Pollard's Entropy Condition, Stochastic Equicontinuity, and its implications for Weak Convergence and M-Estimation Problems

Let $(\mathcal{X}, S, \mu)$ be a probability space, and let $\mathcal{F}$ be a class of real-valued functions with domain $\mathcal{X}$ that is measurable for the sigma-algebra S. Let F be the envelope for $\mathcal{F}$. The envelope is assumed to satisfy $\mu F^2 < \infty$. We equip $\mathcal{F}$ with the pseudometric $d_\mu : \mathcal{F} \times \mathcal{F} \to \mathbb{R}_+$ given by

$$d_\mu(f, g) = \frac{\left(\mu\left|f - g\right|^2\right)^{1/2}}{\mathcal{T}}, \quad \text{where} \quad \mathcal{T}^2 = \mu F^2.$$

The class $\mathcal{F}$ is Euclidean if there exist constants $A, V > 0$ which do not depend on μ such that:

$$N(\epsilon, d_\mu, \mathcal{F}) \leq A\epsilon^{-V} \quad \text{for all } 0 < \epsilon \leq 1.$$

$N(\epsilon, d_\mu, \mathcal{F})$ is the covering number we defined in lecture.

As it turns out, if the class of subgraphs generated by $\mathcal{F}$ is a polynomial class of sets, then $\mathcal{F}$ is Euclidean. This is a sufficient, but not necessary condition.

A class of functions $\mathcal{F}$ is manageable if there exists a deterministic function $\lambda(\cdot)$, called the "capacity bound" such that

$$\text{(i)} \int_0^1 \sqrt{\log \lambda(\epsilon)} d\epsilon < \infty; \quad \text{and} \quad \text{(ii)} \; D(\epsilon, d_\mu, \mathcal{F}) \leq \lambda(\epsilon) \; \forall \; 0 < \epsilon \leq 1$$

$D(\epsilon, d_\mu, \mathcal{F})$ is the packing number we defined in lecture. Clearly, all Euclidean classes are manageable since $D(2\epsilon, d_\mu, \mathcal{F}) \leq N(\epsilon, d_\mu, \mathcal{F})$ and therefore

$$\int_0^1 \sqrt{\log D(\epsilon, d_\mu, \mathcal{F})} d\epsilon \leq \int_0^1 \sqrt{-V \log(\epsilon/2) + \log(A)} d\epsilon < \infty.$$

The class of functions $\mathcal{F}$ satisfies Pollard's Entropy Condition if

$$\int_0^1 \sup_\mu \sqrt{\log N(\epsilon, d_\mu, \mathcal{F})} d\epsilon < \infty \quad \text{for some envelope } F.$$

The search for the supremum is done over the set of all measures that concentrate on a finite set. $N(\epsilon, d_\mu, \mathcal{F})$ is the covering number we defined. Euclidean classes of functions also satisfy Pollard's entropy condition.

Sufficient Conditions for a class $\mathcal{F}$ to satisfy Pollard's Entropy Condition. In practice, we do not verify that the subgraphs generated by $\mathcal{F}$ is of polynomial class. We rely on sufficient conditions, easy to verify. We first introduce the concept of Total Variation:

Total Variation: Let (X, d) be a metric space. A functon $\gamma : [a, b] \to X$ is of bounded variation if there exists M such that for each partition $\mathcal{P} = \{a = t_0 < t_1 < \cdots < t_n = b\}$ of $[a, b]$,

$$v(\gamma, \mathcal{P}) = \sum_{k=1}^{n} d\big(\gamma(t_k), \gamma(t_{k-1})\big) \leq M.$$

The total variation V_γ of γ is defined by

$$V_\gamma = \sup \{v(\gamma, \mathcal{P}) : \ \mathcal{P} \text{ is a partition of } [a, b].\}$$

A function γ is of bounded variation if and only if it can be expressed as the difference between two monotonic functions. In addition, all smooth functions are of bounded variation. We will describe three types of functions that satisfy Pollard's entropy condition.

Type I Class of Functions

A class $\mathcal{F}$ is type-I class if it is of the form

$$\mathcal{F} = \left\{ f : f(x) = h(x'\theta) \; \forall \; x \in \mathcal{X}, \text{ where } \theta \in \Theta \subset \mathbb{R}^k \text{ (bounded), and } h \in V_k, \right\}$$

where V_k is a set of functions from $\mathbb{R} \to \mathbb{R}$ with total variation $\leq K < \infty$. Thus, Type-I functions are transformations of linear indices with bounded variation. This type is useful to deal with M-estimation problems with nondifferentiable objective functions.

Type II Class of Functions

A class $\mathcal{F}$ is of Type-II class if

$$\mathcal{F} = \left\{ f : f(x) = f(x, \theta), \text{ where } \theta \in \Theta \subset \mathbb{R}^k \text{ (bounded), and} \right.$$

every $f \in \mathcal{F}$ satisfies a Lipschitz condition of the form

$$\left| f(\cdot, \theta_1) - f(\cdot, \theta_2) \right| \leq B(\cdot) \left\| \theta_1 - \theta_2 \right\| \quad \forall \; \theta_1, \theta_2 \in \Theta.$$

and some function $B : \mathcal{X} \to \mathbb{R}$ with $E[B(X)^2] < \infty$

Type III Class of Functions

We will discuss them later. They are relevant in semiparametric estimation.

Theorem (Theorem 2 in Andrews 1994, Handbook of Econometrics) If $\mathcal{F}$ is class I or II, then Pollard's entropy condition holds with envelopes:

$$F(\cdot) = \max\left\{1, \sup_{f\in\mathcal{F}} \left|f(\cdot)\right|\right\} \quad \text{for Type-I class,}$$

$$F(\cdot) = \max\left\{1, \sup_{f\in\mathcal{F}} \left|f(\cdot)\right|, B(\cdot)\right\} \quad \text{for Type-II class.}$$

There is an additional (by perhaps now redundant) result from Pakes and Pollard (1989):

Let $\mathcal{F}$ be a class of real-valued functions defined on $\mathcal{X}$ indexed by a bounded subset $T \subset \mathbb{R}^d$. If there exists a constant C such that

$$\left|f(x,t) - f(x,t')\right| \leq c\left|t - t'\right| \quad \forall\, x \in \mathcal{X},\ t, t' \in T,$$

then $\mathcal{F}$ is Euclidean for a constant envelope. (Note that this class of functions is more restrictive than Type-II).

Extending the properties of Euclidean, Manageable, etc. to Sequences of classes $\{\mathcal{F}_n\}$.

The previous concepts can be extended to sequences of classes of functions in a straightforward way. For example, the sequence of classes $\{\mathcal{F}_n\}$ is Euclidean for envelope F (which does not depend on n) if there exist $A, V > 0$ not depending on μ such that

$$\sup_n N(\epsilon, d_\mu, \mathcal{F}_n) \leq A\epsilon^{-V} \quad \forall\, 0 < \epsilon \leq 1$$

The concepts of manageability and Pollard's entropy condition are extended in an analogous way (by taking the corresponding suprema over the sequence $\mathcal{F}_n$).

We can finally go back to our empirical process....

Back to our Empirical Process

We defined it by

$$\nu_N(\tau) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \big[m(X_i, \tau) - E\big[m(X_i, \tau)\big]\big], \quad \tau \in (T, d) \text{ (pseudo-metric space).}$$

$\nu_N(\tau)$ is a sequence of classes of functions. Everything will be determined by the properties of the following class of functions:

$$\mathcal{F}_m = \big\{m(\cdot, \tau) : \tau \in T\big\}$$

Theorem 1 in Andrews (1994) essentially states that if $\mathcal{F}_m$ is Euclidean, then $\nu_N(\tau)$ is also Euclidean, and if $\mathcal{F}_m$ satisfies Pollard's entropy condition, then so does $\nu_N(\tau)$. A consequence of Pollard's entropy condition is that the empirical process $\nu_N(\tau)$ will also satisfy Stochastic Equicontinuity, which we define next (before stating the theorem formally).

Stochastic Equicontinuity of an Empirical Process

As a preamble, we first define the concept of Outer Expectation (and Outer Probability), and the concept of Weak Convergence.

Outer Expectation: This concept helps us deal with processes that are not measurable. We start with a probability space $(\Omega, S, \mathcal{P})$. For each bounded, real-valued function $H : \Omega \to \mathbb{R}$, we define the outer expectation E^*H as

$$E^*(H) = \inf\Big\{ E(h) : H \leq h, \text{ such that } h \text{ is measurable and integrable.} \Big\}$$

Weak Convergence: Consider a sequence of processes $\nu_N(\tau)$ and a process $\nu(\tau)$ indexed by the same pseudometric space (T, d_T). Let $B(T)$ be the space of functions where $\nu_N(\tau)$ lives. Let $\mathcal{U}\big(B(T)\big)$ be the space of all bounded, uniformly continuous transformations defined on $B(T)$. Then, the sequence of processes $\nu_N(\tau)$ converges weakly to the process $\nu(\tau)$ if

$$E^*\big[f\big(\nu_N(\cdot)\big)\big] \longrightarrow E\big[f\big(\nu(\cdot)\big)\big] \quad \forall\, f \in \mathcal{U}\big(B(T)\big)$$

We denote this as

$$\nu_N(\cdot) \Rightarrow \nu(\cdot)$$

Note that in the definition, we assume the limiting process $\nu(\cdot)$ to be measurable.

Why do we care about measurability? The processes $\nu_N(\cdot)$ and $\nu(\cdot)$ are both random functions of τ. Coming up with a probability space in which the process $\nu_N(\cdot)$ is measurable is a difficult (many times impossible) task. Measurability of a random function is a delicate issue. This is very different from talking about the measurability of $\nu_N(\tau_0)$ and $\nu(\tau_0)$ for a fixed value of the index at τ_0.

When dealing with the processes $\nu_N(\cdot)$ and $\nu(\cdot)$ we use outer-expectations because requiring measurability of the process $\nu_N(\cdot)$ is too restrictive even in simple cases. The typical example of this is:

$$T = [0, 1], \quad m(X_i, \tau) = \mathbb{1}\{X_i \leq \tau\}.$$

The question we have to address is: In what measure space is the resulting process $\nu_N(\cdot)$ measurable?

The most obvious measure space that comes to mind is given by

$$\big(D[0,1], \mathcal{B}(D[0,1]), d\big),$$

where $D[0,1]$ is the space of functions that are right-continuous and whose left limits always exist (the space of cadlag functions), and $\mathcal{B}(D[0,1])$ is the Borel-sigma field in $D[0,1]$. d stands for the uniform metric

$$d(x,y) = \sup_{t\in[0,1]} \big|x(t) - y(t)\big|$$

The definition of open sets, etc. is based upon this metric. We need this to find $\mathcal{B}(D[0,1])$. Well, it turns out that our apparently innocuous process $\nu_N(\cdot)$ is not measurable in $\big(D[0,1], \mathcal{B}(D[0,1])\big)$. A solution in this particular case is to drop the uniform metric in favor of the so-called Skorohod metric. Unfortunately, no such alternatives exist for slightly more complicated processes. This is why we focus on outer expectations. We do not worry about this for the limiting process $\nu(\cdot)$ because the results that we will review focus on limiting processes that are uniformly continuous w.p.1. This would be enough to solve any measurability concerns.

Stochastic Equicontinuity

We will focus on the following definition (see two other alternative, equivalent definitions in Andrews 1994, Handbook of Econometrics): The process $\nu_N(\cdot)$ is stochastically equicontinuous if for any $\varepsilon > 0$ and any $\eta > 0$, there exists $\delta > 0$ such that

$$\overline{\lim_{N\to\infty}} \mathrm{P}^* \left[\sup_{\rho(\tau_1,\tau_2)<\delta} \left\| \nu_N(\tau_1) - \nu_N(\tau_2) \right\| > \eta \right] < \varepsilon$$

$\overline{\lim}$ denotes "upper limit", which means that $\overline{\lim} T_n = c$ if c is greater than all but a finite number of terms of $\{T_n\}$, all of which are equal to c.

