Introduction to Empirical Processes and Semiparametric Inference Lecture 24: Regularity, Efficiency and Testing

Michael R. Kosorok, Ph.D.

Professor and Chair of Biostatistics

Professor of Statistics and Operations Research

University of North Carolina-Chapel Hill

Consequences of Non-Regularity

Let T_n be asymptotically linear for $\psi(P)$, with influence function $\check{\psi}_P$.

If T_n is not regular, there exists a function $\tilde{g} \in \dot{\mathcal{P}}_P$ such that for each $a \in \mathbb{R}$, there exists and a sequence of contiguous, one-dimensional submodels P_n, for which

$$\sqrt{n}(T_n(h) - \psi(P_n)(h)) \overset{P_n}{\rightsquigarrow} \mathbb{G}\check{\psi}_P(h) + aP\tilde{g}^2. \qquad (1)$$

This means that T_n has a serious defect.

Specifically, we see from (1) that for any $M < \infty$ and $\epsilon > 0$, we can alter a to generate a one-dimensional submodel $\{P_n\}$ for which

$$\text{pr}\left(\left\|\sqrt{n}(T_n - \psi(P_n))\right\|_{\mathcal{H}} > M\right) > 1 - \epsilon.$$

Thus the estimator T_n has arbitrarily poor performance for certain submodels which are represented by $\dot{\mathcal{P}}_P$.

Hence regularity is not just a mathematically convenient definition, but it reflects, even in infinite-dimensional settings, a certain intuitive reasonableness about T_n.

This does not mean that nonregular estimators are never useful, because they can be, especially when the parameter $\psi(P)$ is not $\sqrt{n}$-consistent.

Nevertheless, regular estimators are very appealing when they are available.

More on Efficiency

The next result assures us that Hadamard differentiable functions of efficient estimators are also asymptotically efficient:

THEOREM 1. *Assume that $\psi : \mathcal{P} \mapsto \mathbb{B}$ is differentiable at P relative to the tangent space $\dot{\mathcal{P}}_P$—with derivative $\dot{\psi}_P g$, for every $g \in \dot{\mathcal{P}}_P$, and efficient influence function $\tilde{\psi}_P$—and takes its values in a subset $\mathbb{B}_\phi$.*

Suppose also that $\phi : \mathbb{B}_\phi \subset \mathbb{B} \mapsto \mathbb{E}$ is Hadamard differentiable at $\psi(P)$ tangentially to

$$\mathbb{B}_0 \equiv \overline{\textit{lin}}\, \dot{\psi}_P(\dot{\mathcal{P}}_P).$$

Then $\phi \circ \psi : \mathcal{P} \mapsto \mathbb{E}$ is also differentiable at P relative to $\dot{\mathcal{P}}_P$.

If T_n is a sequence of estimators with values in $\mathbb{B}_\phi$ that is efficient at P for estimating $\psi(P)$, then $\phi(T_n)$ is efficient at P for estimating $\phi \circ \psi(P)$.

The following very useful theorem completely characterizes efficient estimators of Euclidean parameters:

THEOREM 2. *Let T_n be an estimator for a parameter $\psi : \mathcal{P} \mapsto \mathbb{R}^k$, where ψ is differentiable at P relative to the tangent space $\dot{\mathcal{P}}_P$ with k-variate efficient influence function $\tilde{\psi}_P \in L_2^0(P)$.*

Then the following are equivalent:

(i) T_n is efficient at P relative to $\dot{\mathcal{P}}_P$, and thus the limiting distribution of $\sqrt{n}(T_n - \psi(P))$ is mean zero normal with covariance $P[\tilde{\psi}_P \tilde{\psi}'_P]$.

(ii) T_n is asymptotically linear with influence function $\tilde{\psi}_P$.

The next theorem we present endeavors to extends the above characterization of efficient estimators to more general parameter spaces of the form $\ell^\infty(\mathcal{H})$:

THEOREM 3. *Let T_n be an estimator for a parameter $\psi : \mathcal{P} \mapsto \ell^\infty(\mathcal{H})$, where ψ is differentiable at P relative to the tangent space $\dot{\mathcal{P}}_P$ with efficient influence function $\tilde{\psi}_P : \mathcal{H} \mapsto L_2^0(P)$.*

Let $\mathcal{F} \equiv \{\tilde{\psi}_P(h) : h \in \mathcal{H}\}$.

