2 | Asymptotic Inference for (Finite-Dimensional) Parametric Models

2.1 REGULAR PARAMETRIC MODELS IN THE I.I.D. CASE

We shall review some of the basic results in asymptotic inference, particularly estimation, for regular parametric models. The statements and conditions are essentially those of Le Cam (1956), (1969), (1970), and Hájek (1970), (1972), but the basic heuristic goes back to Fisher (1922), (1925).

Let μ be a fixed σ-finite measure on $(\mathbf{X}, \mathcal{B})$, and let $\mathbf{M}_\mu$ be all probability measures P on $(\mathbf{X}, \mathcal{B})$ dominated by μ; i.e., $\mathbf{M}_\mu = \{P \in \mathbf{M} : P << \mu\}$. Then, as usual, suppose that $X_1, \ldots, X_n$ are i.i.d. with common distribution $P \in \mathbf{P}$, where $\mathbf{P}$ is dominated by μ. Recall that we loosely defined $\mathbf{P}$ to be parametric or finite-dimensional if we could write

$$\mathbf{P} = \{P_\theta : \theta \in \Theta\},$$

where

(i) Θ is a "nice" subset of R^k.

(ii) The parametrization $\theta \to P_\theta$ is "smooth."

Let

$$p(\theta) = p(\cdot, \theta) = \frac{dP_\theta}{d\mu}(\cdot), \qquad l(\theta) = \log p(\theta), \tag{1}$$

be the *density* and *log-likelihood* of P_θ respectively.

Convention. If $h(\theta)$ is a function on $\mathbf{X}$ for fixed θ, then $h(x, \theta)$ denotes its value at x.

The facts according to Fisher (for $k = 1$) are:

(iii) If some estimate T_n satisfies $\mathbf{L}_\theta(\sqrt{n}(T_n - \theta)) \to N(0, \sigma^2(\theta))$ as $n \to \infty$ for all θ, then

$$\sigma^2(\theta) \geq I^{-1}(\theta) \qquad \text{where } I(\theta) = E_\theta\left(\frac{\partial}{\partial\theta} l(\theta)\right)^2.$$

(iv) If θ is identifiable, the maximum likelihood estimate $\hat{\theta}$ solving

$$\sum_{i=1}^{n} \frac{\partial}{\partial\theta} l(X_i, \hat{\theta}) = 0$$

is efficient. That is,

$$\mathbf{L}_\theta\left(\sqrt{n}(\hat{\theta} - \theta)\right) \to N(0, I^{-1}(\theta)) .$$

Of course, these "facts" are not quite right. Some uniformity in the convergence in law in (iii) has to be required to avoid "superefficiency," and some modification of maximum likelihood is often needed to obtain efficiency. Note that the inequality in (iii) can be viewed as an asymptotic version of the well-known Cramér-Rao (information) inequality for unbiased estimation of θ. Fisher (1925, section 7, pages 710, 711) gave a heuristic argument for (iii) and (iv) apparently based on multinomial distributions. A rephrasing of that argument appears in Fisher (1956, section VI.3, pages 147–150); see Savage (1976) for further discussion.

To obtain the elegant statement of correct versions of (iii) and (iv) that we give in sections 3 and 5, it is convenient to view $\mathbf{P}$ as a subset of $L_2(\mu)$ via the embedding $P \to s$, where

$$p \equiv \frac{dP}{d\mu}, \qquad s \equiv \sqrt{p}, \qquad p(\theta) \to s(\theta) .$$

Both this embedding and that of $\mathbf{P}$ into $L_1(\mu)$ via $P \to p$ endow $\mathbf{P}$ with the same topology, convergence in total variation; see (A.6.3). One advantage of this embedding is that we can replace awkward conditions on pointwise differentiability of $\mathbf{l}$ and integrability assumptions by the natural condition of Fréchet (Hellinger) differentiability of the map $\theta \to s(\theta)$. More significantly it enables us (in section 2.4) to give a geometric formulation of the theory, which then extends fairly readily to the semiparametric case. In the following, elements of R^k are written as column vectors, $|\cdot|$ is the Euclidean norm, and $\|\cdot\|$ is the Hilbert norm in $L_2(\mu)$:

$$\|f\|^2 \equiv \int f^2 \, d\mu .$$

Definition 1. θ_0 is a *regular point* of the parametrization $\theta \to P_\theta$ if θ_0 is an interior point of Θ, and

(i) The map $\theta \to s(\theta)$ from Θ to $L_2(\mu)$ is Fréchet differentiable at θ_0: there exists a vector $\dot{s}(\theta_0) \equiv (\dot{s}_1(\theta_0), \ldots, \dot{s}_k(\theta_0))^T$ of elements of $L_2(\mu)$ such that

$$\|s(\theta_0 + h) - s(\theta_0) - \dot{s}^T(\theta_0)h\| = o(|h|) \qquad \text{as } h \to 0 . \tag{2}$$

(ii) The $k \times k$ matrix $\int \dot{s}(\theta_0)\dot{s}^T(\theta_0)\, d\mu$ is nonsingular.

This is exactly as in example A.5.2; see section A.5 for more on Fréchet derivatives.

Definition 2. A parametrization $\theta \to P_\theta$ is *regular* if:

(i) Every point of Θ is regular.

(ii) The map $\theta \to \dot{s}_i(\theta)$ is continuous from Θ to $L_2(\mu)$ for $i = 1, \ldots, k$.

Note that (i) implies that Θ is open. We also note that by proposition A.5.1 Fréchet differentiability of $\theta \to s(\theta)$ implies continuity of this map, and hence, by (A.6.3), the continuity of the more familiar map $\theta \to p(\theta)$ from Θ to $L_1(\mu)$. Define the *Fisher information matrix* of θ by

$$I(\theta) = 4\int \dot{s}(\theta)\dot{s}^T(\theta)\, d\mu \,. \tag{3}$$

We call $\mathbf{P}$ a *regular parametric model* if it has a regular parametrization. In such models the "niceness" of Θ is evident and the "smoothness" of the parametrization is made precise. They are the objects we shall study.

Define the *score function* $\dot{\mathbf{l}}$ of an observation by

$$\dot{\mathbf{l}}(\theta) = 2\frac{\dot{s}(\theta)}{s(\theta)} 1_{[s(\theta)>0]} = \frac{\dot{p}(\theta)}{p(\theta)} 1_{[p(\theta)>0]} \,, \tag{4}$$

where

$$\dot{p}(\theta) = 2s(\theta)\dot{s}(\theta) \,. \tag{5}$$

If θ is a regular point, then $|\dot{\mathbf{l}}(\theta)| \in L_2(P_\theta)$, and the more usual definition of the Fisher information matrix for θ is given by

$$I(\theta) = \int \dot{\mathbf{l}}(\theta)\dot{\mathbf{l}}^T(\theta)\, dP_\theta \,. \tag{6}$$

By proposition A.5.3.F the two definitions of the Fisher information matrix given in (3) and (6) agree.

The following proposition gives sufficient conditions for regularity of a parametric model in terms of ordinary differentiability of the likelihood.

Proposition 1. Suppose Θ is open and for all θ:

(i) $p(x,\theta)$ is continuously differentiable in θ for (μ) almost all x with gradient $\dot{p}(\theta)$.

(ii) $|\dot{\mathbf{l}}(\theta)| \in L_2(P_\theta)$ with $\dot{\mathbf{l}}(\theta)$ as in (4).

(iii) $I(\theta)$ defined in (6) is nonsingular and continuous in θ.

Then, if we define

$$\dot{s}(\theta) = \tfrac{1}{2}p^{-1/2}(\theta)\dot{p}(\theta)1_{[p(\theta)>0]} = \tfrac{1}{2}s(\theta)\dot{\mathbf{l}}(\theta)1_{[p(\theta)>0]} \,, \tag{7}$$

the parametrization $\theta \to P_\theta$ is regular with $\dot{s}(\theta)$ from (7) as Fréchet derivative of $s(\theta)$.

Proof. Note that $\dot{p}(\theta)$ vanishes (μ) almost everywhere outside $A(\theta) = [p(\theta) > 0]$ because of (i), and that, hence, the definition of $\dot{p}(\theta)$ in (5) is consistent with (7). It can be verified by (i) that for (μ) almost all x,

$$s(x,\theta+h) - s(x,\theta) = \int_0^1 \frac{1}{2}p^{-1/2}(x,\theta+\lambda h)h^T\dot{p}(x,\theta+\lambda h)\, d\lambda \,, \tag{a}$$

and hence

(b) $$(s(x, \theta + h) - s(x, \theta))\, 1_{A(\theta)}(x) = \frac{1}{2} p^{-1/2}(x, \theta) h^T \dot{p}(x, \theta)\, 1_{A(\theta)}(x) + o(|h|)$$

holds. By (a), (ii), and (iii) it follows that

$$\int |s(x, \theta + h) - s(x, \theta)|^2 \, d\mu(x)$$

$$\leq \frac{1}{4} \int_0^1 h^T \int \dot{p}(x, \theta + \lambda h) \dot{p}^T(x, \theta + \lambda h) p^{-1}(x, \theta + \lambda h) \, d\mu(x)\, h \, d\lambda$$

(c) $$= \frac{1}{4} h^T I(\theta) h + o(|h|^2)$$

$$= \frac{1}{4} h^T \int_{A(\theta)} \dot{p}(\theta) \dot{p}^T(\theta) p^{-1}(\theta) \, d\mu(x)\, h + o(|h|^2).$$

Assume without loss of generality that $h/|h|$ converges. Using (b) and (c) and applying lemma A.7.5 to $|h|^{-1}(s(\theta + h) - s(\theta))\, 1_{A(\theta)}$, we obtain

(d) $$\int_{A(\theta)} |s(x, \theta + h) - s(x, \theta) - \frac{1}{2} s^{-1}(x, \theta) h^T \dot{p}(x, \theta)|^2 \, d\mu(x) = o(|h|^2),$$

which combined with (c) yields

(e) $$\int_{\mathbb{X} - A(\theta)} |s(x, \theta + h) - s(x, \theta)|^2 \, d\mu(x) = o(|h|^2).$$

This implies the Fréchet differentiability of s at θ with derivative (7). This yields, in view of (iii), the nonsingularity of $I(\theta)$. Hence every θ is regular. From (i) and (iii)

$$\lim_{h \to 0} \dot{s}_i(x, \theta + h) = \dot{s}_i(x, \theta), \qquad (\mu) \text{ a.e. } x \in A(\theta),$$

and

$$\limsup_{h \to 0} \int_{A(\theta)} \dot{s}_i^2(x, \theta + h) \, d\mu(x)$$

$$\leq \limsup_{h \to 0} \int \dot{s}_i^2(x, \theta + h) \, d\mu(x)$$

$$= \int \dot{s}_i^2(x, \theta) \, d\mu(x) = \int_{A(\theta)} \dot{s}_i^2(x, \theta) \, d\mu(x)$$

follow. Continuity of $\theta \to \dot{s}(\theta)$ is obtained from this by the same argument as before. □

Example 1. Exponential family.

Suppose Θ is open and $\{P_\theta\}$ is a curved exponential family, $p(x, \theta) = \exp(c(\theta) T(x) - d(\theta))$. If c has a differential and $c(\Theta)$ is contained in the interior of the natural parameter space of the exponential family, then equation (2) applies. □

Example 2. Translation model with known shape f.

Suppose P_θ is a one-dimensional translation parameter family $p(x,\theta) = f(x-\theta)$, θ real. Then hypothesis (i) of proposition 1 is equivalent to continuity of the derivative f', ruling out the double exponential for instance. However, regularity of the model holds if and only if f is absolutely continuous with Radon-Nikodym derivative f' and $\int [((f')^2/f)(x)]\,dx < \infty$ (Hájek and Šidák (1967, page 211), or corollary A.5.1), and hence the double exponential is regular. This example shows that only the requirement of *continuous differentiability* in (i) of proposition 1 is too strong. □

Example 3. Weibull translation model.

