3 | Information Bounds for Euclidean Parameters in Infinite-Dimensional Models

3.1 INTRODUCTION AND OVERVIEW

Our point of view is that of Stein (1956) as developed by Koshevnik and Levit (1976), Levit (1975), (1978), Pfanzagl and Wefelmeyer (1982), and Begun, Hall, Huang, and Wellner (1983). We consider a general $\mathbf{P} \subset \mathbf{M}$, a parameter $\nu : \mathbf{P} \to R^m$ which we wish to estimate, and a fixed $P_0 \in \mathbf{P}$. Recall that for a fixed sample space $(\mathbf{X}, \mathcal{A})$, $\mathbf{M}$ is the set of all probability measures dominated by the σ-finite measure μ.

Definition 1. We call any subset $\mathbf{Q}$ of $\mathbf{P}$ which has a regular Euclidean parametrization a *regular parametric submodel* of $\mathbf{P}$. $P_0 \in \mathbf{P}$ is regular if it belongs to a regular parametric submodel of $\mathbf{P}$.

For example, $\{ N(\mu,\sigma^2) : \mu \in R,\ \sigma^2 > 0 \}$ is a regular parametric submodel of {all absolutely continuous symmetric distributions on R }.

We say that an estimate T is *locally regular on* $\mathbf{P}$ if it is locally regular on all regular parametric submodels $\mathbf{Q}$ of $\mathbf{P}$. *Local* and *uniform Gaussian regularity* and *linearity* are defined similarly.

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

Suppose $m = 1$. If T is locally Gaussian regular as an estimate of ν on $\mathbf{P}$ with variance $\Sigma(P_0, T)$, and $\mathbf{Q}$ is a regular parametric submodel which contains P_0, then

$$(1) \qquad \Sigma(P_0, T) \ \geq\ I^{-1}(P_0 \mid \nu, \mathbf{Q})$$

if ν considered as a function of the parameter indexing $\mathbf{Q}$ is smooth enough so that $I^{-1}(P_0 \mid \nu, \mathbf{Q})$ is defined. It is then natural to define

$$(2) \qquad I^{-1}(P_0 \mid \nu, \mathbf{P}) \ =\ \sup\{ I^{-1}(P_0 \mid \nu, \mathbf{Q}) : \text{regular parametric } \ \mathbf{Q} \subset \mathbf{P} \}$$

as a measure of the best we can achieve in estimating ν over $\mathbf{P}$.

Definition (2) creates no contradictions if $\mathbf{P}$ is itself a regular parametric model for which $I^{-1}(P_0 \mid \nu, \mathbf{P})$ is defined in the sense of (2.3.1). Indeed, by theorems 2.5.1 and 2.5.2, if $\mathbf{X}$ is Euclidean and ν is identifiable, we can then construct an efficient estimate T^* of ν on $\mathbf{P}$. Such an estimate is Gaussian regular on any regular parametric submodel $\mathbf{Q}$ of $\mathbf{P}$. Therefore,

$$I^{-1}(P_0 \mid \nu, \mathbf{P}) = \Sigma(P_0, T^*) \geq I^{-1}(P_0 \mid \nu, \mathbf{Q}).$$

Since $\mathbf{P}$ is a regular parametric submodel of itself, definitions (2.3.1) and (2) are consistent.

Our main goal can now be stated as exhibiting or showing how to construct estimates T of ν which

(i) are locally Gaussian regular on $\mathbf{P}$;

(ii) have $\Sigma(P_0, T) = I^{-1}(P_0 \mid \nu, \mathbf{P})$.

A further goal is to construct locally Gaussian regular estimates T which satisfy (ii) for all $P_0 \in \mathbf{P}$.

In fact, there may *not* exist any locally regular T which achieve the bounds. Ritov and Bickel (1990) give an example in which $I^{-1}(P_0 \mid \nu, \mathbf{P})$ is well defined and positive, but no $\sqrt{n}$-consistent, much less locally regular, estimate of ν exists.

In this chapter we shall give conditions on $\mathbf{P}$ and ν which enable us, in principle, to find regular parametric submodels $\mathbf{Q}$ for which $I^{-1}(P_0 \mid \nu, \mathbf{P})$ is assumed or at least a sequence $\{\mathbf{Q}_j\}$ of such models such that $I^{-1}(P_0 \mid \nu, \mathbf{Q}_j) \to I^{-1}(P_0 \mid \nu, \mathbf{P})$. Note that such a least favorable $\mathbf{Q}$ will typically not be unique and will depend on P_0, ν, $\mathbf{P}$. We are then able to define I^{-1} for $m > 1$ as well. In chapter 5 we extend these notions to infinite-dimensional ν. In chapters 4 and 6 we discuss the construction of different types of semiparametric models and apply the methods of chapters 3 and 5 to a large number of examples.

Exhibition of $\mathbf{Q}$ or $\{\mathbf{Q}_j\}$ still leaves the major problem of constructing or exhibiting T which satisfy (i) and (ii). In particular models, locally regular T have often been proposed on other grounds and can be checked against (ii). Examples appear in this chapter and chapters 4–6. Various general methods of estimation are discussed and applied in chapter 7.

To calculate least favorable $\mathbf{Q}$ we pick up the theme of section 2.4. We characterize $\mathbf{Q}$ and associated efficient score and influence functions by projection on tangent spaces, appropriately defined closed linear subspaces of $L_2(\mu)$ or $L_2(P_0)$ associated with P_0 and $\mathbf{P}$. Tangent spaces are introduced in section 2. Their calculation is discussed there and for special cases throughout chapters 3–6. The extension of the ideas of section 2.4 is the topic of sections 3 and 4.

The machinery we develop in theorems 3.3.1 and 3.4.1 is sufficient but not necessary to achieve our goals of obtaining T satisfying (i) and (ii). These methods of finding $\mathbf{Q}$ can and shall be used heuristically without necessarily checking regularity conditions. The $\mathbf{Q}$ so obtained if it is a regular parametric submodel can always be used in providing a bound for the asymptotic variance of any locally regular estimate. This point of view is developed in the final example of section 4.

We conclude this section with a cautionary example showing that I^{-1} can be infinite, implying that there are parameters not estimable at rate $n^{-1/2}$.

There is an enormous literature on the subject of best rates of convergence and how to achieve them in various contexts. The geometry behind the spectrum of such rates is explored by Donoho and Liu (1987), (1991a,b), who also put the $n^{-1/2}$-rate domain in context.

Example 1. Density estimation.

Let $\mathbf{P}$ consist of all distributions on R with continuous densities. Let

$$\nu(P) = p(0),$$

where p is the density of P. Then

$$I^{-1}(P \mid \nu, \mathbf{P}) = \infty \tag{3}$$

for all P. Simple nonparametric restrictions on p such as bounds on higher derivatives, analyticity do not help. This is in line with various results in density estimation showing only rates of convergence slower than $\sqrt{n}$ are possible in such models. Surprisingly, however, there does exist a nonparametric model $\mathbf{P}$ (Ibragimov and Has'minskii (1982)) for which I^{-1} is finite. We establish (3) for $p(0) > 0$. Define a parametric submodel, valid for $|\eta| < 1$ by

$$p(x, \eta) = p(x)(1 + \eta h(x)), \tag{4}$$

where

$$\sup_x |h(x)| \le 1, \qquad \int h(x) p(x)\, dx = 0\,. \tag{5}$$

Then

$$\nu(P_\eta) = p(0)(1 + \eta h(0))\,.$$

It is easy to check regularity of $\mathbf{Q} = \{P_\eta\}$ by, e.g., proposition 2.1.1 and calculate

$$I^{-1}(P_0 \mid \nu, \{P_\eta\}) = p^2(0) h^2(0) / \int h^2(x) p(x)\, dx\,.$$

It is also easy to exhibit h_j satisfying (5) and $h_j(0) = 1$ with $\int h_j^2(x) p(x)\, dx \to 0$ as $j \to \infty$, so that the sup in (2) is ∞. Bounds on derivatives of any order on $p(\cdot, \eta)$ only translate into smoothness conditions on h for small $|\eta|$ and thus (3) holds quite generally. This method can be extended to yield optimal rates of convergence; see, e.g., Ibragimov and Has'minskii (1981, section IV.5, pp. 237–240). □

Notation. Quantities such as $I(P \mid \nu, \mathbf{P})$, $\tilde{\mathbf{l}}(\cdot, P \mid \nu, \mathbf{P})$ occur frequently in the sequel. We will often suppress one or more of their arguments, the state of nature, P, the parameter ν, or the model $\mathbf{P}$, when convenient.

3.2 TANGENT SPACES

In this section we will study $\mathbf{P}$ as a subset $\mathbf{S}$ of the Hilbert space $L_2(\mu)$ and as a subset of $L_2(P_0)$, via the correspondences $P \longleftrightarrow s$ and $P \longleftrightarrow r$ of chapter 2. Our aim is to identify regular parametric submodels and score functions, as geometrical objects. We need some definitions whose geometrical origins are evident.

Let $\mathbf{H}$ be a Hilbert space with norm $\|\cdot\|$ and inner product $<\cdot,\cdot>$, and $\mathbf{V}$ a subset of $\mathbf{H}$. We establish several conventions:

Convention 1. If W is a subset or vector of elements of $\mathbf{H}$, let $lin(W)$, $\overline{W}$, and $[W]$ denote the linear span of W, the closure of W, and the closed linear span of W respectively.

Convention 2. If we write

$$a_n = b_n + O(\varepsilon_n),$$

where $a_n, b_n \in \mathbf{H}$, $\varepsilon_n \geq 0$ we mean $\|a_n - b_n\| = O(\varepsilon_n)$. The same convention applies to o notation.

Definition 1. $\mathbf{V}$ is a (k-dimensional) *surface* in $\mathbf{H}$ if it can be represented as the image of the open unit sphere in R^k under a continuously Fréchet differentiable map which is of rank k. That is, we can write

$$\mathbf{V} = \{v(\eta) : |\eta| < 1\},$$

where

(i) $v(\eta + \Delta) = v(\eta) + \Delta^T \dot{v}(\eta) + o(|\Delta|)$,

(ii) $\dot{v} = (\dot{v}_1, \dots, \dot{v}_k)^T \in \mathbf{H}^k$,

(iii) $\dim[\dot{v}] = k$.

Any surface or curve (k = 1) has many representations (parametrizations); however if $\eta \to v(\eta)$ and $\gamma \to g(\gamma)$ are two representations of $\mathbf{V}$, $g(\gamma_0) = v(\eta_0)$, and $v^{-1}(v(\eta_0)) = \eta_0$, then $\dot{v}(\eta_0) = M\dot{g}(\gamma_0)$ where $M_{k\times k}$ is nonsingular. This follows by an application of the chain rule and inverse function theorem (e.g. Dieudonné (1960, theorem 10.2.5)). Therefore $[\dot{v}(\eta_0)] = [\dot{g}(\gamma_0)]$ is independent of the parametrization. We call it the *tangent space* of $\mathbf{V}$ at $v(\eta_0)$ and write it $\dot{\mathbf{V}}$ (or $\dot{\mathbf{V}}(v_0)$ when we need to identify $v_0 = v(\eta_0)$). For convenience we will identify a surface or curve with a representation $v(\cdot)$ and for a curve call $\dot{v}(\eta)$ its tangent at $v(\eta)$.

If we take $\mathbf{H} = L_2(\mu)$ and $\mathbf{V} = \mathbf{S}$ our discussion and example A.2.1 lead to:

Proposition 1.

A. $\mathbf{P} = \{P_\theta : \theta \in \Theta \subset R^k\}$ is a regular parametric model and Θ is a surface in R^k if and only if $\mathbf{S}$ is a k-dimensional surface in $L_2(\mu)$.

B. In this case we can represent the projection operator onto $\dot{\mathbf{S}}$ by

(1) $$\Pi(h \mid \dot{\mathbf{S}}) = 4 < h, \dot{s} >^T I^{-1} \dot{s},$$

where we define $\dot{s} = (\dot{s}_1, \dots, \dot{s}_k)^T$ and hence $< h, \dot{s} > = (< h, \dot{s}_1 >, \dots, < h, \dot{s}_k >)^T$.

Now consider the image of $\mathbf{S}$ under the mapping

(2) $$s \to r = 2(\frac{s}{s_0} - 1)1_{[s_0 > 0]},$$

which maps s_0 into zero. This is a subset of $L_2(P_0)$. The mapping (2) may not be one-to-one, and the location of the image obviously depends on P_0. Nevertheless because all information bound calculations are local we do not lose anything by identifying this image with $\mathbf{P}$. By a familiar abuse of notation, we even call this set $\mathbf{P}$. Then

$$\dot{\mathbf{P}} = [\frac{2\dot{s}}{s_0}] = [\dot{\mathbf{l}}_1, \dots, \dot{\mathbf{l}}_k] \tag{3}$$

by (2.1.4). Moreover

$$\Pi_0(h \mid \dot{\mathbf{P}}) = \langle h, \dot{\mathbf{l}} \rangle_0^T I^{-1} \dot{\mathbf{l}}$$

where $\dot{\mathbf{l}} = (\dot{\mathbf{l}}_1, \dots, \dot{\mathbf{l}}_k)^T$, $\langle h, \dot{\mathbf{l}} \rangle_0$ is now the vector of inner products and the subscript 0 refers to operations in the Hilbert space $L_2(P_0)$. Note that for any $\dot{\mathbf{S}}$ and corresponding $\dot{\mathbf{P}}$, the projection operators in the two Hilbert spaces are related by the identities

$$\Pi_0(h \mid \dot{\mathbf{P}}) = s_0^{-1} \Pi(h s_0 \mid \dot{\mathbf{S}}) \qquad \text{for } h \in L_2(P_0), \tag{4a}$$

$$\Pi(t \mid \dot{\mathbf{S}}) = s_0 \Pi_0(\frac{t}{s_0} \mid \dot{\mathbf{P}}) \qquad \text{for } t \in L_2(\mu). \tag{4b}$$

Note that (4a) and (4b) are equivalent to the isomorphism of $\dot{\mathbf{P}}$ and $\dot{\mathbf{S}}$.

We return to the general nonparametric case.

Definition 2. If $\mathrm{v}_0 \in \mathbf{V} \subset \mathbf{H}$, let the *tangent set* at v_0 be the union of all the (one-dimensional) tangent spaces of curves $\mathbf{C} \subset \mathbf{V}$ passing through v_0, and denote it $\dot{\mathbf{V}}^0$. We call the closed linear span $[\dot{\mathbf{V}}^0]$ of the tangent set $\dot{\mathbf{V}}^0$ the *tangent space* of $\mathbf{V}$, and denote it by $\dot{\mathbf{V}}$.

If $\mathbf{V}$ is a k-dimensional surface, $\dot{\mathbf{V}}^0 = \dot{\mathbf{V}}$ and $\dot{\mathbf{V}}$ agrees with our previous definition. In general, $\dot{\mathbf{V}}^0$ is a union of one-dimensional spaces. For example, let

$$\begin{aligned}\mathbf{P} = \{&N(\mu, 1) : \mu \in R\} \\ &\cup \{(1-\varepsilon)N(0,1) + \varepsilon\, U(-\tfrac{1}{2}, \tfrac{1}{2}) : |\varepsilon| < \phi(\tfrac{1}{2})/(1-\phi(\tfrac{1}{2}))\}\end{aligned}$$

where $U(-\frac{1}{2}, \frac{1}{2})$ is the uniform distribution on $(-\frac{1}{2}, \frac{1}{2})$. At $P_0 = N(0,1)$, $\dot{\mathbf{P}}^0$ is the union of the linear space generated by the identity and that generated by $x \to 1_{[-1/2, 1/2]}(x)/\phi(x) - 1$.