Then the following are equivalent:

(a) T_n is efficient at P relative to $\dot{\mathcal{P}}_P$ and at least one of the following holds:

(i) T_n is asymptotically linear.

(ii) $\mathcal{F}$ is P-Donsker for some version of $\tilde{\psi}_P$.

(b) For some version of $\tilde{\psi}_P$, T_n is asymptotically linear with influence function $\tilde{\psi}_P$ and $\mathcal{F}$ is P-Donsker.

(c) T_n is regular and asymptotically linear with influence function $\check{\psi}_P$ such that $\{\check{\psi}_P(h) : h \in \mathcal{H}\}$ is P-Donsker and $\check{\psi}_P(h) \in \dot{\mathcal{P}}_P$ for all $h \in \mathcal{H}$.

The theorem gives us several properties of efficient estimators that can be useful for a number of things, including establishing efficiency.

In particular, conclusion (c) tells us that a simple method for establishing efficiency of T_n requires only that

- T_n be asymptotically linear
- with an influence function that is contained in a Donsker class
- for which the individual components $\check{\psi}_P(h)$ are contained in the tangent space for all $h \in \mathcal{H}$.

The theorem also tells us that if T_n is efficient, only one of (i) or (ii) in (a) is required and the other will follow.

This means, for example, that if T_n is efficient and $\mathcal{F}$ is not P-Donsker for any version of $\tilde{\psi}_P$, then T_n must not be asymptotically linear.

Also note that the requirement that $\mathcal{F}$ is P-Donsker collapses to requiring that $\|\tilde{\psi}_P\|_{P,2} < \infty$ when $\mathcal{H}$ is finite, and we are therefore in the setting of Theorem 2.

However, such a requirement is not needed in the statement of Theorem 2 since $\|\tilde{\psi}_P\|_{P,2} < \infty$ automatically follows from the required differentiability of ψ when $\psi \in \mathbb{R}^k$.

This follows since the Riesz representation theorem assures us that $\tilde{\psi}_P$ is in the closed linear span of $\dot{\mathcal{P}}_P$ which is a subset of $L_2(P)$.

The following somewhat deep theorem is useful in applications and tells us that pointwise efficiency implies uniform efficiency under weak convergence.

THEOREM 4. *Let T_n be an estimator for a parameter $\psi : \mathcal{P} \mapsto \ell^\infty(\mathcal{H})$, where ψ is differentiable at P relative to the tangent space $\dot{\mathcal{P}}_P$ with efficient influence function $\tilde{\psi}_P : \mathcal{H} \mapsto L_2^0(P)$.*

The following are equivalent:

(a) T_n is efficient for $\psi(P)$.

(b) $T_n(h)$ is efficient for $\psi(P)(h)$, for every $h \in \mathcal{H}$, and $\sqrt{n}(T_n - \psi(P))$ is asymptotically tight under P.

The proof of this theorem makes use of the following deep lemma:

LEMMA 1. *Suppose that $\psi : \mathcal{P} \mapsto \mathbb{D}$ is differentiable at P relative to the tangent space $\dot{\mathcal{P}}_P$ and that $d'T_n$ is asymptotically efficient at P for estimating $d'\psi(P)$ for every d' in a subset $\mathbb{D}' \subset \mathbb{D}^*$ which satisfies*

$$\|d\| \quad \leq \quad c \sup_{d' \in \mathbb{D}',\, \|d'\| \leq 1} |d'(d)|, \tag{2}$$

for some constant $c < \infty$.

Then T_n is asymptotically efficient at P provided $\sqrt{n}(T_n - \psi(P))$ is asymptotically tight under P.