Suppose that $\theta = (\alpha,\beta,\gamma) \in \Theta = \{\theta : \alpha > 0,\ \beta > 0,\ \gamma \in R\}$ and

$$p(x,\theta) = \frac{\beta}{\alpha}\left(\frac{x-\gamma}{\alpha}\right)^{\beta-1} \exp\left(-\left(\frac{x-\gamma}{\alpha}\right)^{\beta}\right) 1_{(\gamma,\infty)}(x),$$

the Weibull translation model. This model is *not* regular, but the restricted model $\mathbf{P} = \{P_\theta : \theta \in \Theta_0\} \subset \Theta$, is regular. □

Example 4. Three-parameter lognormal model.

Suppose that $Y \sim N(\mu,\sigma^2)$ and $X \equiv \gamma + \exp(Y)$. This model, with $\theta = (\mu,\sigma^2,\gamma) \in \Theta = R \times R^+ \times R$,

$$p(x,\theta) = \frac{1}{\sigma(x-\gamma)}\phi\left(\frac{\log(x-\gamma)-\mu}{\sigma}\right) 1_{(\gamma,\infty)}(x)$$

where ϕ is the standard normal density function, and $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$, is a lognormal translation family. It is a regular model, but yields unbounded likelihood functions, and hence the method of maximum likelihood fails; see, e.g., Hill (1963), Cohen and Whitten (1980), and Griffiths (1980) for solutions. In spite of this, the methods to be discussed in section 2.5 based on scores yield efficient estimates. □

The first equality in (4) suggests another useful local embedding of $\mathbf{P}$ into $L_2(P_{\theta_0})$ (for θ_0 fixed) given by

$$P_\theta \longleftrightarrow r(\theta) \equiv 2\left(\frac{s(\theta)}{s(\theta_0)} - 1\right) 1_{\{s(\theta_0)>0\}}.$$

Regularity at θ_0 is equivalent to (see proposition A.5.3.E and F):

(ia) $\theta \to r(\theta)$ is Fréchet differentiable at θ_0 (in $L_2(P_{\theta_0})$) with derivative $\dot{\mathbf{l}}(\theta_0)$.

(ib) $P_{\theta_0+h}(s(\theta_0) = 0) = o(|h|^2)$.

(ii) $I(\theta_0) = \int \dot{\mathbf{l}}(\theta_0)\dot{\mathbf{l}}^T(\theta_0)\,dP_{\theta_0}$ is nonsingular.

The function $r(\theta)$, which belongs to $L_2(P_{\theta_0})$, is a useful proxy for $\mathbf{l}(\theta)$, which may not belong to $L_2(P_{\theta_0})$.

Regularity of θ is enough to guarantee a score function identity which is basic to the Cramér-Rao information bound calculation.

$$\int \dot{\mathbf{l}}(\theta)\,dP_\theta = 0 \tag{8}$$

or, equivalently, by (4) and (5),

$$(9) \qquad \langle \dot{s}_i(\theta), s(\theta) \rangle = 0 \qquad \text{for } i = 1, \ldots, k ,$$

where $\langle f, g \rangle \equiv \int fg \, d\mu$ is the inner product in $L_2(\mu)$.

Proof of (9). By hypothesis

$$(a) \qquad \langle s(\theta), s(\theta) \rangle = 1 \qquad \text{for all} \quad \theta .$$

Fréchet-differentiating with respect to θ_i yields $\langle \dot{s}_i(\theta), s(\theta) \rangle = 0$. □

Further,

$$(10) \qquad \dot{s}(\theta) = \dot{s}(\theta) 1_{[s(\theta) > 0]} \quad \text{a.e. } \mu .$$

This follows since

$$\begin{aligned} &\int (s(\theta + h) - s(\theta) - \dot{s}^T(\theta) h \, 1_{[s(\theta) > 0]})^2 \, d\mu \\ &\quad = \int (s(\theta + h) - s(\theta) - \dot{s}^T(\theta) h)^2 \, 1_{[s(\theta) > 0]} \, d\mu \\ &\quad + \int s^2(\theta + h) \, 1_{[s(\theta) = 0]} \, d\mu \\ &\quad = o(|h|^2) \end{aligned}$$

by (2) and (ib).

The fundamental consequence of regularity at a point θ is the *local asymptotic normality* (LAN) of the model given by the following basic proposition. Define the log-likelihood of $(X_1, \ldots, X_n)$ by

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

and the score function by

$$(11) \qquad S_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{l}(X_i, \theta) .$$

Proposition 2. Suppose that $\mathbf{P} = \{ P_\theta : \theta \in \Theta \}$ is a regular parametric model, and write

$$(12) \qquad L_n(\theta + \frac{t}{\sqrt{n}}) - L_n(\theta) = t^T S_n(\theta) - \frac{1}{2} t^T I(\theta) t + R_n(\theta, t) .$$

Then $R_n(\theta, t) \to 0$ in P_θ probability uniformly for $\theta \in K$ compact $\subset \Theta$ and $|t| \leq M$; i.e., for any compact set $K \subset \Theta$, $0 < M < \infty$, and $\varepsilon > 0$,

$$(13) \qquad \sup_{|t| \leq M} \sup_{\theta \in K} P_\theta(|R_n(\theta, t)| > \varepsilon) \to 0 \qquad \text{as} \quad n \to \infty .$$

Moreover,

$$(14) \qquad \mathbf{L}_\theta(S_n(\theta)) \to N(0, I(\theta))$$

uniformly in $\theta \in K$ for compact $K \subset \Theta$, where $N(\mu, \Sigma)$ is the multivariate normal distribution with mean μ and covariance matrix Σ. Finally, uniformly in

$\theta \in K$, $|t| \leq M$,

$$(15) \qquad S_n(\theta + \frac{t}{\sqrt{n}}) - S_n(\theta) + I(\theta)t \to_{P_\theta} 0 \quad \text{as} \quad n \to \infty .$$

Remark. Note that the left side of (12) is a well-defined extended real-valued random variable under P_θ. For a careful definition of the uniform convergence in (14) see definition 2.2.3.

Proof. The proof of this uniform version of the LAN property is given in appendix 9 along with other contiguity theory facts and proofs. The proofs of (13) and (14) are very similar to the proof of Le Cam's second lemma, and are based on the proof given by Ibragimov and Has'minskii (1981, theorem II.1.2, page 119); equation (15) has been noted in (6.43) of Bickel (1982). Under stronger conditions this proposition follows easily by Taylor expansion of L_n to two terms. See, e.g., Lehmann (1983, proof of theorem 6.2.3, page 415). □

An important property that follows from this basic proposition is contiguity of the product measures corresponding to θ and $\theta + t/\sqrt{n}$ at regular θ. We recall the definition and basic properties of contiguity; also see section A.9.

Definition 3. Two sequences of probability measures $\{P_n\}$, $\{Q_n\}$, each pair defined on the same space, $(\mathbf{X}_n, \mathcal{B}_n)$ are called *contiguous*, and we write $\{P_n\} \lhd\rhd \{Q_n\}$ if $P_n(A_n) \to 0$ if and only if $Q_n(A_n) \to 0$, for $A_n \in \mathcal{B}_n$.

Proposition 3. If $\theta \to P_\theta$ is a regular parametrization and $\theta_n \to \theta$, then $\{P^n_{\theta_n + t_n/\sqrt{n}}\}$ and $\{P^n_{\theta_n}\}$ are contiguous for any bounded sequence $\{t_n\}$.

Proof. $\{P^n_{\theta_n + t_n/\sqrt{n}}\} \lhd \{P^n_{\theta_n}\}$ follows from (12), (14), and corollary A.9.1 of Le Cam's first lemma. To prove $\{P^n_{\theta_n}\} \lhd \{P^n_{\theta_n + t_n/\sqrt{n}}\}$, use (12) and (14) together with Le Cam's lemma A.9.3 and corollary A.9.1. □

Proposition 2 indicates that whatever be M finite, the log-likelihood of the local model $\{P_{\theta + t/\sqrt{n}} : |t| \leq M\}$ is approximated in a weak sense, for n large, by that of the model $\{Q_t : |t| \leq M\}$, where we observe

$$S_n(\theta) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{l}(X_i, \theta)$$

distributed under Q_t as $N(I(\theta)t, I(\theta))$ as if θ is known but the local deviation t is not. This weak type of approximation has been shown to have profound consequences by Le Cam in a series of papers starting in 1956, following the lead of Wald (1943). We explore some of these consequences concentrating on point estimation.

2.2 REGULAR ESTIMATES OF EUCLIDEAN PARAMETERS

Let $\nu : \mathbf{P} \to R^m$ be a Euclidean parameter, where $\mathbf{P}$ is a general (not necessarily parametric) model. An estimate of ν is any (measurable) m-vector $T_n(X_1, \ldots, X_n)$ which depends only on $(X_1, \ldots, X_n)$. We study the limiting behavior of sequences $\{T_n(X_1, \ldots, X_n)\}_{n \geq 1}$ whose members are related in some natural way in the expectation that the limiting behavior will be an approximation to that for fixed n and P. We call such sequences estimates also.

The first property a reasonable sequence of estimates should have is *consistency*:

$$P(|T_n - \nu(P)| \geq \varepsilon) \to 0 \qquad \text{for all } P \in \mathbf{P} \text{ and for all } \varepsilon > 0.$$

In this work we focus on how close estimates can get to ν on the $n^{-1/2}$ scale. So we need to define $\sqrt{n}$-*consistency*:

$$\lim_{M \to \infty} \limsup_{n \to \infty} P(n^{1/2} |T_n - \nu(P)| \geq M) = 0,$$

or, in the familiar O_p, o_p notation,

$$T_n - \nu(P) = O_p(n^{-1/2}).$$

It is desirable to have properties such as consistency, $\sqrt{n}$-consistency, hold as uniformly in $\mathbf{P}$ as possible. Otherwise "n large" depends on the unknown P, and we cannot even in principle specify what an adequate sample size is. Unfortunately, uniformity in consistency and other properties is often unachievable on all of $\mathbf{P}$ and has to be replaced by uniformity on "small" subsets of $\mathbf{P}$.

In particular, we metrize $\mathbf{P}$ with the variational distance (A.6.1) and require uniformity on compact subsets of $\mathbf{P}$.

Definition 1. $T = \{T_n\}$ is *uniformly consistent* if, for every $\varepsilon > 0$,

$$\sup\{P(|T_n - \nu(P)| \geq \varepsilon) : P \in \mathbf{K}\} \to 0 \qquad \text{as} \qquad n \to \infty$$

for all compact subsets $\mathbf{K}$ of $\mathbf{P}$.

Definition 2. $T = \{T_n\}$ is *uniformly* $\sqrt{n}$-*consistent* if

$$\lim_{M \to \infty} \limsup_{n \to \infty} \sup\{P(n^{1/2} |T_n - \nu(P)| \geq M) : P \in \mathbf{K}\} = 0$$

for all compact subsets $\mathbf{K}$ of $\mathbf{P}$.

Here are some desirable properties which are stronger than $\sqrt{n}$-consistency, but which are often possessed by procedures used in practice.

Convention. Write E_P and $\mathbf{L}_P$ for expectations and distributions of measurable functions of $(X_1, \ldots, X_n)$ under P. In parametric models shorten E_{P_θ} and $\mathbf{L}_{P_\theta}$ to E_θ, and $\mathbf{L}_\theta$. Let $\mathbf{K}$ denote an arbitrary (pre-) compact subset of $\mathbf{P}$.

Definition 3. $T = \{T_n\}$ is a *uniformly regular* estimate of $\nu(P) \in R^m$ if $\{\mathbf{L}_P(\sqrt{n}(T_n - \nu(P)))\} \equiv \{\mathbf{L}_P(Z_n)\}$ converge uniformly on compact subsets $\mathbf{K}$ of $\mathbf{P}$. That is,

$$\sup_{P \in \mathbf{K}} |E_P g(Z_n) - E_P g(Z)| \to 0$$

for all g continuous and bounded on R^m, all $\mathbf{K}$, and a fixed family $\Lambda = \{\mathbf{L}_P(Z)\} = \{\mathbf{L}_P\}$ of probability distributions on R^m.

We usually ask for more.