Note that, in general, $\dot{\mathbf{V}}^0$ is linear if any two curves contained in $\mathbf{V}$ passing through v_0 are both contained in some surface contained in $\mathbf{V}$ (at least in a neighborhood of v_0). The following proposition will be very useful in later chapters in calculating tangent spaces $\dot{\mathbf{S}}$ and $\dot{\mathbf{P}}$.

Proposition 2. If $\overline{\dot{\mathbf{V}}^0}$ is linear, then $\overline{\dot{\mathbf{V}}^0} = \dot{\mathbf{V}}$.

Proof. The inclusion $\overline{\dot{\mathbf{V}}^0} \subset \dot{\mathbf{V}}$ is trivial since $\dot{\mathbf{V}}^0 \subset lin(\dot{\mathbf{V}}^0)$, and hence

$\overline{\dot{\mathbf{V}}^0} \subset \overline{lin(\dot{\mathbf{V}}^0)} = [\dot{\mathbf{V}}^0] = \dot{\mathbf{V}}$. To prove the reverse inclusion, note that if $\overline{\dot{\mathbf{V}}^0}$ is linear, then $\overline{\dot{\mathbf{V}}^0} \supset lin(\dot{\mathbf{V}}^0)$. Hence $\overline{\dot{\mathbf{V}}^0} \supset \overline{lin(\dot{\mathbf{V}}^0)} = \dot{\mathbf{V}}$. □

In all examples of interest to us, the closures $\overline{\dot{\mathbf{P}}^0}$ and $\overline{\dot{\mathbf{S}}^0}$ of $\dot{\mathbf{P}}^0$ and $\dot{\mathbf{S}}^0$ are linear spaces; thus proposition 2 will be of frequent use in the subsequent developments. An important feature of general $\dot{\mathbf{S}}$ and $\dot{\mathbf{P}}$ is given by:

Proposition 3. If $t \in \dot{\mathbf{S}}$

$$< t, s_0 > = \int t s_0 \, d\mu = 0,$$

and, if $h \in \dot{\mathbf{P}}$,

$$< h, 1 >_0 = \int h \, dP_0 = 0.$$

For $t \in \dot{\mathbf{S}}^0$, $h \in \dot{\mathbf{P}}^0$, these are just restatements of (2.1.8)–(2.1.9). Note that (4a) and (4b) continue to hold for general $\dot{\mathbf{S}}$ and $\dot{\mathbf{P}}$. We continue to deal with both of these isomorphic spaces because each has its advantages; $\dot{\mathbf{S}}$ in examples such as 3.3.2, $\dot{\mathbf{P}}$ in most of the examples we deal with.

Calculation of Tangent Spaces

If h is tangent at P_0 and $h \longleftrightarrow \{P_\eta : |\eta| < 1\}$, then h is the derivative in P_0-measure of $log(dP_\eta / d\mu)$ at $\eta = 0$. Therefore (by, e.g., Loève (1977, section 9.1, page 153)),

$$h = \lim_j \left\{ \log(dP_{\eta_j}/d\mu) - \log(dP_0/d\mu) / \eta_j \right\} \quad \text{a.s.} \quad P_0$$

for some sequence $\eta_j \to 0$. This suggests that we approach the calculation of $\dot{\mathbf{P}}$ heuristically by

(i) Calculating $(\partial/\partial\eta)\log(dP_\eta/d\mu)\,|_{\eta=0}$ for smooth families $\{P_\eta\}$.

(ii) Intersecting the set of all functions obtained in (i) with $L_2(P_0)$ and forming the closed linear span of the result.

Although this closed linear space may be larger than $\dot{\mathbf{P}}$ and, conceivably, since h need only be a derivative in measure, could be smaller, it coincides with $\dot{\mathbf{P}}$ in most of the substantive examples we study. The class of $\{P_\eta\}$ which need be considered can often be reduced. For instance, suppose $\mathbf{P} = \{P_G : G \in \mathbf{G}\}$ and $\mathbf{G}$ is a convex subset of a linear space. Then the closed linear span of

$$\left\{ \frac{\partial}{\partial\eta} \log p(\,\cdot\,, (1-\eta)G_0 + \eta G_1)|_{\eta=0} : G_1 \in \mathbf{G} \right\}$$

usually agrees with $\dot{\mathbf{P}}$ at P_{G_0}.

Here is another useful principle. If $\mathbf{P} = \{P_{(\alpha,\beta)} : \alpha \in A, \beta \in B\}$, $P_0 = P_{(\alpha_0,\beta_0)}$, and $\dot{\mathbf{P}}_1$ is the tangent space at P_0 to the set $\mathbf{P}_1 = \{P_{(\alpha,\beta_0)} : \alpha \in A\}$ and $\dot{\mathbf{P}}_2$ is the tangent space at P_0 to $\mathbf{P}_2 = \{P_{(\alpha_0,\beta)} : \beta \in B\}$, then $\dot{\mathbf{P}}$ typically agrees with $\dot{\mathbf{P}}_1 + \dot{\mathbf{P}}_2$. This principle is discussed further below in the context of the symmetric location problem. All of these heuristics are used at least implicitly in our subsequent calculations of tangent spaces.

Here are the first important examples in which $\dot{\mathbf{S}} = \overline{\dot{\mathbf{S}}^0}$ and $\dot{\mathbf{P}} = \overline{\dot{\mathbf{P}}^0}$ can be calculated explicitly. We calculate either $\dot{\mathbf{S}}$ and $\Pi(\,\cdot\,|\,\dot{\mathbf{S}})$ and then pass to $\dot{\mathbf{P}}$ and $\Pi_0(\,\cdot\,|\,\dot{\mathbf{P}})$ via (3) and (4) or calculate in the opposite direction, depending on which option is most convenient.

Example 1. All probabilities dominated by μ, $\mathbf{P} = \mathbf{M}_\mu$.

In this case, by proposition 3,

$$\dot{\mathbf{P}} \subset \{h \in L_2(P_0) : \int h\, dP_0 = 0\} \equiv L_2^0(P_0)\,. \tag{5}$$

In fact equality holds in (5).

The claimed identity follows if we exhibit a dense subset of $\{h \in L_2(P_0) : \int h\, dP_0 = 0\}$ which is contained in $\dot{\mathbf{P}}^0$. In particular, consider h such that h is bounded and $\int h\, dP_0 = 0$. To see that this set is dense, approximate general h with $\int h\, dP_0 = 0$ by

$$h_M - \int h_M\, dP_0\,,$$

where $h_M = h\, 1_{[\,|h| \le M]}$. For such h define

$$p(\eta) = \exp[\eta h - b(\eta)]\, p_0\,, \qquad \eta \in R\,, \tag{6}$$

where

$$b(\eta) = \log \int \exp[\eta h]\, p_0\, d\mu\,.$$

Now $\eta \to p(\eta) \in \mathbf{P}$ is an exponential family, and, for $h \neq 0$, regular with

$$\dot{s}(0) = \frac{1}{2} h\, s_0\,, \tag{7}$$

since $\dot{b}(0) = \int h\, dP_0 = 0$. Consequently equality holds in (5) in view of (2.1.4). Equivalently

$$\dot{\mathbf{S}} = \{t \in L_2(\mu) : t = t\, 1_{[s_0 > 0]}\,,\ < t, s_0 > \,=\, 0\}\,. \tag{8}$$

For future reference we note that, by (A.2.16),

$$\begin{aligned} \Pi(t\,|\,\dot{\mathbf{S}}) &= t\, 1_{[s_0 > 0]} - < t, s_0 > s_0 \quad \text{and} \\ \Pi_0(h\,|\,\dot{\mathbf{P}}) &= h - \int h\, dP_0\,. \end{aligned} \tag{9}$$

It is not necessary to restrict attention to bounded h for this example (and the next). Given $h \in L_2(P_0)$, $\int h\, dP_0 = 0$, let $\psi : R \to (0, \infty)$ be a bounded continuously differentiable function with bounded derivative ψ' and with $\psi(0) = \psi'(0) = 1$ and ψ'/ψ bounded; for example $\psi(x) = 2(1 + e^{-2x})^{-1}$. Define

$$p^*(\eta) = p_0\, \psi(\eta h) \ / \int \psi(\eta h)\, dP_0\,.$$

Now, $p^*(0) = p_0$, and since

$$\frac{\partial}{\partial\eta}\int \psi(\eta h)\, dP_0 \mid_{\eta=0} = \int h\, dP_0 = 0 ,$$

we have

$$\frac{\partial}{\partial\eta}\log p^*(\eta)\mid_{\eta=0} = h .$$

It is easy to see, using proposition 2.1.1, that for $h \neq 0$ and ε sufficiently small $\{P_\eta :\ dP_\eta/d\mu = p^*(\eta),\ |\eta|<\varepsilon\}$ is a regular parametric model with score function h, and we again arrive at equality in (5). Note that this second construction in fact shows that $\dot{\mathbf{P}}^0 = \dot{\mathbf{P}}$. □

Example 2. All probabilities symmetric about θ_0 fixed.

If $s(\eta) = p^{1/2}(\eta)$ given by (6) is symmetric about θ_0 for all η, so is the L_2-limit of $(s(\eta) - s(0))/\eta$ since the L_2-limit is an a.e. (μ) limit for some sequence $\eta_j \to 0$. The second construction of example 1 now yields

$$\begin{aligned}\dot{\mathbf{P}} &= \{ h \in L_2(P_0) : \int h\, dP_0 = 0,\ h(x) = h(2\theta_0 - x) \quad \text{a.s.} \quad P_0 \} \\ &= \dot{\mathbf{P}}^0 .\end{aligned}$$

Here we have

(10) $$\Pi_0(h \mid \dot{\mathbf{P}})(x) = \frac{1}{2}(h(x) + h(2\theta_0 - x)) - \int h\, dP_0$$

and

$$\Pi(t \mid \dot{\mathbf{S}}) = \frac{1}{2}(t(x) + t(2\theta_0 - x))\, 1_{[s_0 > 0]} - \langle t, s_0 \rangle s_0 .$$

It is easy to verify (10), using proposition A.3.1, upon noting that

$$\{ h : h \text{ symmetric about } \theta_0 \} = \{ h : h \text{ measurable } \mathcal{B}_0 \},$$

where $\mathcal{B}_0$ is the $\sigma-$ field induced by the function $x \to |x - \theta_0|$. □

Example 3. Constraint defined models.

This is an example where it is convenient to first calculate $\dot{\mathbf{S}}$. Suppose that

(11) $$\mathbf{P} = \{ P << \mu : \gamma_i(P) = 0,\ i = 1,\ldots, r \}.$$

Suppose γ_i is pathwise Fréchet differentiable with derivative $\dot{\gamma}_i$ when viewed as a map from $\mathbf{S}$ to R. That is, abusing notation in a familiar way, as $s \in \mathbf{S}$ tends to s_0,

(12) $$\gamma_i(s) = \gamma_i(s_0) + \langle \dot{\gamma}_i(s_0), s - s_0 \rangle + o(\| s - s_0 \|)$$

with $\dot{\gamma}_i(s_0) \in L_2(\mu)$. An example of such a constraint for P absolutely continuous Lebesgue and supported on a fixed compact is knowledge that the variance of X is σ_0^2, so that

$$\gamma(P) = Var_P(X) - \sigma_0^2 .$$

A similar, but more complicated constraint is studied in section 3.3.

Note that

$$\dot{\mathbf{S}} \subset \{t \in L_2^0(\mu) : t = t\, 1_{[s_0 > 0]}\,,\ < t, \dot{\gamma}_i(s_0) > \ = 0,\ i = 1,\dots, r\},\tag{13}$$

since for any curve $s(\eta) = s_0 + \eta\, t + o(\eta) \in \mathbf{S}$,

$$0 = \gamma_i(s(\eta)) - \gamma_i(s_0) = \eta < \dot{\gamma}_i(s_0), t > \ + \ o(\eta).$$

Equality in (13) holds under further conditions. Here is the case $r = 1$.

Suppose $s \to \dot{\gamma}(s)$ is continuous at s_0 and $\dot{\gamma}(s_0) \notin [s_0]$. There exists $t_1 \in L_2(\mu)$ such that

$$< t_1, s_0 > \ = 0, \qquad < t_1, \dot{\gamma}(s_0) > \ = 1.\tag{14}$$

Given t such that

$$< t, s_0 > \ = 0, \qquad < t, \dot{\gamma}(s_0) > \ = 0\,,$$

let

$$p(\eta, \varepsilon) = p_0\, \psi\left(\varepsilon \frac{t_1}{s_0} + \eta \frac{t}{s_0}\right) \ / \ b(\eta, \varepsilon)$$

where ψ is as in example 1,

$$b(\eta, \varepsilon) = \int p_0\, \psi\left(\varepsilon \frac{t_1}{s_0} + \eta \frac{t}{s_0}\right) d\mu\,.$$

But, if $s(\eta, \varepsilon) = p^{1/2}(\eta, \varepsilon)$ we have, by $< t_1, s_0 > \ = \ < t, s_0 > \ = 0$,

$$s(\eta, \varepsilon) = s(0) + \varepsilon t_1 + \eta t + o(\varepsilon) + o(\eta)\,.$$

Then, by (12)

$$\begin{aligned} \gamma(s(\eta,\varepsilon)) &= \ < \dot{\gamma}(s_0), \varepsilon t_1 + \eta t > \ + \ o(\varepsilon) + o(\eta) \\ &= \varepsilon + o(\varepsilon) + o(\eta)\,. \end{aligned}$$

Consequently $\varepsilon \to \gamma(s(\eta,\varepsilon))$ has a root $\varepsilon(\eta)$ with

$$\varepsilon(\eta) = o(\eta)\,.$$

Then for $|\eta|$ small, $\eta \to s(\eta, \varepsilon(\eta))$ is a curve in $\mathbf{S}$ with the required tangent t, and equality in (13) and $\dot{\mathbf{S}} = \overline{\dot{\mathbf{S}}^0}$ follow.

Here, by example A.2.1 and formula (A.2.11), if $t\, 1_{[s_0 > 0]} = t$,

$$\begin{aligned} \Pi(t \mid \dot{\mathbf{S}}) &= t - \Pi(t \mid [s_0, \dot{\gamma}(s_0)]) \\ &= t - < t, s_0 > s_0 - \Pi(t \mid [\dot{\gamma}(s_0) - < \dot{\gamma}(s_0), s_0 > s_0]) \\ &= t - < t, s_0 > s_0 - c(t)(\dot{\gamma}(s_0) - < \dot{\gamma}(s_0), s_0 > s_0), \end{aligned}\tag{15}$$

where

$$c(t) = \frac{< t, \dot{\gamma}(s_0) - < \dot{\gamma}(s_0), s_0 > s_0 >}{\| \dot{\gamma}(s_0) - < \dot{\gamma}(s_0), s_0 > s_0 \|^2} .$$

It follows that $\dot{\mathbf{P}} = \overline{\dot{\mathbf{P}}^0} = \{h \in L_2(P_0) : \int h \, dP_0 = 0, \int h \, \dot{\lambda} \, dP_0 = 0\}$ and

$$\Pi_0(h \mid \dot{\mathbf{P}}) = h - \int h \, dP_0 - \frac{Cov_0(h, \dot{\lambda})}{Var_0(\dot{\lambda})}(\dot{\lambda} - \int \dot{\lambda} \, dP_0), \tag{16}$$

where $\dot{\lambda} \equiv \dot{\gamma}(s_0)/2s_0$.