Stochastic Equicontinuity is a tremendously useful property. It can yield weak convergence, and it can also help us find the asymptotic distribution of M-estimators very quickly.

Weak Convergence and Stochastic Equicontinuity

The following result can be found in Andrews (1994).

Suppose:

(i) (T, ρ) is a totally bounded pseudometric space.

(ii) Finite-dimensional convergence in distribution holds. That is: For all finite subsets $(\tau_1, \tau_2, \ldots, \tau_J) \in T$, the vector $\big(\nu_N(\tau_1), \nu_N(\tau_2), \ldots, \nu_N(\tau_J)\big)$ converges in distribution.

(iii) $\nu_N(\cdot)$ is stochastically equicontinuous.

Then, there exists a Borel-measurable stochastic process $\nu(\cdot)$ with sample paths that are uniformly continuous w.p.1, such that $\nu_N(\cdot) \Rightarrow \nu(\cdot)$. Conversely, if $\nu_N(\cdot) \Rightarrow \nu(\cdot)$, with $\nu(\cdot)$ satisfying the conditions described above, then (ii) and (iii) hold.

Weak convergence is a very powerful result, for example, because any continuous mapping (functional) of $\nu_N(\cdot)$ will also converge to the corresponding one for $\nu(\cdot)$. For example,

$$\varphi\big(\nu_N(\cdot)\big), \quad \sup_{\tau} \big|\nu_N(\tau)\big|, \quad \int_T \nu_N(\tau) d\tau$$

(where $\varphi(\cdot)$ is a continuous transformation) would converge weakly to

$$\varphi\big(\nu(\cdot)\big), \quad \sup_{\tau} \big|\nu(\tau)\big|, \quad \int_T \nu(\tau) d\tau$$

provided that these objects are well-defined. If we characterize the limiting process $\nu(\cdot)$, we would be able to do inference on these functionals.

Next, we analyze the relationship between stochastic equicontinuity and M-estimation.

Stochastic equicontinuity and M-estimators

Take $\tau \in T \subset \mathbb{R}^k$ and let

$$\overline{m}_N(\tau) = \frac{1}{N}\sum_{i=1}^{N} m(X_i, \tau)$$

Suppose an estimator $\widehat{\tau}$ satisfies

$$\sqrt{N}\overline{m}_N(\widehat{\tau}) = o_p(1), \quad \text{or equivalently,} \quad \overline{m}_N(\widehat{\tau}) = o_p\big(1/\sqrt{N}\big).$$

Suppose $\widehat{\tau}$ is also consistent for τ_0 (the true parameter value), so that $\rho\big(\widehat{\tau}, \tau_0\big) \xrightarrow{p} 0$. Let $\lambda(\tau) \equiv E\big[\overline{m}_N(\tau)\big]$, then τ_0 is defined by $\lambda(\tau_0) = 0$. Suppose $\lambda(\cdot)$ is a smooth function that admits the following Taylor approximation

$$0 = \lambda(\tau_0) = \lambda(\widehat{\tau}) + \nabla_\tau \lambda(\widetilde{\tau})(\tau_0 - \widehat{\tau}),$$

where $\widetilde{\tau}$ is between $\widehat{\tau}$ and τ_0. Suppose $\nabla_\tau \lambda(\tau)$ is invertible in a neighborhood of τ_0, and this inverse is a continuous there. This would be enough to yield, for sufficiently large N:

$$\sqrt{N}(\widehat{\tau} - \tau_0) = \nabla_\tau \lambda(\widetilde{\tau})^{-1}\sqrt{N}\lambda(\widehat{\tau})$$

with $\nabla_\tau \lambda(\widetilde{\tau})^{-1} \xrightarrow{p} \nabla_\tau \lambda(\tau_0)^{-1} \equiv M^{-1}$.

Therefore, the asymptotic distribution of $\sqrt{N}(\widehat{\tau} - \tau_0)$ will depend on the properties of $\sqrt{N}\lambda(\widehat{\tau})$. we have

$$\sqrt{N}\lambda(\widehat{\tau}) = \underbrace{\sqrt{N}\big(\lambda(\widehat{\tau}) - \overline{m}_N(\widehat{\tau})\big)}_{\equiv \nu_N(\widehat{\tau})} + \underbrace{\sqrt{N}\overline{m}_N(\widehat{\tau})}_{= o_p(1)}$$

The relevant empirical process is

$$\nu_N(\tau) = -\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \big[m(X_i, \tau) - E[m(X_i, \tau)]\big].$$

Adding and subtracting $\nu_N(\tau_0)$, we get

$$\sqrt{N}\lambda(\widehat{\tau}) = \nu_N(\widehat{\tau}) - \nu_N(\tau_0) + \nu_N(\tau_0) + o_p(1).$$

We will assume that the relevant conditions are satisfied, so that $\nu_N(\tau_0) \xrightarrow{d} \mathcal{N}(0, S)$. In this case, we would have $\sqrt{N}\lambda(\widehat{\tau}) \xrightarrow{d} \mathcal{N}(0, S)$ if

$$\nu_N(\widehat{\tau}) - \nu_N(\tau_0) = o_p(1).$$

We have the following claim...

Claim: Given that $\widehat{\tau}$ is consistent, $\nu_N(\widehat{\tau}) - \nu_N(\tau_0) = o_p(1)$ if $\nu_N(\cdot)$ is stochastically equicontinuous.

Proof: Fix any $\eta > 0$ and consider the following probability (recall that we may have to use outer probabilities P^* here, if the process is not measurable) for a given δ

$$\Pr\Big(\big|\nu_N(\widehat{\tau}) - \nu_N(\tau_0)\big| > \eta, \rho(\widehat{\tau}, \tau_0) \leq \delta\Big) + \Pr\big(\rho(\widehat{\tau}, \tau_0) > \delta\big)$$

These probabilities constitute an upper bound for $\Pr\Big(\big|\nu_N(\widehat{\tau}) - \nu_N(\tau_0)\big| > \eta\Big)$ for any value of δ. To see this, note that $\delta \to \infty$, these probabilities become approximately equal to $\Pr\Big(\big|\nu_N(\widehat{\tau}) - \nu_N(\tau_0)\big| > \eta\Big)$, and if $\delta \to 0$, they become approximately equal to 1. In particular, this means that for any $\eta > 0$, we can find a $\delta > 0$ such that

$$\begin{aligned}\overline{\lim_{N\to\infty}} P\Big(\big|\nu_N(\widehat{\tau}) - \nu_N(\tau_0)\big| > \eta\Big) \leq & \overline{\lim_{N\to\infty}} P\Big(\big|\nu_N(\widehat{\tau}) - \nu_N(\tau_0)\big| > \eta, \rho(\widehat{\tau}, \tau_0) \leq \delta\Big) \\ & + \underbrace{\overline{\lim_{N\to\infty}} P\big(\rho(\widehat{\tau}, \tau_0) > \delta\big)}_{=o(1), \text{ by consistency of } \widehat{\tau}.}\end{aligned}$$

Now, notice that

$$\left|\nu_N(\widehat{\tau}) - \nu_N(\tau_0)\right| > \eta \text{ and } \rho(\widehat{\tau}, \tau_0) \leq \delta \underline{\text{ only if }} \left[\sup_{\rho(\tau,\tau_0)\leq\delta} \left|\nu_N(\tau) - \nu_N(\tau_0)\right|\right] > \eta.$$

Therefore

$$\overline{\lim_{N\to\infty}} P\Big(\left|\nu_N(\widehat{\tau})-\nu_N(\tau_0)\right| > \eta, \rho(\widehat{\tau}, \tau_0) \leq \delta\Big) \leq \overline{\lim_{N\to\infty}} P\left[\sup_{\rho(\tau,\tau_0)\leq\delta} \left|\nu_N(\tau)-\nu_N(\tau_0)\right| > \eta\right]$$

This yields

$$\overline{\lim_{N\to\infty}} P\Big(\left|\nu_N(\widehat{\tau})-\nu_N(\tau_0)\right| > \eta\Big) \leq \overline{\lim_{N\to\infty}} P\left[\sup_{\rho(\tau,\tau_0)\leq\delta} \left|\nu_N(\tau)-\nu_N(\tau_0)\right| > \eta\right] + o(1)$$

If $\nu_N(\cdot)$ is stochastically equicontinuous, then the first probability on the right-hand side goes to zero for any $\delta > 0$. Therefore $\overline{\lim_{N\to\infty}} P\Big(\left|\nu_N(\widehat{\tau}) - \nu_N(\tau_0)\right| > \eta\Big) \to 0$. Since this holds for an arbitrary $\eta > 0$, we get $\nu_N(\widehat{\tau})-\nu_N(\tau_0) = o_p(1)$ and therefore $\sqrt{N}\lambda(\widehat{\tau}) = \nu_N(\tau_0) + o_p(1)$, and

$$\sqrt{N}\lambda(\widehat{\tau}) \xrightarrow{d} \mathcal{N}(0, S)$$

Therefore, the asymptotic distribution of $\widehat{\tau}$ is given by

$$\sqrt{N}(\widehat{\tau} - \tau_0) = \nabla_\tau \lambda(\widetilde{\tau})^{-1} \nu_N(\tau_0) + o_p(1) \xrightarrow{d} \mathcal{N}(0, M^{-1} S M^{-1}),$$

where $\nabla_\tau \lambda(\tau_0)^{-1} \equiv M^{-1}$.

Stochastic equicontinuity and Semiparametric Estimation $(\Diamond)$

A number of semiparametric estimation problems involve a first-stage estimator $\widehat{\tau}$ for an infinite-dimensional parameter (an unknown function, or functional). Using this plug-in, the econometrician estimates a finite-dimensional parameter θ by satisfying (with respect to $\widehat{\theta}$) a set of asymptotic pseudo-first order conditions

$$\sqrt{N}\overline{m}_N(\widehat{\theta}, \widehat{\tau}) = o_p(1), \quad \text{with} \quad \overline{m}_N(\theta, \widehat{\tau}) = \frac{1}{N} \sum_{i=1}^{N} m(X_i; \theta, \widehat{\tau}).$$

In the context of semiparametric estimation problems, with a nonparametric first-stage estimator for an infinite-dimensional parameter, we will focus here on smooth objective functions (the extension to left-and-right differentiable functions would not be so difficult, given our previous discussions regarding the CLAD estimator. Extensions to discontinuous objective functions are something else entirely, we will not cover them here due to lack of time).