Definition 4. T is a *uniformly Gaussian regular* estimate of $\nu(P)$ if it is uniformly regular and for each P the limit $\mathbf{L}_P = \mathbf{L}_P(Z)$ of $\mathbf{L}_P(Z_n) = \mathbf{L}_P(\sqrt{n}(T_n - \nu(P)))$ is m-variate Gaussian with mean 0 and covariance matrix which we denote $\Sigma(P,T)$; i.e., $Z \sim N(0, \Sigma(P,T))$.

A further structural property which aids our understanding of how T acts and connects this theory with that of robust estimation is given by the next definition.

Definition 5. T is an *asymptotically linear estimate* of ν if there exists

$$\psi : \mathbf{X} \times \mathbf{P} \to R^m$$

such that for all $P \in \mathbf{P}$

$$|\psi(\cdot, P)| \in L_2(P),$$

$$\int \psi(x, P)\, dP = 0, \tag{1}$$

$$T_n = \nu(P) + n^{-1} \sum_{i=1}^{n} \psi(X_i, P) + o_p(n^{-1/2}).$$

Then

$$\Sigma(P,T) = \int \psi\psi^T(x, P)\, dP. \tag{2}$$

We call the function $\psi(\cdot, P)$ (which is unique a.e. P) the *influence function* of T. The observation $X_i = x$, to first order, contributes $n^{-1}\psi(x, P)$ to the error $T_n - \nu(P)$, in essential agreement with the notions of Hampel (1974). If (1) holds for just $P = P_0$, then we call T asymptotically linear at P_0.

Suppose that ν is a parameter defined on all of $\mathbf{M} = \{$ all probability distributions on $\mathbf{X}\}$ satisfying the following regularity conditions:

(i) For all $P_0 \in \mathbf{M}$, ν is continuously Fréchet differentiable at P_0 with respect to d_K given by (A.6.8).

(ii) For all $P_0 \in \mathbf{P}$, the derivative $\dot{\nu}$ has the representation

$$\dot{\nu}(P_0)(P) = \int \psi(x, P_0)\, dP(x),$$

where ψ is continuous, bounded in x, continuous in P_0 with respect to the total variation metric d_v, and

$$\int \psi(x, P_0)\, dP_0(x) = 0.$$

Then, if $\mathbb{P}_n$ is the empirical distribution, the estimator $\nu(\mathbb{P}_n)$ of $\nu(P)$ is asymptotically linear with influence function ψ. Moreover, ψ is the Gâteaux derivative of ν on $\mathbf{M}$ at P:

$$\lim_{\varepsilon \to 0} \frac{\nu((1-\varepsilon)P + \varepsilon Q) - \nu(P)}{\varepsilon} = \frac{\partial}{\partial \varepsilon} \nu((1-\varepsilon)P + \varepsilon Q) \Big|_{\varepsilon = 0} = \int \psi(x, P)\, dQ. \tag{3}$$

By taking Q to be point mass at x we recapture $\psi(x, P)$, making our definition agree with that of Hampel (1974). Proofs of these and stronger results may be found for instance in Huber (1981, section 2.5).

Definition 6. T is a *uniformly asymptotically linear* estimate of $\nu(P)$ if (1) holds uniformly on compact subsets of $\mathbf{P}$; i.e. the $o_p(n^{-1/2})$ in (1) holds uniformly in $P \in \mathbf{K}$.

Existence of estimates of ν possessing these properties places strong restrictions on ν and the limit laws of the estimates, as we see in the following proposition. Metrize the set Λ of probability measures on R^m by the Prohorov metric d_{pr} given in (A.6.10) so that the topology is that of convergence in law, but retain the variational distance topology on $\mathbf{P}$.

Proposition 1. Suppose that T is an estimate of ν.

A. If T is uniformly consistent (on compact subsets of $\mathbf{P}$), then the map $P \to \nu(P)$ is continuous from $\mathbf{P}$ to R^m.

B. If T is uniformly regular with limit laws $\{\mathbf{L}_P\} \subset \Lambda$, then the map $P \to \mathbf{L}_P$ is continuous from $\mathbf{P}$ to Λ.

C. If T is uniformly Gaussian regular, then the map $P \to \Sigma(P,T)$ from $\mathbf{P}$ to the set of nonnegative definite symmetric $m \times m$ matrices is continuous.

D. If T is uniformly asymptotically linear, and $P \to \Sigma(P,T)$ (given by (2)) is continuous, then T is uniformly Gaussian regular with covariance matrix $\Sigma(P, T)$.

Proof. Without loss of generality, take $\mathbf{P}$ to be compact.

A. Define maps g_n, g from $\mathbf{P}$ to Λ by

$$g_n(P) = \mathbf{L}_P(T_n),$$

$$g(P) = \delta_{\nu(P)} \equiv \text{point mass at } \nu(P).$$

Claim A is equivalent to continuity of g. The maps g_n are continuous for each n, since by (A.6.3) and (A.6.5) $d_\nu(P_m, P) \to 0$ implies $d_\nu(P_m^n, P^n) \to 0$ for the product measures P_m^n, P^n. Therefore A follows from

(a) $$\sup\{d_{pr}(g_n(P), g(P)) : P \in \mathbf{P}\} \to 0.$$

But by a theorem of Strassen (1965), if $P(|T_n - \nu(P)| \geq \varepsilon) \leq \varepsilon$, then $d_{pr}(g_n(P), g(P)) \leq \varepsilon$; see Theorem A.6.1. Since

$$\sup\{P(|T_n - \nu(P)| \geq \varepsilon) : P \in \mathbf{P}\} \to 0,$$

(a) and claim A follow.

B. Define

$$h_n(P) = \mathbf{L}_P(\sqrt{n}(T_n - \nu(P))),$$

$$h(P) = \mathbf{L}_P.$$

Since uniform regularity implies uniform consistency (on compacts), g is continuous and hence so is h_n. Since regularity corresponds to uniform convergence of h_n to h, claim B also follows.

C. This follows from B since continuity of h is equivalent to continuity of $P \to \Sigma(P,T)$.

D. Since the uniform asymptotic linearity of T and the boundedness of the eigenvalues of $\Sigma(P,T)$ imply uniform consistency of T, g is continuous. We will verify the continuity of $P \to \mathbf{L}_P(\psi(X_1, P))$ on $\mathbf{P}$ with the total variation distance. Let $\{P_m\}$ and P satisfy $d_v(P_m, P) \to 0$. It suffices to show that there exists a subsequence $\{P_n\}$, say, with $\mathbf{L}_{P_n}(\psi(X_1, P_n)) \to \mathbf{L}_P(\psi(X_1, P))$. Choose $\{P_n\}$ such that $nd_v(P_n, P) \to 0$. Then contiguity and uniform asymptotic linearity imply that

$$n^{-1/2} \sum_{i=1}^{n} (\psi(X_i, P_n) + v(P_n) - \psi(X_i, P) - v(P)) = o_p(1),$$

and by lemma A.7.6 this yields

$$\mathbf{L}_P(\psi(X_1, P_n) + v(P_n)) \to \mathbf{L}_P(\psi(X_1, P) + v(P)).$$

Since $d_v(P_n, P) \to 0$, and g is continuous, $v(P_n) \to v(P)$, and hence $\mathbf{L}_{P_n}(\psi(X_1, P_n)) \to \mathbf{L}_P(\psi(X_1, P))$. The continuity of $P \to \mathbf{L}_P(\psi(X_1, P))$ follows. This continuity and that of $P \to \Sigma(P, T)$ imply that the maps

$$P \to \mathbf{L}_P\left(\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi(X_i, P)\right)$$

converge uniformly to $P \to N(0, \Sigma(P, T))$. This can be verified by checking the Lindeberg condition; see (A.7.5). The result follows. □

The notions of uniform regularity and uniform Gaussian regularity we have introduced are suitable for parametric models but typically are somewhat too strong for the natural nonparametric and semiparametric models we are interested in. The following notion of local regularity at a point of a parametric model $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$ is a useful stepping stone to a suitable definition of regular estimates in nonparametric models. It is due to Hájek (1970), (1972) and implicit in the work of Le Cam.

Definition 7. $T = \{T_n\}$ is *locally regular* at P_{θ_0} if, whenever $\sqrt{n}\,|\theta_n - \theta_0|$ stays bounded,

$$\mathbf{L}_{\theta_n}(\sqrt{n}(T_n - v(P_{\theta_n}))) \to \mathbf{L}_0,$$

where $\mathbf{L}_0$ does not depend on $\{\theta_n\}$. If $\mathbf{L}_0$ is Gaussian, we say that T_n is *locally Gaussian regular*, while *local asymptotic linearity* corresponds to uniformity of the expansion (1) for θ depending on n with $\sqrt{n}\,|\theta - \theta_0|$ bounded.

Necessary and sufficient conditions for an asymptotically linear estimator to be locally regular will be given in section 2.4.

Convention. We will frequently abbreviate *locally regular* to just *regular*, in keeping with most of the recent literature.

Of course, there are estimators which are not uniformly regular or even locally regular. The most well known of these is the "superefficient" estimator of Hodges; see, e.g., Lehmann (1983, pages 405, 407–408). Here is another example.

Example 1. Stein's estimator of a normal mean.

Suppose that $X_1, \ldots, X_n$ are i.i.d. $N_k(\theta, I)$ in R^k, $k \geq 3$. Consider the estimator

$$\hat{\theta}_n \equiv \left(1 - \frac{k-2}{n|\overline{X}|^2}\right)\overline{X}$$

of $\theta \in R^k$. If $\theta_0 = 0$ and $\theta_n \equiv tn^{-1/2}$ with $t \in R^k$, then, since $\mathbf{L}_{\theta_n}(\sqrt{n}(\overline{X} - \theta_n)) = \mathbf{L}_0(X_1) = N_k(0, I)$ under P_{θ_n},

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n - \theta_n) &= \sqrt{n}(\overline{X} - \theta_n) \\ &\quad - \frac{k-2}{|\sqrt{n}(\overline{X} - \theta_n) + t|^2}\{\sqrt{n}(\overline{X} - \theta_n) + t\} \\ &=_d X_1 - \frac{k-2}{|X_1 + t|^2}(X_1 + t)\end{aligned}$$

for every $n \geq 1$, the distribution of which is dependent on t. Hence $\hat{\theta}_n$ is not locally regular at $\theta_0 = 0$. □

There also exist estimators which are (uniformly or locally) regular, but are *not* (uniformly or locally) Gaussian regular. Here are two examples.

Example 2. Minimum Kolmogorov distance estimator of center of symmetry.

Suppose that $X_1, \ldots, X_n$ are i.i.d. with distribution function $F_\theta(x) = F_0(x - \theta)$ where F_0 is continuous and symmetric about 0: $1 - F_0(x) = F_0(-x)$ for all x. Thus $X_1 - \theta, \ldots, X_n - \theta$ are i.i.d. F_0 and hence both

$$\frac{1}{n}\sum_{i=1}^{n} 1_{[X_i - \theta \leq x]} = \boldsymbol{F}_n(x + \theta) \to_{a.s.} F_0(x)$$

and

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} 1_{[\theta - X_i \leq x]} &= 1 - \boldsymbol{F}_n(\theta - x-) \\ &\to_{a.s.} 1 - F_0(-x) = F_0(x),\end{aligned}$$

where

$$\boldsymbol{F}_n(x) \equiv n^{-1}\sum_{i=1}^{n} 1_{[X_i \leq x]}$$

is the empirical df of the X's. This suggests estimation of θ as the value of t which minimizes some measure of distance between the empirical df of $X_1 - t, \ldots, X_n - t$ and the empirical df of $t - X_1, \ldots, t - X_n$. The minimum Kolmogorov-Smirnov distance estimator $\hat{\theta}_{KS}$ of θ is any value of t which minimizes

$$D_n(t) \equiv \sup_x |F_n(x+t) - (1 - F_n(t-x-))| .$$

It was shown by Rao, Schuster, and Littell (1975) that

$$\mathbf{L}_\theta(\sqrt{n}(\hat{\theta}_{KS} - \theta)) \to \mathbf{L}(Z),$$

where $\mathbf{L}(Z)$ is *not* Gaussian, but is expressible in terms of a Brownian bridge process. Since the estimator is translation equivariant, the convergence is trivially uniform in θ. If the Kolmogorov-Smirnov (supremum) distance is replaced by an L_2-distance, then the resulting estimator is uniformly Gaussian regular; see, e.g., Shorack and Wellner (1986, page 759, exercise 22.5.3). □

Example 3. Bickel-Hodges estimate of center of symmetry.