Proof of Theorem 4:

- That (a) implies (b) is obvious.
- Now assume (b), and let $\mathbb{D} = \ell^\infty(\mathcal{H})$ and $\mathbb{D}'$ be the set of all coordinate projections $d \mapsto d_h^* d \equiv d(h)$ for every $h \in \mathcal{H}$.
- Since the uniform norm on $\ell^\infty(\mathcal{H})$ is trivially equal to $\sup_{d' \in \mathbb{D}'} |d'd|$ and all $d' \in \mathbb{D}'$ satisfying $\|d'\| = 1$, Condition (2) is easily satisfied.
- Since $\sqrt{n}(T_n - \psi(P))$ is asymptotically tight by assumption, all of the conditions of Lemma 1 are satisfied.
- Hence T_n is efficient, and the desired conclusions follow.$\Box$

We close this section with an interesting corollary of Lemma 1 that provides a remarkably simple connection between marginal and joint efficiency on product spaces:

THEOREM 5. *Suppose that* $\psi_j : \mathcal{P} \mapsto \mathbb{D}_j$ *is differentiable at* P *relative to the tangent space* $\dot{\mathcal{P}}_P$, *and suppose that* $T_{n,j}$ *is asymptotically efficient at* P *for estimating* $\psi_j(P)$, *for* $j = 1, 2$.

Then $(T_{n,1}, T_{n,2})$ *is asymptotically efficient at* P *for estimating* $(\psi_1(P), \psi_2(P))$.

Proof:

- Let $\mathbb{D}'$ be the set of all maps $(d_1, d_2) \mapsto d_j^* d_j$ for $d_j^* \in \mathbb{D}_j$ and j equal to either 1 or 2.
- Note that by the Hahn-Banach theorem (see Corollary 6.7 of Conway, 1990), $\|d_j\| = \sup\{|d_j^* d_j| : \|d_j^*\| = 1, d_j^* \in \mathbb{D}_j^*\}$.
- Thus the product norm $\|(d_1, d_2)\| = \|d_1\| \vee \|d_2\|$ satisfies Condition (2) of Lemma 1 with $c = 1$.
- Hence the desired conclusion follows.□

Thus marginal efficiency implies joint efficiency even though marginal weak convergence does not imply joint weak convergence!

This is not quite so surprising as it may appear at first.

Consider the finite-dimensional setting where $\psi_j(P) \in \mathbb{R}$ for $j = 1, 2$.

If $T_{n,j}$ is efficient for $\psi_j(P)$, for each $j = 1, 2$, then Theorem 2 tells us that $(T_{n,1}, T_{n,2})$ is asymptotically linear with influence function $(\tilde{\psi}_{1,P}, \tilde{\psi}_{2,P})$.

Thus the limiting joint distribution will in fact be the optimal bivariate Gaussian distribution.

The above theorem can be viewed as an infinite-dimensional generalization of this finite-dimensional phenomenon.

Optimality of Tests

In this section, we study testing of the null hypothesis

$$H_0 : \psi(P) \quad \leq \quad 0 \tag{3}$$

versus the alternative $H_1 : \psi(P) > 0$ for a one-dimensional parameter $\psi(P)$.

The basic conclusion we will endeavor to show is that a test based on an asymptotically optimal estimator for $\psi(P)$ will, in a meaningful way, be asymptotically optimal.

Note that null hypotheses of the form $H_{01} : \psi(P) \leq \psi_0$ can trivially be rewritten in the form given in (3) by replacing $P \mapsto \psi(P)$ with $P \mapsto \psi(P) - \psi_0$.

For dimensions higher than one, coming up with a satisfactory criteria for optimality of tests is difficult and we will not pursue the higher dimensional setting here.

For a given model $\mathcal{P}$ and measure P on the boundary of the null hypothesis where $\psi(P) = 0$, we are interested in studying the "local asymptotic power" in a neighborhood of P.

These neighborhoods are of size $1/\sqrt{n}$ and are the appropriate magnitude when considering sample size computation for $\sqrt{n}$ consistent parameter estimates.

Consider for example the univariate normal setting where the data are i.i.d. $N(\mu, \sigma^2)$.

A natural choice of test for $H_0 : \mu \leq 0$ versus $H_1 : \mu > 0$ is the indicator of whether

$$T_n = \sqrt{n}\frac{\bar{x}}{s_n} > z_{1-\alpha},$$

where

- $\bar{x}$ and s_n are the sample mean and standard deviation from an i.i.d. sample $X_1, \ldots, X_n$,
- z_q is the qth quantile of a standard normal, and
- α is the chosen size of this one-sided test.

