

Introduction to Empirical Process Methods

Myung Hwan Seo

EC485, Michelmas term, 2011¹

¹Cf. Office Hour: 14:30-15:30 Friday; S580.

Contents

1	Introduction	4
1.1	M-estimation	4
1.2	Some Testing Problems	9
1.3	Nonparametric Estimation	12
1.4	Semiparametric Estimation	13
2	Stochastic Equicontinuity and Weak Convergence	14
3	Entropy Conditions	22
3.1	Verification of Entropy Conditions	27
3.2	Permanence Property	32
4	Extensions	39
4.1	Maximal Inequality	39
4.2	Continuous Mapping Theorems	41

5	Statistical Applications	43
5.1	M-estimation	43

References

- [1] Andrews (1993) “An introduction to Econometric applications of Empirical process theory for dependent random variable” *Econometric Reviews*, 12, 183-216.
- [2] Andrews (1994) “Empirical process methods in econometrics” in Handbook of Econometrics IV.
- [3] Billingsley (1968) *Convergence of Probability Measures*.
- [4] Dehling et al. (2002) Empirical process techniques for dependent data.
- [5] Kosorok (2008) *Introduction to empirical processes and semiparametric inference*.
- [6] Newey and McFadden (1994) “Large Sample Estimation and Hypothesis Testing” in Handbook of Econometrics IV.
- [7] van der Vaart & Wellner (1996) *Weak convergence and empirical processes*.

1 Introduction

We begin with examples, for which the empirical process methods prove particularly useful.

1.1 M-estimation

The M-estimation stands for the Maximum likelihood like estimation, that is, by maximizing a sample mean of an unknown function of data indexed by a parameter, say, $f(x, \theta)$, x is a data point and θ is the unknown parameter

$$\max_{\theta} \frac{1}{n} \sum_{i=1}^n f(x_i; \theta).$$

For example, the nonlinear least squares, maximum likelihood estimation, and many others belong to this class. A standard approach is to linearize the objective function by taking Taylor series expansion but many interesting cases have non-differentiable objective functions and even discontinuous ones.

If the objective function can be approximated by a quadratic function, even if it is not differentiable, we can still derive the asymptotic normality of the estimator at the standard \sqrt{n} convergence rate. We call this class of estimation problems “regular M-estimation” and otherwise “non-regular M-estimation”. There are numerous examples for the former case including quantile regression (LAD regression); censored regression; truncated regression; method of simulated moments estimator for multinomial probit etc. And examples for the latter case include the linear regression with change points, the threshold regression, the maximum score estimation, and so on.

Example 1 (*Least Absolute Deviation regression*)

$$\begin{aligned}y_i &= x_i' \beta + \varepsilon_i \\ \text{med}(\varepsilon_i | x_i) &= 0.\end{aligned}$$

And

$$\hat{\beta} = \arg \min_{\beta} \frac{1}{n} \sum_{i=1}^n |y_i - x_i' \beta|.$$

This is a special case of the quantile regression, which is introduced by Koenker and Bassett (1978) and uses the check function

$$\rho_q(u) = (q - 1 \{u < 0\}) u,$$

as the criterion function. Note that $q = 1/2$ yield the LAD estimator.

Example 2 (*The threshold regression model*)

$$y_i = x_i' \beta_1 + \varepsilon_i \quad \text{if } q_i \leq \gamma$$

$$y_i = x_i' \beta_2 + \varepsilon_i \quad \text{if } q_i > \gamma$$

for which we consider the least square estimator

$$\left(\hat{\beta}, \hat{\gamma} \right) = \arg \min_{\beta, \gamma} \frac{1}{n} \sum_{i=1}^n (y_i - x_i' \beta_1 \mathbf{1}\{q_i \leq \gamma\} - x_i' \beta_2 \mathbf{1}\{q_i > \gamma\})^2.$$

For a given γ , the model is linear and thus

$$\hat{\beta}(\gamma) = (X_\gamma' X_\gamma)^{-1} X_\gamma' Y,$$

where X_γ is the matrix stacking $(x_i' \mathbf{1}\{q_i \leq \gamma\}, x_i' \mathbf{1}\{q_i > \gamma\})$ and Y stacks y_i for $i = 1, \dots, n$. See e.g. Hansen (2000).

Example 3 (*Maximum Score Estimation*) Consider the binary response model

$$\begin{aligned}y_i^* &= x_i' \beta + \varepsilon_i \\y_i &= 1 \{y_i^* > 0\},\end{aligned}$$

where we observe (y_i, x_i) . The maximum score estimator (Manski 1975, 85) is defined as

$$\hat{\beta} = \operatorname{argmax}_{\beta} \sum_{i=1}^n (2y_i - 1) 1 \{x_i' \beta > 0\},$$

where β subjects to certain normalization restriction such as $|\beta| = 1$.¹ Kim and Pollard (1990) derived its asymptotic distribution.

¹Manski (1985) shows that the estimator can be viewed as the LAD estimator such that $\sum_{i=1}^n |\operatorname{sgn}(y_i) - \operatorname{sgn}(x_i' \beta)|$.

1.2 Some Testing Problems

Testing problem where some of the parameters are not identified under the null hypothesis, e.g., testing for the variable relevance in non-linear regression, for the presence of certain types of nonlinearity, and for the presence of structural break etc.

Or nonparametric specification test; specification tests for conditional moment restriction etc. That is, the nature of the testing is an infinite-dimensional restriction.

Example 4 (*Testing for threshold effect*) Consider the previous threshold regression model. It is common that the presence of threshold effect is tested against the simpler linear regression and the corresponding null hypothesis of interest is

$$\mathcal{H}_0 : \beta_1 = \beta_2,$$

for any $\gamma \in \Gamma$. Note that this is composite null due to γ , which is not identified under the null.

Example 5 (*Testing for structural break*)

$$y_t = x_t' \beta_1 + \varepsilon_t \quad \text{if } t/n \leq \tau$$

$$y_t = x_t' \beta_2 + \varepsilon_t \quad \text{if } t/n > \tau$$

and the null hypothesis of interest is

$$\mathcal{H}_0 : \beta_1 = \beta_2,$$

for any $\tau \in (0, 1)$.

1.3 Nonparametric Estimation

The kernel based estimation will be studied in a separate topic more intensively. Here we consider a simple example.

