrix}\tau^2 & c \\ c & \sigma^2\end{pmatrix}\right).$$

Then

(24)
$$\mathbf{L}(\,(T_n, \log L_n)^T \mid Q_n\,) \to \mathbf{L}(\,(T + c, \log L + \sigma^2)^T) \sim N_2\left(\begin{pmatrix}\mu + c \\ +\sigma^2/2\end{pmatrix}, \begin{pmatrix}\tau^2 & c \\ c & \sigma^2\end{pmatrix}\right).$$

Remark. If T_n is asymptotically linear and $\log L_n$ is asymptotically linear, for example as in (14) or (20), then verification of (23) is straightforward via the multivariate central limit theorem.

Proofs

Proof of lemma 3.5.(Le Cam's first lemma) Let $B_n \in \mathbf{A}_n$ with $P_n(B_n) \to 0$. By the Neyman - Pearson lemma there is a critical function $\phi_n \equiv 1[L_n > k_n] + \gamma_n 1[L_n = k_n]$ such that $P_n(\phi_n) = \alpha_n \equiv P_n(B_n) \to 0$ and

$$Q_n(B_n) \le Q_n(\phi_n)\,.$$

But for any fixed $0 < y < \infty$

$$\begin{aligned} Q_n(B_n) \le Q_n(\phi_n) &= Q_n(\phi_n 1_{[L_n \le y]}) + Q_n(\phi_n 1_{[L_n > y]}) \\ &\le y\,P_n(\phi_n) + Q_n(1_{[L_n > y]}) \\ &= y\,P_n(\phi_n) + 1 - Q_n(L_n \le y) \\ &= y\,P_n(\phi_n) + 1 - P_n(L_n 1_{[L_n \le y]})\,. \qquad \text{(a)} \end{aligned}$$

Let $\epsilon > 0$ and choose y to be a continuity point of $\mathbf{L}(L)$ such that $1 - E(L 1_{[L \le y]}) < \epsilon/2$; this is possible since $E(L) = 1$ by hypothesis. Then $\mathbf{L}(L_n \mid P_n) \to \mathbf{L}(L)$ implies that $P_n(L_n 1_{[L_n \le y]}) \to E(L 1_{[L \le y]})$ and hence $1 - P_n(L_n 1_{[L_n \le y]}) < \epsilon$ for $n \ge N_1$. Since $P_n(B_n) \to 0$ we also have $y\,P_n(B_n) < \epsilon$ for $n \ge$ some N_2 and hence it follows from (a) that $Q_n(B_n) < 2\epsilon$ for $n \ge \max\{N_1, N_2\}$. $\square$

Proof of lemma 3.9. (Le Cam's 4th lemma) First note that for $B_n \in \mathbf{A}_n$ we have

$$
\begin{aligned}
Q_n(B_n) &= \int 1_{B_n}\, dQ_n \\
&= \int 1_{B_n \cap [p_n = 0]}\, dQ_n + \int 1_{B_n \cap [p_n > 0]}\, L_n\, dP_n \\
&= \int 1_{B_n \cap [p_n = 0]}\, dQ_n + \int 1_{B_n}\, L_n\, dP_n \\
\text{(a)} \qquad &\le Q_n(p_n = 0) + \int 1_{B_n}\, L_n\, dP_n \\
\text{(b)} \qquad &\ge \int 1_{B_n}\, L_n\, dP_n \ .
\end{aligned}
$$

Thus if $P_n(B_n) \to 0$, L_n is uniformly integrable and $Q_n(p_n = 0) \to 0$, then $Q_n(B_n) \to 0$ by (a) and proposition 3.8, so $\{Q_n\} \lhd \{P_n\}$.

Conversely, if $\{Q_n\} \lhd \{P_n\}$ so that $Q_n(B_n) \to 0$, then (b) implies that $\int 1_{B_n} L_n\, dP_n \to 0$ so (ii) of proposition 3.8 holds. Part (i) of proposition 1 holds trivially since $P_n(L_n) = \int L_n\, dP_n = \int 1_{[p_n > 0]}\, dQ_n \le 1$, and therefore $\{L_n\}$ is uniformly integrable with respect to $\{P_n\}$ by proposition 3.8 Since $P_n(p_n = 0) = 0$, contiguity implies that $Q_n(p_n = 0) \to 0$. □

Proof of lemma 3.10. The following proof is from Hájek and Sidák (1967).