This argument can be extended to $r > 1$ if the maps $s \to \dot{\gamma}_i(s)$, $i = 1, \ldots, r$, are continuous at s_0 and $\dot{\gamma}_i(s_0)$, $i = 1, \ldots, r$, and s_0 are linearly independent. □

Characterizing tangent spaces for semiparametric models is much harder. If $\mathbf{P} = \{P_{(\theta, G)} : \theta \in \Theta, G \in \mathbf{G}\}$, let

$$\mathbf{P}_1(G_0) = \{P \in \mathbf{P} : \theta \in \Theta, \ G = G_0\},$$

$$\mathbf{P}_2(\theta_0) = \{P \in \mathbf{P} : G \in \mathbf{G}, \ \theta = \theta_0\}$$

be the two sections of $\mathbf{P}$ at $(\theta_0, G_0) \in \Theta \times \mathbf{G}$.

Define $\dot{\mathbf{P}}_1$, $\dot{\mathbf{P}}_2$ to be the tangent spaces corresponding to $\mathbf{P}_1$, $\mathbf{P}_2$. By definition, $\dot{\mathbf{P}}_1$, $\dot{\mathbf{P}}_2$ are both contained in $\dot{\mathbf{P}}$. Therefore

$$\dot{\mathbf{P}} \supset \dot{\mathbf{P}}_1 + \dot{\mathbf{P}}_2 . \tag{17}$$

Formally we expect the inclusion to be an equality. To see this, suppose

$$p(\eta) = p(\theta(\eta), G(\eta))$$

represents a curve through P_0. Differentiating pointwise formally we get

$$\frac{1}{p_0} \frac{\partial}{\partial \eta} p(\eta) |_{\eta=0} = \frac{1}{p_0} \frac{\partial}{\partial \eta} p(\theta_0, G(\eta)) |_{\eta=0} + \frac{\partial \theta(\eta)}{\partial \eta} \frac{1}{p_0} \frac{\partial}{\partial \theta} p(\theta, G_0) |_{\theta = \theta_0} .$$

The right side formally belongs to $\dot{\mathbf{P}}_1 + \dot{\mathbf{P}}_2$, and so we expect that

$$\dot{\mathbf{P}} = \dot{\mathbf{P}}_1 + \dot{\mathbf{P}}_2 . \tag{18}$$

In fact it is possible that $\dot{\mathbf{P}}_1 + \dot{\mathbf{P}}_2$ is not closed if both $\dot{\mathbf{P}}_1$ and $\dot{\mathbf{P}}_2$ are infinite-dimensional; see section A.4. We can and do verify (18) and $\dot{\mathbf{P}} = \dot{\mathbf{P}}^0$ for the symmetric location model.

Example 4. The symmetric location model.

We make $P \longleftrightarrow (\theta, G)$ where θ is the center of symmetry of P and G is the distribution of $X - \theta$, symmetric about zero. We restrict G so that its density $G' = g$ is absolutely continuous with derivative g' and the Fisher information for location I_g is finite:

$$I_g = \int \frac{[g']^2}{g}(x) \, dx .$$

If $s(\eta)$ is a curve passing through s_0, we write

$$s(\eta) = s_\eta(\,\cdot\, - \theta(\eta))\,.$$

Without loss of generality take $\theta(0) = 0$ and let $t = \dot{s}(0)$ so that

(19) $$s_\eta(\,\cdot\, - \theta(\eta)) - s_0(\,\cdot\,) = \eta t + o(\eta)\,.$$

By corollary A.5.1

$$\dot{\mathbf{S}}_1 = [s_0'] = \dot{\mathbf{S}}_1^0\,,$$

where $s_0' = -g' g^{-1/2}/2$, and by example 2 and the second construction of example 1,

$$\dot{\mathbf{S}}_2 = \{ t \in L_2(\mu) : t \text{ symmetric about } 0\,,\ t 1_{[s_0 > 0]} = t\,,\ \langle t, s_0 \rangle = 0 \}$$

$$= \dot{\mathbf{S}}_2^0 \supset \{ t \in \dot{\mathbf{S}}_2 : t \text{ absolutely continuous with derivative } t' \in L_2(\mu)\}.$$

Of course, μ is Lebesgue measure here. Now (18) is equivalent to $\dot{\mathbf{S}} \subset \dot{\mathbf{S}}_1 + \dot{\mathbf{S}}_2$, which holds if,

(20) $$t \in \dot{\mathbf{S}}_1 + \dot{\mathbf{S}}_2$$

for all $t \in \dot{\mathbf{S}}^0$. Note that $\dot{\mathbf{S}}_1 + \dot{\mathbf{S}}_2$ is closed since $\dot{\mathbf{S}}_1$ is one-dimensional.

Proof of (20). Now

(a) $$\theta(\eta) \to 0$$

by weak convergence of $P_{(\theta(\eta), G_\eta)}$ to $P_{(0, G_0)}$, and

(b) $$s_\eta = s_0(\,\cdot\, + \theta(\eta)) + \eta t + o(\eta)\,,$$

in the L_2 sense. To see (b), note that

$$\| s_\eta - s_0(\,\cdot\, + \theta(\eta)) - \eta t \|$$

$$= \| s_\eta(\,\cdot\, - \theta(\eta)) - s_0 - \eta t(\,\cdot\, - \theta(\eta)) \|$$

$$\le \| s_\eta(\,\cdot\, - \theta(\eta)) - s_0 - \eta t \| + |\eta| \, \| t - t(\,\cdot\, - \theta(\eta)) \|$$

$$= o(\eta)$$

by (19), (a), and the L_2–continuity theorem A.1.1 applied to t. Also

(c) $$\dot{\mathbf{S}}_2^\perp = [s_0] + \mathbf{L}_1 + \mathbf{L}_2\,,$$

where

$$\mathbf{L}_1 = \{ h \in L_2(\mu) : h(x) = -h(-x) \text{ a.e. } \mu \},$$

$$\mathbf{L}_2 = \{ h \in L_2(\mu) : h 1_{[s_0 > 0]} = 0 \}\,.$$

Therefore, since

$$\Pi(h \mid \mathbf{L}_1)(x) = \frac{1}{2}(h(x) - h(-x))$$

by (A.2.5) (note that $h(x) - \Pi(h \mid \mathbf{L}_1) = \frac{1}{2}(h(x) + h(-x)) \perp \mathbf{L}_1$),

$$\Pi\left(\frac{s_\eta(\,\cdot\,) - s_0(\,\cdot\, + \theta(\eta))}{\eta} \mid \mathbf{L}_1\right)$$

$$= \frac{s_\eta(\,\cdot\,) - s_\eta(-\,\cdot\,) - s_0(\,\cdot\, + \theta(\eta)) + s_0(-\,\cdot\, + \theta(\eta))}{2\eta}$$

$$\text{(d)} \qquad = -\frac{s_0(\,\cdot\, + \theta(\eta)) - s_0(\,\cdot\, - \theta(\eta))}{2\eta}$$

by symmetry of s_η and s_0. By (b) and (d)

$$\text{(e)} \qquad D_\eta \equiv \left\| \frac{s_0(\,\cdot\, + \theta(\eta)) - s_0(\,\cdot\, - \theta(\eta))}{2\eta} \right\| = O(\|t\|) = O(1)\,.$$

But by corollary A.5.1 again

$$\text{(f)} \qquad D_\eta = \frac{1}{2}\left|\frac{\theta(\eta)}{\eta}\right| I^{1/2}(G_0)(1 + o(1))\,,$$

and comparison of (e) and (f) yields $|\theta(\eta)/\eta| = O(1)$. Since $t \perp s_0$ and $t = t1_{[s_0 > 0]}$, (c), (b), (d), (f), and example A.3.1 yield

$$\Pi(t \mid \dot{\mathbf{S}}_2^{\perp}) = \Pi(t \mid \mathbf{L}_1)$$

$$= -\frac{s_0(\,\cdot\, + \theta(\eta)) - s_0(\,\cdot\, - \theta(\eta))}{2\eta} + o(1)$$

$$= -\frac{\theta(\eta)}{\eta} s_0'(\,\cdot\,) + o(1)\,,$$

and hence

$$\Pi(t \mid \dot{\mathbf{S}}_2^{\perp}) \in \dot{\mathbf{S}}_1\,.$$

Thus

$$t = \Pi(t \mid \dot{\mathbf{S}}_2^{\perp}) + \Pi(t \mid \dot{\mathbf{S}}_2) \in \dot{\mathbf{S}}_1 + \dot{\mathbf{S}}_2\,. \qquad \square$$

3.3 INFORMATION BOUNDS VIA DERIVATIVES OF FUNCTIONS: THE NONPARAMETRIC APPROACH

In this section we show how to obtain information bounds and efficient influence functions for parameters which are explicitly defined as functions on $\mathbf{P}$. In the next section we consider parameters which are defined implicitly through a parametrization. The approaches are equivalent, but each has its own advantages depending on the example considered.

Suppose that $\mathbf{P}$ is a regular parametric, semiparametric, or nonparametric model, and that $\nu : \mathbf{P} \to R^m$ is a Euclidean parameter. Let ν also denote the corresponding map from $\mathbf{S}$ to R^m given by $\nu(s(P)) = \nu(P)$.

Definition 1. Let $m = 1$. ν is *pathwise differentiable* on $\mathbf{S}$ at s_0 if there exists a bounded linear functional $\dot{\nu}(s_0) : \dot{\mathbf{S}} \to R$ such that

(1) $\nu(s(\eta)) = \nu(s_0) + \eta\, \dot{\nu}(s_0)(t) + o(|\eta|)$

for any curve $s(\cdot)$ in $\mathbf{S}$ which passes through $s_0 = s(0)$ and has $\dot{s}(0) = t$. We define the bounded linear functional $\dot{\nu}(P_0) : \dot{\mathbf{P}} \to R$ by

(2) $\dot{\nu}(P_0)(h) = \dot{\nu}(s_0)(\frac{1}{2} h\, s_0)$.

Then (1) holds if and only if

(3) $\nu(P_\eta) = \nu(P_0) + \eta\, \dot{\nu}(P_0)(h) + o(|\eta|)$,

where $\{P_\eta\}$ is the curve in $\mathbf{P}$ corresponding to $\{s(\eta)\}$ in $\mathbf{S}$ and $h = 2t/s_0$; see (3.2.3). We call (3) *pathwise differentiability* of ν on $\mathbf{P}$ at P_0. For convenience in what follows we ignore the dependence of $\dot{\nu}(P_0)$ on P_0 and write $\dot{\nu}(h)$ for $\dot{\nu}(P_0)(h)$, and likewise $\dot{\nu}(t)$ for $\dot{\nu}(s_0)(t)$.

The functional $\dot{\nu}$ is by (3) uniquely defined on $\dot{\mathbf{P}}^0$ and hence on $\dot{\mathbf{P}}$. Often the parameter ν is described most naturally as the restriction to $\mathbf{P}$ of a parameter ν_e defined on a larger model $\mathbf{M}_0 \supset \mathbf{P}$. Suppose ν_e is pathwise differentiable on $\mathbf{M}_0$ with derivative $\dot{\nu}_e : \dot{\mathbf{M}}_0 \to R$. Necessarily

(4) $\dot{\nu}_e = \dot{\nu}$ on $\dot{\mathbf{P}}$.

By the Riesz representation theorem (see example A.1.8) there exists a unique element of $\dot{\mathbf{P}}$ which, abusing notation we also call $\dot{\nu}$, such that

(5) $\dot{\nu}(h) = \langle \dot{\nu}, h \rangle_0$

for all $h \in \dot{\mathbf{P}}$. Strictly speaking the element is $\dot{\nu}^T(1)$ where $\dot{\nu}^T : R \to \dot{\mathbf{P}}$ is the adjoint of $\dot{\nu} : \dot{\mathbf{P}} \to R$. If $\dot{\nu}_e \in \dot{\mathbf{M}}_0$ is similarly defined, (4) implies that

(6) $\dot{\nu} = \Pi_0(\dot{\nu}_e \mid \dot{\mathbf{P}})$.

This follows since for all $h \in \dot{\mathbf{P}}$,

$$\begin{aligned}\langle \dot{\nu} - \Pi_0(\dot{\nu}_e \mid \dot{\mathbf{P}}),\ h \rangle_0 &= \langle \dot{\nu} - \dot{\nu}_e,\ h \rangle_0 \\ &= \dot{\nu}(h) - \dot{\nu}_e(h) = 0 \quad \text{by (4)}.\end{aligned}$$

Viewed as an element of $L_2(P_0)$, $\dot{\nu}_e$ is called a gradient of ν by Koshevnik, Levit, and Pfanzagl, among others, and $\dot{\nu}$ is the canonical gradient. Figure 1 gives the geometry.

Note that $\dot{\nu}_e \in \dot{\mathbf{M}}$ implies

(7) $\langle \dot{\nu}_e, 1 \rangle_0 = 0$.

Of course this is true for $\dot{\nu}$ as well. As we shall see, the information bounds and efficient influence function are determined by the canonical gradient, but computation of $\dot{\nu}$ is usually done via (6). We illustrate this with:

Example 1. The symmetric location model.

We use the assumptions and notation of example 3.2.4. The parameter $\nu(P)$ we want to consider is the center of symmetry. There are many representations

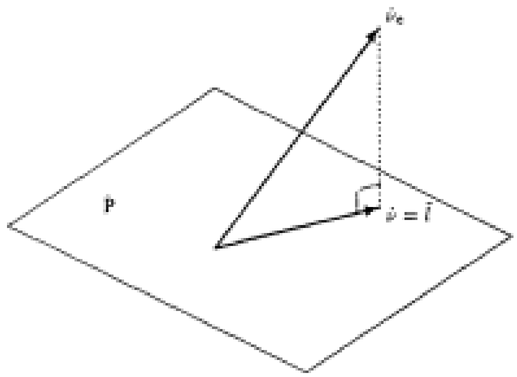

FIGURE 1. Projection of pathwise derivatives.

of ν which can be thought of as parameters on larger models, for instance, the median of P defined as $(F^{-1}(\frac{1}{2}+) + F^{-1}(\frac{1}{2}))/2$ where F is the distribution function corresponding to P and $F^{-1}(t) \equiv \inf\{x : F(x) \geq t\}$. A particularly convenient representation is given by

$$(8) \qquad \nu(P) = \int_0^1 F^{-1}(t)\, w(t)\, dt,$$

where w is symmetric about 1/2, is infinitely differentiable, vanishes off $[1/4, 3/4]$, say, and $\int_0^1 w(s)\, ds = 1$. Formula (8) evidently defines a parameter ν_e on $\mathbf{M}_\mu$ with μ Lebesgue measure which agrees with the center of symmetry on $\mathbf{P}$.