Example 6 (*Empirical Distribution Function*) *Let*

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n 1\{X_i \leq x\}.$$

It follows from the standard LLN and CLT that it is consistent for the true distribution function and

$$\sqrt{n} \left(\hat{F}_n(x) - F(x) \right) \xrightarrow{d} \mathcal{N} \left(0, F(x) (1 - F(x)) \right).$$

What about as a function defined on \mathcal{X} ?

1.4 Semiparametric Estimation

Adaptive estimation; feasible GLS; Partially linear regression; GMM with conditional moment restriction;

Example 7 (GLS) *For the linear regression model*

$$\begin{aligned}y_i &= x_i' \beta + \varepsilon_i \\ \mathbb{E}(\varepsilon_i | x_i) &= 0 \\ \sigma_i^2 &= \mathbb{E}(\varepsilon_i^2 | x_i),\end{aligned}$$

the feasible GLS estimator is given by

$$\hat{\beta} = \left(\sum_{i=1}^n \frac{x_i x_i'}{\hat{\sigma}_i^2} \right)^{-1} \sum_{i=1}^n \frac{x_i y_i}{\hat{\sigma}_i^2},$$

where $\hat{\sigma}_i$ is the Nadaraya-Watson estimator of σ_i .

2 Stochastic Equicontinuity and Weak Convergence

Weak convergence concerns the convergence of a sequence of stochastic processes and stochastic equicontinuity of the sequence is essential to it.

- A stochastic process is an indexed collection of random variables $\{X(\tau) : \tau \in \mathcal{T}\}$.

It can be viewed as

- $X : \Omega \mapsto \mathcal{B} = \{b : \mathcal{T} \mapsto \mathbb{R}\}$
- $X : \Omega \times \mathcal{T} \mapsto \mathbb{R}$
- for a fixed ω , $X_\omega : \mathcal{T} \mapsto \mathbb{R}$ is called a *sample path*
- for a fixed τ or (τ_1, \dots, τ_k) , $(X(\tau_1), \dots, X(\tau_k))$ is called a *marginal*, and its distribution is finite-dimensional distribution (*fidi*).
- if all the marginals are Gaussian, the process is called a Gaussian.
- \mathbb{R} may be replaced with \mathbb{R}^s for some s to consider a vector of stochastic processes.

- An empirical process is a sequence of normalized and centered sums of functions of random variables indexed by an index set \mathcal{T} , that is,

$$\nu_n(\tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (m(W_i, \tau) - \text{Em}(W_i, \tau)) \text{ for } \tau \in \mathcal{T}.$$

1. a collection of \mathbb{R} -valued functions $\mathcal{M} = \{m(w, \tau) : w \in \mathcal{W}, \tau \in \mathcal{T}\}$: defined on a sample space \mathcal{W} and indexed by $\tau \in \mathcal{T}$.
2. the index set \mathcal{T} is equipped with *pseudometric* ρ and is often totally bounded².
3. for each n , it is a stochastic process, i.e.,
 - (a) for each τ , $\nu_n(\tau)$ is a random variable
 - (b) for each realization $\{W_t\}_{t=1}^n$, $\nu_n(\tau)$ is a mapping from \mathcal{T} to \mathbb{R} ,

²The pseudometric ρ satisfies (i) $\rho(\tau_1, \tau_1) = 0$, (ii) $\rho(\tau_1, \tau_2) = \rho(\tau_2, \tau_1)$, (iii) $\rho(\tau_1, \tau_2) \leq \rho(\tau_1, \tau_3) + \rho(\tau_2, \tau_3)$. And (\mathcal{T}, ρ) is *totally bounded* if for any $\varepsilon > 0$ there is finite number of ε balls that cover \mathcal{T} .

4. $\nu_n(\tau)$ takes its sample paths in $\ell^\infty(\mathcal{T})$, which is the class of bounded functions on \mathcal{T} into \mathbb{R} , and is equipped with the uniform metric

$$d(b_1, b_2) = \|b_1 - b_2\|_{\mathcal{T}} := \sup_{\tau \in \mathcal{T}} |b_1(\tau) - b_2(\tau)|,$$

for $b_1, b_2 \in \ell^\infty(\mathcal{T})$.

5. This can be expanded to accomodate vector version by replacing \mathbb{R} with \mathbb{R}^s .

Some notational convention in the literature: \mathbf{P} and \mathbb{P}_n for the true distribution and empirical distribution of the sample, respectively; $\mathbf{P}f = \int_{\mathcal{W}} f(w) d\mathbf{P}(w)$ and $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - \mathbf{P})$. Then,

$$\mathbb{G}_n f = \frac{1}{\sqrt{n}} \sum_{i=1}^n (f(W_i) - \mathbf{P}f), \quad f \in \mathcal{F},$$

is an empirical process. Then, we may write $\mathbb{G}_n m_\tau$ for $\nu_n(\tau)$, $\tau \in \mathcal{T}$.

Definition 1 (*weak convergence* \Rightarrow) we write

$$\nu_n(\cdot) \Rightarrow \nu(\cdot)$$

if

$$\mathbb{E}^* f(\nu_n(\cdot)) \rightarrow \mathbb{E} f(\nu(\cdot)) \quad \forall f \in \mathcal{U}(\ell^\infty(\mathcal{T})),$$

where $\mathcal{U}(\ell^\infty(\mathcal{T}))$ is the collection of all real functions f that are defined on $\ell^\infty(\mathcal{T})$ and bounded and uniformly continuous.

- Remarks

1. Suppose the limit process is *tight*³ i.e. $O_p(1)$. Then, weak convergence holds iff ν_n is *uniformly tight*, or *asymptotically tight*, and all the marginals converge in distribution.

³A process X is *tight* if $\forall \varepsilon > 0, \exists$ a compact set $K \subset B(\mathcal{T})$ such that $\Pr\{X \in K\} > 1 - \varepsilon$. And a sequence of processes X_n is *uniformly tight* if \exists compact K such that $\Pr\{X_n \in K\} > 1 - \varepsilon$, for all n .