For any function h with second derivative h'' we have

$$
\begin{aligned}
h(x) &= h(x_0) + (x - x_0)\, h'(x_0) \\
&\quad + \frac{1}{2}(x - x_0)^2 \int_0^1 2\,(1 - \lambda)\, h''(x_0 + \lambda\,(x - x_0))\, d\lambda
\end{aligned}
$$

by integration by parts. Thus for $h(x) = \log(1 + x)$,

$$
\log(1 + x) = x - \frac{1}{2}x^2 \int_0^1 \frac{2\,(1-\lambda)}{(1 + \lambda\, x)^2}\, d\lambda \ .
$$

Thus, with T_{ni} as in (11)

$$
\begin{aligned}
\log\Big(\frac{g_{ni}}{f_{ni}}(X_i)\Big) &= 2\log\Big(1 + \frac{1}{2}T_{ni}\Big) \\
&= T_{ni} - \frac{1}{4}T_{ni}^2 \int_0^1 \frac{2\,(1-\lambda)}{(1 + \frac{1}{2}\lambda T_{ni})^2}\, d\lambda
\end{aligned}
$$

and

$$
\text{(a)} \qquad \log(L_n) = W_n - \frac{1}{4}\sum_{i=1}^n T_{ni}^2 \int_0^1 \frac{2\,(1-\lambda)}{(1 + \frac{1}{2}\lambda T_{ni})^2}\, d\lambda \ .
$$

Set $T_{ni}^{\delta} \equiv T_{ni} 1_{[|T_{ni}| \le \delta]}$ for $\delta > 0$. From the normal convergence criteria (see e.g. Loéve (1963), page 316), $\mathbf{L}(W_n \mid P_n) \to N(-\frac{1}{4}\sigma^2, \sigma^2)$ and the UAN condition (10) hold if and only if

(c) $$\sum_{i=1}^{n} P_n(|T_{ni}| > \delta) \to 0 ,$$

(d) $$\sum_{i=1}^{n} E(T_{ni}^{\delta}) \to -\frac{1}{4}\sigma^2 ,$$

and

(e) $$\sum_{i=1}^{n} Var(T_{ni}^{\delta}) \to \sigma^2 .$$

where all expectations and variances are under P_n. Note that $\int_0^1 2(1-\lambda)\, d\lambda = 1$ and

$$P_n \{ \max_{1 \le i \le n} |T_{ni}| > \delta \} \le \sum_{i=1}^{n} P_n(|T_{ni}| > \delta) \to 0 \qquad \text{by (c)} .$$

Thus for any $0 < \eta < 1$ there is an $N = N(\eta)$ such that, for $n \ge N$, $P_n(S_n) > 1 - \eta$ on $S_n \equiv [\max_{1 \le i \le n} |T_{ni}| \le \eta]$. It follows that, on S_n

$$\sup_{\lambda} \max_{1 \le i \le n} |(1 + \frac{1}{2}\lambda T_{ni})^{-2} - 1| \le 8\eta$$

and hence

$$\max_{1 \le i \le n} | \int_0^1 \frac{2(1-\lambda)}{(1 + \frac{1}{2}\lambda T_{ni})^2}\, d\lambda - 1 | \le 8\eta .$$

Also, since $T_{ni} = T_{ni}^{\eta}$ for $i = 1, \cdots, n$, on S_n

$$| \sum_{i=1}^{n} T_{ni}^2 \int_0^1 \frac{2(1-\lambda)}{(1 + \frac{1}{2}\lambda T_{ni})^2}\, d\lambda - \sum_{i=1}^{n} T_{ni}^2 | \le 8\eta \sum_{i=1}^{n} T_{ni}^2 = 8\eta \sum_{i=1}^{n} (T_{ni}^{\eta})^2$$

so that

$$\left| \frac{\sum_{i=1}^{n} T_{ni}^2 \int_0^1 \frac{2(1-\lambda)}{(1 + \frac{1}{2}\lambda T_{ni})^2}\, d\lambda}{\sum_{i=1}^{n} (T_{ni}^{\eta})^2} - 1 \right| \le 8\eta \qquad \text{on } S_n .$$