We claim that ν_e is pathwise differentiable on $\mathbf{M}_\mu$, and at P_0 has gradient

$$(9) \qquad \dot{\nu}_e(x) = -\int_0^1 (1_{[u \geq F_0(x)]} - u)\, w(u)\, dF_0^{-1}(u).$$

To see this, write

$$\nu_e(P) = \int_{-\infty}^{\infty} x w(F(x))\, dF(x).$$

Then

$$\begin{aligned} \nu_e(P_\eta) - \nu_e(P_0) = & \int_{-\infty}^{\infty} x\, w(F_0(x))(p_\eta(x) - p_0(x))\, dx \\ & + \int_{-\infty}^{\infty} x\, (w(F_\eta(x)) - w(F_0(x)))\, dF_0(x) \\ & + \int_{-\infty}^{\infty} x\, (w(F_\eta(x)) - w(F_0(x)))(p_\eta(x) - p_0(x))\, dx\, . \end{aligned}$$

If $s_\eta = s_0 + \eta s_0 h/2 + o(\eta)$ for $h \in L_2(P_0)$, then

$$\int |p_\eta - p_0 - \eta h p_0|(x)\, dx = o(\eta)$$

and

$$\sup_x | F_\eta(x) - F_0(x) - \eta \int_{-\infty}^x h(y) p_0(y)\, dy | = o(\eta).$$

Since $w(F_0(x))$ vanishes outside $\{x : 1/4 \le F_0(x) \le 3/4\}$ and $\sup_x | F_\eta(x) - F_0(x) | \to 0$ it is easy to see that

$$\begin{aligned} \nu_e(P_\eta) - \nu_e(P_0) &= \eta \{ \int_{-\infty}^{\infty} x \left[w(F_0(x)) h(x) \right. \\ &\qquad + \left. w'(F_0(x)) \left(\int_{-\infty}^{x} h(y) p_0(y)\, dy \right) \right] p_0(x)\, dx \} \\ &\qquad + o(\eta^2) \\ &= -\eta \int_{-\infty}^{\infty} \left(\int_x^{\infty} w(F_0(y))\, dy \right) h(x) p_0(x)\, dx + O(\eta^2). \end{aligned}$$

We deduce (recalling our convention given in (7)) that

$$\dot{\nu}_e(x) = - \int_x^{\infty} w(F_0(y))\, dy + E_0 \left(\int_X^{\infty} w(F_0(y))\, dy \right)$$

and (9) follows. If $P_0 \in \mathbf{P}$, say $F_0(x) = G_0(x - \theta)$, then it is easy to see (using $\int (1 - 2u) w(u)\, du = 0$) that

(10) $$\dot{\nu}_e(\theta + x) = - \dot{\nu}_e(\theta - x).$$

For simplicity and without loss of generality suppose $\theta = 0$. Then from example 3.2.4 we know that

$$\dot{\mathbf{P}} = \left[\frac{g_0'}{g_0} \right] + \dot{\mathbf{P}}_2,$$

where

$$\dot{\mathbf{P}}_2 = \{ h \in L_2(G_0) : h \ \text{ symmetric about } \ 0, \ \int h\, dG_0 = 0 \}.$$

Let $\psi_0 \equiv - g_0'/g_0$. By (10), $\dot{\nu}_e \perp \dot{\mathbf{P}}_2$, and hence

(11) $$\dot{\nu} = \Pi_0(\dot{\nu}_e \mid [\psi_0]) = \frac{< \dot{\nu}_e, \psi_0 >_0}{\| \psi_0 \|_0^2} \psi_0.$$

A simple calculation yields $< \dot{\nu}_e, \psi_0 >_0 = 1$, in agreement, as we shall see, with (3.4.5). □

Now suppose $\nu = (\nu_1, \ldots, \nu_m)^T$ is m-dimensional. If each ν_i is pathwise differentiable with derivative $\dot{\nu}_i = \dot{\nu}_i(P_0)$, we call ν *pathwise differentiable on* $\mathbf{P}$ *at* P_0 with derivative $\dot{\nu} = (\dot{\nu}_1, \ldots, \dot{\nu}_m)^T$. The remarks we have just made extend obviously to $m > 1$.

For a regular k-dimensional parametric model $\mathbf{Q} = \{ P_\theta : \theta \in \Theta \}$, pathwise differentiability of $\nu(P_\theta) = q(\theta) : R^k \to R^m$ at $P_0 = P_{\theta_0}$ holds if q is differentiable at θ_0 with total derivative matrix $\dot{q}_{m \times k}$. To see this, note that $\dot{\mathbf{Q}} = [\dot{\mathbf{l}}(\theta_0)] \equiv [\dot{\mathbf{l}}]$, so all $h \in \dot{\mathbf{Q}}$ are of the form $h = a^T \dot{\mathbf{l}}$ for some $a \in R^k$.

But with $p_0(\eta) = p(\theta_0 + \eta a)$, $a \in R^k$, $p_0(\eta)$ has tangent $a^T \dot{\mathbf{l}}$, and since

$$(12) \qquad < I^{-1}(\theta_0)\dot{\mathbf{l}}, \dot{\mathbf{l}}^T >_0 \; = \; J \; \equiv \; \text{identity matrix} ,$$

it follows that

$$\nu(P_0(\eta)) = q(\theta_0 + \eta a)$$
$$= q(\theta_0) + \eta\, \dot{q}(\theta_0) a + o(|\eta|)$$
$$(13) \qquad = \nu(P_0) + \eta < \dot{q}(\theta_0) I^{-1}(\theta_0)\dot{\mathbf{l}}, \dot{\mathbf{l}}^T a >_0 + o(|\eta|) .$$

Comparison of (3), (5), and (13) yields

$$(14) \qquad \dot{\nu} = \dot{q}(\theta_0) I^{-1}(\theta_0)\dot{\mathbf{l}}(\theta_0) = \tilde{\mathbf{l}}(\cdot, P_0 \mid \nu, \mathbf{Q}) ,$$

the efficient influence function for ν in $\mathbf{Q}$ as defined in (2.3.2), and hence

$$(15) \qquad < \dot{\nu}, \dot{\nu}^T >_0 = \dot{q}(\theta_0) I^{-1}(\theta_0) \dot{q}^T(\theta_0) = I^{-1}(P_0 \mid \nu, \mathbf{Q}) ,$$

the information bound for ν in $\mathbf{Q}$. (Note that $\dot{\nu}^T$ here is a row vector of elements of $L_2(P_0)$ rather than the adjoint of the mapping $\dot{\nu} : \dot{\mathbf{P}} \to R^m$.)

Theorem 1. Let ν be pathwise differentiable on $\mathbf{P}$ at P_0 with $m = 1$. For any regular parametric submodel $\mathbf{Q}$ for which $I^{-1}(P_0 \mid \nu, \mathbf{Q})$ is defined, the efficient influence function $\tilde{\mathbf{l}}(\cdot, P_0 \mid \nu, \mathbf{Q})$ satisfies

$$(16) \qquad \tilde{\mathbf{l}}(\cdot, P_0 \mid \nu, \mathbf{Q}) = \Pi_0(\dot{\nu}(P_0) \mid \dot{\mathbf{Q}}) ,$$

and consequently

$$(17) \qquad I^{-1}(P_0 \mid \nu, \mathbf{Q}) = \| \Pi_0(\dot{\nu}(P_0) \mid \dot{\mathbf{Q}}) \|_0^2 \le \| \dot{\nu}(P_0) \|_0^2$$

with equality if and only if

$$(18) \qquad \dot{\nu}(P_0) \in \dot{\mathbf{Q}} .$$

Proof. Fix $P_0 \in \mathbf{P}$. For any regular k-dimensional parametric submodel $\mathbf{Q} = \{Q_\eta : \eta \in H\}$ with $Q_0 = P_0 \in \mathbf{P}$ for which $I^{-1}(P_0 \mid \nu, \mathbf{Q})$ is defined, $\eta \to \nu(s(\eta))$ is differentiable at $\eta = 0$, and by the argument leading to (14), the efficient influence function for ν in $\mathbf{Q}$ is

$$(a) \qquad \tilde{\mathbf{l}}(\cdot, P_0 \mid \nu, \mathbf{Q}) = \Pi_0(\dot{\nu}(P_0) \mid \dot{\mathbf{Q}}) .$$

Equation (16) has been proved, and (17) follows from it by Pythagoras since $\dot{\mathbf{Q}} \subset \dot{\mathbf{P}} \subset L_2(P_0)$. □

If $\mathbf{Q}$ is a regular parametric submodel of $\mathbf{P}$ satisfying (18) and hence giving equality in (17), we call $\mathbf{Q}$ *least favorable*. It follows immediately from theorem 1 that under the condition $\dot{\nu}(P_0) \in \dot{\mathbf{P}}^0$ there exists a least favorable $\mathbf{Q}$ which may be chosen one-dimensional.

These results extend to $m \ge 1$ as follows. Let

$$(19) \qquad \tilde{\mathbf{l}}_j = \dot{\nu}_j(P_0) , \qquad j = 1, \ldots, m ,$$
$$\tilde{\mathbf{l}} = (\tilde{\mathbf{l}}_1, \ldots, \tilde{\mathbf{l}}_m)^T .$$

Corollary 1. Suppose that:

(i) ν is pathwise differentiable on $\mathbf{P}$ at P_0.

(ii) $\mathbf{Q}$ is any regular parametric submodel for which $I^{-1}(P_0 \mid \nu, \mathbf{Q})$ is defined.

Then the efficient influence function $\tilde{\mathbf{l}}(\cdot, P_0 \mid \nu, \mathbf{Q})$ for ν in $\mathbf{Q}$ satisfies (16) and, consequently,

$$
(20) \qquad I^{-1}(P_0 \mid \nu, \mathbf{Q}) = \langle \Pi_0(\dot{\nu}(P_0) \mid \dot{\mathbf{Q}}), \Pi_0(\dot{\nu}^T(P_0) \mid \dot{\mathbf{Q}}) \rangle_0 \\ \leq \langle \tilde{\mathbf{l}}, \tilde{\mathbf{l}}^T \rangle_0
$$

in the order $A \leq B$ if $B - A$ is nonnegative definite. Equality holds if and only if

$$
(21) \qquad \tilde{\mathbf{l}}_j \in \dot{\mathbf{Q}}, \qquad j = 1, \dots, m,
$$

and, in that case, the efficient influence function $\tilde{\mathbf{l}}(\cdot, P_0 \mid \nu, \mathbf{Q})$ equals $\tilde{\mathbf{l}}$. Furthermore, if

(iii) $[\tilde{\mathbf{l}}_j : j = 1, \dots, m] \subset \dot{\mathbf{P}}^0$,

then for each $a \in R^m$ there exists a one-dimensional regular parametric submodel $\mathbf{Q}$ such that

$$
(22) \qquad a^T I^{-1}(P_0 \mid \nu, \mathbf{Q}) a = a^T E(\tilde{\mathbf{l}} \tilde{\mathbf{l}}^T) a .
$$

Proof. Apply theorem 1 to $\nu(P) \equiv \sum_{i=1}^m a_i \nu_i(P)$ and note that $\Pi_0(\sum_{i=1}^m a_i \dot{\nu}_i(P_0) \mid \dot{\mathbf{Q}}) = \sum_{i=1}^m a_i \Pi_0(\dot{\nu}_i(P_0) \mid \dot{\mathbf{Q}})$. □

A least favorable model in the sense of corollary 1 will exist under our assumptions if any m curves intersecting at s_0 are all contained in an m-dimensional surface passing through s_0 at least in a neighborhood of s_0. If a least favorable submodel exists, there is an m-dimensional least favorable submodel.

As we have seen in theorem 1, condition (iii) yields, for $m = 1$, a least favorable parametric model $\mathbf{Q}$ in which the information and efficient influence function for ν can naturally be identified with the corresponding quantities for $\mathbf{P}$. We extend this idea for general m to situations in which no least favorable $\mathbf{Q}$ exists, but a least favorable sequence $\{\mathbf{Q}_j\}$ does.

Suppose that ν is pathwise differentiable as in theorem 1 and corollary 1, and

$\overline{\text{(iii)}}$ $[\dot{\nu}_j(P_0) : j = 1, \dots, m] \subset \overline{\dot{\mathbf{P}}^0}$.

Under these conditions we define efficient influence functions and the information matrix for ν.

Definition 2. We call $\tilde{\mathbf{l}} = \tilde{\mathbf{l}}(\cdot, P_0 \mid \nu, \mathbf{P})$ defined by (19) the *efficient influence function* for ν:

$$
(23) \qquad \tilde{\mathbf{l}} = \tilde{\mathbf{l}}(\cdot, P_0 \mid \nu, \mathbf{P}) = \dot{\nu}(P_0).
$$

The *information* $I(P_0 \mid \nu, \mathbf{P})$ for ν in $\mathbf{P}$ is defined as the inverse of

(24) $$I^{-1}(P_0 \mid \nu, \mathbf{P}) = \langle \tilde{\mathbf{l}}, \tilde{\mathbf{l}}^T \rangle_0 = E_0(\tilde{\mathbf{l}}\tilde{\mathbf{l}}^T) .$$

Definition 3. If $\hat{\nu}_n$ is a locally Gaussian regular (on $\mathbf{P}$) estimate of ν with $\Sigma(P_0, \hat{\nu}_n) = I^{-1}(P_0 \mid \nu, \mathbf{P})$ we say $\hat{\nu}_n$ is *efficient at* P_0. If $\hat{\nu}_n$ is efficient at all regular P_0, then we say $\hat{\nu}_n$ is *efficient*.

If an efficient $\hat{\nu}_n$ exists, its influence function must be $\tilde{\mathbf{l}}$, and conversely any locally Gaussian regular and linear estimate with influence function $\tilde{\mathbf{l}}$ is efficient. See also proposition 1 below.

Corollary 2. Suppose that conditions (i) and $\overline{\text{(iii)}}$ hold for $\nu: \mathbf{P} \to R^m$. Let $q: R^m \to R^d$ and assume that q is continuously differentiable with derivative $\dot{q}$ (at $\nu(P_0)$). Then $q(\nu(P))$ also satisfies (i) and $\overline{\text{(iii)}}$ and has influence function

(25) $$\tilde{\mathbf{l}}(\,\cdot\,, P_0 \mid q(\nu), \mathbf{P}) = \dot{q}(\nu(P_0))\tilde{\mathbf{l}}(\,\cdot\,, P_0 \mid \nu, \mathbf{P}) .$$

Proof. Pathwise differentiability of $q(\nu(s))$ with derivative $\dot{q}(\nu(s_0))\dot{\nu}(s_0)$ follows from the chain rule for Fréchet derivatives, and the conclusion follows from theorem 1. □

Note that corollary 2 is just a generalization of (2.3.2). Also, note that $\overline{\text{(iii)}}$ holds for *all* ν satisfying (i) if $\overline{\dot{\mathbf{P}}^0} = \dot{\mathbf{P}}$, or, equivalently (see proposition 3.2.2), if $\overline{\dot{\mathbf{P}}^0}$ is a linear space.

The following theorem justifies definitions 2 and 3.

Theorem 2. Suppose that $\nu: \mathbf{P} \to R^m$ is pathwise differentiable on $\mathbf{P}$ at P_0, $\overline{\text{(iii)}}$ holds, and T_n is a locally regular estimate of ν with corresponding limit law $\mathbf{L}_{P_0}$. Then:

A. $\mathbf{L}_{P_0}$ can be represented as the convolution of a $N(0, I^{-1}(P_0 \mid \nu, \mathbf{P}))$ distribution with another distribution on R^m. More generally,

(26) $$\mathbf{L}_{P_0}\begin{pmatrix} \sqrt{n}(T_n - \nu(P_0)) - \dfrac{1}{\sqrt{n}}\displaystyle\sum_{i=1}^n \tilde{\mathbf{l}}(X_i, P_0 \mid \nu, \mathbf{P}) \\ \dfrac{1}{\sqrt{n}}\displaystyle\sum_{i=1}^n \tilde{\mathbf{l}}(X_i, P_0 \mid \nu, \mathbf{P}) \end{pmatrix} \to \mathbf{L}\begin{pmatrix} \Delta_0 \\ Z_0 \end{pmatrix}$$

where $Z_0 \sim N(0, I^{-1}(P_0 \mid \nu, \mathbf{P}))$ and Δ_0 is independent of Z_0. More generally still, if $h \in (\overline{\dot{\mathbf{P}}^0})^l$ then

(27) $$\mathbf{L}_{P_0}\begin{pmatrix} \sqrt{n}(T_n - \nu(P_0)) - \dfrac{1}{\sqrt{n}}\displaystyle\sum_{i=1}^n \tilde{\mathbf{l}}(X_i, P_0 \mid \nu, \mathbf{P}) \\ \dfrac{1}{\sqrt{n}}\displaystyle\sum_{i=1}^n h(X_i) \end{pmatrix} \to \mathbf{L}\begin{pmatrix} \Delta_0 \\ W_0 \end{pmatrix}$$

where $W_0 \sim N(0, Ehh^T)$ and Δ_0 is independent of W_0.