2. The tightness is closely related to the continuity of the sample paths. In fact, the asymptotic tightness is equivalent to the stochastic equicontinuity given the marginal convergence. Roughly speaking, if the function is continuous and the index set is bounded, the the function will be bounded.
3. If the limit process is Gaussian and tight, which is mostly the case due to the CLT, then the index set (\mathcal{T}, ρ_p) is totally bounded and almost all sample paths are uniformly continuous (wrt ρ_p), where

$$\rho_p(\tau_1, \tau_2) = \mathbb{E}^{1/p} (\nu(\tau_1) - \nu(\tau_2))^p.$$

The reverse is also true.

4. In general, measurability is an issue but we do not discuss it here.
5. A class of functions $\mathcal{F} = \{f\}$ is sometimes called **P**-Donsker if $\mathbb{G}_n f$ converges weakly.
6. We may also be specific about the space, in which the empirical process ν_n takes its values, by writing $\nu_n \Rightarrow \nu$ in $\ell^\infty(\mathcal{T})$.

Definition 2 (*stochastic equicontinuity*) A sequence of stochastic processes $\{\nu_n(\cdot) : n \geq 1\}$ is stochastically (or asymptotically) equicontinuous if $\forall \varepsilon > 0$ and $\eta > 0$, $\exists \delta > 0$ and n_0 such that

$$\sup_{n \geq n_0} \Pr \left\{ \sup_{\rho(\tau_1, \tau_2) < \delta} \|\nu_n(\tau_1) - \nu_n(\tau_2)\| > \eta \right\} < \varepsilon,$$

for all large n .

- Remarks

1. asymptotically the process ν_n takes its sample paths with probability arbitrarily close to one on a class of functions, which is equicontinuous.
2. it is a **necessary** condition of weak convergence to a limit process, which is uniformly continuous in τ with probability one.
3. equivalently, if for any $\delta_n \rightarrow 0$,

$$\sup_{\rho(\tau_1, \tau_2) < \delta_n} |\nu_n(\tau_1) - \nu_n(\tau_2)| \xrightarrow{P} 0.$$

(\Rightarrow) $|\delta_n| < \delta$ for all large n . So follows the implication. (\Leftarrow) by negation.

4. equivalently, if for any $\hat{\tau}_{1n}$ and $\hat{\tau}_{2n}$ such that $\rho(\hat{\tau}_{1n}, \hat{\tau}_{2n}) \xrightarrow{P} 0$,

$$\nu_n(\hat{\tau}_{1n}) - \nu_n(\hat{\tau}_{2n}) \xrightarrow{P} 0.$$

(\Rightarrow) One can find a sequence $\delta_n \rightarrow 0$ s.t. $\Pr\{\rho(\hat{\tau}_{1n}, \hat{\tau}_{2n}) < \delta_n\} \rightarrow 1$. So follows the implication. (\Leftarrow) by negation.

5. an examples $\{m(w, \tau) = w'\tau\}$ with $\mathcal{T} \subset \mathbb{R}^k$ and a counter example

$$\{m(w, \tau) = 1 \{w \in \tau\}\}$$

with \mathcal{T} being the collection of all Borel sets in \mathcal{W} and $\rho = L_r(\mathbf{P})$ with continuously distributed \mathbf{P} will be elaborated on in the class.

The relation between weak convergence and stochastic equicontinuity is summarized in the following proposition (e.g. Andrews p.2251).

Proposition 1 *Assume that (\mathcal{T}, ρ) is totally bounded. Then, the following two are equivalent*

(a) *fidi holds; ν_n is stochastically equicontinuous*

(b) *$\nu_n \Rightarrow \nu : \Omega \mapsto \ell^\infty(\mathcal{T})$, whose sample paths are uniformly continuous (wrt ρ) w.p.1*

- Conditions for stochastic equicontinuity (Theorem 1 & 4 of Andrews)
 1. an entropy condition (with or w/o bracketing): a measure of complexity of \mathcal{M} .
 2. a moment condition ($2 + \delta$ moment) for an envelope function (e.g. $\bar{M}(\cdot) = \sup_{\tau \in \mathcal{T}} |m(\cdot, \tau)| \vee 1$)
 3. conditions on dependence (while we focus on iid case mainly, some introductions for dependent cases will be given).
- P-Glivenko-Cantelli: a class of functions which yields ULLN. a weaker entropy condition is required.

3 Entropy Conditions

Let \mathcal{F} and Q indicate a collection of generic functions f and a probability measure on the sample space \mathcal{W} , respectively. We follow the convention that

$$Qf = \int_{\mathcal{W}} f(w) dQ(w).$$

Let \mathcal{F} be a subclass of $\mathcal{L}_2(Q)$, the class of square integrable functions wrt Q , where $L_2(Q)$ -norm is also denoted by

$$\|f\|_{Q,2} = \left[\int f^2 dQ \right]^{1/2}.$$

We drop the subscript Q when obvious. Also let F be an envelope function of \mathcal{F} .

Definition 3 For any $\varepsilon > 0$, the cover (covering) number $N(\varepsilon, \mathcal{F}, \|\cdot\|)$ is the smallest value of n for which there exist functions $f_1, \dots, f_n \in \mathcal{F}$ such that $\min_{j \leq n} \|f - f_j\| < \varepsilon$ for any $f \in \mathcal{F}$.

That is, the cover number is the minimum number of ε balls that cover \mathcal{F} .

Typically, $\|\cdot\| = \|f\|_{Q,2}$. So it is denoted by $N_2(\varepsilon, Q, \mathcal{F})$ in Andrews.

Definition 4 The bracket $[l, u]$ is $\{f \in \mathcal{F} : l \leq f \leq u\}$. An ε -bracket is a bracket $[l, u]$ with $\|l - u\| < \varepsilon$. The bracketing number is the minimum number of ε -brackets needed to cover \mathcal{F} . It is denoted by $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$.

It is always the case that the bracketing number only uses the L_p -norm, $\|\cdot\|_p$, with the true distribution \mathbf{P} and for $p \geq 2$. So comes Andrews' notation $N_p^B(\varepsilon, \mathbf{P}, \mathcal{F})$.

As $f \in [l, u]$ implies that f belongs to the $\varepsilon/2$ ball of $(l + u)/2$, the covering number is smaller than the bracketing number. the bracket is sort of *uniform*.⁴

The brackets themselves need not be in \mathcal{F} .