Thus in order to prove the lemma it suffices to show that

(f) $$\sum_{i=1}^{n} (T_{ni}^{\eta})^2 \;\to_{P_n}\; \sigma^2 \,.$$

To prove (f) it suffices, by Chebychev's inequality, to show that

(g) $$\sum_{i=1}^{n} E[(T_{ni}^{\eta})^2] \;\to\; \sigma^2$$

and

(h) $$\lim_{\eta \to 0} \limsup_{n \to \infty} \sum_{i=1}^{n} Var[(T_{ni}^{\eta})^2] \;=\; 0\,.$$

But by virtue of (e), (g) is equivalent to

(i) $$\sum_{i=1}^{n} [E(T_{ni}^{\eta})]^2 \;\to\; 0\,.$$

We first prove (i) and hence (g): If $\eta > 2$, then $T_{ni}^{\eta} \le T_{ni}$ since $T_{ni} \ge -2$ a.s. by definition of T_{ni}. Therefore

$$E(T_{ni}^{\eta}) \;\le\; E\,T_{ni} \;=\; 2\,E\{\,\frac{g_{ni}^{1/2}}{f_{ni}^{1/2}}(X_i)\,\} \;-\; 2 \;\le\; 0$$

by Jensen's inequality. Thus for $\eta > 2$

$$\sum_{i=1}^{n} (\,-\,E\,T_{ni}^{\eta}\,)^2 \;\le\; \max_{1 \le i \le n}\,(\,-\,E\,T_{ni}^{\eta}\,) \sum_{i=1}^{n} (\,-\,E(T_{ni}^{\eta})) \;\to\; 0$$

since

$$\sum_{i=1}^{n} (\,-\,E(T_{ni}^{\eta})) \;\to\; \frac{1}{4}\,\sigma^2 \qquad \text{by (d)}$$

and

$$\max_{1 \le i \le n}\,(\,-\,E\,T_{ni}^{\eta}\,) \;\to\; 0 \qquad \text{by the UAN condition (10)}\,.$$

Now note that if (i) holds for any $\eta > 2$, it holds for all $\eta > 0$: Since

$$\sum_{i=1}^{n} E[(T_{ni}^{\eta})^2] \;\le\; \sum_{i=1}^{n} E[(T_{ni}^{\gamma})^2] \qquad \text{for } \eta < \gamma$$

and, by (e) both $\sum_{i=1}^{n} Var(\,T_{ni}^{\eta}\,) \to \sigma^2$ and $\sum_{i=1}^{n} Var(\,T_{ni}^{\gamma}\,) \to \sigma^2$, it follows that

$$\begin{aligned} \sum_{i=1}^{n} [E(T_{ni}^{\eta})]^2 \;&=\; \sum_{i=1}^{n} \{\,E[(T_{ni}^{\eta})^2] \;-\; Var(T_{ni}^{\eta})\,\} \\ &\le\; \sum_{i=1}^{n} \{\,E[(T_{ni}^{\gamma})^2] \;-\; Var(T_{ni}^{\gamma}) \;+\; Var(T_{ni}^{\gamma}) \;-\; Var(T_{ni}^{\eta})\,\} \end{aligned}$$

$$= \sum_{i=1}^{n} [E(T_{ni}^{\gamma})]^2 \; + \; \sum_{i=1}^{n} Var(T_{ni}^{\gamma}) \; - \; \sum_{i=1}^{n} Var(T_{ni}^{\eta})$$
$$\to \; 0 \; + \; \sigma^2 \; - \; \sigma^2 \; = \; 0 \; ,$$

completing the proof of (i) and hence (g).

To prove (h), note that

$$\sum_{i=1}^{n} Var[(T_{ni}^{\eta})^2] \; \le \; \sum_{i=1}^{n} E[(T_{ni}^{\eta})^4\,] \; \le \; \eta^2 \sum_{i=1}^{n} E[(T_{ni}^{\eta})^2] \; .$$

Then, by (g)

$$\limsup_{n \to \infty} \sum_{i=1}^{n} Var[(T_{ni}^{\eta})^2] \;\; \le \;\; \eta^2\, \sigma^2 \; ,$$

and hence (h) holds. □

Proof of corollary 3.11. Let

$$T_{ni} \; \equiv \; 2\,\{ \, \frac{f_n^{1/2}(X_i)}{f^{1/2}(X_i)} \; - \; 1 \, \} \, , \qquad i \; = \; 1, \cdots, n, \quad n \; \ge \; 1 \, .$$