We assume $\overline{m}_N(\cdot, \tau)$ to be a smooth function of θ. A first-order Taylor approximation yields

$$o_p(1) = \overline{m}_N(\widehat{\theta}, \widehat{\tau}) = \overline{m}_N(\theta_0, \widehat{\tau}) + \nabla_\theta \overline{m}_N(\widetilde{\theta}, \widehat{\tau})(\widehat{\theta} - \theta_0)$$

We assume that $\nabla_\theta \overline{m}_N(\widetilde{\theta}, \widehat{\tau})^{-1} \xrightarrow{p} M^{-1}$. This allows us to express (for sufficiently large N)

$$\sqrt{N}(\widehat{\theta} - \theta_0) = -\nabla_\theta \overline{m}_N(\widetilde{\theta}, \widehat{\tau})^{-1} \sqrt{N} \overline{m}_N(\theta_0, \widehat{\tau}) + o_p(1). \tag{1}$$

Let $\lambda(\theta, \tau) = E[\overline{m}_N(\theta, \tau)]$. The relevant empirical process here will be

$$\nu_N(\theta, \tau) = \sqrt{N}\Big(\overline{m}_N(\theta, \tau) - \lambda(\theta, \tau)\Big)$$

We have

$$\begin{aligned}
\sqrt{N}\overline{m}_N(\theta_0, \widehat{\tau}) &= \nu_N(\theta_0, \widehat{\tau}) + \sqrt{N}\lambda(\theta_0, \widehat{\tau}) \\
&= \nu_N(\theta_0, \widehat{\tau}) - \nu_N(\theta_0, \tau_0) + \nu_N(\theta_0, \tau_0) + \sqrt{N}\lambda(\theta_0, \widehat{\tau})
\end{aligned}$$

The term $\nu_N(\theta_0, \tau_0)$ will have an asymptotically normal distribution. How about the term $\sqrt{N}\lambda(\theta_0, \widehat{\tau})$? A simple case is one in which τ is some unknown function of, say, Z_i and

$$\overline{m}_N(\theta_0, \widehat{\tau}) = \frac{1}{N} \sum_{i=1}^{N} m(X_i, \theta_0, \widehat{\tau}(Z_i)).$$

In this case,

$$\lambda(\theta_0, \widehat{\tau}) = \frac{1}{N}\sum_{i=1}^{N} E\big[m(X_i, \theta_0, \widehat{\tau}(Z_i))\big]$$

$$= \underbrace{\frac{1}{N}\sum_{i=1}^{N} E\big[m(X_i, \theta_0, \tau_0(Z_i))\big]}_{=0} + \frac{1}{N}\sum_{i=1}^{N} \nabla_\tau \Big(E\big[m(X_i, \theta_0, \tau_0(Z_i))\big]\Big)\big(\widehat{\tau}(Z_i) - \tau_0(Z_i)\big)$$

$$+ \frac{1}{2N}\sum_{i=1}^{N} \big(\widehat{\tau}(Z_i) - \tau_0(Z_i)\big)' \nabla_{\tau\tau'} \Big(E\big[m(X_i, \theta_0, \widetilde{\tau}(Z_i))\big]\Big)\big(\widehat{\tau}(Z_i) - \tau_0(Z_i)\big)$$

Suppose we show that

$$\frac{1}{N}\sum_{i=1}^{N} \Big\| \nabla_{\tau\tau'} \Big(E\big[m(X_i, \theta_0, \widetilde{\tau}(Z_i))\big]\Big) \Big\| = O_p(1), \quad \sup_i \Big\| \widehat{\tau}(Z_i) - \tau_0(Z_i) \Big\| = o_p(N^{-1/2}).$$

Then, we have

$$\sqrt{N}\lambda(\theta_0, \widehat{\tau}) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \nabla_\tau \Big(E\big[m(X_i, \theta_0, \tau_0(Z_i))\big]\Big)\big(\widehat{\tau}(Z_i) - \tau_0(Z_i)\big) + o_p(1)$$

As it turns out, in many situations, and under appropriate assumptions regarding the way in which $\widehat{\tau}(\cdot)$ is estimated, the term

$$\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \nabla_{\tau} \Big(E\big[m(X_i, \theta_0, \tau_0(Z_i))\big] \Big) \big(\widehat{\tau}(Z_i) - \tau_0(Z_i) \big)$$

can be expressed as a "U-statistic" (a higher order summation) which has an asymptotically normal distribution plus a term that is $o_p(1)$. In this case, $\sqrt{N}\lambda(\theta_0, \widehat{\tau}) \xrightarrow{d} \mathcal{N}(0, B)$. Going back, we have

$$\begin{aligned} \sqrt{N}\overline{m}_N(\theta_0, \widehat{\tau}) &= \nu_N(\theta_0, \widehat{\tau}) + \sqrt{N}\lambda(\theta_0, \widehat{\tau}) \\ &= \nu_N(\theta_0, \widehat{\tau}) - \nu_N(\theta_0, \tau_0) + \nu_N(\theta_0, \tau_0) + \sqrt{N}\lambda(\theta_0, \widehat{\tau}). \end{aligned}$$

As before, if the process $\nu_N(\cdot)$ is stochastically equicontinuous, we will have $\nu_N(\theta_0, \widehat{\tau}) - \nu_N(\theta_0, \tau_0) = o_p(1)$, and if the above conditions are met, we will have

$$\sqrt{N}\overline{m}_N(\theta_0, \widehat{\tau}) \xrightarrow{d} \mathcal{N}(0, C)$$

where C would reflect also the asymptotic covariance between $\nu_N(\theta_0, \tau_0)$ and $\sqrt{N}\lambda(\theta_0, \widehat{\tau})$. Our semiparametric estimator $\widehat{\theta}$ will have an asymptotic distribution

$$\sqrt{N}\big(\widehat{\theta} - \theta_0\big) \xrightarrow{d} \mathcal{N}(0, M^{-1}CM^{-1})$$

where $M^{-1} = \text{plim } \nabla_{\theta}\overline{m}_N(\theta_0, \widehat{\tau})^{-1}$

Theorem: Pollard's Entropy Condition and Stochastic Equicontinuity
Consider the process

$$\nu_N(\tau) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \big[m(X_i, \tau) - E[m(X_i, \tau)] \big]$$

Let

$$\mathcal{F}_m = \big\{ m(\cdot, \tau); \tau \in T \big\}$$

If the class $\mathcal{F}_m$ satisfies Pollard's entropy condition with some envelope $\overline{M}(\cdot)$ such that

$$\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} E\big[\overline{M}(X_i)^{2+\delta} \big] \quad \text{for some } \delta > 0,$$

and if the random variables $\{X_i\}$ are m-dependent (this includes independence as a special case, but also covers a family of time series), then the process $\nu_N(\cdot)$ is stochastically equicontinuous.

Thus, we can save pages of proofs for the asymptotic distribution of an estimator if we only show that the function $m(\cdot, \tau)$ belongs to one of the classes described above!

What is the connection between the entropy of a process and stochastic equicontinuity?

This discussion centers on Chapter 3 in Pollard (1990). We define first an the concept of Orlicz Norm:

Orlicz Norm: Suppose Φ is a convex, increasing function in $\mathbb{R}_+$ with $0 \leq \Phi(0) \leq 1$. The resulting Orlicz norm of a random variable Z, denoted by $\|Z\|_\Phi$ is

$$\|Z\|_\Phi = \inf\left\{c > 0 : \ E\left[\Phi(|Z|/c)\right] \leq 1\right\}$$

Loosely speaking, we can think of the Orlicz norm as being $E\left[\Phi(Z)\right]$ when this expectation exists.

The results that follow focus on the Orlicz norm that results from the function

$$\Psi(x) = \frac{1}{5}\exp(x^2).$$

we will denote the resulting Orlicz norm by $\|\cdot\|_\Psi$

The next lemma provides a very useful result to us:

Lemma 3.2 in Pollard (1990)
For any finite collection of m random variables $Z_1, \ldots, Z_m$,

$$\left\| \max_{i \leq m} \left| Z_i \right| \right\|_\Psi \leq \sqrt{2 + \log(m)} \cdot \max_{i \leq m} \left\| Z_i \right\|_\Psi$$

The next lemma builds on this result, and provides a hint of the link between stochastic equicontinuity and entropy (more precisely, packing numbers).

Lemma 3.4 in Pollard (1990)
If a stochastic process $Z(\cdot)$ has continuous sample paths and satisfies

$$\left\| Z(s) - Z(t) \right\|_\Psi \leq d(s,t) \; \forall \; s, t \in T,$$

and if there exists $t_0 \in T$ such that $\sup_{t \in T} d(t, t_0) = \delta$ for some δ, then

$$\left\| \sup_T \left| Z(t) - Z(s) \right| \right\|_\Psi \leq \sum_{i=0}^{\infty} \frac{\delta}{2^i} \sqrt{2 + \log D\big(\delta / 2^{i+1}, d, T\big)},$$

where D stands for packing number as before.

Proof: The proof involves a "chaining" argument, which is a "discretization" reminiscent of Huber's (1967) grid-construction. Since the results in Andrews (1994) deal with bounded pseudo-metric spaces, we can safely focus on the case $\delta < \infty$ (the case $\delta = \infty$ is trivial).

Step 1.- Denote

$$\delta_i \equiv \frac{\delta}{2^i}, \text{ for } i = 0, 1, 2, \ldots.$$

Now, construct a sequence of maximal sets

$$t_0 \equiv T_0, T_1, T_2, \ldots,$$

where each $T_i \subseteq T$ is the maximal (i.e, the largest) set in T with the property that

$$d(s,t) > \delta_i \quad \text{if } s, t \in T_i \text{ and } s \neq t.$$

Maximality implies that if $s \notin T_i$, then $\min_{t \in T_i} d(s,t) \leq \delta_i$. Note that, by definition of packing numbers, we must have $\#T_i = D(\delta_i, d, T)$.

Step 2.- For $i = 1, 2, \ldots,$ let $S_i \equiv T_i \notin T_{i-1}$. Approximate $\sup_T \left|Z(t) - Z(t_0)\right|$ with $\max_{t \in S_m} \left|Z(t) - Z(t_0)\right|$ for some S_m. This is a discretization of the problem because T_m and S_m are discrete sets. Note that $\#S_i \leq D(\delta_i, d, T)$.

Take any $t \in S_m$. By construction, for any such t we can always find a sequence of points leading from t to t_0,

$$t \equiv t_m, t_{m-1}, t_{m-2}, \ldots, t_1, t_0$$

such that: (i) $t_i \in S_i$ for each element in the sequence, and (ii) $d(t_i, t_{i-1}) \leq \delta_i$. Using this sequence, a triangle inequality yields

$$\begin{aligned} \max_{S_m} \left|Z(t) - Z(t_0)\right| &\leq \max_{S_m} \sum_{i=1}^{m} \left|Z(t_i) - Z(t_{i-1})\right| \\ &\leq \sum_{i=1}^{m} \max_{S_i} \left|Z(t_i) - Z(t_{i-1})\right|. \end{aligned}$$

Any well-defined norm must preserve this inequality. Going back to our Orlicz-Norm $\|\cdot\|_\Psi$, this yields (using a further round of the triangle inequality)

$$\left\|\max_{S_m}\left|Z(t)-Z(t_0)\right|\right\|_\Psi \leq \sum_{i=1}^{m}\left\|\max_{S_i}\left|Z(t_i)-Z(t_{i-1})\right|\right\|_\Psi$$

Step 3.- Using the bound in Lemma 3.2 of Pollard (1990) (mentioned above), we get

$$\begin{aligned}\left\|\max_{S_i}\left|Z(t_i)-Z(t_{i-1})\right|\right\|_\Psi &\leq \sqrt{2+\log\big(\#S_i\big)}\times\left\|\max_{S_i}\left|Z(t_i)-Z(t_{i-1})\right|\right\|_\Psi \\ &\leq \sqrt{2+\log D(\delta_i,t,T)}\times\max_{S_i} d(t_i,t_{i-1}) \\ &\leq \sqrt{2+\log D(\delta_i,t,T)}\times\delta_i.\end{aligned}$$

Therefore,

$$\left\|\max_{S_m}\left|Z(t)-Z(t_0)\right|\right\|_\Psi \leq \sum_{i=1}^{m}\sqrt{2+\log D(\delta_i,t,T)}\times\delta_i$$

Continuity of sample paths and a monotone-convergence argument allow us to link this result back to $\left\|\sup_T\left|Z(t)-Z(t_0)\right|\right\|_\Psi$ by taking

$$\lim_{m\to\infty}\left\|\max_{S_m}\left|Z(t)-Z(t_0)\right|\right\|_\Psi.$$

We get

$$\lim_{m\to\infty}\left\|\max_{S_m}\left|Z(t)-Z(t_0)\right|\right\|_{\Psi}\leq\sum_{i=1}^{\infty}\delta_i\sqrt{2+\log D(\delta_i,t,T)}.$$

To link this back to the definition of Entropy, we simply need to find an upper bound for $\sum_{i=1}^{\infty}\delta_i\sqrt{2+\log D(\delta_i,t,T)}$ in terms of an integral.