⁴Let $\mathcal{F}_{1,\varepsilon} = \{f \in \mathcal{F} : \|f - f_1\| < \varepsilon\}$. But, it is possible that $\left\| \sup_{f \in \mathcal{F}_{1,\varepsilon}} |f - f_1| \right\| > \varepsilon$.

Definition 5 *The $(L_2(Q))$ ε -entropy is $\log N_2(\varepsilon, Q, \mathcal{F})$.*

The log of the bracketing number is called *entropy with bracketing*.

- Two types of entropy conditions for weak convergence

– *Pollard's entropy condition* is given by

$$\int_0^1 \sup_Q \left(\log N_2 \left(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q) \right) \right)^{1/2} d\varepsilon < \infty, \quad (1)$$

where the supremum is taken over all finitely discrete measures Q on \mathcal{W} . This condition is also referred to as the *uniform entropy condition*.

⁵

– *Ossiander's L_p entropy condition* for some $p \geq 2$:

$$\int_0^1 \left(\log N_p^B (\varepsilon, \mathbf{P}, \mathcal{F}) \right)^{1/2} d\varepsilon < \infty.$$

typically with $p = 2$. Also called *bracketing (integral) condition*.

⁵The presence of $\|F\|_{Q,2}$ can be detrimental if it is smaller than 1. However, it is known that we can choose F s.t. $\|F\|_{Q,2} > 1$ for all Q if there exists a square integrable envelop. (VW p.133)

– Roughly speaking, we need the entropy by at the rate of

$$O\left(\left(\frac{1}{\varepsilon}\right)^{2-\delta}\right),$$

for some $\delta > 0$.

- entropy conditions for **P**-Glivenko-Cantelli

1. for any $\varepsilon > 0$ and $M > 0$, the class $\mathcal{F}_M = \{f1_{\{F \leq M\}}, f \in \mathcal{F}\}$ satisfies

$$\log N(\varepsilon, \mathcal{F}_M, L_1(\mathbb{P}_n)) = o_p(n).$$

This also implies uniform L_1 convergence.

2. for any $\varepsilon > 0$,

$$N_{[]}(\varepsilon, \mathcal{F}, L_1(\mathbf{P})) < \infty.$$

3.1 Verification of Entropy Conditions

We provide several classes of functions on \mathcal{W} that satisfy either of the entropy conditions.

1. Vapnik-Červonenkis (VC)⁶ classes of functions:

(a) have the covering numbers uniformly bounded by

$$\sup_Q N\left(\varepsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q)\right) \leq K\varepsilon^{-r(V-1)},$$

⁶Let \mathcal{C} be a collection of subsets in \mathcal{X} . Say that \mathcal{C} *picks out* a certain subset of $\{x_1, \dots, x_n\}$ if this can be formed as a set of the form $C \cap \{x_1, \dots, x_n\}$ for a $C \in \mathcal{C}$. Now, the collection \mathcal{C} is said to *shatter* $\{x_1, \dots, x_n\}$ if any of its subset can be picked out by \mathcal{C} . The *VC-index* $V(\mathcal{C})$ is the smallest n for which no set of size n is shattered by \mathcal{C} . In case of a class of functions, we consider the *subgraph* of functions,

$$\{(x, t) : t < f(x)\}.$$

If the index is finite, we call it a *VC-class* or *VC-subgraph class*. For example, consider $\mathcal{C} = \{(-\infty, c], c \in \mathbb{R}\}$ and it cannot shatter any two-point set $\{x_1, x_2\}$ since it fails to pick out the largest of the two.

where the supremum is taken over all possible probability, $F = \sup |f| \vee 1$, and V is the VC index.

- (b) A finite-dimensional vector space of functions, e.g. $\{f(x) = x'\theta : \theta \in \mathbb{R}^d\}$, with VC index smaller than equal to $d + 2$.
- (c) $\{f(x) = h(x'\theta)\}$ for some h that is monotonic or of a finite total variation, e.g. the indicator, sign function, the identity function, $h(z) = z1\{z > 0\}$, etc.
- (d) the indicator functions of sets that constitute a VC-class, and a collection of functions $c1\{a < x \leq b\}$, with $a, b \in \mathbb{R}$, and $c > 0$.
- (e) Box-Cox family of transformations $\{f_\lambda : \mathbb{R}^+ \rightarrow \mathbb{R} : \lambda \in \mathbb{R} \setminus \{0\}\}$ where $f_\lambda(x) = (x^\lambda - 1) / \lambda$.
- (f) $\mathcal{F} \vee \mathcal{G}$, $\mathcal{F} \wedge \mathcal{G}$, $\{\mathcal{F} > 0\}$, $\mathcal{F} + g$, $\mathcal{F} \cdot g$, $\mathcal{F} \circ \psi$, $\phi \circ \mathcal{F}$ for a monotone ϕ . This can be more relevant for dependent data.

2. Finite-dimensional τ :

(a) Lipschitz in parameters: for some L^p -bounded function B on \mathcal{W}

$$|f(\cdot, \tau_1) - f(\cdot, \tau_2)| \leq B(\cdot) |\tau_1 - \tau_2| \text{ for all } \tau_1, \tau_2 \in \mathcal{T},$$

where $|\cdot|$ is the Euclidean norm. Then,

$$N_{\square} \left(2\varepsilon \|B\|_p, \mathcal{F}, \|\cdot\|_p \right) \leq N(\varepsilon, \mathcal{T}, |\cdot|),$$

the latter is polynomial in ε^{-1} given boundedness of \mathcal{T} .

(b) L^p -continuous functions such that

$$\mathbf{P}^{1/p} \left[\sup_{\tau_1: |\tau_1 - \tau| < \delta} |f_{\tau_1} - f_{\tau}|^p \right] \leq C\delta^\psi,$$

$\forall \tau \in \mathcal{T} \subset \mathbb{R}^d$, and $\forall \delta > 0$ in a neighborhood of 0, for some ψ . Then,

with $F = \sup_{\tau} |f_{\tau}|$ and $p \geq 2$,

$$N_{\square} \left(\varepsilon, \mathcal{F}, \|\cdot\|_p \right) \leq C\varepsilon^{-d/\psi}.$$

Proof. see Andrews (1994).

3. Smooth function class on a **bounded** $\bar{\mathcal{W}}$: the class \mathcal{F} of all bounded functions on a bounded, convex, nonempty set $\bar{\mathcal{W}} \subset \mathbb{R}^s$ that possess uniformly bounded partial derivatives up to order $[\alpha]$ for some $\alpha > 0$ and whose highest derivatives are Lipschitz of order $\alpha - [\alpha]$.