For any $\mu > 0$, T_n diverges to infinity with probability 1.

However, if $\mu = k/\sqrt{n}$ for some finite k, then $T_n \rightsquigarrow N(k, 1)$.

Thus we can derive non-trivial power functions only for shrinking "contiguous alternatives" in a $1/\sqrt{n}$ neighborhood of zero.

In this case, since $\tilde{\psi}_P = X$ and the corresponding one-dimensional submodel $\{P_t\}$ must satisfy

$$\left.\frac{\partial\psi(P_t)}{\partial t}\right|_{t=1} = k,$$

we know that the score function g corresponding to $\{P_t\}$ must be $g(X) = kX/\sigma$.

Hence, in this example, we can easily express the local alternative sequence in terms of the score function rather than k.

Thus it makes sense in general to study the performance of tests under contiguous alternatives defined by one-dimensional submodels corresponding to score functions.

Accordingly, for a given element g of a tangent set $\dot{\mathcal{P}}_P$, let $t \mapsto P_{t,g}$ be a one-dimensional submodel which is differentiable in quadratic mean at P with score function g along which ψ is differentiable, i.e.,

$$\frac{\psi(P_{t,g}) - \psi(P)}{t} \to P[\tilde{\psi}_P g],$$

as $t \downarrow 0$.

For each such g for which $P[\tilde{\psi}_P g] > 0$, we can see that when $\psi(P) = 0$, the submodel $\{P_{t,g}\}$ belongs to $H_1 : \psi(P) > 0$ for all sufficiently small $t > 0$.

We will therefore consider power over contiguous alternatives of the form $\{P_{h/\sqrt{n},g}\}$ for $h > 0$.

Before continuing, we need to define a *power function*.

For a subset $\mathcal{Q} \subset \mathcal{P}$ containing P, a power function $\pi : \mathcal{Q} \mapsto [0, 1]$ at level α is a function on probability measures that satisfies $\pi(Q) \leq \alpha$ for all $Q \in \mathcal{Q}$ for which $\psi(Q) \leq 0$.

We say that a sequence of power function $\{\pi_n\}$ has asymptotic level α if $\limsup_{n\to\infty} \pi_n(Q) \leq \alpha$ for every $Q \in \mathcal{Q} : \psi(Q) \leq 0$.

The power function for a level α hypothesis test of H_0 is the probability of rejecting H_0 under Q.

Hence statements about power functions can be viewed as statements about hypothesis tests.

Here is our main result:

THEOREM 6. *Let $\psi : \mathcal{P} \mapsto \mathbb{R}$ be differentiable at P relative to the tangent space $\dot{\mathcal{P}}_P$ with efficient influence function $\tilde{\psi}_P$, and suppose $\psi(P) = 0$.*

Then, for every sequence of power functions $P \mapsto \pi_n(P)$ of asymptotic level α tests for $H_0 : \psi(P) \leq 0$, and for every $g \in \dot{\mathcal{P}}_P$ with $P[\tilde{\psi}_P g] > 0$ and every $h > 0$,

$$\limsup_{n\to\infty} \pi_n(P_{h/\sqrt{n},g}) \leq 1 - \Phi\left[z_{1-\alpha} - h\frac{P[\tilde{\psi}_P g]}{\sqrt{P[\tilde{\psi}_P^2]}}\right].$$

This is a minor modification of Theorem 25.44 in Section 25.6 of van der Vaart (1998) and the proof is given therein.

While there are some differences in notation, the substantive modification is that

- van der Vaart requires the power functions to have level α for each n,
- whereas our version only require the levels to be asymptotically α.

This does not affect the proof, and we omit the details.

An advantage of this modification is that

- it permits the use of approximate hypothesis tests,
- such as those which depend on the central limit theorem,
- whose level for fixed n may not be exactly α
- but whose asymptotic level is known to be α.

An important consequence of Theorem 6 is that a test based on an efficient estimator T_n of $\psi(P)$ can achieve the given optimality.

To see this, let S_n^2 be a consistent estimator of the limiting variance of $\sqrt{n}(T_n - \psi(P))$, and let $\pi_n(Q)$ be the power function defined as the probability that $\sqrt{n}T_n/S_n > z_{1-\alpha}$ under the model Q.