By definition of δ, we have $x<\delta\Rightarrow\ D(x,t,T)\geq 2$. Now, it is easy to verify that $\sqrt{2+\log(1+D)}<2.2\log(D)$ for $D\geq 2$. Using the fact that $\delta_i=2(\delta_i-\delta_{i+1})$, we get

$$\sum_{i=1}^{\infty}\delta_i\sqrt{2+\log D(\delta_i,t,T)}\leq 4.4\sum_{i=1}^{\infty}\big(\delta_i-\delta_{i+1}\big)\sqrt{\log D(\delta_i,t,T)}$$

By construction, $D(x,t,T)$ is a decreasing function of x. Therefore,

$$\big(\delta_i-\delta_{i+1}\big)\sqrt{\log D(\delta_i,t,T)}\leq\int_{\delta_{i+1}}^{\delta_i}\sqrt{\log D(x,t,T)}dx.$$

Therefore,

$$\sum_{i=1}^{\infty} \delta_i \sqrt{2 + \log D(\delta_i, t, T)} \leq 4.4 \sum_{i=1}^{\infty} \int_{\delta_{i+1}}^{\delta_i} \sqrt{\log D(x, t, T)} dx$$
$$= 4.4 \int_0^{\delta/2} \sqrt{\log D(x, t, T)} dx \leq 4.4 \int_0^{\delta} \sqrt{\log D(x, t, T)} dx.$$

We arrive at an expression of the form $4.4 \int_0^1 \sqrt{\log D(x, t, T)} dx$ by normalizing the pseudometric d by $\sup_{t \in T} d(t, t_0)$

U-Statistics and U-Processes

Let P be a distribution on a set S, let $Z_1, \ldots, Z_n$ an iid sample from P. Let f denote a real-valued function defined on $S^k = \underbrace{S \otimes S \cdots \otimes S}_{\text{k factors}}$ with $k \geq 1$. We define the **U-statistic of order *k*** by

$$U_{n,k} f = (n_k)^{-1} \sum_{\mathbb{I}_k} f(Z_{i_1}, \ldots, Z_{i_k})$$

where $(n)_k = n \times (n-1) \times \cdots \times (n-k+1)$, and $\mathbb{I}_k$ is the set of all $(n)_k$ ordered k-tuples of distinct integers from the set $\{1, \ldots, n\}$. Simply put, $(n)_k$ denotes the number of permutations of n elements, taking k at a time.

If $f(Z_{i_1}, \ldots, Z_{i_k})$ is symmetric in its arguments, then we can re-express $U_{n,k} f$ as

$$U_{n,k} = \binom{n}{k}^{-1} \sum_{i_1 < i_2 < \cdots < i_k} f(Z_{i_1}, \ldots, Z_{i_k})$$

where the summation index $i_1 < i_2 < \cdots < i_k$ signifies the set of all combinations of n elements, taking k at a time.

Consider the following functional notation: Take $k = 3$, then

$$f(P, s, t) = E\big[f(z_1, z_2, z_3)\big|z_2 = s, z_3 = t\big];$$
$$f(P, s, P) = E\big[f(z_1, z_2, z_3)\big|z_2 = s\big];$$
$$Qf = E\big[f(z_1, z_2, z_t)\big]$$

Note that Q is the product measure $Q = \underbrace{P \otimes \cdots \otimes P}_{\text{k factors}}$.

Suppose now that the function f is such that under the product measure $Q = P \otimes \cdots \otimes P$, the conditional expectation of f given any $k-1$ of its k arguments is identically zero. Then we say that f is ***P*-degenerate**, and that $\boldsymbol{U_{n,k}f}$ **is P-degenerate**.

Hoeffding Decomposition

Let f, P and Q be as described above. If $Qf < \infty$, then there exist real-valued functions $f_1, \ldots, f_k$ such that for each j, f_j is $P-$degenerate on S_j and

$$U_{n,k}f = Qf + P_n f_1 + \sum_{j=2}^{k} U_{n,j} f_j$$

where, for each z in S,

$$f_1(z) = f(z, P, \ldots, P) + f(P, z, P, \ldots, P) + \cdots + f(P, \ldots, P, z) - kQf$$

Asymptotic Properties of U-statistics

Hoeffding's decomposition will prove to be a very useful result because, as we shall see, degenerate U-statistics of higher order converge to zero "really fast". To see this, we characterize the expression for the variance of a U-statistic.

Variance of a U-statistic

Without loss of generality, consider a symmetric U-statistic of order k, based on a symmetric function $f(x_1, x_2, \ldots, x_k)$ which satisfies

$$E\big[f(X_1, \ldots, X_k)^2\big] < \infty.$$

Let

$$f_c(x_1, x_2, \ldots, x_c) \equiv f(x_1, x_2, \ldots, x_c, \underbrace{P, P, \ldots, P}_{\text{k-c times}})$$

$$= E\big[f(x_1, x_2, \ldots, x_c, X_{c+1}, \ldots, X_k)\big|X_1 = x_1, X_2 = x_2, \ldots, X_c = x_c\big].$$

Define $\zeta_0 = 0$ and, for $1 \leq c \leq k$, let

$$\zeta_c = \text{Var}\big[f_c(X_1, X_2, \ldots, X_c)\big].$$

It is intuitively clear that

$$0 = \zeta_0 \leq \zeta_1 \leq \cdots \leq \zeta_k = \text{Var}\big[f(X_1, \ldots, X_k)^2\big].$$

Lemma (Variance of a U-statistic)
The variance of the m-th order U-statistic

$$U_{n,k}f = \binom{n}{k}^{-1} \sum_{i_1<i_2<\cdots<i_k} f\big(X_{i_1}, X_{i_2}, \ldots, X_{i_k}\big)$$

is given by

$$\mathrm{Var}\big[U_{n,k}f\big] = \binom{n}{k}^{-1} \sum_{c=1}^{k} \binom{k}{c}\binom{n-k}{k-c}\zeta_c$$

Suppose our U-statistic $U_{n,k}$ is such that $0 \equiv \zeta_0 = \zeta_1 = \cdots = \zeta_{c-1} = 0 < \zeta_c$. The previous formula immediately yields

$$\mathrm{Var}\big[U_{n,k}f\big] = \frac{c!\binom{k}{c}^2}{n^c}\zeta_c + O\big(n^{-c-1}\big)$$

Letting $\theta \equiv E[U_{n,k}f]$, this yields

$$\mathrm{Var}\Big[n^{c/2}\big(U_{n,k}f - \theta\big)\Big] \longrightarrow c!\binom{k}{c}^2\zeta_c,$$

which suggests that the random variable $n^{c/2}\big(U_{n,k}f - \theta\big)$ converges in distribution to a nondegenerate distribution.

If $U_{n,k}f$ is a degenerate U-statistic, all of the above simply implies

$$U_{n,k}f = O_p(n^{-k/2})$$

Going back to the Hoeffding decomposition

$$U_{n,k}f = Qf + P_n f_1 + \sum_{j=2}^{k} U_{n,j} f_j = Qf + O_p(n^{-1/2}) + O_p(n^{-1}),$$

where $f_1(z) = f(z, P, \ldots, P) + f(P, z, P, \ldots, P) + \cdots + f(P, \ldots, P, z) - kQf$. If f is symmetric, the previous results immediately imply that if $E[f(X_1, \ldots, X_k)^2] < \infty$, and if $\zeta_1 > 0$:

$$\sqrt{n}\Big(U_{n,k} - Qf\Big) = \frac{k}{\sqrt{n}} \sum_{i=1}^{n} \Big(E[f(X_{i_1}, X_{i_2}, \ldots, X_{i_k}) \big| X_{i_1}] - Qf\Big) + \sqrt{n} O_p(n^{-1})$$
$$\xrightarrow{d} \mathcal{N}\big(0, k^2 \zeta_1\big).$$

This result is known as the **Central Limit Theorem for U-statistics** (See Theorem 5.5.1(A) in Serfling's book).

If $\zeta_1 = 0$ (so that $E[f(X_{i_1}, X_{i_2}, \ldots, X_{i_k})\big|X_{i_1}] = Qf$ w.p.1), but $\zeta_c > 0$ for some $c \geq 2$, then $n^{c/2}\Big(U_{n,k} - Qf\Big)$ will converge in distribution (possibly to a non-normal law).

Ahn and Powell's Projection Theorem for U-statistics

Useful in semiparametric settings, where the function f depends on n itself, basically through the presence of a bandwidth sequence that goes to zero as $n \to \infty$. Suppose f_n is symmetric. The Hoeffding decomposition is

$$\begin{aligned} U_{n,k} f_n &= Q f_n + P_n f_{1_n} + \sum_{j=2}^{k} U_{n,j} f_{j_n} \\ &= \frac{k}{n} \sum_{i=1}^{n} \Big(E\big[f_n\big(X_{i_1}, X_{i_2}, \ldots, X_{i_k}\big) \big| X_{i_1} \big] - Q f_n \Big) + O_p(n^{-1}\zeta_{1_n}) \end{aligned}$$

Recall from above that $0 \equiv \zeta_{0_n} \leq \zeta_{1_n} \leq \cdots \leq \zeta_{k_n} = \text{Var}\big[f_n\big(X_{i_1}, X_{i_2}, \ldots, X_{i_k}\big)\big]$.

Therefore, if $\mathrm{Var}\big[f_n\big(X_{i_1}, X_{i_2}, \ldots, X_{i_k}\big)\big]\big/\sqrt{n} \longrightarrow 0$, we will have

$$\sqrt{n}\big(U_{n,k} f_n - Q f_n\big) = \frac{k}{\sqrt{n}} \sum_{i=1}^{n} \Big(E\big[f_n\big(X_{i_1}, X_{i_2}, \ldots, X_{i_k}\big)\big|X_{i_1}\big] - Q f_n \Big) + o_p(1).$$

U-Processes

We define a U-process in the exact analogous way as an empirical process, except that we now deal with a "higher-order" summation of terms.

As before, we have a class of functions $\mathcal{F}$ which produces the U-process. The properties of the process can be determined by those of the class $\mathcal{F}$. A lot of good things happen if $\mathcal{F}$ has the same nice properties as we examined before: Euclidean or, more generally, manageable.

Moment Maximal Inequalities for U-Processes

We will only cite two corollaries of his main result here, which are used to prove the asymptotic normality of the Maximum Rank Correlation estimator:

Lemma

Let $\mathcal{F}$ be a class of zero-mean functions f on S^k, $k \geq 1$. If $\mathcal{F}$ is Euclidean for a constant envelope, then

$$\sup_{\mathcal{F}} \left| U_{n,k} f \right| = O_p(1/\sqrt{n}).$$

Lemma ♣

Let $\mathcal{F}$ be a class of P-degenerate functions on S^k, $k \geq 1$. If

(i) *$\mathcal{F}$ contains the zero function.*

(ii) *$\mathcal{F}$ is Euclidean for the constant envelope F,*

then

(a) $\sup_{\mathcal{F}} \left| n^{k/2} U_{n,k} f \right| = O_p(1)$.

(b) $\sup_{\mathcal{F}} \left| n^{k/2-\gamma} U_{n,k} f \right| \longrightarrow 0$ *almost surely.*

These maximal-inequality results (and other more elaborate extensions) prove to be extremely useful in semiparametric estimation asymptotics. We illuminate this by studying the Maximum-Rank Correlation estimator.