B. $\Delta_0 = 0$ if and only if T_n is asymptotically linear with influence function $\tilde{\mathbf{l}}$, and then T_n is efficient.

The extension (27) of (26) was noticed by Pfanzagl (1989).

Proof. Note that if T_n is locally regular on $\mathbf{P}$, it is locally regular on any one-dimensional submodel $\mathbf{Q}$. Suppose $\dot{\mathbf{l}}$ is the score function for such a model $\{P_\gamma : |\gamma| < 1\}$; that is,

(a) $$s(P_\gamma) = s(P_0) + \frac{1}{2}\gamma \dot{\mathbf{l}} s_0 + o(|\gamma|) .$$

Let $\nu(P_\gamma) \equiv q(\gamma)$. By (i),

(b) $$q(\gamma) = q(0) + \gamma E(\tilde{\mathbf{l}}\dot{\mathbf{l}}) + o(\gamma),$$

where $\tilde{\mathbf{l}}$ is defined by (19). Here and in the sequel, expectations are under $P \equiv P_0$.

Let

(c) $$U_n = \sqrt{n}(T_n - \nu(P)) ,$$

(d) $$V_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\mathbf{l}}(X_i) .$$

Then, by theorem 2.3.1,

$$\mathbf{L}(U_n, V_n) \to \mathbf{L}(U, V)$$

and using (a) and (b), identity (d) in the proof of theorem 2.3.1 becomes, for all $a \in R^m$, $t \in R$,

(e) $$E\exp[ia^TU + tV - \frac{1}{2}t^2E(\dot{\mathbf{l}}^2)] = \exp[ia^TE(\tilde{\mathbf{l}}\dot{\mathbf{l}})t]E\exp[ia^TU] .$$

By $\overline{\text{(iii)}}$, for each $b \in R^m$, $c \in R^l$, there exists a sequence $\{\dot{\mathbf{l}}_{bcj}\}$ of score functions of regular one-dimensional models such that

(f) $$\| b^T\mathbf{l}^* + c^Th - \dot{\mathbf{l}}_{bcj} \|_0 \to 0 \quad \text{as } j \to \infty,$$

where

(g) $$\mathbf{l}^* \equiv I(P \mid \nu, \mathbf{P})\tilde{\mathbf{l}} .$$

Let

$$V_{nj} \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\mathbf{l}}_{bcj}(X_i)$$

and let $\mathbf{L}(U, V_j)$ denote the limit of $\mathbf{L}(U_n, V_{nj})$. Now let

$$V_n = n^{-1/2} \sum_{i=1}^n \mathbf{l}^*(X_i) , \qquad W_n = n^{-1/2} \sum_{i=1}^n h(X_i) .$$

Suppose, for a subsequence which we identify with $\{n\}$, $\mathbf{L}(U_n, V_n, W_n) \to \mathbf{L}(U, V, W)$. Such a subsequence always exists since $\{U_n\}$, $\{V_n\}$, $\{W_n\}$ are tight. Since

$$\sup_n E(b^T V_n + c^T W_n - V_{nj})^2 = \| b^T \dot{\mathbf{l}}^* + c^T h - \dot{\mathbf{l}}_{bcj} \|_0^2 \rightarrow 0$$

as $j \rightarrow \infty$, we must have

$$\mathbf{L}(U, b^T V + c^T W) = \lim_{j \rightarrow \infty} \mathbf{L}(U, V_j).$$

Take $\dot{\mathbf{l}} = \dot{\mathbf{l}}_{bcj}$, $t = 1$, $V = V_j$ in (e). Since

$$E e^{V_j} = \exp[\tfrac{1}{2} E(\dot{\mathbf{l}}_{bcj}^2)] \rightarrow \exp[\tfrac{1}{2} E(b^T \dot{\mathbf{l}}^* + c^T h)^2]$$

$$= E e^{b^T V + c^T W},$$

we can pass to the limit as $j \rightarrow \infty$ to get

(h) $$E \exp\{ ia^T U + b^T V + c^T W - \tfrac{1}{2} b^T I b - b^T E \dot{\mathbf{l}}^* h^T c - \tfrac{1}{2} c^T E h h^T c \}$$

$$= \exp[ia^T b + ia^T E \tilde{\mathbf{l}} h^T c] E \exp[ia^T U],$$

where $I \equiv I(P_0 \mid \nu, \mathbf{P})$. Use analytic continuation in (h) and take $b = -i I^{-1} a$, $c = id$ to get, after some algebra,

(i) $$E \exp[ia^T(U - I^{-1} V) + id^T W]$$

$$= E \exp[ia^T U + \tfrac{1}{2} a^T I^{-1} a] \exp[-\tfrac{1}{2} d^T E h h^T d].$$

Since

$$\mathbf{L}(U - I^{-1} V, W)$$

$$= \lim_n \mathbf{L}(\sqrt{n}(T_n - \nu(P_0)) - n^{-1/2} \sum_{i=1}^n \tilde{\mathbf{l}}(X_i), n^{-1/2} \sum_{i=1}^n h(X_i)),$$

the various parts of the theorem follow from (i) in the same manner as theorem 2.3.1 followed from (e) of that proof. □

Using theorem 2 we strengthen corollary 1 along the lines of proposition 2.4.3.

Proposition 1. Suppose that T_n is an asymptotically linear estimate of $\nu : \mathbf{P} \rightarrow R^m$ at P_0, with influence function ψ. Then:

A. T_n is regular at P_0 if and only if ν is pathwise differentiable with derivative $\dot{\nu}$ and

(28) $$\psi - \dot{\nu} = \psi - \tilde{\mathbf{l}} \perp \dot{\mathbf{P}}.$$

B. Let T_n be regular at P_0 and suppose that $\overline{\text{(iii)}}$ holds. Then T_n is efficient at P_0 if and only if $\psi \in \dot{\mathbf{P}}^m$, and then $\psi = \tilde{\mathbf{l}}$.

The geometrical content of this proposition is apparent from figure 2.

Remark 1. Suppose $T_n = \nu_e(\mathbb{P}_n)$ for an extension $\nu_e : \mathbf{M}_0 \rightarrow R$ where

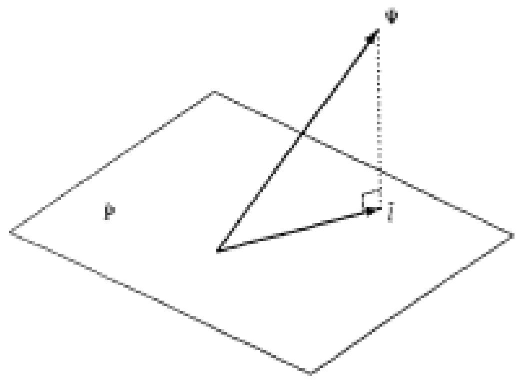

FIGURE 2. Projection of influence functions.

$\mathbf{M}_0 \supset \mathbf{P} \cup \{\text{discrete distributions}\}$. If ν_e is pathwise differentiable and Gâteaux differentiable on $\mathbf{M}_0$, as in (2.2.3), a formal computation suggests that T_n is asymptotically linear with influence function ψ such that

$$\dot{\nu}_e = \psi .$$

By proposition 1, this implies that T_n is regular.

Proof of proposition 1. The idea for the proof of A is the same as that for the proof of proposition 2.4.3.A. Let $\{P_\eta\}$ be an l-dimensional regular parametric submodel with score function $h \in (\dot{\mathbf{P}}^0)^l$ at $\eta = 0$. Let $P_n = P_{\eta_n}$, $\eta_n = t_n/\sqrt{n}$, $t_n \to t$. By the asymptotic linearity of $\{T_n\}$ at P_0 and proposition 2.1.2,

(a) $$\mathbf{L}_{P_0}\begin{pmatrix}\sqrt{n}(T_n - \nu(P_0)) \\ L_n(\eta_n) - L_n(0)\end{pmatrix} \to N\left(\begin{pmatrix}0 \\ -\Sigma_{22}/2\end{pmatrix}, \Sigma\right),$$

with

(b) $$\Sigma = [\Sigma_{ij}] , \quad \Sigma_{11} = E\psi\psi^T , \quad \Sigma_{12} = E\psi h^T t , \quad \Sigma_{22} = t^T E h h^T t .$$

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

(c) $$\mathbf{L}_{P_n}(\sqrt{n}(T_n - \nu(P_0))) \to N(\Sigma_{12}, \Sigma_{11}) .$$

Assume now that T_n is regular. Then

(d) $$\mathbf{L}_{P_n}(\sqrt{n}(T_n - \nu(P_n))) \to N(0, \Sigma_{11}) ,$$

and from (c) and (d) we conclude that

(e) $$\sqrt{n}(\nu(P_n) - \nu(P_0)) \to \Sigma_{12} = E\psi h^T t$$

or

$$\nu(P_{\eta_n}) = \nu(P_0) + \langle \psi, h^T \rangle_0 \eta_n + o(|\eta_n|)$$

for any sequence η_n as above. Now take $l = 1$. Given any sequence of reals, $\tilde{\eta}_m \to 0$, if we take $n_m = [\tilde{\eta}_m^{-2}]$, the integer part of $\tilde{\eta}_m^{-2}$, then $\tilde{\eta}_m = c_m / \sqrt{n_m}$ where $c_m \to 1$. This implies that ν is pathwise differentiable with derivative $\dot{\nu}$ satisfying

$$\dot{\nu}(h) = \langle \psi, h \rangle_0$$

for all $h \in \dot{\mathbf{P}}^0$. Consequently, $\dot{\nu}$ can be defined on all of $\dot{\mathbf{P}}$ by

$$\text{(f)} \qquad \dot{\nu}(h) = \langle \psi, h \rangle_0 = \langle \dot{\nu}, h \rangle_0 \qquad \text{with} \quad \dot{\nu} \in \dot{\mathbf{P}} .$$

But (f) implies (28).

On the other hand, if ν is pathwise differentiable and (28) holds, then (e) is valid, which together with (c) implies (d). This completes the proof of A.

For the proof of B, note that A implies that (i) holds. First suppose that T_n is efficient at P_0. Then definition 3 and theorem 2.A imply that $\Delta_0 = 0$, and hence T_n has influence function $\psi = \tilde{\mathrm{l}}$ by theorem 2.B.

Now suppose that $\psi \in \dot{\mathbf{P}}^m$. Then $\psi - \tilde{\mathrm{l}} \in \dot{\mathbf{P}}^m$. Together with (28) this yields $\psi = \tilde{\mathrm{l}}$, and hence efficiency of T_n at P_0. □

If we examine the proof of this convolution theorem 2, we see that we can weaken the definition of $\dot{\mathbf{P}}^0$, $\dot{\mathbf{S}}^0$, and the corresponding $\dot{\mathbf{P}}$, $\dot{\mathbf{S}}$, and still obtain the conclusions of the theorem. Specifically, let

$$\dot{\mathbf{P}}_w^0 = \{ h \in L_2(P_0) : r(\gamma) = \gamma h + o(\gamma) \}$$

for some mapping $\gamma \to P_\gamma$, $|\gamma| < 1$, and $r(\cdot)$ as defined in (3.2.2). That is, we replace the requirement of continuous Fréchet differentiability of $\gamma \to r(\gamma)$ by Fréchet differentiability at $\gamma = 0$. Evidently $\dot{\mathbf{P}}_w^0 \supset \dot{\mathbf{P}}^0$, and if we define $\dot{\mathbf{P}}_w = [\dot{\mathbf{P}}_w^0]$, we have $\dot{\mathbf{P}}_w \supset \dot{\mathbf{P}}$. If we replace $\dot{\mathbf{P}}$ by $\dot{\mathbf{P}}_w$ in (19), (23), and so forth, theorem 2 continues to hold.

Here are three examples in which we exhibit efficient estimates.

Example 2. Estimation of the mean, P unconstrained.

Suppose that μ concentrates on an interval $[-M, M]$. We want to estimate

$$\nu(P) = \int x \, dP(x) .$$

Identify ν on $\mathbf{S}$ as

$$\nu(s) = \int x s^2(x) \, d\mu(x) .$$

Fix s_0. Then ν is pathwise differentiable, and using (7) in the form $\dot{\nu}(s_0) \perp s_0$ we obtain

$$\text{(29)} \qquad \dot{\nu}(s_0)(x) = 2 s_0(x)(x - E_0 X) \quad \text{a.e. } \mu .$$

To see (29), check

$$\text{(30)} \qquad \nu(s) - \nu(s_0) - \langle \dot{\nu}, s - s_0 \rangle = \int (x - E_0 X)(s - s_0)^2(x) \, d\mu(x)$$
$$= O(\| s - s_0 \|^2) .$$

Hence by (3.2.8), (23), and (29) it follows that

$$\tilde{\mathbf{l}}(x, P_0) = x - E_0 X .$$

Not surprisingly, the sample mean is efficient:

$$\Sigma(P_0, \overline{X}) = \int \tilde{\mathbf{l}}^2(x)\, dP_0(x) .$$

More generally, suppose that **P** satisfies

$$\sup_{P \in \mathbf{P}} E_P X^2 < \infty . \tag{31}$$

Then $\nu(P) \equiv \int x\, dP(x) = E_P X$ is pathwise differentiable on **P** with $\dot{\nu}(s_0)(x) = 2 s_0(x)(x - E_0 X)$ or $\dot{\nu}(P_0) = x - E_0 X = \tilde{\mathbf{l}}(x, P_0)$ by proposition A.5.2. The mean continues to be efficient and locally regular in view of Le Cam's third lemma A.9.3. □

Example 3. Estimation of the mean, P constrained.

Suppose that $\mathbf{X} = [-M, M]$, ν is as in example 1, but now as in example 3.2.3, we constrain **P** by fixing the coefficient of variation at $\sqrt{c_0}$. This is equivalent to

$$\gamma(P) = \int x^2\, dP(x) - (1 + c_0)\nu^2(P) = 0 .$$

The mean ν is, of course, pathwise differentiable on the smaller tangent space of this model, and $\dot{\nu}(P_0)$ is given by (6), where $\dot{\nu}_e(P_0)(x) = x - E_0 X$ by example 1. Check that

$$\dot{\gamma}(P)(x) = (x^2 - \int x^2\, dP(x)) - 2(1 + c_0)\nu(P)(x - \nu(P)) ,$$

where $P \to \dot{\gamma}(P)$ is continuous at P_0. We can apply example 3.2.3, which shows that $\dot{\mathbf{P}} = \{h \in L_2^0(P) :\ h \perp \dot{\gamma}(P)\}$. Thus (23), (3.2.16), and algebra yield

$$\begin{aligned} \tilde{\mathbf{l}}(x, P_0) &= \Pi_0(\dot{\nu}_e \mid \dot{\mathbf{P}}) \\ &= \dot{\nu}_e(P_0) - \frac{\langle \dot{\nu}_e(P_0), \dot{\gamma}(P_0) \rangle_0}{\| \dot{\gamma}(P_0) \|_0^2} \dot{\gamma}(P_0) \\ &= \left[1 + 2(1 + c_0)(E_0 X) a(P_0)\right](x - E_0 X) \\ &\quad - a(P_0)(x^2 - E_0 X^2) \end{aligned} \tag{32}$$

where

$$a(P) \equiv \frac{E(\dot{\nu}(P)\dot{\gamma}(P))}{E(\dot{\gamma}^2(P))} = \frac{Cov[X - EX, X^2 - 2(1 + c_0)(EX)X]}{Var[X^2 - 2(1 + c_0)(EX)X]} . \tag{33}$$

Consider the estimate

(34) $$\hat{\nu}_n = \overline{X} + a(\mathbb{P}_n)\gamma(\mathbb{P}_n),$$

where $\mathbb{P}_n$ is the empirical df. It is straightforward to show that $\hat{\nu}_n$ is uniformly Gaussian regular and efficient. For this model, $\hat{\nu}_n$ improves $\overline{X}$:

$$\Sigma(P_0, \hat{\nu}) = \Sigma(P_0, \overline{X}) - a^2(P_0)Var_0\left[X^2 - 2(1+c_0)(E_0X)X\right]$$
$$= I^{-1}(P_0 \mid \nu, \mathbf{P}).$$

Estimates such as (34) were introduced by Levit (1975). Alternative procedures are discussed in Haberman (1984) and Sheehy (1987), (1988).