Suppose that $X_1, \ldots, X_n$ are i.i.d. with symmetric distribution as in example 2, and let $\hat{\theta}_n \equiv \text{med}\{\frac{1}{2}(X_{(i)} + X_{(n-i+1)}) : 1 \le i \le n\}$. Then Bickel and Hodges (1967) show that

$$\mathbf{L}_\theta(\sqrt{n}(\hat{\theta}_n - \theta)) \to \mathbf{L}(Z),$$

where $\mathbf{L}(Z)$ is again not Gaussian, but is expressible in terms of a Brownian motion process. If the estimator is instead the Hodges-Lehmann estimator $\tilde{\theta}_n = \text{med}\{\frac{1}{2}(X_{(i)} + X_{(j)}) : 1 \le i, j \le n\}$, then $\tilde{\theta}_n$ is a uniformly Gaussian regular estimator of θ; see section 2.5. □

2.3 THE INFORMATION BOUND AND THE HÁJEK-LE CAM CONVOLUTION AND ASYMPTOTIC MINIMAX THEOREMS

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

(2) $\tilde{\mathbf{l}}(\cdot, P \mid \nu, \mathbf{P}) = \dot{q}(\theta) I^{-1}(\theta) \dot{\mathbf{l}}(\theta)$, 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 1. $I^{-1}(P \mid \nu, \mathbf{P})$ and $\tilde{\mathbf{l}}(\cdot, P \mid \nu, \mathbf{P})$ 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

$$Q_\gamma = P_{\theta(\gamma)}.$$

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 $\tilde{\mathbf{I}}$. □

Theorem 1. (Convolution Theorem) Suppose T is a uniformly regular estimate of ν with corresponding limit law $\mathbf{L}_\theta$. Then:

A. $\mathbf{L}_\theta$ is representable as the convolution of a $N(0, I^{-1}(P_\theta \mid \nu, \mathbf{P}))$ distribution with that of another m-vector Δ_θ; $\mathbf{L}_\theta = \mathbf{L}(Z_\theta + \Delta_\theta)$, where $Z_\theta \sim N(0, I^{-1}(P_\theta \mid \nu, \mathbf{P}))$ and Δ_θ are independent. More generally,

$$\mathbf{L}_\theta \begin{pmatrix} \sqrt{n}(T_n - q(\theta)) - \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\mathbf{l}}(X_i, P_\theta \mid \nu, \mathbf{P}) \\ \frac{1}{\sqrt{n}} \sum_{i=1}^n \tilde{\mathbf{l}}(X_i, P_\theta \mid \nu, \mathbf{P}) \end{pmatrix} \to \mathbf{L}_\theta \begin{pmatrix} \Delta_\theta \\ Z_\theta \end{pmatrix}. \tag{3}$$

B. If the map $\theta \to \dot{q}(\theta)$ is continuous, then (3) holds uniformly on compact subsets $\mathbf{K}$ of Θ.

C. Moreover, if $\theta \to \dot{q}(\theta)$ is continuous, then $\mathbf{L}_\theta = N(0, I^{-1}(P_\theta \mid \nu, \mathbf{P}))$ for all θ if and only if T is uniformly asymptotically linear with efficient influence function $\tilde{\mathbf{l}} = \tilde{\mathbf{l}}(\cdot, P_\theta \mid \nu, \mathbf{P})$.

Since it is intuitively clear that an estimator $T = \{T_n\}$ with limit law $\mathbf{L}_\theta = \mathbf{L}(Z_\theta) = N(0, I^{-1}(P_\theta \mid \nu, \mathbf{P}))$ and $\Delta_\theta \equiv 0$ is "less spread out" than an estimator with nondegenerate Δ_θ, and since C implies that all uniformly Gaussian regular estimates T with $\Sigma(P_\theta, T) = I^{-1}(P_\theta \mid \nu, \mathbf{P})$ are asymptotically equivalent, it is reasonable to use the result of theorem 1 to define efficiency.

Definition 1. If $T = \{T_n\}$ is a uniformly Gaussian regular estimator of $\nu(P_\theta) = q(\theta)$ with $\Sigma(P_\theta, T) = I^{-1}(P_\theta \mid \nu, \mathbf{P})$, we say that T is *uniformly efficient*. If this holds with "uniformly" replaced by "locally," then we say that T is *locally efficient* or just *efficient*.

Proof of theorem 1. Fix θ. Recall that $S_n(\theta) = n^{-1/2} \sum_{i=1}^n \dot{\mathbf{l}}(X_i, \theta)$. Regularity of $\{T_n\}$ and regularity of the model $\mathbf{P}$ imply that

$$(U_n, V_n) \equiv (\sqrt{n}(T_n - q(\theta)), S_n(\theta))$$

are marginally convergent in law, hence marginally tight, and hence jointly tight, since marginal tightness implies joint tightness.

Fix a subsequence $\{n'\}$. By Prohorov's theorem A.7.5, there exists a further

subsequence $\{n''\}$ such that $\mathbf{L}_\theta(U_{n''}, V_{n''})$ has a joint limit $\mathbf{L}(U, V)$ where $V \sim N(0, I(\theta))$. For convenience of notation, we now denote the subsequence $\{n''\}$ by $\{n\}$. Let

$$W_n = L_n(\theta + \frac{t}{\sqrt{n}}) - L_n(\theta) .$$

Then $\mathbf{L}_\theta(U_n, W_n)$ has limit $\mathbf{L}(U, t^T V - \frac{1}{2} t^T I(\theta) t)$ by the LAN property, proposition 2.1.2. Next, since the map $\theta \to P_\theta$ is continuous at regular points θ, the regularity of $\{T_n\}$ implies by proposition 2.2.1.B that, for all t,

$$\mathbf{L}_{\theta + t/\sqrt{n}}(\sqrt{n}(T_n - q(\theta + \frac{t}{\sqrt{n}}))) \to \mathbf{L}_\theta = \mathbf{L}(U) .$$

Hence, by differentiability of q,

(a) $$\mathbf{L}_{\theta + t/\sqrt{n}}(U_n) \to \mathbf{L}(U + \dot{q}(\theta) t) .$$

By (a)

(b) $$E_{\theta + t/\sqrt{n}}(\exp[\, ia^T U_n \,]) \to \exp[\, ia^T \dot{q}(\theta) t \,] E \exp[\, ia^T U \,] .$$

On the other hand, by contiguity,

(c) $$E_{\theta + t/\sqrt{n}}(\exp[\, ia^T U_n \,]) = E_\theta(\exp[\, ia^T U_n + W_n \,]) + o(1) .$$

Finally, note that $\{\, |\exp[\, ia^T U_n + W_n \,]| \,\} = \{\, \exp(W_n) \,\}$ is a uniformly integrable sequence of variables by proposition 2.1.2 and lemma A.7.2. Combining (b) and (c) we arrive, again by lemma A.7.2, at the identity, valid for all $a \in R^m$, $t \in R^k$,

(d) $$E \exp[\, ia^T U + t^T V - \tfrac{1}{2} t^T I(\theta) t \,] = \exp[\, ia^T \dot{q}(\theta) t \,] E \exp[\, ia^T U \,] .$$

Both sides of (d), for fixed a, are functions of t analytic on $\mathbf{C}^k$, where $\mathbf{C}$ is the complex plane. Therefore by analytic continuation, identity (d) holds for $t^T = -i(a^T - b^T)\dot{q}(\theta) I^{-1}(\theta)$ as well. Then (d) becomes

(e) $$E \exp[\, ia^T (U - \dot{q}(\theta) I^{-1}(\theta) V) + i b^T \dot{q}(\theta) I^{-1}(\theta) V \,]$$
$$= E \exp[\, ia^T U + \tfrac{1}{2} a^T \dot{q}(\theta) I^{-1}(\theta) \dot{q}^T(\theta) a \,]$$
$$\cdot \exp[\, -\tfrac{1}{2} b^T \dot{q}(\theta) I^{-1}(\theta) \dot{q}^T(\theta) b \,] .$$

Note that formula (e) is the characteristic function of the limit law of $(U_{n''} - \dot{q}(\theta) I^{-1}(\theta) V_{n''}, \dot{q}(\theta) I^{-1}(\theta) V_{n''})$ and is the same for every choice of initial subsequence $\{n'\}$. Consequently the full sequence, which is the left side of (3), has limit law with characteristic function (e). For $b = 0$, (e) reduces to

(f) $$E \exp[ia^T (U - \dot{q}(\theta) I^{-1}(\theta) V)]$$
$$= E \exp[ia^T U + \tfrac{1}{2} a^T \dot{q}(\theta) I^{-1}(\theta) \dot{q}^T(\theta) a] .$$

Together, (e) and (f) yield

(g) $$E\exp[ia^T(U - \dot{q}(\theta)I^{-1}(\theta)V) + ib^T\dot{q}(\theta)I^{-1}(\theta)V]$$
$$= E\exp[ia^T(U - \dot{q}(\theta)I^{-1}(\theta)V)]\exp[-\tfrac{1}{2}b^T\dot{q}(\theta)I^{-1}(\theta)\dot{q}^T(\theta)b],$$

which is a product of a function of a with a function of b. The function of b is the characteristic function of a $N(0, I^{-1}(P_\theta \mid \nu, \mathbf{P}))$ distribution. From (g) with $b = a$, the first part of A follows. But (g) also yields (3).

To prove B and C, we copy the proof leading to (g), replacing θ by a sequence $\{\theta_n\}$ tending to θ, e.g.,

(a′) $$\mathbf{L}_{\theta_n + t/\sqrt{n}}(U_n) \to \mathbf{L}(U + \dot{q}(\theta)t).$$

Note that for (a′) the continuous differentiability of q is used. In this way, (g) holds, and hence (3), with θ on the left side replaced by θ_n, is proved.

If $\mathbf{L}_\theta = N(0, I^{-1}(P_\theta \mid \nu, \mathbf{P}))$, then part B of the theorem shows that $\Delta_\theta = 0$ a.s. and hence that T is uniformly asymptotically linear with efficient influence function $\tilde{\mathbf{I}}$. Conversely, if T is uniformly asymptotically linear with influence function $\tilde{\mathbf{I}}$, then proposition 2.2.1.D implies uniform Gaussian regularity, since $\theta \to I^{-1}(P_\theta \mid \nu, \mathbf{P})$ is continuous by (1), the continuity of $\theta \to \dot{q}(\theta)$, and the regularity of the model. □

Information inequality. If T is uniformly Gaussian regular, then

(4) $$\Sigma(P_\theta, T) \geq I^{-1}(P_\theta \mid \nu, \mathbf{P})$$

in the order on nonnegative definite matrices. Equality holds if and only if T is uniformly efficient.

Proof. By the convolution theorem

$$\Sigma(P_\theta, T) = I^{-1}(P_\theta \mid \nu, \mathbf{P}) + E(\Delta_\theta - E\Delta_\theta)(\Delta_\theta - E\Delta_\theta)^T.$$

If equality holds, Δ_θ is constant. Since the asymptotic mean of $\sqrt{n}(T_n - \nu)$ is zero, we must have $\Delta_\theta = 0$. □

Note that, if q is the identity,

$$I^{-1}(P_\theta \mid \nu, \mathbf{P}) = I^{-1}(\theta),$$

and (4) is the usual (asymptotic) information bound for θ.

Definition 2. A function $l : R^m \to R^+$ is called *bowl-shaped* if
$l(x) = l(-x)$, and $\{x : l(x) \leq c\}$ is convex for every $c \geq 0$.
Bowl-shaped functions l generate loss functions and risks via $E_\theta l(\sqrt{n}(T_n - q(\theta)))$.

Asymptotic optimality theorem. If T is uniformly regular and l is bowl-shaped, then

(5) $$\liminf_{n \to \infty} E_\theta l(\sqrt{n}(T_n - q(\theta))) \geq El(Z_\theta),$$

where $Z_\theta \sim N(0, I^{-1}(P_\theta \mid \nu, \mathbf{P}))$.