Heuristics of Asymptotic Normality of Maximum Rank Correlation (MRC) Estimator

Suppose all we know about Y and X is that $E[Y|X]$ is a monotonic, strictly increasing transformation $F_0(\cdot)$of the linear index $X'\beta$. The exact functional form of $F_0(\cdot)$ is unknown except for some invertibility assumptions.

Consider the objective function given by the following U-statistic:

$$G_n(\beta) = (n)_2^{-1} \sum_{i \neq j} \mathbb{1}\{Y_i > Y_j\} \mathbb{1}\{X_i'\beta > X_j'\beta\}$$

The maximizer is Han's Maximum Rank Correlation (MRC) estimator.

It bears that name because the estimator maximizes the correlation between the relative ranks of $Y_1, Y_2, \ldots, Y_n$ and the linear indices $X_1, X_2, \ldots, X_n$.

Please read the handout titled "Moment Maximal Inequalities for U-processes and Asymptotic Normality of Maximum Rank Correlation Estimator" to see, step-by-step, the proof that $\widehat{\beta}$ is $\sqrt{n}$-consistent, asymptotically normal.

Key to the asymptotic distribution results of the MRC estimator is a stronger version of Lemma ♣ (above), which characterizes a special case in which the $O_p(\cdot)$ can be replaced with $o_p(\cdot)$:

Lemma (Sherman)
Suppose all the conditions of the "Moment Maximal Inequalities Lemma" (above) hold and suppose that there exists $\beta_0 \in \Theta$ such that $f(\cdot, \beta_0) \equiv 0$. If the parameterization is $\mathcal{L}^2(Q)$-continuous at β_0, that is, if

$$\int \left| f(\cdot, \beta) \right|^2 dQ \longrightarrow 0 \quad \text{as} \quad \beta \longrightarrow \beta_0$$

then

$$n^{k/2} U_{n,k} f(\cdot, \beta) = o_p(1)$$

uniformly over $o_p(1)$ neighborhoods of β_0.

Asymptotic Distribution Results for Degenerate U-Statistics

From above, we concluded that if $U_{n,k}f$ is a degenerate U-statistic of order k, then $n^{k/2}U_{n,k}$ should converge in distribution to a nondegenerate law. We present two results for second-order U-statistics:

Serfling (1980), Theorem 5.5.2

Let $\{X_i\}_{i=1}^n$ be an iid sample with $X_i \sim F$. Let $h(X_{i_1}, \ldots, X_{i_k})$ be symmetric in all its k arguments, satisfying $E\big[h(X_{i_1}, \ldots, X_{i_k})\big] = 0$ and $E\big[h^2(X_{i_1}, \ldots, X_{i_k})\big] < \infty$. Let $h_2(x_1, x_2) = E\big[h(X_{i_1}, \ldots, X_{i_k})\big|X_{i_1} = x_1, X_{i_2} = x_2\big]$. For any squared-integrable function g define the linear operator $Ag(x) = \int_{-\infty}^{\infty} h_x(x,u)g(u)dF(u)$. For this operator, define the associated eigenvalues $\lambda_1, \lambda_2, \ldots$ to be the real numbers associated to the distinct solutions $g_1, g_2, \ldots,$ of the equation $Ag - \lambda g = 0$. Let $h_1(x) = E\big[h(X_{i_1}, \ldots, X_{i_k})\big|X_{i_1} = x\big]$. Then, if $\text{Var}\big[h_1(X_i)\big] = 0$, then

$$nU_{n,k}h \xrightarrow{d} \frac{k(k-1)}{2}\mathcal{Y},$$

where $\mathcal{Y}$ is a random variable of the form

$$\mathcal{Y} = \sum_{j=1}^{\infty} \lambda_j\big(\chi^2_{1_j} - 1\big), \quad \text{where } \chi^2_{1_1}, \chi^2_{1_2}, \ldots \text{ are independent } \chi^2_1 \text{ r.vs}$$

The following theorem focuses on cases where the function $h(\cdot)$ depends on the sample size n. It presents a result for degenerate, second-order U-statistics

Hall (1984), Theorem 1

Assume $h_n(X_i, X_j)$ is symmetric, and $E\big[h_n(X_i, X_j)\big|X_i\big] = 0$ almost surely, and $E\big[h_n^2(X_i, X_j)\big] < \infty$ for each n. Define $G_n(x_1, x_2) = E\big[h_n(X_i, x_1)h_n(X_i, x_2)\big]$. If

$$\frac{E\big[G_n^2(X_i, X_j)\big] + n^{-1}E\big[h_n^4(X_i, X_j)\big]}{E\big[h_n^2(X_i, X_j)\big]} \to 0,$$

then $\quad n \cdot U_{n,2}h_n \xrightarrow{d} \mathcal{N}\Big(0, 2E\big[h_n^2(X_i, X_j)\big]\Big)$.

The previous result is especially useful when the function $h_n(\cdot)$ involves a bandwidth sequence. Often, choosing the appropriate rate of convergence for such bandwidth will ensure the condition in the theorem is met. Extensions of this result to higher-order U-statistics can be found in Fan and Li (1996) and the references cited there.

Kernel-based Nonparametric Estimation

In a nutshell, this is an estimation technique that is very popular to estimate unknown densities or conditional moments that depend on a vector of random variables. In particular, it helps us deal with continuously-distributed random vectors. The aim is to estimate these unknown functions pointwise, for a given realization of these r/v's. Throughout we assume an iid sample $\left\{X_i\right\}_{i=1}^{n}$.

Density estimation

Consider the case of a real-valued random variable X, continuously distributed with density $f_X(\cdot)$ and support $\mathbb{S}(X)$. Suppose $f_X(\cdot)$ is continuously differentiable and uniformly bounded in $\mathbb{R}$ (we will add more smoothness assumptions soon). Let $K : \mathbb{R} \to \mathbb{R}$ denote a symmetric function that satisfies (we will add more conditions soon):

$$\int_{-\infty}^{\infty} K(\psi)d\psi = 1, \quad \int_{-\infty}^{\infty} \left|\psi K(\psi)\right| d\psi < \infty, \quad \sup_{\psi\in\mathbb{R}} \left|K(\psi)\right| \leq \overline{K} < \infty$$

Let h_n denote a bandwidth sequence satisfying $h_n \to 0$ (we will add more conditions soon).

Fix a value $x \in \mathbb{S}(X)$. A kernel estimator for $f_X(x)$ is given by

$$\widehat{f}_X(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h_n}\right).$$

With our assumptions so far, this density estimator is biased, but its bias disappears asymptotically:

$$E\Big[\widehat{f}_X(x)\Big] = \frac{1}{h_n} E\left[K\left(\frac{X_i - x}{h_n}\right)\right] = \frac{1}{h_n}\int_{-\infty}^{\infty} K\left(\frac{u-x}{h_n}\right) f_X(u)du.$$

A change of variable helps here. Let $\psi \equiv \frac{u-x}{h_n}$. The integral becomes (recall that $K(\cdot)$ is symmetric)

$$\begin{aligned} E\Big[\widehat{f}_X(x)\Big] &= \int_{-\infty}^{\infty} K(\psi) f_X(h_n\psi + x)d\psi = \int_{-\infty}^{\infty} K(\psi)\Big[f_X(x) + h_n\psi f_X(\widetilde{h}_n\psi + x)\Big]d\psi \\ &= f_X(x) + h_n\int_{-\infty}^{\infty} \psi K(\psi) f_X(\widetilde{h}_n\psi + x)d\psi = f_X(x) + \underbrace{h_n O(1)}_{=o(1)} \longrightarrow f_X(x). \end{aligned}$$

Ultimately, we will need to address much more than this. Next, let us take a look at $\text{Var}\left[\widehat{f}_X(x)\right]$:

$$\text{Var}\left[\widehat{f}_X(x)\right] = \frac{1}{n} \times \frac{1}{h_n^2}\text{Var}\left[K\left(\frac{X_i - x}{h_n}\right)\right] = \underbrace{\frac{1}{nh_n} \times \frac{1}{h_n}\text{Var}\left[K\left(\frac{X_i - x}{h_n}\right)\right]}_{\text{Convenient grouping}}.$$

we have

$$\text{Var}\left[K\left(\frac{X_i - x}{h_n}\right)\right] = E\left[K\left(\frac{X_i - x}{h_n}\right)^2\right] - \left\{E\left[K\left(\frac{X_i - x}{h_n}\right)^2\right]\right\}^2$$

$$= h_n \int_{-\infty}^{\infty} K^2(\psi) f_X(h_n\psi + x)d\psi - h_n^2\left\{\int_{-\infty}^{\infty} K(\psi) f_X(h_n\psi + x)d\psi\right\}^2 = h_n O(1),$$

The last equality is obtained by adding the requirement that $\int K^2(\psi)d\psi < \infty$. Therefore

$$\text{Var}\left[\widehat{f}_X(x)\right] = \frac{1}{nh_n}O(1).$$

Consistency of $\widehat{f}_X(x)$ will immediately require as a necessary condition that:

$$nh_n \to \infty.$$

The usual Chebyshev-Inequality argument will yield

$$\widehat{f}_X(x) \xrightarrow{p} E\big[\widehat{f}_X(x)\big] \xrightarrow{p} f_X(x) \quad \text{if } nh_n \to \infty.$$

How about establishing a rate of convergence, or equivalently, a Central Limit Theorem? The previous results will yield (via Liapunov's Central Limit Theorem)

$$\sqrt{nh_n}\Big(\widehat{f}_X(x) - E\big[\widehat{f}_X(x)\big]\Big) \xrightarrow{d} \mathcal{N}\left(0, f_X(x)\int K^2(\psi)d\psi\right)$$

In order to obtain the stronger result $\sqrt{nh_n}\Big(\widehat{f}_X(x) - f_X(x)\Big)$, we will need to ensure that $\sqrt{nh_n}\Big(E\big[\widehat{f}_X(x)\big] - f_X(x)\Big) \longrightarrow 0$. A way to achieve this is through the use of **"higher order"** or **"bias-reducing kernels"**, plus **additional smoothness assumptions** about $f_X(\cdot)$.

The symmetric (around zero) function $K : \mathbb{R} \to \mathbb{R}$ is a bias-reducing kernel of order M if:

$$\int_{-\infty}^{\infty} K(\psi)d\psi = 1, \quad \int_{-\infty}^{\infty} \psi^j K(\psi)d\psi \text{ for } j = 1, \ldots, M-1,$$

and $\int_{-\infty}^{\infty} \left|\psi^M K(\psi)\right| d\psi < \infty$. A bias-reducing kernel must take negative values. A quick way to construct one would be, for example, to combine a symmetric, continuous pdf with a polynomial. For example, a fifth-order bias reducing kernel:

$$K(\psi) = \left[a_0 + a_1\psi^2 + a_2\psi^4\right]\phi(\psi)$$

the coefficients a_0, a_1 and a_2 would be picked to force $\int_{-\infty}^{\infty} K(\psi)d\psi = 1$, $\int_{-\infty}^{\infty} \psi^2 K(\psi)d\psi = 0$ and $\int_{-\infty}^{\infty} \psi^4 K(\psi)d\psi = 0$. We will have $\int_{-\infty}^{\infty} \psi^j K(\psi)d\psi = 0$ for any odd-number j by virtue of the symmetry of $K(\psi)$.