- The entropy with bracketing is $O\left(\left(\frac{1}{\varepsilon}\right)^{s/\alpha}\right)$, where the constant depends on the smoothness α , size of \mathcal{W} in terms of \mathbf{P} , and the dimension s of w . Note that higher dimension requires smoother functions as we need $s/\alpha < 2$.
- Extensions to the class of functions defined on an **unbounded** \mathcal{W} can be done using partitions $\mathcal{W} = \sum_{j=1}^{\infty} \mathcal{W}_j$ of sample space. See Corollary 2.7.4 of VW for an bound for the entropy with bracketing, which in particular demands $\sum_{j=1}^{\infty} \mathbf{P}(\mathcal{W}_j)^{\frac{V}{V+p}} < \infty$, where $V = s/\alpha$ and p from L_p .
 - Example: bounded and smooth functions on \mathbb{R} with some $\alpha > 1/2$. The moment condition of the envelope implies the summability

condition. I.e., we need $\sum_{j=1}^{\infty} \mathbf{P}(\mathcal{W}_j)^k < \infty$ for some $k < 1/2$, which follows from $\mathbf{E}|w_i|^{2+\delta} < \infty$ for some $\delta > 0$.

- Also see the subsequent section for more discussion.
 - type III, V, and VI in Andrews for slightly different exposition.
4. Convex function class \mathcal{F} : the collection of all convex and Lipschitz functions $f : C \mapsto [0, 1]$ defined on a compact, convex subset $C \subset \mathbb{R}^s$ such that $|f(x) - f(y)| \leq L|x - y|$ for all x, y . Then,

$$\log N(\varepsilon, \mathcal{F}, \|\cdot\|_{\infty}) \leq K \left(\frac{1}{\varepsilon}\right)^{s/2}$$

5. The set \mathcal{F} of all monotone functions $f : \mathbb{R} \rightarrow [0, 1]$ satisfies

$$\log N_{\square}(\varepsilon, \mathcal{F}, L_r(Q)) \leq K \left(\frac{1}{\varepsilon}\right),$$

for every Q , $r \geq 1$, and a constant K that depends only on r . And it can be extended to the functions into \mathbb{R} using a permanence property.

3.2 Permanence Property

Various mix and match of different types of functions still satisfy the entropy conditions. These are verified only for independent (or m-dependent) case.

- Permanence of \mathbf{P} -Donsker: Consider $\mathcal{F}_1, \dots, \mathcal{F}_k$ with F_1, \dots, F_k , and $\phi(\mathcal{F}_1, \dots, \mathcal{F}_k)$, such that

$$|\phi(f_1, \dots, f_k) - \phi(g_1, \dots, g_k)|^2 \leq \sum_{i=1}^k (f_i(w) - g_i(w))^2, \quad (2)$$

for every $f = (f_1, \dots, f_k)$ and $g = (g_1, \dots, g_k) \in \mathcal{F} = \mathcal{F}_1 \times \dots \times \mathcal{F}_k$ and $w \in \mathcal{W}$. If $\mathcal{F}_1, \dots, \mathcal{F}_k$ are Donsker, then so is $\phi(\mathcal{F}_1, \dots, \mathcal{F}_k)$, provided that $\|\phi(f)\|_2 < \infty$ for some $f \in \mathcal{F}$.

- Proof. See VW Theorem 2.10.6.
- For example, $\mathcal{F}_1 \vee \mathcal{F}_2$, $\mathcal{F}_1 \wedge \mathcal{F}_2$, $\mathcal{F}_1 + \mathcal{F}_2$, \mathcal{F}^{-1} , etc.
- This Lipschitz condition is somewhat strong and can be viewed as a *uniform* Lipschitz condition. This does not allow for $\mathcal{F}_1 \mathcal{F}_2$, for

instance, unless both are uniformly bounded. That is, by Jensen's inequality,

$$\begin{aligned} |f_1 f_2 - g_1 g_2|^2 &= |f_1 f_2 - g_1 f_2 + g_1 f_2 - g_1 g_2|^2 \\ &\leq 2 \left(f_2^2 |f_1 - g_1|^2 + g_1^2 |f_2 - g_2|^2 \right) \\ &\leq C \left(|f_1 - g_1|^2 + |f_2 - g_2|^2 \right), \end{aligned}$$

for some $C < \infty$, provided that \mathcal{F}_1 and \mathcal{F}_2 are uniformly bounded.

- The same remark applies to $g\mathcal{F}$.
- A modification is plausible. Let

$$|\phi(f_1, \dots, f_k) - \phi(g_1, \dots, g_k)|^2 \leq \sum_{i=1}^k L_i^2(w) (f_i(w) - g_i(w))^2, \quad (3)$$

where $L_i \mathcal{F}_i$ is Donsker. The last condition needs to be verified directly.

- Permanence of uniform-entropy:

1. Let \mathcal{F} and \mathcal{G} satisfy the uniform entropy condition with envelopes F , and G . Then, so does each of the following classes with envelopes in parentheses: $\mathcal{F} \cup \mathcal{G}$ ($F \vee G$), $\mathcal{F} + \mathcal{G}$ ($F + G$), $\mathcal{F} \vee \mathcal{G}$ ($F \vee G$), $\mathcal{F} \wedge \mathcal{G}$ ($F \vee G$), $\mathcal{F}\mathcal{G}$ ($(F \vee 1)(G \vee 1)$).
2. Consider $\mathcal{F}_1, \dots, \mathcal{F}_k$ with F_1, \dots, F_k , and $\phi(\mathcal{F}_1, \dots, \mathcal{F}_k)$, which is Lipschitz of orders $\alpha_1, \dots, \alpha_k \in (0, 1]$, in the sense that

$$|\phi \circ f(w) - \phi \circ g(w)|^2 \leq \sum_{i=1}^k L_{\alpha_i}^2(w) |f_i - g_i|^{2\alpha_i}(w),$$

for all $f = (f_1, \dots, f_k)$ and $g = (g_1, \dots, g_k) \in \mathcal{F} = \mathcal{F}_1 \times \dots \times \mathcal{F}_k$ and $w \in \mathcal{W}$. Then,

$$\begin{aligned} & \int_0^\delta \sup_Q \sqrt{\log N\left(\varepsilon \|L_\alpha \cdot F^\alpha\|_{Q,2}, \phi(\mathcal{F}), L_2(Q)\right)} d\varepsilon \\ & \leq \sum_{i=1}^k \int_0^{\delta^{1/\alpha_i}} \sup_Q \sqrt{\log N\left(\varepsilon \|F_i\|_{Q,2\alpha_i}, \mathcal{F}_i, L_{2\alpha_i}(Q)\right)} \varepsilon^{\alpha_i-1} d\varepsilon, \end{aligned}$$

where $L_\alpha \cdot F^\alpha = 2 \left(\sum_{i=1}^k L_{\alpha,i}^2 F_i^{2\alpha_i} \right)^{1/2}$.