Example 1. Quadratic loss.
$l(x) = x^2$.
The risk of T_n is n times its mean square error. □

Example 2. Zero-one loss.

$l(x) \equiv 1 - 1_C(x)$, where C is a bounded symmetric convex set. The risk of T_n is the probability that the confidence region $T_n + C/\sqrt{n}$ does not cover $\nu(P_\theta) = q(\theta)$. □

Proof of the asymptotic optimality theorem. Define

$$l_k(x) = 2^{-k} \sum_{i=1}^{k2^k} (1 - 1_{C_{ik}}(x)),$$

with $C_{ik} \equiv \{ y : l(y) \leq i2^{-k} \}$. Note that $l_k \uparrow l$ as $k \to \infty$ and that l_k is bowl-shaped and continuous Lebesgue a.e. since the boundary of a convex set is a Lebesgue null set. Since T is uniformly regular, by the convolution theorem its limit law $\mathbf{L}_\theta$ is absolutely continuous and can be represented as that of $Z_\theta + \Delta_\theta$ where Δ_θ is independent of Z_θ. The boundedness and a.e. continuity of l_k then yield

$$\liminf_{n \to \infty} E_\theta l(\sqrt{n}(T_n - q(\theta))) \geq E l_k(Z_\theta + \Delta_\theta).$$

Apply Anderson's theorem (Anderson (1955)) and the independence of Z_θ and Δ_θ to conclude

$$E l_k(Z_\theta + \Delta_\theta) \geq E l_k(Z_\theta).$$

The theorem follows by monotone convergence. □

Note. The Hájek–Le Cam convolution theorem, the information inequality, and the asymptotic optimality theorem hold if uniform (Gaussian) regularity and uniform efficiency are replaced by their local versions.

An extension of the asymptotic optimality theorem which applies to *all* (rather than just regular) estimates T is the

Local asymptotic minimax theorem. If $\{ T_n \}$ is any sequence of estimates, then

$$\lim_{M \to \infty} \liminf_{n \to \infty} \sup \{ E_{\theta'} l(\sqrt{n}(T_n - q(\theta'))) : \sqrt{n} |\theta' - \theta| \leq M \} \geq El(Z_\theta). \tag{6}$$

Proof. See Ibragimov and Has'minskii (1981, remark II.12.2). □

In this book we restrict ourselves to regular estimates and shall not refer to this result further.

2.4 NUISANCE PARAMETERS, ADAPTATION, AND SOME GEOMETRY

As we saw in chapter 1, parameters ν are often defined implicitly through a parametrization $\theta \to P_\theta$, where $\theta^T = (\nu^T, \eta^T)$, $\nu \in N \subset R^m$, $\eta \in H \subset R^{k-m}$, where ν is the parameter of interest and η is a *nuisance parameter*. 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 com-

paring the information bounds and efficient influence functions for ν at P_{θ_0} in $\mathbf{P}_1(\eta_0)$ and $\mathbf{P}$.

As before, we let $<\cdot,\cdot>_0$ 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

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

where

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

By (2.3.1) and (2.3.2), the information bound for estimating ν in $\mathbf{P}$ is $I^{11} = I_{1\cdot 2}^{-1}$ and the efficient influence function for ν in $\mathbf{P}$ is

$$(3) \qquad \begin{aligned} \tilde{\mathbf{l}}_1 &= I^{11}\dot{\mathbf{l}}_1 + I^{12}\dot{\mathbf{l}}_2 \\ &= I_{1\cdot 2}^{-1}(\dot{\mathbf{l}}_1 - I_{12} I_{22}^{-1}\dot{\mathbf{l}}_2) \qquad \text{by (1)} \\ &\equiv I_{1\cdot 2}^{-1}\dot{\mathbf{l}}_1^*. \end{aligned}$$

Since

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

we see that (3) 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 $I_{11\cdot 2} = E_0(\dot{\mathbf{l}}_1^*\dot{\mathbf{l}}_1^{*T})$, and $\dot{\mathbf{l}}$ replaced by

$$(4) \qquad \dot{\mathbf{l}}_1^* \equiv \dot{\mathbf{l}}_1 - I_{12} I_{22}^{-1}\dot{\mathbf{l}}_2.$$

We therefore call $\dot{\mathbf{l}}_1^*$ 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

$$I_{11}^{-1}\,\dot{\mathbf{l}}_1 \,. \tag{5}$$

From the block matrix formulas relating $[I_{ij}]$ and $[I^{ij}]$, we can derive some important relations between these quantities. First note from (1) and (2) that

$$(I^{11})^{-1} = I_{11\cdot 2} = I_{11} - I_{12}I_{22}^{-1}I_{21}\,, \tag{6}$$

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

$$I^{11} = I_{11\cdot 2}^{-1} = I_{11}^{-1} + I^{12}(I^{22})^{-1}I^{21}\,, \tag{7}$$

so not knowing η increases the information bound (inverse information) by $I^{12}(I^{22})^{-1}I^{21}$. Moreover, from (6),

$$I_{11\cdot 2} = I_{11} \qquad \text{and} \qquad I_{11\cdot 2}^{-1} = I_{11}^{-1} \tag{8}$$

if and only if

$$I_{12} = 0\,. \tag{9}$$

In this case it also follows from (3), (4), and (8) that

$$\tilde{\mathbf{l}}_1 = I_{11}^{-1}\,\dot{\mathbf{l}}_1 \qquad \text{and} \qquad \mathbf{l}_1^{*} = \dot{\mathbf{l}}_1\,. \tag{10}$$

Definition 1. $\{\hat{\nu}_n\}$ 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 (9) and (10), a necessary condition for the existence of adaptive estimates in regular parametric models is

$$I_{12}(\theta) = 0 \qquad \text{for all } \theta\,. \tag{11}$$

Cox and Reid (1987) discuss reparametrization of η to achieve (11); see also Kass (1989, section 2.1.4, and pages 201, 202). In any case, adaptation is very much a feature of the parametrization, as the following examples show.

Example 1. Normal location-scale.

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 or not we know the variance. □

Example 2. 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

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

We can think of $\dot{\mathbf{l}}_1^*$ 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 2 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.

Some Geometry

The efficient influence function $\tilde{\mathbf{l}}_1$ and efficient score function $\dot{\mathbf{l}}_1^*$ 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 (4) the efficient score function $\dot{\mathbf{l}}_1^*$ 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 sections 3 and 4 of chapter 3.

Proposition 1.

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

See figures 1 and 3.

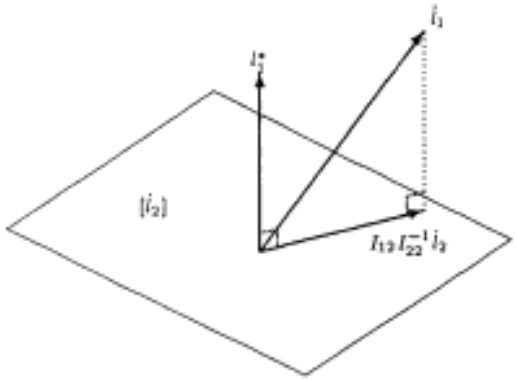

FIGURE 1. Projection of score functions.

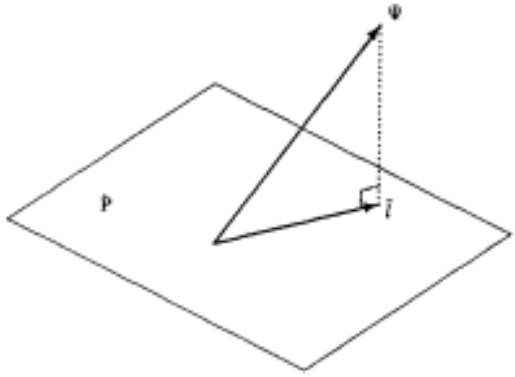

FIGURE 2. Projection of influence functions.

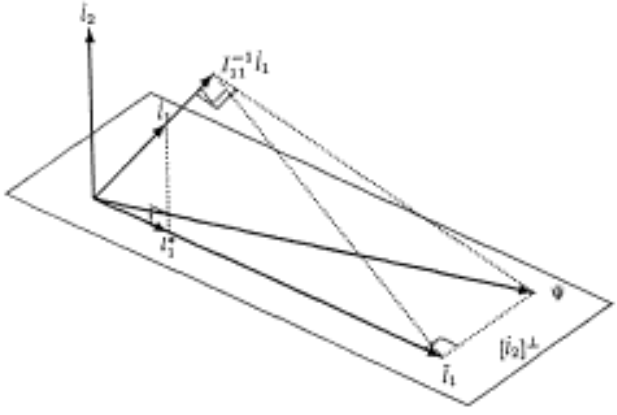

FIGURE 3. Score and influence function projections.

Here is another relationship between the influence and score functions of $\mathbf{P}_1(\eta_0)$ and its companion $\mathbf{P}_2(\nu_0) = \{P_{(\nu_0,\eta)} : \eta \in H\}$. 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

$$\tilde{\mathbf{l}}_1 = I_{11}^{-1}\dot{\mathbf{l}}_1 - I_{11}^{-1}I_{12}\tilde{\mathbf{l}}_2 . \tag{12}$$

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 (12) provides an immediate proof, by orthogonality of $\tilde{\mathbf{l}}_2$ to $[\dot{\mathbf{l}}_1]$, of the formula

$$I_{11\cdot 2}^{-1} = I_{11}^{-1} + I_{11}^{-1}I_{12}I_{22\cdot 1}^{-1}I_{21}I_{11}^{-1} , \tag{13}$$

which is another way of writing (7).

Proof of (12). From (1),

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

Table 1 summarizes the efficient score functions, efficient influence functions, information, and inverse information for the two models $\mathbf{P}$ and $\mathbf{P}_1(\eta_0)$.

Name	Notation	Model: $\mathbf{P}$	Model: $\mathbf{P}_1(\eta_0)$
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	$I(P \mid \nu, \cdot)$	$E\mathbf{l}_1^*\mathbf{l}_1^{*T} = I_{11} - I_{12}I_{22}^{-1}I_{21} = I_{11\cdot 2}$	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

We now use several examples to illustrate the relationships between score functions, efficient score functions, and efficient influence functions given in proposition 1 and table 1. In our first two examples, the efficient score functions are intuitively plausible.

Example 3. The bivariate normal distribution.

Suppose that $X = (X_1, X_2) \sim N_2(\theta, \Sigma)$, where $\theta = (\nu, \eta) \in R^2$, and Σ is the covariance matrix with ones on the diagonal and correlation ρ; we will suppose here that ρ is known. The joint density is

$$p(x,\theta) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)}\{z_1^2 - 2\rho z_1 z_2 + z_2^2\}\right\},$$

where $z_1 \equiv x_1 - \nu$, $z_2 \equiv x_2 - \eta$, and the model is $\mathbf{P} = \{P_\theta : \theta \in R^2\}$. Hence, the scores for $\nu = \theta_1$ and $\eta = \theta_2$ are

$$\dot{l}_1(x) = \frac{1}{1-\rho^2}\{(x_1 - \nu) - \rho(x_2 - \eta)\},$$

$$\dot{l}_2(x) = \frac{1}{1-\rho^2}\{(x_2 - \eta) - \rho(x_1 - \nu)\}.$$

It is straightforward to calculate

$$I(\theta) = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix}.$$

Hence the efficient score for $\nu = \theta_1$ in the presence of the nuisance parameter $\eta = \theta_2$ is

$$\dot{l}_1^*(x) = x_1 - \nu,$$

and the information for ν is $I(P_\theta | \nu, \mathbf{P}) = 1$. Of course the information bound is achieved by the sample mean, $\hat{\nu} = \overline{X}_1 \equiv n^{-1}\sum_{i=1}^n X_1^{(i)}$; here $X^{(1)}, \ldots, X^{(n)}$ denote the observations which are i.i.d. as X. Note that indeed $\dot{l}_1^* \perp \dot{l}_2$.