Suppose we strengthen our assumptions and assume that $f_X(\cdot)$ is M-times differentiable with bounded derivatives. An M^{th}-order Taylor approximation yields

$$\begin{aligned} E\Big[\widehat{f}_X(x)\Big] &= \int_{-\infty}^{\infty} K(\psi)\Big[f_X(x) + \sum_{j=1}^{M-1}(h_n\psi)^j f_X^{(j)}(x) + (h_n\psi)^j f_X^{(M)}(\widetilde{h}_n\psi + x)\Big]d\psi \\ &= f_X(x) + h_n^M O(1). \end{aligned}$$

Therefore, if $nh_n \to \infty$, and $\sqrt{nh_n}h_n^M \to 0$, we will have

$$\sqrt{nh_n}\Big(\widehat{f}_X(x) - f_X(x)\Big) \xrightarrow{d} \mathcal{N}\left(0, f_X(x)\int K^2(\psi)d\psi\right).$$

Next, conditional moment estimators...

Conditional Moment Estimators

Let us continue for the moment assuming that $X \in \mathbb{R}$ (real-valued). Suppose Y is also a real-valued random variable (the extension to multivariate Y is immediate, since we would apply the following analysis element-wise to each component of Y).

Suppose $(Y_i, X_i)_{i=1}^n \sim F_{Y,X}(\cdot,\cdot)$ iid. Let $E[Y_i|X_i = x] \equiv \mu(x)$ be well-defined for $x \in \mathbb{S}(X)$. We have

$$\mu(x) = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy = \frac{\int_{-\infty}^{\infty} y f_{Y,X}(y,x)dy}{f_X(x)} \equiv \frac{R(x)}{f_X(x)}.$$

We estimate $\mu(x)$ by using kernel-weighed analog objects to those described above.

$$\widehat{R}(x) = \frac{1}{nh_n}\sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h_n}\right), \quad \widehat{f}_X(x) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h_n}\right), \quad \widehat{\mu}(x) = \frac{\widehat{R}(x)}{\widehat{f}_X(x)}$$

We would also arrive at this expression if we choose $\widehat{\mu}(x)$ to minimize:

$$\sum_{i=1}^{n}\Big(Y_i - \widehat{\mu}(x)\Big)^2 K\left(\frac{X_i - x}{h_n}\right) \quad \text{(a kernel-weighed sum of squared residuals)}$$

We have already studied the properties of $\widehat{f}_X(\cdot)$. Those of $\widehat{R}(x)$ can be studied in the exact same way. First, note that

$$E\big[\widehat{R}(x)\big] = \frac{1}{h_n} E\left[\mu(X_i) K\left(\frac{X_i - x}{h_n}\right)\right] = \int_{-\infty}^{\infty} K(\psi)\mu\big(h_n\psi + x\big) f_X\big(h_n\psi + x\big) d\psi.$$

Now, in addition to smoothness assumptions for $f_X(\cdot)$, we will need smoothness assumptions for $\mu(\cdot)$. If both $\mu(\cdot)$ and $f_X(\cdot)$ are M-times differentiable with bounded derivatives, and M^{th}-order bias reducing kernel will yield (using a Taylor approximation)

$$E\big[\widehat{R}(x)\big] = f_X(x)\mu(x) + h_n^M O(1) \longrightarrow f_X(x)\mu(x) \equiv R(x)$$

Let $\mu_2(x) \equiv E\big[Y_i^2 \big| X_i = x\big]$. Then, if $nh_n \to \infty$ and $\sqrt{nh_n} h_n^M \to 0$, we will have

$$\sqrt{nh_n}\Big(\widehat{R}(x) - R(x)\Big) \xrightarrow{d} \mathcal{N}\left(0, \mu_2(x) f_X(x) \int K^2(\psi) d\psi\right)$$

To find the asymptotic distribution of $\widehat{\mu}(x)$, take a simple, second-order Taylor approximation:

$$\begin{aligned}\widehat{\mu}(x) &= \mu(x) + \frac{\widehat{R}(x) - R(x)}{f_{_X}(x)} - \mu(x)\left[\frac{\widehat{f}_{_X}(x) - f_{_X}(x)}{f_{_X}(x)}\right] + O_p\left(\frac{1}{nh_{_n}}\right) \\ &= \mu(x) + \frac{\widehat{R}(x) - \mu(x)\widehat{f}_{_X}(x)}{f_{_X}(x)} + O_p\left(\frac{1}{nh_{_n}}\right).\end{aligned}$$

Therefore,

$$\widehat{\mu}(x) - \mu(x) = \frac{1}{nh_{_n}}\sum_{i=1}^{n}\frac{\big[Y_i - \mu(x)\big]}{f_{_X}(x)}K\left(\frac{X_i - x}{h_{_n}}\right) + O_p\left(\frac{1}{nh_{_n}}\right).$$

If $\mu(x)$ and $f_{_X}(x)$ are M-times differentiable with bounded derivatives, and we use an M^{th}-order bias reducing kernel, we have

$$\frac{1}{\sqrt{nh_{_n}}}\sum_{i=1}^{n}\frac{\big[Y_i - \mu(x)\big]}{f_{_X}(x)}K\left(\frac{X_i - x}{h_{_n}}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2(x)}{f_{_X}(x)}\int_{-\infty}^{\infty}K^2(\psi)d\psi\right)$$

where $\sigma^2(x) = \text{Var}\big[Y_i\big|X_i = x\big]$.

Therefore

$$\sqrt{nh_n}\Big(\widehat{\mu}(x) - \mu(x)\Big) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma^2(x)}{f_X(x)} \int_{-\infty}^{\infty} K^2(\psi) d\psi\right)$$

Higher-dimensional vectors and the "Curse of Dimensionality"

Suppose now that $\boldsymbol{X}_i \in \mathbb{R}^L$ is a vector of jointly continuously distributed rv's. A way to deal with this is to use an L^{th}-dimensional kernel $K : \mathbb{R}^L \to \mathbb{R}$. The simplest such kernel would be multiplicative:

$$K(\Psi) = K(\psi_1) \times K(\psi_2) \times \cdots \times \times K(\psi_L).$$

Density estimation

We would estimate $f_X(\boldsymbol{x})$ by

$$\widehat{f}_X(\boldsymbol{x}) = \frac{1}{nh_n^L} \sum_{i=1}^{n} K\left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{h_n}\right)$$

Its expectation would be given by the L^{th}-dimensional integral

$$E\big[\widehat{f}_{X}(\boldsymbol{x})\big] = \frac{1}{h_{n}^{L}} \int K\left(\frac{\boldsymbol{u}-\boldsymbol{x}}{h_{n}}\right) f_{X}(\boldsymbol{u}) d\boldsymbol{u} = \int K\big(\boldsymbol{\Psi}\big) f_{X}\big(h_{n}\boldsymbol{\Psi}+\boldsymbol{x}\big) d\boldsymbol{\Psi},$$

where $\boldsymbol{\Psi} = \big(\psi_1, \psi_2, \ldots, \psi_L\big)$, with $\psi_j = \big(u_j - x_j\big)/h_{n}$. An M^{th}-order bias-reducing kernel is now a symmetric (around zero) function $K : \mathbb{R}^{L} \rightarrow \mathbb{R}$ that satisfies, for any vector $\boldsymbol{\Psi} \equiv \big(\psi_1, \psi_2, \ldots, \psi_L\big)$

$$\int K(\boldsymbol{\Psi}) d\boldsymbol{\Psi} = 1, \quad \int \psi_1^{q_1} \psi_2^{q_2} \cdots \psi_L^{q_L} K(\boldsymbol{\Psi}) d\boldsymbol{\Psi} = 0 \; \forall \; \{q_j\}_{j=1}^{L} : \sum_{j=1}^{L} q_j \leq M-1$$

and $\displaystyle\int \big\|\boldsymbol{\Psi}\big\|^{M} \big|K(\boldsymbol{\Psi})\big| d\boldsymbol{\Psi} < \infty.$

If $f_{X}(\cdot)$ has bounded cross-partial derivatives up to order M, then we will have an extension of the result in the one-dimensional case:

$$E\big[\widehat{f}_{X}(\boldsymbol{x})\big] = f_{X}(\boldsymbol{x}) + h_{n}^{M} O(1)$$

The "curse of dimensionality" becomes apparent when we analyze $\mathrm{Var}\big[\widehat{f}_{X}(\boldsymbol{x})\big]$.

We will have

$$\text{Var}\big[\widehat{f}_X(\boldsymbol{x})\big] = \frac{1}{nh_n^L} O(1).$$

Therefore, under the smoothness conditions above, and using a bias-reducing kernel of order M:

$$\text{if } nh_n^L \to \infty, \text{ and } \sqrt{nh_n^L}h_n^M \to 0, \sqrt{nh_n^L}\Big(\widehat{f}_X(\boldsymbol{x}) - f_X(\boldsymbol{x})\Big) \xrightarrow{d} \mathcal{N}\big(0, V\big).$$

The rate of convergence of $\widehat{f}_X(\boldsymbol{x})$ decreases with L.

Conditional Moment Estimators

Suppose $Y_i \in \mathbb{R}$. We want to estimate $\mu(\boldsymbol{x}) = E\big[Y_i \big| \boldsymbol{X}_i = \boldsymbol{x}\big]$. We proceed as before, by generalizing what we did in one dimension

$$\widehat{R}(\boldsymbol{x}) = \frac{1}{nh_n^L}\sum_{i=1}^n Y_i K\left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{h_n}\right), \quad \widehat{f}_X(\boldsymbol{x}) = \frac{1}{nh_n}\sum_{i=1}^n K\left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{h_n}\right), \quad \widehat{\mu}(\boldsymbol{x}) = \frac{\widehat{R}(\boldsymbol{x})}{\widehat{f}_X(\boldsymbol{x})}$$

Under the smoothness assumptions, we can obtain a generalization of the one-dimensional result

$$\widehat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x}) = \frac{1}{nh_n^L}\sum_{i=1}^n \frac{\big[Y_i - \mu(\boldsymbol{x})\big]}{f_X(\boldsymbol{x})} K\left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{h_n}\right) + O_p\left(\frac{1}{nh_n^L}\right). \qquad (2)$$

Uniform Linear-representation Results and Semiparametric Estimation

Suppose that we are able to establish that, for any compact set $\mathcal{X} \in \mathbb{S}(X)$ such that $\inf_{x \in \mathcal{X}} f_X(x) \geq \underline{f} > 0$,

$$\widehat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x}) = \frac{1}{n h_n^L} \sum_{i=1}^{n} \frac{\big[Y_i - \mu(\boldsymbol{x})\big]}{f_X(\boldsymbol{x})} K\left(\frac{\boldsymbol{X}_i - \boldsymbol{x}}{h_n}\right) + \xi_n(\boldsymbol{x}),$$

where $\sup_{x \in \mathcal{X}} \big|\xi_n(\boldsymbol{x})\big| = O_p\Big(\frac{1}{n^{1-\delta} h_n^L}\Big)$ for any $\delta > 0$.

Let us go back to section $(\lozenge)$ ("Stochastic equicontinuity and semiparametric estimation"). Consider the estimator $\widehat{\theta}$ that satisfies $\widehat{\theta} \xrightarrow{p} \theta_0$ and

$$\sqrt{N}\overline{m}_N(\widehat{\theta}, \widehat{\mu}) = o_p(1), \quad \text{with} \quad \overline{m}_N(\theta, \widehat{\mu}) = \frac{1}{N} \sum_{i=1}^{N} m(X_i; \theta, \widehat{\mu}(W_i)) \mathbb{1}\{W_i \in \mathcal{W}\}.$$

denote $\mu_0(w) = E\big[Y_j \big| W_j = w\big]$ (we use the subscript $\mu_0(\cdot)$ to follow the notation used in the empirical-process sections), and the set $\mathcal{W}$ is such that

$$\widehat{\mu}(\boldsymbol{w}) - \mu_0(\boldsymbol{w}) = \frac{1}{N h_N^L} \sum_{i=1}^{N} \frac{\big[Y_i - \mu_0(\boldsymbol{w})\big]}{f_W(\boldsymbol{w})} K\left(\frac{\boldsymbol{W}_i - \boldsymbol{w}}{h_N}\right) + \xi_N(\boldsymbol{w}),$$

where $\sup_{w \in \mathcal{W}} \left| \xi_N(\boldsymbol{w}) \right| = O_p\left(\frac{1}{N^{1-\delta} h_N^L} \right)$ for any $\delta > 0$.