Note that (all expectations and variances being calculated under P with density f)

$$E(T_{ni}) \; = \; 2\{ \int f_n^{1/2} f^{1/2} \, d\mu \; - \; 1 \} \; = \; - \, H^2(f_n, f)$$

where

$$H^2(f_n, f) \; \equiv \; \int (f_n^{1/2} \; - \; f^{1/2})^2 \, d\mu \; = \; ||\, f_n^{1/2} \; - \; f^{1/2}\, ||_2^2$$

and

$$Var(T_{ni}) \; = \; 4 \int (f_n^{1/2} \; - \; f^{1/2})^2 \, d\mu \; - \; [E(T_{ni})]^2$$
$$= \; 4\, H^2(f_n, f) \; - \; H^4(f_n, f) \, .$$

Therefore, since the hypothesis $||\sqrt{n}\,(f_n^{1/2} \; - \; f^{1/2}) \; - \; \delta\, ||_2 \; \to \; 0$ implies $n\, ||f_n^{1/2} \; - \; f^{1/2}\, ||_2^2 \; \to \; ||\delta||_2^2$ and $||f_n^{1/2} \; - \; f^{1/2}|| \; \to \; 0$, the W_n of lemma 2 has

(a) $$E(W_n) \; = \; - \, n\, H^2(f_n, f) \; \to \; - \, ||\delta||_2^2$$

and

$$Var(W_n) \; = \; 4\, n\, H^2(f_n, f) \; - \; n\, H^2(f_n, f)\, H^2(f_n, f)$$
$$\to \; 4\, ||\delta||_2^2 \, .$$

Note that since

$$\epsilon\, P\Big(|\frac{f_n}{f}(X_i) - 1| \ge \epsilon \Big) \le E\,|\frac{f_n}{f}(X) - 1|$$
$$= E\Big(|\frac{f_n^{1/2}}{f^{1/2}} - 1|\,|\frac{f_n^{1/2}}{f^{1/2}} + 1| \Big)$$
$$\le \|f_n^{1/2} - f^{1/2}\|_2\, \|f_n^{1/2} + f^{1/2}\|_2$$
$$\to 0$$

uniformly in $1 \le i \le n$ as $n \to \infty$, the UAN condition (10) holds. Furthermore,

$$Var\Big\{ W_n - \frac{2}{\sqrt{n}} \sum_{i=1}^n \frac{\delta}{f^{1/2}} \Big\}$$
$$= 4\, n\, Var\Big\{ \frac{f_n^{1/2}}{f^{1/2}}(X_1) - 1 - n^{-1/2} \frac{\delta}{f^{1/2}}(X_1) \Big\}$$
$$= 4\, n\, \|(f_n^{1/2} - f^{1/2}) - n^{-1/2}\delta\|_2^2 - n\, H^2(f_n, f)\, H^2(f_n, f)$$
$$= 4\, \|\sqrt{n}(f_n^{1/2} - f^{1/2}) - \delta\|_2^2 - n\, H^4(f_n, f)$$

(c) $$\to 0$$

and hence

$$E\Big\{ W_n - \|\delta\|_2^2 - \frac{2}{\sqrt{n}} \sum_{i=1}^n \frac{\delta}{f^{1/2}} + 2\|\delta\|_2^2 \Big\}^2$$
$$= Var\Big\{ W_n - \frac{2}{\sqrt{n}} \sum_{i=1}^n \frac{\delta}{f^{1/2}} \Big\}$$
$$+ \Big\{ E\Big(W_n - \frac{2}{\sqrt{n}} \sum_{i=1}^n \frac{\delta}{f^{1/2}} + \|\delta\|_2^2 \Big) \Big\}^2$$
$$= o(1) + (\, E(W_n) + \|\delta\|_2^2\,)^2 \qquad \text{by (c)}$$
$$\to 0 + 0 = 0 \qquad \text{by (a).}$$

Thus

(d) $$W_n - \|\delta\|_2^2 - \Big\{ \frac{2}{\sqrt{n}} \sum_{i=1}^n \frac{\delta}{f^{1/2}} - 2\|\delta\|_2^2 \Big\} = o_P(1)\,.$$