- The Lipschitz condition here is weaker than (2). Instead, we need moment condition for $L_\alpha \cdot F^\alpha$.
- As a consequence, $\mathcal{FG} = \{fg : f \in \mathcal{F}, g \in \mathcal{G}\}$, \mathcal{F}^{-1} , $|\mathcal{F}|$, etc also satisfy the entropy condition, with corresponding conditions on the envelope.

- Permanence of bracketing entropy:
 1. Let \mathcal{F} and \mathcal{G} satisfy the bracketing entropy condition with envelopes F , and G . Then, so does each of the following classes with envelopes in parentheses: $\mathcal{F} \cup \mathcal{G}$ ($F \vee G$), $\mathcal{F} + \mathcal{G}$ ($F + G$), $\mathcal{F} \vee \mathcal{G}$ ($F \vee G$), $\mathcal{F} \wedge \mathcal{G}$ ($F \vee G$), $|\mathcal{F}|$ (F).
 2. The class \mathcal{FG} satisfies the bracketing condition with $L_p(\mathbf{P})$ for $p \geq 2$ and envelope GF , if \mathcal{G} and \mathcal{F} satisfy the bracketing condition with $L_\lambda(\mathbf{P})$ and $L_\mu(\mathbf{P})$ for $\lambda, \mu > p$ and $\lambda\mu/(\lambda + \mu) \geq p$, and if $E^{1/\lambda}G(W_i)^\lambda < \infty$ and $E^{1/\mu}F(W_i)^\mu < \infty$.
- A proof of the permanence of bracketing condition for \mathcal{FG} (Andrews 94)
 1. Let $g \in [-b_j + a_j, b_j + a_j]$, and $f \in [-b_i^* + a_i^*, b_i^* + a_i^*]$

$$\begin{aligned}
 |gf - a_j a_i^*| &\leq |gf - g a_i^*| + |g a_i^* - a_j a_i^*| \\
 &\leq G |f - a_i^*| + |a_i^* - f + f| |g - a_j| \\
 &\leq G b_i^* + b_i^* b_j + F b_j.
 \end{aligned}$$

2. The $L_p(\mathbf{P})$ norm of the bound can be shown to be $C\varepsilon$ for some constant $C < \infty$.
3. 1 and 2 imply that the bracketing number of \mathcal{FG} can be bounded by the product of those of \mathcal{F} and \mathcal{G} .

- Partitions of the sample space:

$$\mathcal{W} = \cup_{j=1}^{\infty} \mathcal{W}_j, \quad \mathcal{F}_j = \{f \cdot 1_{\mathcal{W}_j} : f \in \mathcal{F}\}.$$

- if \mathcal{F} is Donsker, each \mathcal{F}_j is Donsker (Lipschitz transformation)
- if each \mathcal{F}_j is Donsker and they become suitably small as $j \rightarrow \infty$, then \mathcal{F} is Donsker. More specifically, if

$$\sum_{j=1}^n c_j < \infty; \quad \mathbf{E} \sup_{f \in \mathcal{F}_j} |\mathbb{G}_n f| \leq C c_j,$$

for a constant C not depending on j nor n .

4 Extensions

4.1 Maximal Inequality

This section presents two moment and tail bounds for the supremum of the empirical process, $\|\mathbb{G}_n\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\mathbb{G}_n f|$. If $\mathcal{F} = \{f : \rho(f, f_1) < \delta\}$ for a given f_1 , then the bound provides an upper bound on the modulus of continuity of $\mathbb{G}_n f$ at f_1 , which is employed to obtain the convergence rate of a finite dimensional parameter estimate.

1. For $p \geq 1$,

$$\|\|\mathbb{G}_n\|_{\mathcal{F}}\|_p \lesssim \sup_Q \int_0^1 \sqrt{1 + \log N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))} d\varepsilon \cdot \|F\|_{2 \vee p}.$$

- 2.

$$\|\|\mathbb{G}_n\|_{\mathcal{F}}\|_1 \lesssim \int_0^1 \sqrt{1 + \log N_{[]}(\varepsilon \|F\|_2, \mathcal{F}, \|\cdot\|_2)} d\varepsilon \cdot \|F\|_2$$

They are rather crude bounds but still useful in many applications.

Under the entropy conditions in Section 3, both integrals are bounded uniformly in n . Thus, the bounds are controlled by the size of the envelope function F .

For the proof, see VW Ch.2.14.

4.2 Continuous Mapping Theorems

Theorem 2 (Extended CMT) *Let $\mathbb{D}_n \subset \mathbb{D}$ and $g_n : \mathbb{D}_n \rightarrow \mathbb{E}$ and $g : \mathbb{D}_0 \rightarrow \mathbb{E}$ satisfy that for any $x_n \rightarrow x$ with $x_n \in \mathbb{D}_n$ and $x \in \mathbb{D}_0 \subset \mathbb{D}$, $g_n(x_n) \rightarrow g(x)$. Let X_n and X be maps with values in \mathbb{D}_n and \mathbb{D}_0 , respectively. Then,*

$$X_n \Rightarrow X \text{ implies that } g_n(X_n) \Rightarrow g(X).$$

Proof. See Theorem 1.11.1 in VW. ■

Theorem 3 (Argmax CMT) *Let X_n, X be stochastic processes indexed by a metric space \mathcal{T} such that $X_n \Rightarrow X$ in $\ell^\infty(K)$ for every compact $K \subset \mathcal{T}$. Suppose that almost all sample paths of X are upper semicontinuous and possesses a unique maximum at a (random) point $\hat{\tau}$, which as a random map in \mathcal{T} is tight. If the sequence $\hat{\tau}_n$ is uniformly tight and satisfies $X_n(\hat{\tau}_n) \geq \sup_{\tau} X_n(\tau) - o_p(1)$, then $\hat{\tau}_n \Rightarrow \hat{\tau}$ in \mathcal{T} .*