The function $m(\cdot)$ was assumed to be smooth enough, so that we could approximate (see Eq. 1)

$$\sqrt{N}(\widehat{\theta} - \theta_0) = -\nabla_\theta \overline{m}_N(\widetilde{\theta}, \widehat{\mu})^{-1} \sqrt{N} \overline{m}_N(\theta_0, \widehat{\mu}) + o_p(1). \tag{3}$$

Let $\lambda(\theta, \mu) = E[\overline{m}_N(\theta, \mu)]$. The relevant empirical process here will be

$$\nu_N(\theta, \mu) = \sqrt{N}\Big(\overline{m}_N(\theta, \mu) - \lambda(\theta, \mu) \Big)$$

We determined that if the process $\nu_N(\theta, \mu)$ was stochastically equicontinuous, then

$$\sqrt{N} \overline{m}_N(\theta_0, \widehat{\mu}) = \nu_N(\theta_0, \mu_0) + \sqrt{N} \lambda(\theta_0, \widehat{\mu})$$

Suppose we show that, for any $\left\{ \widetilde{\mu}(W_i) \right\}_{i=1}^N$ such that $\widetilde{\mu}(W_i)$ is between $\widehat{\mu}(W_i)$ and $\mu_0(W_i)$ for each i, we have

$$\frac{1}{N} \sum_{i=1}^N \left\| \nabla_{\mu\mu'} \Big(E\big[m(X_i, \theta_0, \widetilde{\mu}(W_i)) \big] \Big) \right\| = O_p(1).$$

Suppose also that we choose the bandwidth $h_{_N}$ such that $N^{1/2}\big/\big(N^{1-\delta}h_{_N}^{L}\big) \to 0$ for some $\delta > 0$. This, and the uniform linear representation result from above would yield

$$\sup_i \left\| \widehat{\mu}(W_i) - \mu_0(W_i) \right\|^2 = o_p(N^{-1/2}).$$

Then, we have

$$\sqrt{N}\lambda(\theta_0, \widehat{\mu}) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \nabla_\mu \Big(E\big[m(X_i, \theta_0, \mu_0(W_i))\big] \Big) \big(\widehat{\mu}(W_i) - \mu_0(W_i) \big) + o_p(1)$$

We have to examine the term

$$\begin{aligned} &\frac{1}{N} \sum_{i=1}^{N} \nabla_\mu \Big(E\big[m(X_i, \theta_0, \mu_0(W_i))\big] \Big) \big(\widehat{\mu}(W_i) - \mu_0(W_i) \big) \\ &= \frac{1}{N} \sum_{i=1}^{N} \nabla_\mu \Big(E\big[m(X_i, \theta_0, \mu_0(W_i))\big] \Big) \left[\frac{1}{N h_{_N}^{L}} \sum_{j=1}^{N} \frac{\big[Y_j - \mu_0(W_i)\big]}{f_{_W}(W_i)} K\left(\frac{W_i - W_j}{h_{_N}} \right) \right. \\ &\qquad \left. + \xi_{_N}(W_i) \right] \end{aligned}$$

Abbreviate $\varphi(X_i, W_i) \equiv \nabla_\mu \Big(E\big[m(X_i, \theta_0, \mu_0(W_i))\big] \Big)$. We can split the previous term as

$$\underbrace{\frac{K(0)}{N h_N^L} \times \underbrace{\frac{1}{N}\sum_{i=1}^{N} \varphi(X_i, W_i) \frac{\big[Y_i - \mu_0(W_i)\big]}{f_W(W_i)}}_{=O_p(1)} + \frac{(N-1)}{2N} \times}_{=O_p\left(1/(N h_N^L)\right)}$$

$$\binom{N}{2}^{-1} \sum_{i<j} \left[\varphi(X_i, W_i) \frac{\big[Y_j - \mu_0(W_i)\big]}{h_N^L f_W(W_i)} + \varphi(X_j, W_j) \frac{\big[Y_i - \mu_0(W_j)\big]}{h_N^L f_W(W_j)} \right] K\left(\frac{W_i - W_j}{h_N}\right)$$

$$+ \underbrace{\frac{1}{N}\sum_{i=1}^{N} \varphi(X_i, W_i) \xi_N(W_i)}_{\leq \sup_i \left|\xi_N(W_i)\right| \times \frac{1}{N}\sum_{i=1}^{N} \left|\left|\varphi(W_i, X_i)\right|\right| = o_p(N^{-1/2})}$$

So it all comes down to the symmetric, second-order U-statistic

$$\binom{N}{2}^{-1}\sum_{i<j}\left[\varphi(X_i,W_i)\frac{\left[Y_j-\mu_0(W_i)\right]}{h_N^L f_W(W_i)}+\varphi(X_j,W_j)\frac{\left[Y_i-\mu_0(W_j)\right]}{h_N^L f_W(W_j)}\right]K\left(\frac{W_i-W_j}{h_N}\right)$$

We have

$$E\left[\varphi(X_i,W_i)\frac{\left[Y_j-\mu_0(W_i)\right]}{h_N^L f_W(W_i)}K\left(\frac{W_i-W_j}{h_N}\right)\middle| X_i,W_i,Y_i\right]$$

$$= E\left[\varphi(X_i,W_i)\frac{\left[\mu_0(W_j)-\mu_0(W_i)\right]}{h_N^L f_W(W_i)}K\left(\frac{W_i-W_j}{h_N}\right)\middle| X_i,W_i,Y_i\right]$$

If $\mu_0(\cdot)$ and $f_W(\cdot)$ are M-times differentiable with respect to W with bounded derivatives, then a Taylor expansion and the use of an M^{th}-order kernel will yield

$$E\left[\varphi(X_i,W_i)\frac{\left[Y_j-\mu_0(W_i)\right]}{h_N^L f_W(W_i)}K\left(\frac{W_i-W_j}{h_N}\right)\middle| X_i,W_i\right] = h_N^M\frac{\varphi(X_i,W_i)}{f_W(W_i)}\mathcal{R}_N^1(W_i)$$

where $\sup_i\left|\mathcal{R}_N(W_i)\right| \leq \overline{R}^1 < \infty$

We have

$$E\left[\varphi(X_j,W_j)\frac{\big[Y_i-\mu_0(W_j)\big]}{h_N^L f_W(W_j)}K\left(\frac{W_i-W_j}{h_N}\right)\middle|X_i,W_i,Y_i\right]$$
$$=E\left[\underbrace{E\big[\varphi(X_j,W_j)\big|W_j\big]}_{\equiv\Sigma(W_j)}\frac{\big[Y_i-\mu_0(W_j)\big]}{h_N^L f_W(W_j)}K\left(\frac{W_i-W_j}{h_N}\right)\middle|X_i,W_i,Y_i\right]$$
$$=\int\Sigma(u)\frac{\big[Y_i-\mu_0(u)\big]}{h_N^L f_W(u)}K\left(\frac{W_i-u}{h_N}\right)f_W(u)du$$

If $\Sigma(\cdot)$ is M-times differentiable with bounded derivatives, a Taylor approximation produces

$$E\left[\varphi(X_j,W_j)\frac{\big[Y_i-\mu_0(W_j)\big]}{h_N^L f_W(W_j)}K\left(\frac{W_i-W_j}{h_N}\right)\middle|X_i,W_i,Y_i\right]$$
$$=\Sigma(W_i)\big[Y_i-\mu_0(W_i)\big]+h_N^M\mathcal{R}_N^{(2)}(W_i)$$

The last step before using Hoeffding's decomposition to our advantage is to verify that the variance of our U-statistic disappears sufficiently fast. In order to safely ignore the second-order degenerate U-statistic in the decomposition, it is enough if:

$$\sqrt{N}\frac{1}{N}\mathrm{E}\left\{\left[\varphi(X_i,W_i)\frac{\left[Y_j-\mu_0(W_i)\right]}{h_N^L f_W(W_i)}+\varphi(X_j,W_j)\frac{\left[Y_i-\mu_0(W_j)\right]}{h_N^L f_W(W_j)}\right]^2 K\left(\frac{W_i-W_j}{h_N}\right)^2\right\}$$

goes to zero as $N \longrightarrow \infty$. It all comes down to showing that

$$\frac{1}{\sqrt{N}h_N^L}\mathrm{E}\left\{\frac{1}{h_N^L}\left[\varphi(X_i,W_i)\frac{\left[Y_j-\mu_0(W_i)\right]}{f_W(W_i)}+\varphi(X_j,W_j)\frac{\left[Y_i-\mu_0(W_j)\right]}{f_W(W_j)}\right]^2 K\left(\frac{W_i-W_j}{h_N}\right)^2\right\}$$

goes to zero. As long as the expectations involved exist, we only need $\sqrt{N}h_N^L \longrightarrow \infty$. This strengthens the previous condition that $Nh_N^L \longrightarrow \infty$.

The projection of our U-statistic yields

$$\frac{1}{N}\sum_{i=1}^{N}\nabla_{\mu}\Big(E\big[m(X_i,\theta_0,\mu_0(W_i))\big]\Big)\big(\widehat{\mu}(W_i)-\mu_0(W_i)\big)$$

$$=\frac{1}{N}\sum_{i=1}^{N}\Sigma(W_i)\big[Y_i-\mu_0(W_i)\big]+\underbrace{h_N^M\frac{1}{N}\sum_{i=1}^{N}\big[\mathcal{R}_N^1(W_i)+\mathcal{R}_N^{(2)}(W_i)\big]}_{=O_p(h_N^M)}+o_p(N^{-1/2})$$

therefore, as long as $\sqrt{N}h_N^M \longrightarrow 0$, our semiparametric estimator will satisfy

$$\sqrt{N}\big(\widehat{\theta}-\theta_0\big)=M^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\Big[m\big(X_i,\theta_0,\mu_0(W_i)\big)+\Sigma(W_i)\big[Y_i-\mu_0(W_i)\big]\Big]+o_p(1)$$

$$\equiv\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\psi_i+o_p(1)$$

ψ_i is the influence function for $\widehat{\theta}$.

Many examples of existing semiparametric estimators can be studied in this fashion (or some variation). A nice list of examples can be found in Chapter 5 of Pagan and Ullah's book.

Semiparametric Efficiency Bounds

Suppose we have an estimation problem based on a model that is assumed to satisfy a set of semiparametric assumptions. The sample we observe is assumed to come from a data generating process that satisfies these assumptions.

Being more precise, we can think of the data as being generated by a particular parametric model that satisfies the semiparametric assumptions. That is, the data we observe comes from a (parametric) *submodel* that contains the truth.

Examples of parametric submodels: Suppose we are interested in the parameter $\beta_0 = E[Z]$, where the distribution of Z is unknown. A parametric submodel would consist of any likelihood function $f(z|\theta)$ such that $f(z|\theta_0)$ is the true distribution of Z.

A second example could be a partially linear model, given by

$$Y = X'\beta_0 + g_0(V) + \varepsilon,$$

where both X and V are observable to the econometrician, but the exact functional form of $g_0(\cdot)$ is unknown. If the joint density $h(X, V, \varepsilon)$ is also unknown, then a parametric submodel would be any $h(x, v, \varepsilon; \eta)$ and $g_0(v, \gamma)$. The parameters of any parametric submodel are $\theta = (\beta, \eta, \gamma)$. The true data generating process corresponds to the case $\theta = \theta_0$.