Since

$$\mathbf{L}\Big(\frac{2}{\sqrt{n}} \sum_{i=1}^n \frac{\delta}{f^{1/2}} - 2\|\delta\|_2^2 \,|\, P \Big) \to N\Big(-\frac{1}{2}\|2\,\delta\|_2^2,\ \|2\,\delta\|_2^2 \Big)$$

it follows that

$$\mathbf{L}(\, W_n \;-\; ||\delta||_2^2 \mid P\,) \;\to\; N(\,-\frac{1}{2}\,||2\,\delta||_2^2,\, ||2\,\delta||_2^2\,)$$

and hence, by lemma 2, that

$$\log L_n \;=\; W_n \;-\; ||\delta||_2^2 \;+\; o_P(1) \;=\; \frac{2}{n^{1/2}} \sum_{i=1}^n \frac{\delta}{f^{1/2}}(X_i) \;-\; 2\,||\delta||_2^2 \;+\; o_P(1)\,.$$

□

Proof of lemma 3.14. (Le Cam's 3rd lemma) Since $\mathbf{L}(\,\log L_n \mid P_n\,) \to \mathbf{L}(\,\log L\,) \;=\; N(\,-\frac{1}{2}\,\sigma^2\,,\sigma^2\,) \;=\; \mathbf{L}(\,\sigma Z \;-\; \sigma^2/2\,)$ where $Z \sim N(0,1)$, it follows that $E(L) \;=\; E\exp(\sigma Z \;-\; \sigma^2/2) \;=\; 1$ and thus $\{\,Q_n\,\} \lhd \{\,P_n\,\}$ by Le Cam's first lemma. Hence by lemma 3.9 (Le Cam's 4th lemma), L_n is uniformly integrable and $Q_n(p_n \;=\; 0\,) \to 0$ as $n \to \infty$.

Now let $f : R^2 \to R$ be bounded and continuous. Then

$$E_{Q_n} f(T_n, \log L_n\,)$$

$$= \; E_{Q_n} f(T_n, \log L_n)\{1_{[p_n\,>\,0]} \;+\; 1_{[p_n\,=\,0]}\,\}$$

(a) $$= \; E_{P_n} f(T_n, \log L_n\,)\, L_n \;+\; E_{Q_n} f(T_n, \log L_n\,)\, 1_{[p_n\,=\,0]}$$

(b) $$\to \; E\,[f(T, \log L)\, L]$$

(c) $$= \; E\, f(T \,+\, c, \log L + \sigma^2)$$

where (b) holds since $f(T_n, \log L_n)\, L_n$ is uniformly integrable by uniform integrability of L_n and boundedness of f, and since the second term in (a) is bounded by $||f||_\infty\, Q_n(p_n \;=\; 0) \to 0$. It remains only to establish (c).

To verify (c), note that (23) implies that

(d) $$\mathbf{L}(\,T \mid \log L\,) \;=\; \mathbf{L}(\frac{c}{\sigma^2}\,(\,\log L \;+\; \frac{1}{2}\,\sigma^2) \;+\; \tilde{Z}\,)$$

where $\mathbf{L}(\tilde{Z}) \;=\; N(\mu\,, \sigma^2\,(1 - \rho^2))\,,\; \rho \;=\; c/(\sigma\,\tau)\,,$ is independent of $\log L$, and hence

(e) $$\mathbf{L}(\,T \;+\; c \mid \log L\,) \;=\; \mathbf{L}(\,c \;+\; \frac{c}{\sigma^2}\,(\,\log L \;+\; \frac{1}{2}\,\sigma^2\,) \;+\; \tilde{Z}\,)$$

$$= \; \mathbf{L}(\frac{c}{\sigma^2}\,(\,\log L \;+\; \sigma^2 \;+\; \frac{1}{2}\,\sigma^2\,) \;+\; \tilde{Z}\,)\,.$$

Furthermore

(f) $$L \;=\; exp(\,\log L\,) \;=\; \frac{\text{density of } \; N(+\frac{1}{2}\,\sigma^2, \sigma^2)}{\text{density of } \; N(-\frac{1}{2}\,\sigma^2, \sigma^2)} \quad \text{at} \quad \log L\,.$$