Note that $\hat{\nu}_n$ can be identified with $\psi(\mathbb{P}_n)$ where

(35) $$\psi(P) = \int x\, dP + a(P)\gamma(P).$$

We can think of $\psi(P)$ as being a rather unobvious extension of the mean from our constrained $\mathbf{P}$ to $\mathbf{M}$. It is the "right" extension in the sense that $\dot{\psi} \in \dot{\mathbf{P}}$. □

We can generalize example 1 considerably.

Example 4. d_K**-differentiable parameters on** $\mathbf{M} = \{\text{all } P \text{ on } \mathbf{X}\}$.

Suppose that $\mathbf{X} \subset R$ and that ν is a parameter on $\mathbf{M}$ 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

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

If $\mathbf{X} = [-M, M]$ the mean has these properties as well as the functional $\psi(P)$ of (35).

Let $\mathbb{P}_n$ be the empirical distribution. Then $\nu(\mathbb{P}_n)$ is an efficient estimate of ν. To see this, note that by (i) and (ii)

(36) $$\nu(\mathbb{P}_n) = \nu(P) + n^{-1}\sum_{i=1}^{n}\psi(X_i, P) + o_P(n^{-1/2}).$$

Since ψ is continuous, bounded, and continuous in P, the conditions of proposition 2.2.1.D hold, and $\nu(\mathbb{P}_n)$ is regular. By (3.2.9), $\psi(\cdot, P_0) \in \dot{\mathbf{M}}$, and efficiency follows. Note that we do not really need (i) and (ii) for $\nu(\mathbb{P}_n)$ to be efficient here, but merely $\nu(\mathbb{P}_n)$ asymptotically linear. □

3.4 INFORMATION BOUND CALCULATIONS VIA SCORES: THE SEMIPARAMETRIC APPROACH

In semiparametric models parameters are defined implicitly through the parametrization rather than directly as functions on **P**. Part A of proposition 2.4.1 generalizes to such models and gives a way of computing information and influence functions. It is convenient to present this approach in terms of **P** rather than **S**. Let

(1) $$\mathbf{P} = \{P_{(\nu,G)} : \nu \in N,\ G \in \mathbf{G}\},$$

where N is an open subset of R^m and $\mathbf{G}$ is general. Fix $P_0 = P_{(\nu_0,G_0)}$. In agreement with chapter 2, let $\dot{\mathbf{l}}_1$ denote the vectors of partial derivatives of $\log p(\,\cdot\,,\nu,G)$ with respect to ν evaluated at P_0, and this is just the score function for the parametric model

$$\mathbf{P}_1 = \{P_{(\nu,G_0)} : \nu \in N\}\ .$$

We also let

$$\mathbf{P}_2 = \{P_{(\nu_0,G)} :\ G \in \mathbf{G}\}$$

and follow a similar convention for parametric submodels **Q**.

Theorem 1. Suppose that:

(i) $\mathbf{P}_1$ is regular.

(ii) ν is a 1-dimensional parameter.

Let

(2) $$\mathbf{l}_1^* = \dot{\mathbf{l}}_1 - \Pi_0(\dot{\mathbf{l}}_1 \mid \dot{\mathbf{P}}_2)\ .$$

Then:

A. If $\mathbf{Q} = \{P_{(\nu,G_\gamma)} : \nu \in N,\ \gamma \in \Gamma\}$ is a regular parametric submodel of **P** with $P_0 \in \mathbf{Q}$, then

(3) $$I(P_0 \mid \nu, \mathbf{Q}) \geq \|\mathbf{l}_1^*\|_0^2 = E_0(\mathbf{l}_1^{*2})$$

with equality if and only if $\mathbf{l}_1^* \in \dot{\mathbf{Q}}$.

B. If $\dot{\mathbf{P}} = \dot{\mathbf{P}}_1 + \dot{\mathbf{P}}_2$, $\mathbf{l}_1^* \neq 0$ and the assumptions of theorem 3.3.1 apply to **P** and ν, then the efficient influence function for ν, as defined in section 3, is given by

(4) $$\tilde{\mathbf{l}}_1 = \|\mathbf{l}_1^*\|_0^{-2}\,\mathbf{l}_1^* \equiv I^{-1}(P_0 \mid \nu, \mathbf{P})\mathbf{l}_1^*\ .$$

In the same way as (2.4.4) led to the definition of efficient score function in regular parametric models, this theorem leads to

Definition 1. We call $\mathbf{l}_1^*$ the *efficient score function* for ν in **P**, and write it $\mathbf{l}^*(\cdot, P_0 \mid \nu, \mathbf{P})$.

The geometry of the relationship between $\dot{\mathbf{l}}_1$ and $\mathbf{l}_1^*$ is given in figure 3.

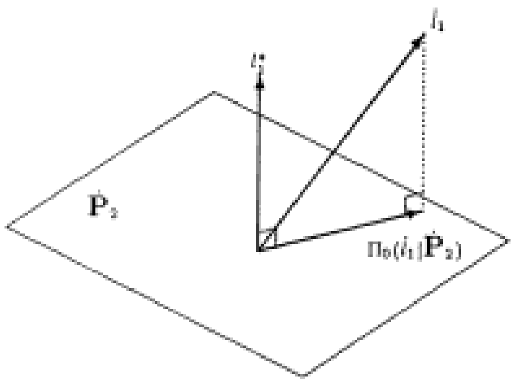

FIGURE 3. Projection of score functions.

Proof of theorem 1.

A. By proposition 2.4.1.A for any regular parametric submodel $\mathbf{Q}$ as above,

$$\begin{aligned} I(P_0 \mid \nu, \mathbf{Q}) &= \| \dot{\mathbf{l}}_1 - \Pi_0(\dot{\mathbf{l}}_1 \mid \dot{\mathbf{Q}}_2) \|_0^2 \\ &= \| \dot{\mathbf{l}}_1 - \Pi_0(\dot{\mathbf{l}}_1 \mid \dot{\mathbf{P}}_2) + \Pi_0(\dot{\mathbf{l}}_1 \mid \dot{\mathbf{P}}_2) - \Pi_0(\dot{\mathbf{l}}_1 \mid \dot{\mathbf{Q}}_2) \|_0^2 \\ &= \| \mathbf{l}_1^* \|_0^2 + \| \Pi_0(\dot{\mathbf{l}}_1 \mid \dot{\mathbf{Q}}_2) - \Pi_0(\dot{\mathbf{l}}_1 \mid \dot{\mathbf{P}}_2) \|_0^2 \end{aligned}$$

since $\mathbf{l}_1^* \perp \dot{\mathbf{P}}_2$ and, since $\dot{\mathbf{Q}}_2 \subset \dot{\mathbf{P}}_2$ and both $\Pi_0(\dot{\mathbf{l}}_1 \mid \dot{\mathbf{P}}_2)$ and $\Pi_0(\dot{\mathbf{l}}_1 \mid \dot{\mathbf{Q}}_2) \in \dot{\mathbf{P}}_2$. The result (3) follows.

Part B of the theorem follows from theorem 3.3.1 and the following proposition, which isolates the essential features of any differentiable function $\nu(P)$ which identifies ν on $\mathbf{P}$ of (1). □

Proposition 1. Suppose that $\mathbf{P}$ is as in (1), that ν is identified by $\nu : \mathbf{P} \to R$ so $\nu(P_{(\nu, G)}) = \nu$ for $P_{(\nu, G)} \in \mathbf{P}$, that $\nu(P)$ is pathwise differentiable on $\mathbf{P}$ with derivative $\dot\nu(P_0)$, and that $\mathbf{P}_1$ is regular. Then

(5) $$\langle \dot\nu(P_0), \dot{\mathbf{l}}_1 \rangle_0 = 1 ,$$

and

(6) $$\dot\nu(P_0) \perp \dot{\mathbf{P}}_2 .$$

If also $\dot{\mathbf{P}} = \dot{\mathbf{P}}_1 + \dot{\mathbf{P}}_2$, then

(7) $$\dot\nu(P_0) = \tilde{\mathbf{l}}(\cdot, P_0 \mid \nu, \mathbf{P}) = \tilde{\mathbf{l}}_1$$

where $\tilde{\mathbf{l}}_1 \equiv I^{-1}(P_0 \mid \nu, \mathbf{P})\mathbf{l}_1^*$ as in (4) with efficient score function $\mathbf{l}_1^*$ given in (2).

Proof. First consider ν restricted to the model $\mathbf{P}_1$. Then for $P_{(\nu, G_0)} \equiv P_\nu \in \mathbf{P}_1$, it follows on the one hand that

(a) $$\nu(P_{(\nu, G_0)}) - \nu(P_0) = \nu - \nu_0 = t$$

by definition of $\nu(P)$, while, on the other hand, pathwise differentiability of ν yields

$$\text{(b)} \qquad \nu(P_{(\nu, G_\nu)}) - \nu(P_0) = t < \dot{\nu}(P_0), \dot{\mathbf{l}}_1 >_0 + o(|t|) .$$

Equality of (a) and (b) for all t yields

$$1 = < \dot{\nu}(P_0), \dot{\mathbf{l}}_1 >_0 ,$$

proving (5). To prove (6), note that $h \in \dot{\mathbf{P}}_2^0$ implies that there exists a one-dimensional regular parametric submodel $\mathbf{Q}_2 = \{ P_{(\nu_0, G_\gamma)} : |\gamma| < 1 \}$ having h as a tangent. Consideration of ν restricted to $\mathbf{Q}_2$ yields, if $P_\gamma \equiv P_{(\nu_0, G_\gamma)}$,

$$0 = \nu(P_\gamma) - \nu(P_0) = \gamma < \dot{\nu}(P_0), h >_0 + o(\gamma)$$

and hence

$$0 = < \dot{\nu}(P_0), h >_0 \qquad \text{for} \quad h \in \dot{\mathbf{P}}_2^0 ,$$

which implies (6).

To prove (7), note that, since $\dot{\nu}(P_0) \in \dot{\mathbf{P}}$,

$$\dot{\nu}(P_0) = a \, \mathbf{l}_1^* + bh,$$

where $h \in \dot{\mathbf{P}}_2$. Here we have used the assumptions that $\dot{\mathbf{P}} = \dot{\mathbf{P}}_1 + \dot{\mathbf{P}}_2$ and $\mathbf{P}_1$ is regular. Then (6) implies that

$$\dot{\nu}(P_0) = a \, \mathbf{l}_1^* ,$$

and (5) then yields

$$1 = a < \mathbf{l}_1^*, \dot{\mathbf{l}}_1 >_0 = a \, \| \mathbf{l}_1^* \|_0^2 .$$

Hence (7) holds. □

Suppose ν is m-dimensional, $m > 1$, with vector efficient score function

$$\text{(8)} \qquad \mathbf{l}_1^* = (\mathbf{l}_{11}^*, \ldots, \mathbf{l}_{1m}^*)^T ,$$

where $\mathbf{l}_{11}^*, \ldots, \mathbf{l}_{1m}^*$ are the score functions of the components of ν given by (2).

Corollary 1. Suppose $\mathbf{P}_1$ is regular.

A. If $\mathbf{Q}$ is a regular parametric submodel as in theorem 1, for which $I^{-1}(P_0 \mid \nu, \mathbf{Q})$ is defined, then

$$\text{(9)} \qquad I(P_0 \mid \nu, \mathbf{Q}) \geq E_0(\mathbf{l}_1^* \mathbf{l}_1^{*T}) ,$$

with equality if and only if $[\mathbf{l}_1^*]$ is contained in $\dot{\mathbf{Q}}$.

B. If $\dot{\mathbf{P}} = \dot{\mathbf{P}}_1 + \dot{\mathbf{P}}_2$ and the conditions of theorem 3.3.2 hold, then the efficient influence function is given by

$$\text{(10)} \qquad \tilde{\mathbf{l}}_1 = I^{-1}(P_0 \mid \nu, \mathbf{P}) \mathbf{l}_1^*$$

and

$$\text{(11)} \qquad I(P_0 \mid \nu, \mathbf{P}) = E_0(\mathbf{l}_1^* \mathbf{l}_1^{*T}) .$$

Proof. Apply proposition 2.4.1.A to $\mathbf{Q} = \{P_{(\nu, G_\gamma)} : \nu \in N,\ \gamma \in \Gamma\}$ to deduce that for all $a \in R^m$,

$$\begin{aligned} a^T I(P_0 \mid \nu, \mathbf{Q}) a &= \| a^T \dot{\mathbf{l}}_1 - \Pi_0 (a^T \dot{\mathbf{l}}_1 \mid \dot{\mathbf{Q}}_2) \|_0^2 \\ &= \| a^T \mathbf{l}_1^* \|_0^2 \\ &\quad + \| a^T (\Pi_0 (\dot{\mathbf{l}}_1 \mid \dot{\mathbf{Q}}_2) - \Pi_0 (\dot{\mathbf{l}}_1 \mid \dot{\mathbf{P}}_2)) \|_0^2 \\ &\geq a^T E_0 (\mathbf{l}_1^* \mathbf{l}_1^{*T}) a , \end{aligned}$$

and A follows.