Proof. See Theorem 3.2.2 in VW. ■

Lemma 4 *Let $Z(\tau) = -\frac{1}{2}\tau'V\tau + W(\tau)$ for $\tau \in \mathbb{R}^k$, where V is p.d. and W is a gaussian process with mean zeros, continuous sample paths, and $\text{var}(Z(t) - Z(s)) \neq 0$ for $t \neq s$. Then, with probability one, its sample path possesses a unique maximum and $Z(t) \rightarrow -\infty$ as $|t| \rightarrow \infty$.*

Proof. See Lemma 2.5 and 2.6 of Kim and Pollard (1990). ■

5 Statistical Applications

5.1 M-estimation

The M-estimator is defined as

$$\hat{\theta} = \operatorname{argmax}_{\theta \in \Theta} \mathbb{M}_n(\theta) := \mathbb{P}_n m_\theta.$$

The asymptotics for $\hat{\theta}$ takes three steps:

1. consistency
2. convergence rate r_n
3. weak convergence of a reparametrized objective function. for every compact K of $h = r_n(\theta - \theta_0)$. Then, the argmax continuous mapping theorem yields the asymptotic distribution of $\hat{h} = r_n(\hat{\theta} - \theta_0)$.

Theorem 5 Any sequence $\tilde{\theta}$ s.t. $\mathbb{M}_n(\tilde{\theta}) \geq \sup_{\theta} \mathbb{M}_n(\theta) - o_p(1)$ is consistent if the following holds:

(i) Uniform convergence: $\mathbb{M}_n(\theta) \xrightarrow{p} \mathbb{M}(\theta)$ uniformly in $\theta \in \Theta$, where \mathbb{M} is deterministic.

(ii) Good separation of θ_0 : $\mathbb{M}(\theta_0) > \sup_{\theta \notin G} \mathbb{M}(\theta)$ for any open set G containing θ_0 .

Proof. See Theorem 3.2.3 in VW. ■

The conditions can be replaced with the following three: 1. $\mathbb{M}_n(\theta)$ converges uniformly to $\mathbb{M}(\theta)$ for every compact $K \subset \Theta$; 2. $\mathbb{M}(\theta)$ is upper semicontinuous with a unique maximum at θ_0 ; 3. $\hat{\theta} = O_p(1)$.

Theorem 6 Suppose that $\hat{\theta} \xrightarrow{p} \theta_0$. Then, $r_n d(\hat{\theta}, \theta_0) = O_p(1)$, provided that
(i) $\forall \theta$ in a neighborhood of θ_0 ,

$$\mathbb{M}(\theta) - \mathbb{M}(\theta_0) \lesssim -d^2(\theta, \theta_0). \quad (4)$$

(ii) $\forall n > 0$, sufficiently small δ ,

$$\mathbb{E} \left[\sup_{d(\theta, \theta_0) < \delta} |\mathbb{G}_n(m_\theta - m_{\theta_0})| \right] \lesssim \phi_n(\delta), \quad (5)$$

where $\phi_n(x)/x^\alpha$ is decreasing for some $\alpha < 2$ and

(iii) $r_n^2 \phi_n(r_n^{-1}) \sim \sqrt{n}$.

Proof. See Theorem 3.2.5 in VW. ■

In this theorem, the function $d(\theta, \theta_0)$ needs not be a distance but any arbitrary function from Θ to the nonnegative values. And \lesssim means “is bounded above up to a universal constant”.

Examples

1. \mathbb{M} is twice continuously differentiable with non-singular second derivative with $d(\theta, \theta_0) = |\theta - \theta_0|$:
 - (a) if $\phi_n(\delta) = \delta$, $r_n = \sqrt{n}$.
 - (b) if $\phi_n(\delta) = \sqrt{\delta}$, then $r_n = n^{1/3}$, e.g. the maximum score estimator.
2. super-consistent estimator such as the threshold estimator: a) $d(\theta, \theta_0) = |\theta_1 - \theta_{10}| + \sqrt{|\theta_2 - \theta_{20}|}$; b) $\phi_n(\delta) = \delta$; c) \mathbb{M} is not differentiable.

Typically, the first condition is easy to check but the second condition, which is a bound on the modulus of continuity of the centered empirical process at θ_0 , is more difficult to check. One way is to apply the maximal inequality in Section 4.1 to

$$\mathcal{F}_\delta = \{m_\theta - m_{\theta_0} : d(\theta, \theta_0) < \delta\},$$

with an envelope function F_δ . Then, $\|F_\delta\|_2$ is typically $\phi_n(\delta)$.

Given the rate result, it is reasonable to focus on compact sets for $h = r_n(\theta - \theta_0)$. That is,

$$r_n^2 \left(\mathbb{M}_n \left(\theta_0 + \frac{h}{r_n} \right) - \mathbb{M}_n(\theta_0) \right) = \frac{r_n^2}{\sqrt{n}} \mathbb{G}_n \left(m_{\theta_0 + \frac{h}{r_n}} - m_{\theta_0} \right) + r_n^2 \mathbf{P} \left(m_{\theta_0 + \frac{h}{r_n}} - m_{\theta_0} \right). \quad (6)$$

1. The convergence of the last term is standard algebra. In particular, if $\mathbf{P}m_\theta$ is twice continuously differentiable with nonsingular second derivative matrix V , then the limit is $\frac{1}{2}h'Vh$.
2. The tightness of the first term, i.e., the empirical process, follows from the same manner as for (5). In other words, apply the maximal inequality in Section 4.1 to

$$\mathcal{F}_{n\delta} = \left\{ \frac{r_n^2}{\sqrt{n}} \left(m_{\theta_0 + \frac{h_1}{r_n}} - m_{\theta_0 + \frac{h_2}{r_n}} \right) : |h_1 - h_2| < \delta \right\}, \quad (7)$$

with a proper envelope function, to conclude the stochastic equicontinuity of the empirical process.