Smooth Parametric Submodel: The exact definition is found in Appendix A. Definition A.1, in Newey (1990) (Journal of Applied Econometrics). Basically, it requires $f(z|\theta)$ to be continuously-differentiable (z denotes observable data), with squared-integrable derivative on an open set Θ. The score and information matrix for a smooth parametric submodel is given by

$$S_\theta, \quad \text{and} \quad E_\theta\left[S_\theta S_\theta'\right]$$

A regular parametric submodel is one that is smooth and has a nonsingular information matrix.

Clearly, one can always express the parameter of interest as $\beta(\theta)$. In the first example, this is $\beta(\theta) = \int z f(z|\theta) dz$, while on the second case, β is simply a subvector of θ.

The asymptotic variance of any semiparametric estimator is no smaller than the supremum of the Cramer-Rao lower bounds ($E_\theta\left[S_\theta S_\theta'\right]^{-1}$) for all parametric submodels. We denote this lower bound by V.

Efficiency rankings is only meaningful for estimators that are regular in a sense shared by MLE:

Regular Estimators: A Local Data Generating Process (LDGP) is one such that, for each sample size n, the data is distributed according to θ_n, where $\sqrt{n}(\theta_n - \theta_0)$ is bounded. An estimator $\widehat{\beta}$ is regular in a parametric submodel if the limiting distribution of $\sqrt{n}\left(\widehat{\beta} - \beta(\theta_n)\right)$ is the same for any LDGP. An estimator is regular if it is regular in any regular parametric submodel.

Theorem 2.1 in Newey (1990) If $\widehat{\beta}$ is regular, then the limiting distribution of $\sqrt{n}(\widehat{\beta} - \beta_0)$ is equal to the distribution of $Y + U$, where Y is distributed as $\mathcal{N}(0, V)$ and U is some random vector independent of Y.

Thus, the asymptotic variance of $\widehat{\beta}$ is $V + E[UU']$. A semiparametric estimator is efficient if its limiting distribution is $N(0, V)$ and it is regular.

Asymptotically Linear Estimator: An estimator $\widehat{\beta}$ is asymptotically linear if it is asymptotically equivalent to a sample average with mean zero:

$$\sqrt{n}\big(\widehat{\beta} - \beta_0\big) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_i + o_p(1), \quad E\big[\psi\big] = 0, \quad E\big[\psi\psi'\big] \text{ finite and nonsingular.}$$

we call ψ the Influence Function of $\widehat{\beta}$.

The efficiency bound applies to regular estimators. The next result provides sufficient and necessary conditions for asymptotically linear estimators.

Theorem 2.2 in Newey (1990) Suppose $\widehat{\beta}$ is an asymptotically linear estimator with influence function ψ, and for all regular parametric submodels $\beta(\theta)$ is differentiable and $E_\theta\big[\|\psi\|^2\big]$ exists and is continuous on a neighborhood of θ_0. Then $\widehat{\beta}$ is regular if and only if, for all regular parametric submodels,

$$\frac{\partial \beta(\theta_0)}{\partial \theta} = E\big[\psi S_\theta'\big].$$

Example: M-Estimators (e.g, Huber 1967) We saw in our extensive discussions about M-estimators that satisfy $\sum_{i=1}^n m(z_i, \widehat{\beta}) = o_p(n^{-1/2})$ that they have influence function given by

$$\psi(z) = -M^{-1} m(z, \beta_0), \quad \text{where} \quad M = \left.\frac{\partial E\big[m(z,\beta)\big]}{\partial \beta}\right|_{\beta=\beta_0}$$

To see that M-estimators are regular, note that the definition of β_0 is that: $E_\theta\big[m(z, \beta(\theta))\big] = 0$. This equality must hold for all $\theta \in \Theta$. Direct differentiation yields

$$M \times \frac{\partial \beta(\theta_0)}{\partial \theta} + E\big[m(z,\beta_0) S_\theta'\big] = 0,$$

or equivalently, $\frac{\partial \beta(\theta_0)}{\partial \theta} = E\big[\psi S_\theta'\big]$. This is the so-called "Generalized Information Matrix Equality".

Efficiency Bounds and Influence Function of Regular Estimators

Fix a regular parametric submodel and suppose $\beta(\theta)$ is differentiable. Let $\widehat{\theta}$ be the MLE, and let $\widehat{\beta} = \beta(\widehat{\theta})$. This way, $\widehat{\beta}$ would be the more efficient estimator of β (for this submodel), and using the Delta Method its variance is given by

$$V_\theta = \frac{\partial \beta(\theta_0)}{\partial \theta} \Big(E\big[S_\theta S_\theta'\big] \Big)^{-1} \frac{\partial \beta(\theta_0)'}{\partial \theta} = E\big[\psi S_\theta'\big] \Big(E\big[S_\theta S_\theta'\big] \Big)^{-1} E\big[S_\theta \psi'\big].$$

This result immediately shows that V_θ is the efficiency bound for any regular estimator of $\beta(\theta)$ because

$$E\big[\psi\psi'\big] - V_\theta = E\Big[\big(\psi - AS_\theta\big)\big(\psi - AS_\theta\big)'\Big]$$

with $A = E\big[\psi S_\theta'\big] \Big(E\big[S_\theta S_\theta'\big] \Big)^{-1}$.

Computing the semiparametric efficiency bound would involve searching for the supremum of V_θ over all regular submodels. We briefly describe the issues involved next...

Finding the Semiparametric Efficiency Bound

We say that $\beta(\theta)$ is a differentiable parameter if it is differentiable for all smooth parametric submodels and if there exists a $q \times 1$ (q is the dimension of β) random vector d such that $E\left[dd'\right]$ is finite, and for all regular parametric submodels:

$$\frac{\partial \beta(\theta_0)}{\partial \theta} = E\left[dS'_\theta\right]$$

From above, we know that if there exists a regular, asymptotically linear estimator $\widehat{\beta}$ with influence function ψ, then the above requirement will be satisfied with $d = \psi$. Note that d is not unique. We can add any u that is orthogonal to S_θ, and it would serve as well. In particular, we can add any constant to d (because S_θ has expected value equal to zero). From our previous discussions, the Cramer-Rao lower bound for a parametric submodel is now given by

$$V_\theta = E\left[dS'_\theta\right]\left(E\left[S_\theta S'_\theta\right]\right)^{-1} E\left[S_\theta d'\right] = E\left[d_\theta d'_\theta\right],$$

where $d_\theta = E\left[dS'_\theta\right]\left(E\left[S_\theta S'_\theta\right]\right)^{-1} S_\theta$. By looking closely at d_θ, it becomes clear that it is the orthogonal projection of d on S_θ (it is the "population regression" of d on S_θ).

So V_θ, the Cramer-Rao lower bound for $\beta(\theta)$ for a parametric submodel is given by the variance of the orthogonal projection of d on S_θ. The semiparametric efficiency bound cannot be smaller than the worst possible case for all (regular) parametric submodels. How can we look for this supremum?

Take a collection of parametric submodels $j = 1, \ldots, J$. Let S_{θ_j} denote the score for the j^{th} submodel. Then, the variance of the projection of d onto any individual S_{θ_j} cannot be greater (in a matrix sense) than the variance of the projection of d onto the linear space spanned by $S_{\theta_1}, S_{\theta_2}, \ldots, S_{\theta_J}$. This is equivalent to adding more variables to a regression: It will never decrease the variance (in a matrix sense) of the resulting predicted values.

Using this intuition, an automatic upper bound for V_θ for any parametric submodel could result from projecting d onto the linear space spanned by a "sufficiently large" number of parametric scores for different parametric submodels. We focus only on smooth parametric submodels (see definition above).

This notion is formalized by taking the orthogonal projection of d on to a Hilbert space. We define the tangent set $\mathcal{S}$ as

$$\mathcal{S} = \left\{ s \in \mathbb{R}^q : \; E\left[\left\|s\right\|^2\right] < \infty, \text{ and } \exists \, \{A_j\}_{j=1}^J : \; \lim_{J\to\infty} E\left[\left\| s - \sum_{j=1}^{J} A_j S_{\theta_j} \right\|^2\right] = 0 \right\}$$

each A_j is a matrix of constants with q rows, conformable with S_{θ_j}. The tangent set $\mathcal{S}$ is the mean-squared closure of all q-dimensional linear combinations of S_θ for smooth parametric submodels.

The orthogonal projection of d on $\mathcal{S}$ is called the efficient score. Denote it by δ, which satisfies (by definition of orthogonal projection)

$$\delta \in \mathcal{S}, \quad E\left[(d-\delta)'s\right] = 0 \quad \text{for all} \quad s \in \mathcal{S}$$

Theorem 3.1 in Newey (1990) Suppose the parameter is differentiable, $\mathcal{S}$ is linear and $E[\delta\delta']$ is nonsingular, where δ is the projection of d on $\mathcal{S}$. Then the semiparametric efficiency bound is $V = E[\delta\delta']$.

Therefore, to find the efficiency bound of a semiparametric estimator, one must first characterize the tangent set $\mathcal{S}$, and then find the projection of d (any d that satisfies the requirement of a differentiable parameter will yield the same projection) on to $\mathcal{S}$.

Characterizing $\mathcal{S}$ will reflect all the restrictions placed on the scores S_θ that are placed by whatever semiparametric assumptions we impose.

If we have a regular, asymptotically linear estimator $\widehat{\beta}$ with influence function ψ and we find that $\psi \in \mathcal{S}$, then this estimator achieves the efficiency bound.

An excellent paper which presents a clear methodology to find efficiency bounds is Severini and Tripathi (2001). Several examples are also included in Newey (1990).

Implications of Efficiency Bounds for Semiparametric Estimators
We begin by stating the next result:

Theorem 3.2 in Newey (1990) If $f(z|\beta)$ is smooth with score S_β, $\mathcal{S}$ is linear, and the residual S of the projection of S_β on $\mathcal{S}$ is such that $E\big[SS'\big]$ is nonsingular, then β is a differentiable parameter and has efficient influence function

$$\big(E[SS']\big)^{-1}S, \quad \text{with } V = \big(E[SS']\big)^{-1}.$$

Infinite efficiency bounds and nonexistence of $\sqrt{n}$-consistent estimators
Chamberlain (1986) showed that if the bound is infinitely large (e.g, if $E[SS']$ is singular, or in general, if the bound involves taking the inverse of a singular matrix), then no $\sqrt{n}$-consistent estimator exists. This could be the result of an ill-defined problem in which the parameter β is nonidentified, but it may also arise in cases where the parameter IS identified, but the semiparametric assumptions are not enough to yield a finite efficiency bound. The best example is Manski's Maximum Score estimator, which focuses on a binary choice problem $Y_i = \mathbb{1}\{X_i'\beta_0 + \varepsilon_i \geq 0\}$ under the sole assumption that ε_i has median zero conditional on X_i. It is known that a $n^{1/3}$-consistent estimator for β exists in this case, and it has a non-normal asymptotic distribution.

Construction of Efficient Estimators

For semiparametric models with a parametric and a nonparametric component, some researchers (see Section 5 in Newey, 1990) have suggested that an efficient estimator can be constructed starting from a $\sqrt{n}$-consistent estimator $\widetilde{\beta}$ by using

$$\widehat{\beta} = \widetilde{\beta} + \frac{1}{n}\sum_{i=1}^{n} \widehat{d}(z_i, \widetilde{\beta}), \quad (\clubsuit)$$

where $\widehat{d}(z_i, \widetilde{\beta})$ is an estimator of the efficient influence function. The key for such a construction to work is that the fact that the efficient score itself is estimated should not affect the asymptotic distribution of $\widehat{\beta}$. This will be true if the estimated efficient score converges at a sufficiently fast rate. Other methods involve nonparametric-type maximum likelihood. See Section 5 in Newey (1990) for the specifics.