Therefore,

$$
\begin{aligned}
E\, f(T, \log L)\, L \;=\;& E\, E(f(T, \log L)\, e^{\log L} \,|\, \log L\,) \\
=\;& E\big\{ e^{\log L}\, E(f(T, \log L)) \,|\, \log L\,)\big\} \\
=\;& E\Big\{ e^{\log L}\, E(f(\frac{c}{\sigma^2}\,(\,\log L \;+\; \sigma^2/2) \;+\; \tilde{Z}\,, \log L\,) \,|\, \log L\,)\Big\} \quad \text{by (d)} \\
=\;& E\Big\{\, E(f(\frac{c}{\sigma^2}\,(\,\log L \;+\; \sigma^2 \;+\; \sigma^2/2) \;+\; \tilde{Z}\,, \log L \;+\; \sigma^2\,) \,|\, \log L\,)\Big\} \\
& \qquad \text{by (f)} \\
=\;& E\Big\{\, E(f(c \;+\; \frac{c}{\sigma^2}\,(\,\log L \;+\; \sigma^2/2) \;+\; \tilde{Z}\,, \log L \;+\; \sigma^2\,) \,|\, \log L\,)\Big\} \\
=\;& E\, f(T \;+\; c, \log L \;+\; \sigma^2\,) \qquad \text{by (e)}\,,
\end{aligned}
$$

which completes the proof of (c). Hence (24) holds. □

3.4

4. Hájek's convolution theorem and Local Asymptotic Minimax theorem

Now we give statements of several convolution and asymptotic minimax theorems. The key hypotheses involved in virtually all the different formulations of these theorems are:

A. *Local asymptotic normality* (LAN) of the local likelihood ratios of the model. A sufficient condition for this is differentiability of the model in an appropriate sense.

B. For the convolution theorems we will also hypothesize *regularity* of the estimators: the only estimators considered will be those for which the local limiting distributions do not depend on the direction (or magnitude) of the approach of the local parameter point to the fixed parameter point under consideration.

C. *Pathwise differentiability* of the parameter being estimated as a function of the underlying $P \in \mathbf{P}$ with $\mathbf{P}$ metrized by the Hellinger metric. This amounts to Hadamard differentiability along the model $\mathbf{P}$.

Our goal in this chapter will be to explain the basic hypotheses required in different settings, and to discuss several useful refinements and extensions of the basic theorems.

For complete proofs we refer the reader to the primary sources: Hajek (1970, 1972), Le Cam (1972), Ibragimov and Has'minskii (1981), van der Vaart (1988), Bickel, Klaassen, Ritov, and Wellner (1993), van der Vaart and Wellner (1991), Millar (1983), and Le Cam (1986).

Convolution and LAM theorems for finite - dimensional parameter spaces

Suppose that $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$, $\Theta \subset R^k$ is a Hellinger - differentiable parametric model. With $l(x;\theta) \equiv \log p(x;\theta)$, let

$$L_n(\theta) = \sum_{i=1}^n l(X_i;\theta)$$

denote the log - likelihood of $X_1, \cdots, X_n$, a sample from $P_{\theta_0} \equiv P_0 \in \mathbf{P}$. Then it is well - known that

$$L_n(\theta_0 + \frac{t}{\sqrt{n}}) - L_n(\theta_0) = S_n(\theta_0)t^T - \frac{1}{2}\, t\, I(\theta_0)t^T + o_{P_0}(1) \tag{1}$$

where

$$S_n(\theta_0) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\mathbf{l}}(X_i;\theta_0)$$

is the score for θ at θ_0 (based on the entire sample $X_1, \cdots, X_n$) and $I(\theta_0)$ is the Fisher information matrix defined in section 2. It follows that

$$(2) \qquad L_n(\theta_0 + \frac{t}{\sqrt{n}}) - L_n(\theta_0) \to_d N(-\frac{1}{2} tI(\theta_0)t^T, tI(\theta_0)t^T)$$

under P_0. This is sometimes called the *Local Asymptotic Normality*, or LAN, condition. It is one key ingredient of the Hájek convolution theorem. The second key ingredient is the following definition of regularity of an estimator sequence T_n :

Definition 4.1. $T = \{T_n\}$ is a *locally regular estimator* of θ at $\theta = \theta_0$ if, for every sequence $\{\theta_n\} \subset \Theta$ with $\sqrt{n}(\theta_n - \theta_0) \to t \in R^k$, under P_{θ_n}

$$\sqrt{n}(T_n - \theta_n) \to_d \mathbb{Z} \quad \text{as} \quad n \to \infty$$

where the law of $\mathbb{Z}$ depends on θ_0 but not on t. Thus the limit distribution of $\sqrt{n}(T_n - \theta_n)$ does not depend on the direction of approach t of θ_n to θ_0.