To prove B we generalize proposition 1. □

Proposition 2. Suppose $\mathbf{P}_1$ is regular, $\nu : \mathbf{P} \to R^m$ is identified as in proposition 1, and that ν is pathwise differentiable on $\mathbf{P}$ with derivative $\dot{\nu}(P_0)_{m \times 1}$ at P_0 and $E_0 \dot{\nu}(P_0) = 0$. Then

(12) $$E_0 (\dot{\nu}(P_0) \dot{\mathbf{l}}_1^T) = J_{m \times m} \equiv \text{the } m \times m \text{ identity}$$

and

(13) $$[\dot{\nu}(P_0)] \perp \dot{\mathbf{P}}_2.$$

If also $\dot{\mathbf{P}} = \dot{\mathbf{P}}_1 + \dot{\mathbf{P}}_2$ holds, then

(14) $$\dot{\nu}(P_0) = \tilde{\mathbf{l}}(P_0 \mid \nu, \mathbf{P}) = \tilde{\mathbf{l}}_1 ,$$

where $\tilde{\mathbf{l}}_1$ is given by (3.3.16) and

(15) $$\tilde{\mathbf{l}}_1 = I^{-1} (P_0 \mid \nu, \mathbf{P}) \mathbf{l}_1^* ,$$

where $I^{-1}(P_0 \mid \nu, \mathbf{P})$ is given in (3.3.17) and

(16) $$I(P_0 \mid \nu, \mathbf{P}) = E_0 (\mathbf{l}_1^* \mathbf{l}_1^{*T}).$$

Proof. Claims (12) and (13) are proved as for proposition 1. Claims (14)–(16) follow by writing $\dot{\nu}(P_0) = A\, \mathbf{l}_1^* + h$, where $A_{m \times m}$ is a matrix and h is an m-vector whose components belong to $\dot{\mathbf{P}}_2$. Then, for any $b \in R^m$,

(a) $$\dot{\nu}^T (P_0) b = \mathbf{l}_1^{*T} A^T b + h^T b .$$

As before, we get from (13) that $h^T b = 0$. Then, by taking the inner product of both sides of (a) with $b^T h$ in $L_2(P_0)$,

$$[E_0 (\dot{\mathbf{l}}_1 \dot{\nu}^T (P_0))] b = E_0 (\dot{\mathbf{l}}_1 \mathbf{l}_1^{*T}) A^T b$$

or from (12) and (13)

$$b = E_0 (\mathbf{l}_1^* \mathbf{l}_1^{*T}) A^T b \qquad \text{for every} \quad b \in R^m .$$

Hence, A and $E_0 (\mathbf{l}_1^* \mathbf{l}_1^{*T})$ are nonsingular and, by (12),

$$A = [E_0 (\mathbf{l}_1^* \mathbf{l}_1^{*T})]^{-1} .$$

So, $\tilde{\mathbf{l}}_1 = \Pi_0 (\dot{\nu}(P_0) \mid \dot{\mathbf{P}}) = [E_0(\mathbf{l}_1^* \mathbf{l}_1^{*T})]^{-1} \mathbf{l}_1^*$, and the proposition follows. □

Now suppose we have Euclidean nuisance parameters present. That is, $\mathbf{P} = \{P_{(\theta, G)} : \theta = (\nu, \eta),\ \nu \in N,\ \eta \in H,\ G \in \mathbf{G}\}$, $N \subset R^m$, $H \subset R^{k-m}$, with N, H open. We follow the usual convention and define

$$
\begin{aligned}
&\mathbf{P}_{12} = \{P_{(\theta, G_0)} :\ \theta \in \Theta\}\,, \\
(17)\qquad &\mathbf{P}_{23} = \{P_{(\nu_0, \eta, G)} :\ \eta \in H\,,\ G \in \mathbf{G}\}\,, \\
&\mathbf{P}_1 = \{P_{(\nu, \eta_0, G_0)} :\ \nu \in N\}\,,
\end{aligned}
$$

and so forth, and follow the same convention for parametric submodels $\mathbf{Q}$. Suppose the parametric model $\mathbf{P}_{12}$ is regular. Let $\dot{\mathbf{l}}_1$, $\dot{\mathbf{l}}_2$ be the score functions for $\mathbf{P}_1$, $\mathbf{P}_2$, and define

$$(18)\qquad \mathbf{l}_1^* \equiv \dot{\mathbf{l}}_1 - \Pi_0(\dot{\mathbf{l}}_1 \mid \dot{\mathbf{P}}_2 + \dot{\mathbf{P}}_3)\,.$$

Note that, since $\dot{\mathbf{P}}_2$ is finite-dimensional, $\dot{\mathbf{P}}_2 + \dot{\mathbf{P}}_3$ is closed, and if $\mathbf{Q} = \{P_{(\theta, G_\gamma)} :\ \theta \in \Theta\,,\ \gamma \in \Gamma\}$ is a regular parametric submodel then

$$\dot{\mathbf{Q}}_{23} = \dot{\mathbf{Q}}_2 + \dot{\mathbf{Q}}_3 \subset \dot{\mathbf{P}}_2 + \dot{\mathbf{P}}_3\,.$$

We can deduce, as in corollary 1, that

$$I(P_0 \mid \nu, \mathbf{Q}) \geq E_0(\mathbf{l}_1^* \mathbf{l}_1^{*T})\,.$$

If further

$$(19)\qquad \dot{\mathbf{P}} = \dot{\mathbf{P}}_1 + \dot{\mathbf{P}}_2 + \dot{\mathbf{P}}_3$$

and the conditions of theorem 3.3.2 hold, then the efficient influence function is given by (10) and (11) with $\mathbf{l}_1^*$ defined in (18). We can use the same argument as in proposition 2.

Note, by applying (A.2.11), that (18) can be calculated by first projecting on $\dot{\mathbf{P}}_2$ and then on the orthocomplement of $\dot{\mathbf{P}}_2$ in $\dot{\mathbf{P}}_2 + \dot{\mathbf{P}}_3$, yielding

$$
\begin{aligned}
(20)\qquad \mathbf{l}_1^* &= \dot{\mathbf{l}}_1 - \Pi_0(\dot{\mathbf{l}}_1 \mid \dot{\mathbf{P}}_2) - \Pi_0(\dot{\mathbf{l}}_1 - \Pi_0(\dot{\mathbf{l}}_1 \mid \dot{\mathbf{P}}_2) \mid (\dot{\mathbf{P}}_2 + \dot{\mathbf{P}}_3) \cap \dot{\mathbf{P}}_2^{\perp}) \\
&= \dot{\mathbf{l}}_1 - \Pi_0(\dot{\mathbf{l}}_1 \mid \dot{\mathbf{P}}_2) - \Pi_0(\dot{\mathbf{l}}_1 \mid (\dot{\mathbf{P}}_2 + \dot{\mathbf{P}}_3) \cap \dot{\mathbf{P}}_2^{\perp})\,,
\end{aligned}
$$

since $\dot{\mathbf{P}}_2 \perp \dot{\mathbf{P}}_2^{\perp}$. Alternatively, we can reverse the roles of $\dot{\mathbf{P}}_2$, $\dot{\mathbf{P}}_3$. In fact, it is often convenient to proceed as follows. Begin by computing the efficient score function of $\theta = (\nu, \eta)$, which we shall call $\mathbf{l}^*$. So

$$(21)\qquad \mathbf{l}^* = \dot{\mathbf{l}} - \Pi_0(\dot{\mathbf{l}} \mid \dot{\mathbf{P}}_3) \equiv \mathbf{l}^*(\,\cdot\,, P_0 \mid \theta, \mathbf{P})\,.$$

If we write $\mathbf{l}^* = (\mathbf{l}_1^*, \ldots, \mathbf{l}_k^*)^T$, it is *not* true, in general, that, say,

$$\mathbf{l}^*(\,\cdot\,, P_0 \mid \nu_1, \mathbf{P}) = \mathbf{l}_1^*\,.$$

However, by reversing the roles of $\dot{\mathbf{P}}_2$ and $\dot{\mathbf{P}}_3$ in (20) and taking $m = 1$ we obtain

$$(22)\qquad \mathbf{l}^*(\,\cdot\,, P_0 \mid \nu_1, \mathbf{P}) = \mathbf{l}_1^* - \Pi_0(\mathbf{l}_1^* \mid [\mathbf{l}_i^* : i \neq 1])\,.$$

The definition 2.4.1 of full adaptation extends naturally to semiparametric models.

Definition 2. If

$$(23)\qquad I(P_0 \mid \nu, \mathbf{P}) = I(P_0 \mid \nu, \mathbf{P}_{12})$$

and there exists a locally Gaussian regular (on $\mathbf{P}$) estimate $\hat{\nu}_n$ which is efficient at P_0, we say that $\hat{\nu}_n$ is *adaptive at* P_0. If $\hat{\nu}_n$ is efficient for all P_0 such that $\mathbf{P}_{12}$ is regular we say $\hat{\nu}_n$ is *adaptive*.

Proposition 3. Under the conditions of corollary 1 the adaptation condition (23) is equivalent to

$$(24)\qquad [\, \dot{\mathbf{l}}_1 - \Pi_0(\dot{\mathbf{l}}_1 \mid \dot{\mathbf{P}}_2)\,] \;\perp\; \dot{\mathbf{P}}_3 \,.$$

Proof. By (11) and (20),

$$I(P_0 \mid \nu, \mathbf{P}) = I(P_0 \mid \nu, \mathbf{P}_{12}) - E_0(WW^T),$$

where

$$W = \Pi_0(\dot{\mathbf{l}}_1 - \Pi_0(\dot{\mathbf{l}}_1 \mid \dot{\mathbf{P}}_2) \mid (\dot{\mathbf{P}}_2 + \dot{\mathbf{P}}_3) \cap \dot{\mathbf{P}}_2^{\perp}) \,.$$

$W = 0$ if and only if (24) holds. □

A useful sufficient condition for (24) is given by:

Corollary 2. If the conditions of corollary 1 hold and $\Pi_0(\dot{\mathbf{l}}_1 \mid \dot{\mathbf{P}}_2) = \Pi_0(\dot{\mathbf{l}}_1 \mid \dot{\mathbf{P}}_3)$, then (23) and (24) hold.

Corollary 3. Let $\mathbf{F}$ be the set of all real-valued functions q of θ such that

(i) q is a function of ν only;

(ii) q has a derivative $\dot{q}_{1 \times k}$.

If $\mathbf{P}_{12}$ is regular, then (24) is equivalent to

$$(25)\qquad (\dot{\mathbf{l}}_1^T, \dot{\mathbf{l}}_2^T) I^{-1}(\theta) \dot{q}^T \;\perp\; \dot{\mathbf{P}}_3$$

for all $q \in \mathbf{F}$.

Proof. The left-hand side of (25) is, by (2.3.2), $\tilde{\mathbf{l}}(\,\cdot\, \mid q, \mathbf{P}_{12})$, which is just a nonsingular transformation of the left side of (24), and hence the conditions are equivalent. □

Examples

Now we illustrate the methods developed in this section by considering two important examples. Many more examples will be given in chapter 4.

Example 1. The symmetric location model.

By example 3.2.4 the conditions of theorem 1 are satisfied here. If $I(G) < \infty$, then, with the notation of example 3.3.1,

$$(26)\qquad \dot{\mathbf{l}}_1(x) = \psi_0(x - \theta_0) = -\frac{g_0'}{g_0}(x - \theta_0)$$

is the efficient score function in $\mathbf{P}_1$. There is no parametric nuisance parameter η and

(27) $$\dot{\mathbf{P}}_3 = \{a(\,\cdot\, - \theta_0) : a \quad \text{symmetric}, \quad E_0 a(X - \theta_0) = 0\}.$$

Since $X - \theta_0$ is symmetric about 0, ψ_0 is odd and

$$E_0 \psi_0 (X - \theta_0) a(X - \theta_0) = 0 \qquad \text{for all} \quad a(\,\cdot\, - \theta_0) \in \dot{\mathbf{P}}_3 .$$

Thus, $\dot{\mathbf{l}}_1 \perp \dot{\mathbf{P}}_3$, $\tilde{\mathbf{l}}_1$ agrees with (3.3.11), and full adaptation to shape is possible in this model. Examples of locally Gaussian regular adaptive estimates have been constructed by Van Eeden (1970), Beran (1974), (1978), Sacks (1975), and Stone (1975) among others. We give a construction in chapter 7. □

Further applications of this method (or at least the point of view) of Begun, et al. (1983) are given in chapter 4. For a systematic development of *score operators* as introduced by Begun, et al. (1983), see sections 5.4 and 5.5. Huang (1982), (1984), Choi (1989), and Choi and Hall (1988) have developed connections between score operators and least favorable submodels as defined in section 3.3. We will not pursue their approach here.

In the following example, and many other examples of interest in chapters 4 and 6, it is difficult to determine $\dot{\mathbf{P}}$ exactly. We now expand the argument in section 3.1 to the effect that such lack of knowledge is not necessarily troublesome for our theory. Suppose that $\mathbf{T}$ is a subspace of $L_2^0(P_0)$ which is a candidate for $\dot{\mathbf{P}}$. Typically we can easily prove either

(28) $$\mathbf{T} \subset \dot{\mathbf{P}}$$

or

(29) $$\mathbf{T} \supset \dot{\mathbf{P}} .$$

Usually the latter inclusion is the most difficult one to establish. For example, if the model involves group structure, mixing, or missing data, then it is easy to verify that $\mathbf{T} \equiv \overline{\mathbf{R}(\dot{\mathbf{l}})} \subset \dot{\mathbf{P}}$ where $\dot{\mathbf{l}}$ is the score operator for the model. (See example 2 below and sections 5.4 and 5.5 for careful definitions and a full development.) But the reverse inclusion $\mathbf{T} \supset \dot{\mathbf{P}}$ is much more difficult; recall the difficulty in showing $\dot{\mathbf{P}} = \dot{\mathbf{P}}_1 + \dot{\mathbf{P}}_2$ in the symmetric location model, example 3.2.4. On the other hand, in the case of constraint-defined models such as in examples 3.2.3, 3.3.3, and section 6.2, (29) is easy to prove for a natural choice of $\mathbf{T}$, but (28) is more difficult and holds only under further hypotheses; recall example 3.2.3.

We argue here that if we are in case (28), have established an information bound based on $\mathbf{T}$, *and* can produce a regular estimator which *achieves* the bound based on $\mathbf{T}$, then proving the reverse inclusion (29) and hence $\dot{\mathbf{P}} = \mathbf{T}$ is largely irrelevant. Here is a simple argument showing why this is true. Note that if $\mathbf{T} \subset \dot{\mathbf{P}}$, then for $h \in L_2^0(P_0)$

(30) $$\| \Pi_0(h \mid \mathbf{T}) \|_0 \le \| \Pi_0(h \mid \dot{\mathbf{P}}) \|_0 .$$

Thus the conjectural or putative information bound $\| \Pi_0(\dot{\nu} \mid \mathbf{T}) \|_0^2$ based on the subspace $\mathbf{T}$ is less than or equal to the honest, true information bound $I^{-1}(\nu \mid P_0, \mathbf{P})$ defined by (3.3.24) which we may not be able to calculate because of lack of knowledge of $\dot{\mathbf{P}}$. Now suppose that $\hat{\nu}_n$ is a Gaussian regular estimator of ν with asymptotic variance $\Sigma(P_0, \hat{\nu}_n)$. Then by the convolution theorem and (30)

$$\Sigma(P_0, \hat{\nu}_n) \geq I^{-1}(\nu \mid P_0, \mathbf{P}) = \| \Pi_0(\dot{\nu} \mid \dot{\mathbf{P}}) \|_0^2 \geq \| \Pi_0(\dot{\nu} \mid \mathbf{T}) \|_0^2 . \tag{31}$$

But if $\hat{\nu}_n$ achieves the bound provided by $\mathbf{T}$; i.e., if

$$\Sigma(P_0, \hat{\nu}_n) = \| \Pi_0(\dot{\nu} \mid \mathbf{T}) \|_0^2 ,$$

then equality must hold throughout (31), and both the bound is sharp and the estimate $\hat{\nu}_n$ is *asymptotically efficient*. Alternatively, if $\hat{\nu}_n$ is asymptotically linear with influence function $\psi \in \mathbf{T}$, then $\psi \in \dot{\mathbf{P}}$ by (28), and by proposition 3.3.1.B, $\hat{\nu}_n$ is again asymptotically efficient. These arguments justify our attitude toward exact determination of the tangent space $\dot{\mathbf{P}}$ in example 2 below and in the examples in chapters 4 and 6 for which the inclusion (28) holds. Of course, the argument should be completed by producing a regular estimator which achieves the bound, as is done in the following example, but not universally for the examples considered in chapters 4 and 6.

Example 2. Cox model without censoring.