Now consider the submodel $\mathbf{P}_1 = \mathbf{P}_1(\eta_0)$ in which η is known; without loss of generality suppose that $\eta_0 = 0$. For this submodel $\mathbf{P}_1$, the score function $\dot{l}_1$ for ν is also the efficient score function $\dot{l}_1^* \equiv \dot{l}_1^*(\cdot, P_\theta | \nu, \mathbf{P}_1)$, so that

$$I(P_\theta | \nu, \mathbf{P}_1) = \frac{1}{1-\rho^2}, \qquad I^{-1}(P_\theta | \nu, \mathbf{P}_1) = 1 - \rho^2.$$

The efficient influence function for ν in the submodel $\mathbf{P}_1$ is

$$\tilde{l}_1(x, P_\theta | \nu, \mathbf{P}_1) = x_1 - \nu - \rho x_2.$$

Thus knowledge of η reduces the information bound for estimation of ν from 1 to $1 - \rho^2$. An estimator achieving this bound (assuming known covariance matrix Σ) is $\hat{\nu}^0 = n^{-1}\sum_{i=1}^n (X_1^{(i)} - \rho X_2^{(i)})$. □

Example 4. The multinomial distribution.

Suppose that $X = (X_1, \ldots, X_{k+1}) \sim Mult_{k+1}(1, (p_1, \ldots, p_{k+1}))$, and let $\theta = (p_1, \ldots, p_k)$ so that $p_{k+1} = 1 - \sum_{i=1}^k p_i = 1 - \sum_{i=1}^k \theta_i$. The density function is

$$p(x,\theta) = \left\{\prod_{i=1}^k \theta_i^{x_i}\right\}(1 - \sum_{i=1}^k \theta_i)^{x_{k+1}}$$

for $x_j \in \{0,1\}$, $j=1, \ldots, k+1$. Hence the scores for θ are easily calculated to be

$$\dot{l}_j(x) = \frac{x_j}{\theta_j} - \frac{x_{k+1}}{p_{k+1}}, \qquad j=1, \ldots, k,$$

and the information matrix for θ is

$$I(P_\theta \mid \theta, \mathbf{P}) \equiv I(\theta) = \operatorname{diag}\left(\frac{1}{\theta}\right) + \left(\frac{1}{p_{k+1}}\right) \underline{1}\,\underline{1}^T,$$

where $\underline{1}$ is a k-vector of 1's. Therefore the information bound is given by

$$I^{-1}(\theta) = \operatorname{diag}(\theta) - \theta\theta^T.$$

Of course, the usual estimator $\hat{\theta}$ given by the first k coordinates of $\hat{p} = n^{-1}\sum_{i=1}^{n} X^{(i)} = \overline{X}$ (where $X^{(1)}, \ldots, X^{(n)}$ are i.i.d. as X) achieves this bound.

Now consider estimation of $\nu = (\theta_1, \ldots, \theta_m)$ with $m < k$. The efficient scores for ν, with $(\theta_{m+1}, \ldots, \theta_k) \equiv \eta$ as nuisance parameters are

$$\mathbf{l}_j^*(x) = \frac{x_j}{\theta_j} - \frac{\sum_{l=m+1}^{k+1} x_l}{\sum_{l=m+1}^{k+1} p_l}, \qquad j = 1, \ldots, m,$$

which can be easily checked via orthogonality. This is the same as the efficient score for ν in the multinomial model for observation of

$$Y = (X_1, \ldots, X_m, \sum_{j=m+1}^{k+1} X_j) \sim Mult_{m+1}(1, (\nu, 1 - \sum_{j=1}^{m} \nu_j)),$$

and corresponds with intuition. Consequently the information bound for estimation of ν is just the upper left corner of $I^{-1}(\theta)$:

$$I^{-1}(\nu) = I^{-1}(P_\theta \mid \nu, \mathbf{P}) = \operatorname{diag}(\nu) - \nu\nu^T.$$

Again the bound is achieved by the natural estimator $\hat{\nu}$ given by the first m coordinates of $\hat{p} = \overline{X}$ (which is exactly equal to the first m coordinates of $\overline{Y}$ where $Y^{(1)}, \ldots, Y^{(n)}$ are defined in terms of $X^{(1)}, \ldots, X^{(n)}$ exactly as Y was defined in terms of X above).

On the other hand, consider the submodel $\mathbf{P}_1 = \mathbf{P}_1(\eta_0)$ in which $\eta = (\theta_{m+1}, \ldots, \theta_k) = \eta_0$ is known. Then the efficient scores $\mathbf{l}_j^*(\,\cdot\,, P_\theta \mid \nu, \mathbf{P}_1)$ equal the original scores $\dot{\mathbf{l}}_j$, $j = 1, \ldots, m$, and the $(m \times m)$-information matrix is given by the upper left corner of $I(\theta)$:

$$I(P_\theta \mid \nu, \mathbf{P}_1) = \operatorname{diag}\left(\frac{1}{\nu}\right) + \frac{1}{p_{k+1}} \underline{1}\,\underline{1}^T.$$

Therefore the information bound is given by

$$I^{-1}(P_\theta \mid \nu, \mathbf{P}) = \operatorname{diag}(\nu) - \frac{1}{c}\nu\nu^T,$$

where $c \equiv 1 - \sum_{j=m+1}^{k} \theta_j = \sum_{i=1}^{m} \nu_i + p_{k+1}$. This bound is achieved by the estimator

$$\hat{\nu}_j^0 = \frac{c\overline{X}_j}{\overline{X}_{k+1} + \sum_{l=1}^{m} \overline{X}_l}, \qquad j = 1, \ldots, m,$$

where $\overline{X}_j = n^{-1}\sum_{i=1}^n X_j^{(i)}, j = 1, \ldots, m, k+1$. This estimator is the maximum likelihood estimator with respect to both the unconditional and conditional (given $X_{m+1}, \ldots, X_k$) likelihoods for the submodel $\mathbf{P}_1$. □

Example 5. Gaussian linear regression model.

Let $X = (Z, Y)$ where Y is scalar, Z and θ are k-vectors, $Y = \theta^T Z + e$, and $e \sim N(0, 1)$ independent of Z. Here $Z \sim H$, and we assume that $E(ZZ^T)$ is nonsingular. This is the model specified by (1.3.6), (1.3.7) with $\sigma^2 = 1$ for simplicity. If $\theta \equiv (\theta_1, \ldots, \theta_k)^T$, identify ν with θ_1, η with $(\theta_2, \ldots, \theta_k)$. Then, if $Z \equiv (Z_1, \ldots, Z_k)^T$,

$$\dot{l} = Ze,$$

and hence

$$\dot{l}_1 = Z_1 e, \qquad \dot{l}_2 = Z_{[2]} e,$$

where $Z_{[2]} \equiv (Z_2, \ldots, Z_k)^T$. Therefore

$$I = E(ZZ^T e^2) = E(ZZ^T), \tag{14}$$
$$\tilde{l} = I^{-1} Ze,$$

and

$$\tilde{l}_1^* = (Z_1 - I_{12} I_{22}^{-1} Z_{[2]})e, \tag{15}$$

where $I_{12} = E(Z_1 Z_{[2]}^T)$, $I_{22} = E(Z_{[2]} Z_{[2]}^T)$, and

$$\tilde{l}_1 = (Z_1 - I_{12} I_{22}^{-1} Z_{[2]})e \,/\, E(Z_1 - I_{12} I_{22}^{-1} Z_{[2]})^2 . \tag{16}$$

If we observe $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$, $\mathbf{Z}_{k\times n} = (Z^{(1)}, \ldots, Z^{(n)})$ where $(Z^{(j)}, Y_j)$ is the jth observation, then the least squares (maximum likelihood) estimate of θ is

$$\hat{\theta} = [\mathbf{Z}\mathbf{Z}^T]^{-1}\mathbf{Z}\mathbf{Y}.$$

Since $n^{-1}\mathbf{Z}\mathbf{Z}^T = n^{-1}\sum_{i=1}^n Z^{(i)}[Z^{(i)}]^T = E(ZZ^T) + O_p(n^{-1/2})$, we see that

$$\hat{\theta} = \theta + n^{-1}\sum_{i=1}^n I^{-1} Z^{(i)} e_i + o_p(n^{-1/2})$$

indeed has the (efficient) influence function $\tilde{l}$. If we replace I_{12}, I_{22} by the corresponding blocks $\hat{I}_{12}$, $\hat{I}_{22}$ in $n^{-1}\mathbf{Z}\mathbf{Z}^T$, we obtain the least squares estimate of ν as

$$\hat{\nu} \equiv \sum_{i=1}^n (Z_1^{(i)} - \hat{I}_{12}\hat{I}_{22}^{-1} Z_{[2]}^{(i)}) Y_i \,/\, \sum_{j=1}^n (Z_1^{(j)} - \hat{I}_{12}\hat{I}_{22}^{-1} Z_{[2]}^{(j)})^2 \tag{17}$$
$$= \nu + n^{-1}\sum_{i=1}^n \tilde{l}_1(Z^{(i)}, Y_i) + o_p(n^{-1/2}).$$

Note that the influence function merely replaces the coefficients of the regression of $Z_1^{(i)}$ on $Z_{[2]}^{(i)}$ based on the n observations by the corresponding population quantities, or, equivalently, the empirical measure orthogonality condition

$$n^{-1}\sum_{i=1}^n (Z_1^{(i)} - \hat{I}_{12}\hat{I}_{22}^{-1} Z_{[2]}^{(i)}) Z_{[2]}^{(i)} = 0 \tag{18}$$

is replaced by the population orthogonality condition

$$E(\mathbf{I}_1^* \dot{\mathbf{l}}_2) = E[(Z_1 - I_{12} I_{22}^{-1} Z_{[2]}) Z_{[2]}] = 0, \tag{19}$$

which corresponds to proposition 1.A. Similarly if η is assumed known, the least squares estimate of ν is

$$\hat{\nu}^* = \sum_{i=1}^{n} Z_1^{(i)} Y_i \Big/ \sum_{j=1}^{n} [Z_1^{(j)}]^2$$

with influence function $[I_{11}]^{-1} \dot{\mathbf{l}}_1$. Also from (17) and (18)

$$\hat{\nu}^* - \hat{\nu} = \sum_{i=1}^{n} \left\{ \frac{Z_1^{(i)}}{| Z_1^{(\cdot)} |^2} - \frac{Z_1^{(i)} - \hat{I}_{12} \hat{I}_{22}^{-1} Z_{[2]}^{(i)}}{| Z_1^{(i)} - \hat{I}_{12} \hat{I}_{22} Z_{[2]}^{(\cdot)} |^2} \right\} \epsilon_i \tag{20}$$

where $| Z_1^{(\cdot)} |$ is the Euclidean norm of $(Z_1^{(1)}, \ldots, Z_1^{(n)})^T$. Now

$$\sum_{i=1}^{n} \left\{ \frac{Z_1^{(i)}}{| Z_1^{(\cdot)} |^2} - \frac{Z_1^{(i)} - \hat{I}_{12} \hat{I}_{22}^{-1} Z_{[2]}^{(i)}}{| Z_1^{(\cdot)} - \hat{I}_{12} \hat{I}_{22}^{-1} Z_{[2]}^{(\cdot)} |^2} \right\} Z_1^{(i)} = 0 \tag{21}$$

by (18). Again this orthogonality relation is the sample version of the population relation

$$E\left(\left(\frac{\dot{\mathbf{l}}_1}{\| \dot{\mathbf{l}}_1 \|_0^2} - \tilde{\mathbf{I}}_1\right) \dot{\mathbf{l}}_1\right) = E\left(\frac{Z_1}{\| Z_1 \|_0^2} - \frac{Z_1 - I_{12} I_{22}^{-1} Z_{[2]}}{\| Z_1 - I_{12} I_{22}^{-1} Z_{[2]} \|_0^2} \right) Z_1 = 0, \tag{22}$$

which corresponds to proposition 1.B. □

Here is another basic example illustrating proposition 1.

Example 6. The bivariate normal distribution, continued.