It is easy to see that this power function has asymptotic power α under the null hypothesis.

The following result shows that this procedure is asymptotically optimal:

LEMMA 2. *Let $\psi : \mathcal{P} \mapsto \mathbb{R}$ be differentiable at P relative to the tangent space $\dot{\mathcal{P}}_P$ with efficient influence function $\tilde{\psi}_P$, and suppose $\psi(P) = 0$.*

Suppose the estimator T_n is asymptotically efficient at P, and, moreover, that $S_n^2 \xrightarrow{\mathsf{P}} P\tilde{\psi}_P^2$.

Then, for every $h \geq 0$ and $g \in \dot{\mathcal{P}}_P$,

$$\limsup_{n\to\infty} \pi_n(P_{h/\sqrt{n},g}) = 1 - \Phi\left(z_{1-\alpha} - h\frac{P[\tilde{\psi}_P g]}{\sqrt{P\tilde{\psi}_P^2}}\right).$$

Proof: By Theorem 2 and Part (i) of Theorem 11.14, we have that

$$\frac{\sqrt{n}T_n}{S_n} \overset{P_n}{\rightsquigarrow} Z + h\frac{P[\tilde{\psi}_P g]}{\sqrt{P\tilde{\psi}_P^2}},$$

where $P_n \equiv P_{h/\sqrt{n},g}$ and Z has a standard normal distribution.

The desired result now follows.□

Consider, for example, the Mann-Whitney test discussed in Section 12.2.2 for comparing two independent samples of respective sample sizes m and n.

Let $\mathbb{F}_m$ and $\mathbb{G}_n$ be the respective empirical distributions with corresponding true distributions F and G which we assume to be continuous.

The Mann-Whitney statistic is

$$T_n = \int_{\mathbb{R}} \mathbb{G}_n(s) d\mathbb{F}_m(s) - 1/2$$

which is consistent for $\psi(P) = \int_{\mathbb{R}} G(s) dF(s) - 1/2$.

We are interested in testing the null hypothesis $H_0 : \psi(P) \leq 0$ versus $H_1 : \psi(P) > 0$.

By Theorem 5, we know that $(\mathbb{F}_m, \mathbb{G}_m)$ is jointly efficient for (F, G).

Moreover, by Lemma 12.3, we know that $(F, G) \mapsto \int_{\mathbb{R}} G(s) dF(s)$ is Hadamard differentiable.

Thus Theorem 1 applies, and we obtain that $\int_{\mathbb{R}} \mathbb{G}_n(s) d\mathbb{F}_n(s)$ is asymptotically efficient for $\int_{\mathbb{R}} G(s) dF(s)$.

Hence Lemma 2 also applies, and we obtain that T_n is optimal for testing H_0, provided it is suitably standardized.

We know from the discussion in Section 12.2.2 that the asymptotic variance of $\sqrt{n}T_n$ is $1/12$.

Thus the test that rejects the null when $\sqrt{12n}T_n$ is greater than $z_{1-\alpha}$ is optimal.

Another simple example is the sign test for symmetry about zero for a sample of real random variables $X_1, \ldots, X_n$ with distribution F that is continuous at zero.

The test statistic is

$$T_n = \int_{\mathbb{R}} \text{sign}(x) d\mathbb{F}_n(x),$$

where $\mathbb{F}_n$ is the usual empirical distribution.

Using arguments similar to those used in the previous paragraphs, it can be shown that T_n is an asymptotically efficient estimator for

$$\psi(P) = \int_{\mathbb{R}} \text{sign}(x) dF(x) = \text{pr}(X > 0) - \text{pr}(X < 0).$$

Thus, by Lemma 2 above, the sign test is asymptotically optimal for testing the null hypothesis H_0 : pr$(X > 0) \leq$ pr$(X < 0)$ versus the alternative H_1 : pr$(X > 0) >$ pr$(X < 0)$.

These examples illustrates the general concept that

- if the parameter of interest is a smooth functional of the underlying distribution functions,
- then the estimator obtained by substituting the true distributions with the corresponding empirical distributions will be asymptotically optimal,
- provided we are not willing to make any parametrically restrictive assumptions about the distributions.