We consider a simple version of example 1.3.7 with $C = \infty$ (no censoring) and $m = k = 1$. We observe (Z,T) with T having conditional hazard function given $Z = z \in R$ specified by

$$\lambda(t \mid z) = r(\nu z)\lambda(t), \tag{32}$$

where $\nu \in R$, $\lambda = g/\overline{G}$, $r(z) = exp(z)$, g is the density of G on $[0, \infty)$, and $\overline{G} = 1 - G$. The distribution H of Z is assumed known with density h with respect to some fixed measure m while, with $\mu \equiv$ Lebesgue measure,

$$\mathbf{G} = \{G << \mu\} .$$

Then, if we write $r = r(\nu z)$,

$$P[T \geq t \mid Z = z] = \overline{G}^{\,r}(t) = [\overline{G}(t)]^r$$

and the density with respect to $\mu \times m$ is

$$f(z, t, \nu, G) = r\, g(t) \overline{G}^{\,r-1}(t)\, h(z). \tag{33}$$

Calculation of scores for ν and "for g" is straightforward: differentiation yields

$$\dot{\mathbf{l}}_1(z,t) = z[1 - \Lambda_r(t)], \tag{34}$$

where

$$\Lambda_r(t) \equiv r(\nu z)\Lambda(t) \equiv r(\nu z)\int_0^t \frac{dG}{1-G} .$$

Similarly, letting $\{g_\eta\}$ be a regular parametric family through $g \equiv g_0$ with

$$a \in \{a \in L_2(G) : \int a\, dG = 0\} \equiv \mathbf{H}_0$$

as score for η, the score (operator) for g is

$$(35)\qquad \dot{\mathbf{l}}_2 a(z,t) = a(t) + (r-1)\frac{\int_t^\infty a(s)\,dG(s)}{\overline{G}(t)}\,.$$

For example,

$$g_\eta = \psi(\eta a)g \;\; / \int \psi(\eta a)\,dG$$

will do, if $\psi : R \to R^+$ is bounded and continuously differentiable with bounded derivative ψ' satisfying $\psi(0) = \psi'(0) = 1$ and with ψ'/ψ bounded; see also example 3.2.1.

Formula (35) defines a linear transformation $\dot{\mathbf{l}}_2$ from functions of t to functions of (z,t) which is well defined on $\mathbf{H}_0$, but is, in general, unbounded, and hence its range is not necessarily in $L_2(P)$. (We will show that $\dot{\mathbf{l}}_2$ is a bounded operator if $\nu \geq 0$, $Z \geq 0$, and $Er(\nu Z) < \infty$, and that this can always be arranged by a reparametrization if (46) holds (Z bounded).) Under the same assumption we can show, by arguing as in example 3.1.1, that $\overline{Range(\dot{\mathbf{l}}_2)} \equiv \overline{\mathbf{R}(\dot{\mathbf{l}}_2)} \subset \dot{\mathbf{P}}_2$. Equality of these two sets has not yet been established.

It will be very helpful to rewrite the scores in (34) and (35) in terms of the operators R and L introduced in section A.1. Let $\mathbf{H} \equiv L_2(G)$, and let $\mathbf{H}_0 \equiv \{ a \in \mathbf{H} : \int a\,dG = 0 \}$ as above. Then $L : \mathbf{H} \to \mathbf{H}_0$ and $R : \mathbf{H}_0 \to \mathbf{H}_0$ are defined by

$$(36)\qquad R\,a(t) \equiv a(t) - \frac{\int_t^\infty a\,dG}{1 - G(t)}$$

$$= -E\{ a(Y) - a(t) \mid Y > t \} \equiv -e_a(t)$$

$$\equiv -\text{ mean residual life of } a(Y) \text{ at } t\,,$$

where $Y \sim G$, and

$$(37)\qquad L\,a(t) \equiv a(t) - \int_0^t a\frac{dG}{1-G} = a(t) - \int_0^t a\,d\Lambda,$$

where $\Lambda(t) \equiv \int_0^t (1-G)^{-1}\,dG$ is the cumulative hazard function corresponding to G. The operators R and L arise as the logarithmic derivatives of the maps $Hg \equiv g/\overline{G} \equiv \lambda$ and $D\lambda \equiv \lambda \exp(-\int_0^{\cdot} \lambda(s)\,ds) \equiv g$, respectively. In view of (32) and (33), it is not at all surprising that the scores (34) and (35) can be expressed in terms of R and L and the operators R_r and L_r corresponding to the conditional (given $Z = z$) df $(1 - G)^r$ with hazard rate $\lambda_r \equiv r\lambda$:

$$(38)\qquad R_r a(t) \equiv a(t) - \frac{\int_t^\infty a\,dF(\,\cdot\mid z)}{1 - F(t\mid z)}$$

$$= -E\{ a(Y) - a(t) \mid Z = z,\; Y > t \}\,,$$

and

$$(39) \qquad L_r a(t) \equiv a(t) - \int_0^t a\, d\Lambda(\,\cdot \mid z\,) = a(t) - r(\nu z)\int_0^t a\, d\Lambda\,.$$

Comparison of (34) with (39) yields

$$(40) \qquad \dot{\mathbf{l}}_1(z,t) = z(L_r 1)(t)\,.$$

Furthermore

$$(41) \qquad \dot{\mathbf{l}}_2\, a(z,t) = (L_r R\, a)(z,t)$$

follows from (35), (36), and (39) since $L \cdot Ra = a$ implies

$$a(t) = Ra(t) - \int_0^t Ra\, d\Lambda$$

or, by definition of R,

$$(42) \qquad \frac{\int_t^\infty a\, dG}{1 - G(t)} = a(t) - Ra(t) = -\int_0^t Ra\, d\Lambda\,.$$

Hence, the right side of (35) equals

$$\begin{aligned} Ra(t) + r\frac{\int_t^\infty a\, dG}{1 - G(t)} &= Ra(t) - r\int_0^t Ra\, d\Lambda \qquad \text{by (42)} \\ &= Ra(t) - \int_0^t Ra\, d\Lambda_r \\ &= L_r Ra(t) \qquad \text{by (39)}\,. \end{aligned}$$

Based on these score calculations, we can take several different approaches to establishing information bounds for estimation of ν. Here is a brief sketch of three different approaches:

Method 1. (Via a candidate efficient estimator). Consider a particular estimator $\hat{\nu}_n$ of ν, for example the Cox (1972) partial likelihood estimator. If we show that the given estimator $\hat{\nu}_n$ is (locally) regular and asymptotically linear with an influence function ψ in the tangent space $\dot{\mathbf{P}}$ of the model, then in view of proposition 3.3.1, $\psi = \tilde{\mathbf{l}}_1$ and $\hat{\nu}_n$ is efficient.

Method 2. (Via orthogonality calculations). This approach proceeds without a candidate efficient estimator. Instead, we try to compute $\mathbf{l}_1^*$, and hence $\tilde{\mathbf{l}}_1$, via orthogonality considerations: we try to find $a^* \in \mathbf{H}_0$ so that

$$(43) \qquad \dot{\mathbf{l}}_1 - \dot{\mathbf{l}}_2 a^* \perp \dot{\mathbf{l}}_2\, a \qquad \text{for all} \quad a \in \mathbf{H}_0\,.$$

If $\dot{\mathbf{P}}_2 = \overline{\mathbf{R}(\dot{\mathbf{l}}_2) \cap L_2^0(P)}$, this gives $\dot{\mathbf{l}}_1 - \dot{\mathbf{l}}_2 a^* \perp \dot{\mathbf{P}}_2$, and $\mathbf{l}_1^* = \dot{\mathbf{l}}_1 - \dot{\mathbf{l}}_2 a^*$ as in (2). If not, we at least obtain a lower bound.

Method 3. (Via inversion of $\dot{\mathbf{l}}_2^T \dot{\mathbf{l}}_2$). This approach involves calculation of the projection $\Pi_0(\dot{\mathbf{l}}_1 \mid \dot{\mathbf{P}}_2)$ by inversion of $\dot{\mathbf{l}}_2^T \dot{\mathbf{l}}_2$ and application of theorem A.2.2.

This was the approach taken for the Cox model (with censoring) by Begun, et al. (1983). Inversion of $\dot{\mathbf{l}}_2^T \dot{\mathbf{l}}_2$ is complicated however, and gives much more than is needed for calculation of a bound for estimation of ν; in fact, this is essentially related to calculation of bounds for estimation of G, a theme which will be pursued in section 5.4. Again the sharpness of the bound depends on the validity of $\dot{\mathbf{P}}_2 = \overline{\mathbf{R}(\dot{\mathbf{l}}_2) \cap L_2^0(P)}$.

We now illustrate methods 1 and 2 in the present case of the Cox model without censoring.

Method 1. (Via estimation).

The well-known partial likelihood estimator $\hat{\nu}_n$ of Cox (1972) is our natural candidate for an efficient estimator. If $EZ^2 r^2(\nu Z)$ is bounded uniformly in a neighborhood of ν_0, $\hat{\nu}_n$ is locally linear and asymptotically normal as can be seen by generalizing the proof of Tsiatis (1981); see example 7.4.4. The influence function of $\hat{\nu}_n$ is $\mathbf{l}_1^* \,||\mathbf{l}_1^*||_0^{-2}$, where

$$(44)\qquad \mathbf{l}_1^*(z,t) = \dot{\mathbf{l}}_1(z,t) - \left(\frac{S_1}{S_0}(t) - r(\nu z)\int_0^t \frac{S_1}{S_0}\,d\Lambda\right),$$

and

$$(45)\qquad S_i(t) = E\{Z^i r(\nu Z)\, 1_{[T \ge t]}\}, \qquad i = 0, 1 .$$

Under the assumption

$$(46)\qquad |Z| \le K < \infty \quad \text{a.s.},$$

we shall show that $\mathbf{l}_1^* = \dot{\mathbf{l}}_1 - \dot{\mathbf{l}}_2 a^*$ for $a^* \in \mathbf{H}_0$, and hence $\mathbf{l}_1^* \in \dot{\mathbf{P}}$. In view of proposition 1, this yields the efficiency of Cox's estimator.

Note that (46) implies $|S_1| \le K S_0$, so S_1/S_0 is bounded, and hence $L(S_1/S_0) \in \mathbf{H}_0$. Since $R \cdot La = a$ for $a \in \mathbf{H} \equiv L_2(G)$ we can write $S_1/S_0 = R \cdot L(S_1/S_0)$. But then, in view of the definition (39) of L_r, $\mathbf{l}_1^*$ in (44) can be written as

$$\begin{aligned}
\mathbf{l}_1^*(z,t) &= \dot{\mathbf{l}}_1(z,t) - L_r\frac{S_1}{S_0}(z,t) \\
&= \dot{\mathbf{l}}_1(z,t) - L_r R(L\frac{S_1}{S_0})(z,t) \\
&= \dot{\mathbf{l}}_1(z,t) - \dot{\mathbf{l}}_2\, a^*(z,t) \qquad \text{by (41)}
\end{aligned}$$

with

$$a^*(t) = L(\frac{S_1}{S_0}) \in \mathbf{H}_0 .$$

This establishes the claim: $\mathbf{l}_1^* \in \dot{\mathbf{P}}$ and hence $\hat{\nu}_n$ is efficient.

Method 2. (Via orthogonality).

We want to find an $a^* \in \mathbf{H}_0$ so that (43) holds. To do this, we first express the scores (40) and (41) in terms of a counting process martingale. Now N_r defined by

$$N_r(t) \equiv 1_{[T \le t]}, \qquad t \ge 0,$$

is a counting process. Conditional on $Z = z$, it has a compensator

$$\begin{aligned} \mathbb{A}_r(t) &= \int_0^t 1_{[T \ge s]}\, d\Lambda_r(s) \\ &= r \int_0^t 1_{[T \ge s]}\, d\Lambda(s); \end{aligned}$$

thus

$$\mathbb{M}_r(t) \equiv N_r(t) - \mathbb{A}_r(t)$$

is a martingale with respect to the σ-fields $\mathcal{F}_t \equiv \sigma\{Z, 1_{[T \le s]} : s \le t\}$. Suppose b is a fixed function. Then, by direct calculation

$$\begin{aligned} \int_0^\infty b(s)\, d\mathbb{M}_r(s) &= b(T) - \int_0^\infty b(s)\, 1_{[T \ge s]}\, d\Lambda_r(s) \\ &= b(T) - \int_0^T b(s)\, d\Lambda_r(s) \\ &= L_r b(T) \qquad \text{by (39)}. \end{aligned} \tag{47}$$

Comparison of (47) with (40) and (41) yields

$$\dot{\mathrm{l}}_1(Z, T) = Z \int_0^\infty d\mathbb{M}_r(s) \tag{48}$$

and

$$\dot{\mathrm{l}}_2\, a(Z, T) = \int_0^\infty Ra(s)\, d\mathbb{M}_r(s). \tag{49}$$

To calculate the efficient score function l_1^* for ν, we want to find a function a^* with $\int a^*\, dG = 0$ so that

$$\mathrm{l}_1^* \equiv \dot{\mathrm{l}}_1 - \dot{\mathrm{l}}_2\, a^* \perp \dot{\mathrm{l}}_2\, a \qquad \text{in} \quad L_2(P)$$

for all functions $a \in \mathbf{H}_0$; i.e.,

$$E\{[\dot{\mathrm{l}}_1 - \dot{\mathrm{l}}_2 a^*]\dot{\mathrm{l}}_2 a\} = 0 \qquad \text{for all } a \in \mathbf{H}_0. \tag{50}$$

This is just as in Begun, et al. (1983), except that here we are working in $L_2(P)$ rather than $L_2(\mu)$ and have replaced A by $\dot{\mathrm{l}}_2$, β^* by a^*, and β by a.

Now we use a bit of martingale theory. By conditioning on Z, the expectation in (50) is easily calculated as the expectation of the predictable covariation process of the martingale transforms in (48) and (49); see, e.g., theorem B.3.1, page 891 of Shorack and Wellner (1986). Thus the left side of (50) equals

$$\begin{aligned} &E\,E\{[\dot{\mathrm{l}}_1 - \dot{\mathrm{l}}_2\, a^*]\dot{\mathrm{l}}_2\, a \mid Z\} \\ &= E\,E\{\textstyle\int (Z - Ra^*) Ra\, 1_{[T \ge s]}\, r(\nu Z)\, d\Lambda(s) \mid Z\} \end{aligned}$$

$$= \int \{ E[Z\, r(vZ)\, 1_{[T \geq s]}] - E[r(vZ)\, 1_{[T \geq s]}]\, Ra^*(s) \}\, Ra(s)\, d\Lambda(s)$$

$$(51) \qquad = \int \{ S_1(s) - S_0(s) Ra^*(s) \}\, Ra(s)\, d\Lambda(s)$$

where S_1 and S_0 are as defined in (45). From (51) it is easy to make the right choice of a^*: set

$$(52) \qquad a^* \equiv L(\frac{S_1}{S_0}) .$$

Since $R \circ L$ = identity by proposition A.1.8, it follows that

$$Ra^* = \frac{S_1}{S_0} ,$$

and hence the integrand of (51) is zero identically, and (50) holds. Thus the efficient score function for estimation of v is

$$\mathbf{l}_1^*(Z, T) = \dot{\mathbf{l}}_1(Z, T) - \dot{\mathbf{l}}_2 a^*(Z, T)$$

$$= \int_0^\infty [Z - \frac{S_1(t)}{S_0(t)}]\, d\mathbb{M}_r(t)$$

$$(53) \qquad = \int_0^\infty [Z - E(Z \mid T = t)]\, d\mathbb{M}_r(t),$$

since

$$(54) \qquad \frac{S_1}{S_0}(t) = E(Z \mid T = t)$$

by straightforward calculations. Hence the information for v is, by an easy martingale calculation

$$I(v) = E[\mathbf{l}_1^*(Z, T)^2]$$

$$= E\,E[\int_0^\infty [Z - E(Z \mid T = t)]^2\, 1_{[T \geq t]}\, r(vZ)\, d\Lambda(t) \mid Z]$$

$$= E \int_0^\infty [Z - E(Z \mid T = t)]^2\, r(vZ)\, \overline{G}(t)^{r(vZ)-1}\, dG(t)$$

$$(55) \qquad = E[Z - E(Z \mid T)]^2 = E\, Var(Z \mid T) .$$

This argument can be extended to the censored case. For more on the Cox model, see examples 4.7.1, 5.5.2, and 7.4.4. For information calculations in a family of models containing the Cox model, see Sasieni (1992). □