Suppose that $X = (X_1, X_2) \sim N_2(\nu, \Sigma)$ where $\nu \in R^2$ and

$$\Sigma = \begin{pmatrix} \eta_1^2 & \rho \eta_1 \eta_2 \\ \rho \eta_1 \eta_2 & \eta_2^2 \end{pmatrix};$$

here $\theta = (\nu_1, \nu_2, \eta_1^2, \eta_2^2, \rho) \in R^2 \times R^{+2} \times (-1, 1)$. The joint density is, with $z_i = z_i(\theta) = (x_i - \nu_i)/\eta_i$, $i = 1, 2$,

$$p(x, \theta) = \frac{1}{2\pi \eta_1 \eta_2 \sqrt{1 - \rho^2}} \exp\left\{ -\frac{1}{2(1 - \rho^2)} \{ z_1^2 - 2\rho z_1 z_2 + z_2^2 \} \right\},$$

and the model is $\mathbf{P} = \{P_\theta : \theta \in R^2 \times R^{+2} \times (-1, 1)\}$, the family of (nondegenerate) bivariate normal distributions with all five parameters unknown. Hence, the scores for θ are

$$\dot{\mathbf{l}}_1(x) = \frac{1}{(1 - \rho^2)\eta_1} \{ z_1 - \rho z_2 \},$$

$$\dot{\mathbf{l}}_2(x) = \frac{1}{(1 - \rho^2)\eta_2} \{ z_2 - \rho z_1 \},$$

$$\dot{\mathbf{l}}_3(x) = -\frac{1}{2\eta_1^2}\left\{1 - \frac{1}{1-\rho^2}\{z_1^2 - \rho z_1 z_2\}\right\},$$

$$\dot{\mathbf{l}}_4(x) = -\frac{1}{2\eta_2^2}\left\{1 - \frac{1}{1-\rho^2}\{z_2^2 - \rho z_1 z_2\}\right\},$$

$$\dot{\mathbf{l}}_5(x) = \frac{1}{(1-\rho^2)^2}\{\rho(1-\rho^2) - \rho\{z_1^2 + z_2^2\} + (1+\rho^2)z_1 z_2\}.$$

It is straightforward to calculate that $[\dot{\mathbf{l}}_1, \dot{\mathbf{l}}_2] \perp [\dot{\mathbf{l}}_3, \dot{\mathbf{l}}_4, \dot{\mathbf{l}}_5]$, and hence the information matrix $I(P_\theta \mid \theta, \mathbf{P})$ is block diagonal: the upper left 2×2 block is

$$I(\nu) = \frac{1}{1-\rho^2}\begin{pmatrix} \frac{1}{\eta_1^2} & -\frac{\rho}{\eta_1\eta_2} \\ -\frac{\rho}{\eta_1\eta_2} & \frac{1}{\eta_2^2} \end{pmatrix},$$

and the lower right 3×3 block for $(\eta_1^2, \eta_2^2, \rho)$ is

$$I(\eta_1^2, \eta_2^2, \rho) = \frac{1}{1-\rho^2}\begin{pmatrix} \frac{2-\rho^2}{4\eta_1^4} & \frac{-\rho^2}{4\eta_1^2\eta_2^2} & \frac{-\rho}{2\eta_1^2} \\ \frac{-\rho^2}{4\eta_1^2\eta_2^2} & \frac{2-\rho^2}{4\eta_2^4} & \frac{-\rho}{2\eta_2^2} \\ \frac{-\rho}{2\eta_1^2} & \frac{-\rho}{2\eta_2^2} & \frac{1+\rho^2}{1-\rho^2} \end{pmatrix}.$$

Thus, knowledge of $(\eta_1^2, \eta_2^2, \rho)$ does not affect how well we can estimate (ν_1, ν_2), and vice versa. Furthermore, $I^{-1}(P_\theta \mid \theta, \mathbf{P})$ is also block diagonal with

$$I^{-1}(P_\theta \mid \nu, \mathbf{P}) = \begin{pmatrix} \eta_1^2 & \rho\eta_1\eta_2 \\ \rho\eta_1\eta_2 & \eta_2^2 \end{pmatrix}$$

and

$$I^{-1}(P_\theta \mid \eta_1^2, \eta_2^2, \rho, \mathbf{P}) = \begin{pmatrix} 2\eta_1^4 & 2\rho^2\eta_1^2\eta_2^2 & \rho(1-\rho^2)\eta_1^2 \\ 2\rho^2\eta_1^2\eta_2^2 & 2\eta_2^4 & \rho(1-\rho^2)\eta_2^2 \\ \rho(1-\rho^2)\eta_1^2 & \rho(1-\rho^2)\eta_2^2 & (1-\rho^2)^2 \end{pmatrix}.$$

These information bounds are achieved by the usual maximum likelihood estimators $(\bar{X}, \hat{\Sigma})$, where

$$\hat{\eta}_j^2 = n^{-1}\sum_{i=1}^n (X_j^{(i)} - \bar{X}_j)^2, \qquad j = 1, 2,$$

$$\hat{\rho} = n^{-1}\sum_{i=1}^n (X_1^{(i)} - \bar{X}_1)(X_2^{(i)} - \bar{X}_2) / (\hat{\eta}_1\hat{\eta}_2),$$

and $X^{(1)}, \ldots, X^{(n)}$ are i.i.d. copies of X. Note that the efficient score function for ν_2 in the model $\mathbf{P}$ is

$$\mathbf{l}_2^*(x, P_\theta \mid \nu_2, \mathbf{P}) = \dot{\mathbf{l}}_2(x) - I_{21} I_{11}^{-1} \dot{\mathbf{l}}_1(x) = \frac{z_2}{\eta_2},$$

and the efficient influence function for ν_2 in the model $\mathbf{P}$ is

$$\tilde{\mathbf{l}}_2(x, P_\theta \mid \nu_2, \mathbf{P}) = \eta_2 z_2 = x_2 - \nu_2.$$

Now consider the submodel $\mathbf{P}_1 = \mathbf{P}_1(\nu_2) \subset \mathbf{P}$ in which the mean ν_2 of X_2 is known to be zero:

$$\mathbf{P}_1 = \{P_\theta \in \mathbf{P} : \nu_2 = 0\}.$$

For this submodel $\mathbf{P}_1$, the results from example 3 generalize straightforwardly: the efficient score function $\mathbf{l}_1^*$ for ν_1 is

$$\mathbf{l}_1^*(x, P_\theta \mid \nu_1, \mathbf{P}_1) = \dot{\mathbf{l}}_1(x, P_\theta \mid \nu_1, \mathbf{P}_1) = \frac{1}{(1-\rho^2)\eta_1}(z_1 - \rho z_2),$$

so that

$$I(P_\theta \mid \nu_1, \mathbf{P}_1) = \frac{1}{(1-\rho^2)\eta_1^2}, \qquad I^{-1}(P_\theta \mid \nu_1, \mathbf{P}_1) = (1-\rho^2)\eta_1^2,$$

and the efficient influence function (for ν_1 in the submodel $\mathbf{P}_1$) is

$$\tilde{\mathbf{l}}_1(x, P_\theta \mid \nu_1, \mathbf{P}_1) = \eta_1(z_1 - \rho z_2) = x_1 - \nu_1 - \rho\frac{\eta_1}{\eta_2} x_2.$$

Thus knowledge of ν_2 reduces the information bound for estimation of ν_1 from η_1^2 to $(1-\rho^2)\eta_1^2$. If η_1^2, η_2^2, ρ are also known, an estimator achieving this bound is

$$\hat{\nu}_1^0 = \frac{1}{n}\sum_{i=1}^{n} (X_1^{(i)} - \rho\frac{\eta_1}{\eta_2} X_2^{(i)}).$$

If η_1^2, η_2^2, ρ are unknown, then replacing them by their natural estimators also yields an efficient estimate in the model $\mathbf{P}_1$.

Similarly, if η_1^2 and η_2^2 are unknown, the information lower bound for estimation of ρ is $(1-\rho^2)^2$. But if $\eta_1^2 = \eta_2^2 = 1$ are both known, then the information lower bound for ρ is $(1-\rho^2)^2/(1+\rho^2)$. □

Proposition 1 can be put in a broader context.

Proposition 2. 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

$$\psi - \tilde{\mathbf{l}}_1 \perp \dot{\mathbf{P}} = [\dot{\mathbf{l}}_1, \dot{\mathbf{l}}_2], \tag{23}$$

or, equivalently, if and only if both

$$\langle \psi, \dot{\mathbf{l}}_1 \rangle_0 = 1 \tag{24}$$

and

$$\psi \perp [\dot{\mathbf{l}}_2]. \tag{25}$$

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

See figures 2 and 3 on page 31.

We note in passing that (24) and (25) are asymptotic versions of the equations leading to the Cramér-Rao information bound. Consider the problem of minimizing $\Sigma(P_\theta, T) = E_0 \psi^2$ subject to (24) and (25). 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 (25) 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 (24) 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 3. 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 $\tilde{\mathbf{l}} \equiv \tilde{\mathbf{l}}(\cdot, P_{\theta_0} \mid \nu, \mathbf{P})$,

$$\psi - \tilde{\mathbf{l}} \perp \dot{\mathbf{P}} = [\dot{\mathbf{l}}_1, \dot{\mathbf{l}}_2], \tag{26}$$

where (26) is equivalent to

$$E_0 \psi \dot{\mathbf{l}}^T = \dot{q}(\theta_0). \tag{27}$$

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

$$\psi = \tilde{\mathbf{l}} = \dot{q}(\theta_0) I^{-1}(\theta_0) \dot{\mathbf{l}}(\theta_0). \tag{28}$$

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

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

and from (c) and (d) we conclude

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

But this implies that q is differentiable at θ_0 with derivative $\dot{q}(\theta_0)$ satisfying (27) and hence (26).

On the other hand, if q is differentiable and (27) 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 (26) holds. Since $\tilde{\mathbf{l}} \in \dot{\mathbf{P}}^{(m)}$, (26) 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 (27) becomes

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

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

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

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

For elementary versions of propositions 2 and 3 under different hypotheses, see Hall and Mathiason (1990, section 3.1).

2.5 CONSTRUCTION OF $\sqrt{N}$-CONSISTENT AND EFFICIENT ESTIMATES

There are many ways of constructing estimates in smooth parametric models which under appropriate regularity conditions are efficient. The most popular construction is Fisher's method of maximum likelihood. Although this method

can fail spectacularly—a famous example is in Kiefer and Wolfowitz (1956, page 905)—there are closely related M-estimate methods which work under minimal conditions. We shall discuss these further below and in considerable detail in chapter 7. Bayes estimates corresponding to smooth prior distributions and bounded bowl-shaped loss functions also work quite generally. The best recent result is due to Ibragimov and Has'minskii (1981, theorem III.3.1, page 185). Other methods include minimum Hellinger distance estimates (see, e.g., Beran (1977b)), and a variety of approaches suitable in particular models such as minimum χ^2 in discrete data models, and L- and R-estimates for location and scale; see Huber (1981, chapter 3), for example.

$\sqrt{n}$-Consistent Preliminary Estimators

We begin by studying the construction of regular, but not in general efficient, minimum distance estimates. These procedures were introduced by Wolfowitz (1957). Using these $\sqrt{n}$-consistent starting points we then show how to construct efficient estimates. There are, of course, many methods of constructing $\sqrt{n}$-consistent estimates. These include the direct efficient constructions as well as other portmanteau methods, such as the method of moments. We specialize to the method of minimum distance because it can be applied to general (not just parametric) models and is simple to describe and analyze, though not necessarily to implement.

Let $\mathbf{P} \subset \mathbf{M}$ be the set of all probability measures on $\mathbf{X}$, and let $\theta : \mathbf{P} \to R^k$ be a parameter. A natural way of constructing estimates of θ is to find a "smooth" extension $\overline{\theta}$ of θ to $\mathbf{M}$ and let

$$T_n = \overline{\theta}(\mathbb{P}_n), \tag{1}$$

where $\mathbb{P}_n$ is the empirical distribution of $X_1, \ldots, X_n$ i.i.d. P_0 given by (A.6.6). One way of obtaining a smooth extension $\overline{\theta}$ of θ is by "minimum distance" as follows. Suppose that:

(D1) ρ is a metric compatible with $\mathbb{P}_n$ in the sense of (A.6.7).

(D2) There exists a map Π: $\mathbf{M} \to \mathbf{P}$ with $\rho(\Pi(Q), Q) = \inf\{\rho(P, Q) : P \in \mathbf{P}\}$ for every $Q \in \mathbf{M}$.