With these two basic ingredients, we can state a simplified version of Hájek's (1970) convolution theorem:

Theorem 4.2. (Hájek). Suppose that (2) holds with $I(\theta_0)$ nonsingular and that $\{T_n\}$ is a regular estimator of θ at θ_0. Then

$$(3) \qquad \mathbb{Z} \stackrel{d}{=} \mathbb{Z}_0 + \Delta_0$$

where $\mathbb{Z}_0 \sim N(0, I^{-1}(\theta_0))$ is independent of Δ_0.

Hájek (1970) proved a somewhat more general theorem based on just the LAN hypothesis (2) using a method based on "Bayesian considerations." A different proof using characteristic functions due to Peter Bickel may be found in Roussas (1972), and also in Bickel, Klaassen, Ritov, and Wellner (1993). This latter type of proof was exploited and developed by R. Beran (1977a, 1977b) in more general settings.

In words, theorem 4.2 says that the limiting distribution of any regular estimator T_n of θ must be at least as "spread out" as the $N(0, I^{-1}(\theta_0))$ distribution of $\mathbb{Z}_0$. Thus an *efficient estimator* is a regular estimator for which the limiting distribution is exactly equal to $\mathbb{Z}_0$. Another way to say this is in terms of the following asymptotic optimality theorem:

Corollary 4.3. (Hájek, 1970). Suppose that $\{T_n\}$ is a locally regular estimator of θ at θ_0 and that $l : R^k \to R^+$ is bowl - shaped: i.e.

(i) $l(x) = l(-x)$.

(ii) $\{x : l(x) \le c\}$ is convex for every $c \ge 0$.

Then

$$\liminf_{n \to \infty} E_{\theta_0} l(\sqrt{n}(T_n - \theta_0)) \geq E\, l(ZZ_0) \tag{4}$$

where $ZZ_0 \sim N(0, I^{-1}(\theta_0))$.

If a supremum over θ in a local neighborhood of θ_0 is added to the left side of (2.2.4), then the same type of statement holds for an *arbitrary* (not necessarily regular) estimator T_n of θ . This is the Hájek - Le Cam asymptotic minimax theorem due to Hájek (1971), and, in a more abstract form, to Le Cam (1971).

Theorem 4.4. (Hájek, 1971). Suppose that (2) holds, that T_n is any estimator of θ , and that l is bowl - shaped. Then

$$\lim_{b \to \infty} \liminf_{n \to \infty} \sup_{\{\theta :\, \sqrt{n}|\theta - \theta_0| \leq b\}} E_\theta l(\sqrt{n}(T_n - \theta)) \geq El(ZZ_0) \tag{5}$$

More generally, suppose that $\{P_\theta^n : \theta \in \Theta ,\ n = 1, 2, \cdots \}$ is a family of distributions indexed by $\Theta \subset R^k$ (not necessarily the distributions of independent observations) satisfying

$$\frac{dP^n_{\theta_0 + u\psi(n)}}{dP^n_{\theta_0}} = \exp\{S_n u^T - \frac{1}{2}|u|^2 + R_n(\theta_0, u)\} \tag{6}$$

where

$$S_n \to_d Z \sim N_k(0, J) \qquad \text{under } P^n_{\theta_0} , \tag{7}$$

J denotes the $k \times k$ identity matrix, $\psi(n) = \psi(n, \theta_0)$ is a $k \times k$ matrix, and

$$R_n(\theta_0, u) \to_p 0 \qquad \text{under } P^n_{\theta_0} . \tag{8}$$

Then, the asymptotic minimax theorem holds in the following strengthened form:

Theorem 4.5. (Ibragimov and Has'minskii, 1981). Suppose that LAN in the form (6) - (8) holds and that l is bowl - shaped. Then for any estimator T_n of θ ,

$$\begin{aligned} &\liminf_{n \to \infty} \sup_{\{\theta :\, \psi^{-1}(n)(\theta - \theta_0) \in K_b\}} E_\theta l(\psi^{-1}(n)(T_n - \theta)) \\ &\qquad \geq \begin{cases} (1 - b^{-1/2})^k\, El(Z) 1_{[Z \in K_{\sqrt{b}}]} \\ \dfrac{1}{2^k} El(Z) 1_{[Z \in K_{b/2}]} \end{cases} \end{aligned} \tag{9}$$

where $Z \sim N_k(0, J)$ and K_b is the cube in R^k with vertices having all coordinates $\pm b$.

Proof. See Ibragimov and Has'minskii (1981), equations (12.18) and (12.19), page 168 and 169. □