(D3) θ is ρ-continuous on $\mathbf{P}$. That is, $\rho(P_m, P) \to 0$ with $P_m, P \in \mathbf{P}$ implies $|\theta(P_m) - \theta(P)| \to 0$.

Then the ρ-extension $\overline{\theta}$ of θ is given by $\overline{\theta}(Q) = \theta(\Pi(Q))$, and (1) becomes

$$T_n = \theta(\Pi(\mathbb{P}_n)). \tag{2}$$

Lemma 1. If (D1)–(D3) hold, then T_n given by (2) is consistent.

Proof. By definition of Π and by (D1)

$$\text{(a)} \qquad \rho(\Pi(\mathbb{P}_n), P_0) \le \rho(\Pi(\mathbb{P}_n), \mathbb{P}_n) + \rho(\mathbb{P}_n, P_0) \le 2\rho(\mathbb{P}_n, P_0) \to_p 0 \quad \text{as} \quad n \to \infty .$$

The lemma follows from (D3). □

Note. Here and in the future we ignore measurability questions involving T_n which are unimportant in practice.

Definition 1. θ is ρ-*Lipschitz* on $\mathbf{P}$ if for every $P_0 \in \mathbf{P}$ there exists $c(P_0) < \infty$, $\varepsilon(P_0) > 0$, such that

$$|\theta(P) - \theta(P_0)| \leq c(P_0)\rho(P, P_0) \qquad \text{if } \rho(P,P_0) \leq \varepsilon(P_0). \tag{3}$$

Of course, ρ-Lipschitz implies ρ-continuity.

Lemma 2. If (D1)–(D2) hold and θ is ρ-Lipschitz on $\mathbf{P}$, then T_n given by (2) is $\sqrt{n}$-consistent.

Proof. By (2), (3) and then (a) of lemma 1,

$$|T_n - \theta(P_0)| \leq c(P_0)\rho(\Pi(\mathcal{P}_n), P_0) \leq 2c(P_0)\rho(\mathcal{P}_n, P_0). \qquad \square$$

We use the approach of lemma 2 to prove the following important theorem, concerning the existence of uniformly $\sqrt{n}$-consistent estimates, due to Le Cam (1956). For a generalization to non-Euclidean sample spaces, see Le Cam (1986, section 17.6).

Theorem 1. If $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$ is a regular parametric model on a Euclidean space $\mathbf{X}$ and θ is identifiable, then there exist uniformly $\sqrt{n}$-consistent estimates of θ.

Proof. It follows from proposition A.5.3.C that, uniformly for $\theta \in K$ compact and uniformly for all Borel sets A,

$$P_{\theta+h}(A) - P_\theta(A) = \int_A \dot{p}^T(x,\theta)h \, d\mu(x) + o(|h|). \tag{a}$$

Hence for any class $\mathbf{A}$ of Borel sets, it follows that

$$\begin{aligned} \text{(b)} \qquad & \liminf |h|^{-1} \sup \left\{ |P_{\theta+h}(A) - P_\theta(A)| : A \in \mathbf{A} \right\} \\ & \geq \inf_{|e|=1} \sup \left\{ \left| \int_A \dot{p}^T(x,\theta)e \, d\mu(x) \right| : A \in \mathbf{A} \right\}. \end{aligned}$$

The right side of (b) vanishes if and only if there exists e such that

$$\int_A \dot{p}^T(x,\theta)e \, d\mu(x) = 0 \qquad \text{for all } A \in \mathbf{A}. \tag{c}$$

If $\mathbf{A}$ is the class of all shifted quadrants, (c) implies

$$\frac{\dot{p}^T(x,\theta)}{p(x,\theta)}e = 0 \quad \text{a.s. } P_\theta, \tag{d}$$

which cannot hold if $I(\theta)$ is nonsingular. Therefore, for a regular parametrization,

$$d_K(P_\theta, P_{\theta_0}) \geq c^{-1}(\theta_0)|\theta - \theta_0| \qquad \text{if} \quad |\theta - \theta_0| \leq \varepsilon(\theta_0). \tag{e}$$

Moreover, the uniformity in (a) and L_1 continuity of $\theta \to \dot{p}(\theta)$ imply that we can bound $c(\theta)$ and $1/\varepsilon(\theta)$ on compacts.

Since Θ is open, there exist compacts K_j with $\Theta = \cup_{j=1}^{\infty} K_j$, $K_{j+1} \supset K_j$, $j \geq 1$. Define T_{nj} to minimize $d_K(P_\theta, \mathbb{P}_n)$ for $\theta \in K_j$. d_K continuity of $\theta \to P_\theta$ on K_j guarantees the existence of T_{nj}. Since θ is identifiable, the map $P_\theta \to \theta$ is d_K continuous on K_j and (e) implies (3) with $\rho = d_K$. As in lemma 2, it follows from the compatibility of d_K with $\mathbb{P}_n$ (see section A.6) that T_{nj} is uniformly $\sqrt{n}$-consistent on K_j. Then let $T_n = T_{nj}$ with $d_K(P_{T_n}, \mathbb{P}_n) \leq n^{-1/4}$ and j minimal. It is easy to see that $P_\theta(T_n = T_{nj_0}) \to 1$, where K_{j_0} is the first K_j such that $\theta \in K_j$. Uniform $\sqrt{n}$-consistency of T_n follows. □

Asymptotically Efficient Estimators

The classical method of estimation in regular parametric models is maximum likelihood. If $\mathbf{P} = \{P_\theta : \theta \in \Theta\}$ is a (regular) parametric model, a *maximum likelihood estimate* $\hat{\theta}_n$ of θ satisfies

$$L_n(\hat{\theta}_n) = \max\{L_n(\theta) : \theta \in \Theta\},$$

where $L_n(\theta)$ is the log-likelihood for θ as defined in (2.1.10). Of course, as noted at the beginning of this section, $\hat{\theta}_n$ may not exist (as in the example of Kiefer and Wolfowitz (1956)), or it may exist, but be inconsistent (see, e.g., Kraft and Le Cam (1956), Ferguson (1982), or Le Cam (1990)). However, the "usual" Cramér-type smoothness assumptions, typically involving boundedness of third derivatives of $\mathbf{l}$, imply that if $\hat{\theta}_n$ is well defined it satisfies

$$n^{-1/2} \sum_{i=1}^{n} \dot{\mathbf{l}}(X_i, \hat{\theta}_n) = S_n(\hat{\theta}_n) = 0, \tag{4}$$

and, if it is consistent, it is also asymptotically efficient; see, e.g., Cramér (1946, section 33.3), Lehmann (1983, sections 6.2 and 6.3), or Ibragimov and Has'minskii (1981, section III.3). (The conditions of the latter authors for asymptotic efficiency of the maximum likelihood estimate are the weakest.)

Even if the maximum likelihood estimate $\hat{\theta}_n$ does not exist, we can define a one-step Newton-Raphson approximate "solution" of (4) by

$$\hat{\theta}_n^{\text{approx}} = \tilde{\theta}_n + \left[-\frac{1}{n} \sum_{i=1}^{n} \ddot{\mathbf{l}}(X_i, \tilde{\theta}_n) \right]^{-1} \frac{1}{n} \sum_{i=1}^{n} \dot{\mathbf{l}}(X_i, \tilde{\theta}_n) \tag{5}$$

(assuming existence of the Hessian matrix $\ddot{\mathbf{l}}$). This is the basis of the construction of an efficient estimator given below which, by avoiding use of $\ddot{\mathbf{l}}$, always works for regular parametric models.

In fact, it follows from the work of Le Cam (1960), (1969), (1970), that regularity of the model together with existence of a $\sqrt{n}$-consistent preliminary estimator is enough to guarantee the existence of an efficient estimator. When the sample space is Euclidean, existence of uniformly $\sqrt{n}$-consistent estimators is guaranteed by theorem 1. Here is an admittedly artificial construction that is motivated by and refines the one-step approximate solution (5).

(i) Construct $\tilde{\theta}_n$ uniformly $\sqrt{n}$-consistent as in theorem 1.

(ii) Form a grid of cubes with sides of length $c\,n^{-1/2}$ over R^k, and, given $\tilde{\theta}_n$, define θ_n^* to be the midpoint of the cube into which $\tilde{\theta}_n$ has fallen (with some consistent rule for the boundaries of cubes); then θ_n^* is also uniformly $\sqrt{n}$-consistent. This discretization was introduced by Le Cam (1956, page 144).

(iii) As in section 2.3, let

$$\tilde{\mathrm{I}}(\cdot, \theta) = \tilde{\mathrm{I}}(\cdot, P_\theta \mid \theta, \mathbf{P}) = I^{-1}(\theta)\dot{\mathrm{I}}(\cdot, \theta),$$

and define

$$\hat{\theta}_n = \theta_n^* + n^{-1}\sum_{i=1}^{n} \tilde{\mathrm{I}}(X_i, \theta_n^*). \tag{6}$$

Theorem 2. If $\mathbf{P}$ is a regular parametric model and if there exists a uniformly (respectively locally) $\sqrt{n}$-consistent estimator $\tilde{\theta}_n$ of θ, then the estimator $\hat{\theta}_n$ given in (6) is a uniformly (respectively locally) efficient estimator of θ.

Proof. Suppose that $\theta_n \to \theta$ for $\theta_n \in \Theta$. In view of theorem 2.3.1 and (2.1.14) of proposition 2.1.2, it suffices to show that, for all $\varepsilon > 0$,

$$\text{(a)} \qquad P_{\theta_n}\Big(\Big|\sqrt{n}(\hat{\theta}_n - \theta_n) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \tilde{\mathrm{I}}(X_i, \theta_n)\Big| \geq \varepsilon\Big) \to 0$$

as $n \to \infty$. Now it follows from the definition (6) of $\hat{\theta}_n$ that for any $M > 0$ the left side of (a) is bounded by

$$\text{(b)} \qquad P_{\theta_n}(A_n) + P_{\theta_n}\Big(\Big|\sqrt{n}(\theta_n^* - \theta_n) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \{\tilde{\mathrm{I}}(X_i, \theta_n^*) - \tilde{\mathrm{I}}(X_i, \theta_n)\}\Big| \geq \varepsilon, A_n^c\Big)$$

$$\text{(c)} \qquad \leq P_{\theta_n}(A_n) + \sum_{\theta_n'} P_{\theta_n}\Big(\Big|\sqrt{n}(\theta_n' - \theta_n) + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \{\tilde{\mathrm{I}}(X_i, \theta_n') - \tilde{\mathrm{I}}(X_i, \theta_n)\}\Big| \geq \varepsilon\Big),$$

where

$$A_n = [\sqrt{n}\,|\theta_n^* - \theta_n| > M],$$

and where the sum in (c) is over all θ_n' in the grid which are at distance at most $Mn^{-1/2}$ from θ_n. Now choose M so large that the limsup on n of the first term is arbitrarily small; this is possible by the uniform $\sqrt{n}$-consistency of $\tilde{\theta}_n$ and the choice of θ_n^*. Since the number of summands in the second term is bounded, it suffices to show that for all θ_n' satisfying $|\theta_n' - \theta_n| \leq Mn^{-1/2}$ we have

$$\text{(d)} \qquad \sqrt{n}(\theta_n' - \theta_n) + I^{-1}(\theta_n')S_n(\theta_n') - I^{-1}(\theta_n)S_n(\theta_n) = o_{P_{\theta_n}}(1).$$

But this follows from (2.1.15) of proposition 2.1.2 and the continuity of $\theta \to I(\theta)$.

For the local version of the theorem, the argument is the same with $\theta_n = \theta + t_n/\sqrt{n}$ with $|t_n|$ bounded. □

The construction (6) uses $I(\theta_n^*)$ as an estimator of $I(\theta)$. The theorem remains valid if $I(\theta)$ is estimated instead by

$$\hat{I}_n = \hat{I}_n(\theta_n^*) = n^{-1} \sum_{i=1}^{n} \dot{l}\dot{l}^T(X_i, \theta_n^*),$$

as is easily proved by use of (d) of the proof of lemma A.9.5 and Chung's uniform law of large numbers, theorem A.7.3.

Another (also artificial) construction of an efficient estimator will be given in section 7.8 as a corollary of a different construction using sample splitting in place of discretization.