3. The *fidic* follows from standard CLTs for triangular arrays. Refer to Lindeberg-Feller or -Levy CLT etc in EC484 lecture note.

Remark 1 *In case that θ_0 lies at the **boundary** of the parametr space Θ , the space for h needs to be restricted accordingly. Suppose $\Theta = [0, \infty)$ and $\theta_0 = 0$. Then, the convergence of the reparametrized process in (6) is on $[0, K]$ for any $K < \infty$. Assuming that the first derivative of $\mathbf{P}m_\theta$ is zero at θ_0 , we can derive the convergence of the process by the same manner as above.*

Example 8 (*Lipschitz in parameter*) Assume that, for every θ_1 and θ_2 in a neighborhood of θ_0 ,

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq \dot{m}(x) |\theta_1 - \theta_2|,$$

and $\mathbf{P}m_\theta$ is twice continuously differentiable at θ_0 with nonsingular second derivative matrix V . Assuming the consistency of $\hat{\theta}$, the rate of convergence can be derived by Theorem 6. In particular, the first condition (4) of the theorem follows from the differentiability assumption and the modulus of continuity (5) can be obtained by applying the maximal inequality in Section 4.1 to the class of functions

$$\mathcal{F}_\delta = \{m_{\theta_1} - m_{\theta_2} : |\theta_1 - \theta_2| < \delta\},$$

with an envelop function

$$F_\delta(x) = \delta \dot{m}(x).$$

This yields the modulus of continuity of $\phi_n(\delta) \sim \|F_\delta\|_2 \sim \delta$ and thus $r_n = \sqrt{n}$. Then, reparametrize the objective function as in (6) and argue that the first term is tight defining $\mathcal{F}_{n\delta}$ as in (7) and the envelop as $F_{n\delta}(x) = \delta \dot{m}(x)$ and applying

the maximal inequality⁷ in Section 4.1. Next, apply the Lindeberg-Levy CLT to specify *fidi*. Finally, the second term converges to $\frac{1}{2}h'Vh$.

With an addition of

$$\mathbf{P} [m_\theta - m_{\theta_0} - (\theta - \theta_0)' \dot{m}_{\theta_0}]^2 = o(|\theta - \theta_0|^2),$$

which is a weak differentiability condition, we can get

$$\sqrt{n}(\hat{\theta} - \theta_0) = -V^{-1}\mathbb{G}_n \dot{m}_{\theta_0} + o_p(1).$$

See Lemma 3.2.21 in VW. Also observe that \dot{m} is not necessarily \dot{m}_{θ_0} .

⁷Note that even the bracketing integral is computed using the envelope of the size $\varepsilon \|F_{n\delta}\|_2$ and $\|F_{n\delta}\|_2 \rightarrow 0$ as $\delta \rightarrow 0$. However, it follows from the Lipschitz property that the bracketing number is proportional to δ and thus the bracketing integral is uniformly bounded over δ .

Example 9 (Revisit to the LAD estimator in Example 1) *Note that*

$$||y - x'\beta_1| - |y - x'\beta_2| | \leq |x| |\beta_1 - \beta_2|, \quad (8)$$

due to the triangular and Cauchy-Schwarz inequalities. Furthermore, $\mathbf{P} |y - x'\beta|$ is twice continuously differentiable at β_0 if the distribution of the regression error has a positive density at its median. See 3.2.23 VW(p.305) also, Koenker's book "quantile regression", which shows that

$$\sqrt{n} \left(\hat{\beta} - \beta_0 \right) \xrightarrow{d} \mathcal{N} \left(0, \frac{1}{4} \mathbf{P} (xx' f(0|x))^{-1} \mathbf{P} (xx') \mathbf{P} (xx' f(0|x))^{-1} \right),$$

where $f(\cdot|x)$ is the conditional density function of $(y - x'\beta_0)$ given x .

Proof. takes three steps as above.

1. Consistency can be argued by

- (a) the convexity of and the uniqueness of the minimizer of $\mathbb{M}(\beta) = \mathbf{E} |y_i - x'_i\beta|$, which can be shown by the graphical illustration under the assumption of $\mathbf{P}(xx') > 0$.

- (b) the Lipschitz property of (8) implies the uniform convergence of $\mathbb{M}_n(\beta) \xrightarrow{p} \mathbb{M}(\beta)$ on any compact set.
2. The convergence rate of \sqrt{n} can be argued as in Example 8. In particular, observe that the second derivative matrix at β_0 is $\mathbf{P}(xx'f(0|x))/2$.
3. For the asymptotic distribution, recall (6).
- (a) Letting $z_{ni}(h) = \sqrt{n}(|e_i - x_i h n^{-1/2}| - |e_i|)$ and $Z \sim \mathcal{N}(0, \mathbf{P}(xx'))$ note that

$$\frac{r_n^2}{\sqrt{n}} \mathbb{G}_n \left(m_{\theta_0 + \frac{h}{r_n}} - m_{\theta_0} \right) = \frac{1}{\sqrt{n}} \sum_{i=1}^n [z_{ni}(h) - \mathbb{E}(z_{ni}(h))] \Rightarrow Zh,$$

for which we apply the Lindeberg-Levy CLT for *fidi* and the tightness follows as in Example 8.

■

Some additional comments follow.

1. It can also be viewed as a VC class with $\beta \in B$, which is a compact subset of \mathbb{R}^k , by noting that

$$\{|y - x'\beta| : \beta \in B\} = \{y - x'\beta : \beta \in B\} \vee \{-y + x'\beta : \beta \in B\}.$$

2. We can do without the compactness of the parameter space, invoking the convexity of the objective function. See e.g. VW problem 4 in p.308. In particular, the convexity and the uniform convergence on any compact B implies that $\hat{\beta} = O_p(1)$.
3. Applying the Lindeberg-Levy CLT⁸, the following is useful⁹. By the second order Taylor series approximation, one can show that

$$\mathbb{E} (|e_i - x'_i\delta| - |e_i|)^2 \mathbf{1}_{\{|e_i| \leq |x'_i\delta|\}} = o(|\delta|^2),$$

⁸ If $\mathbb{E}x_{ni} = 0$, $\sum_{i=1}^n \text{var}(x_{ni}) = 1$, and $n\mathbb{E}(x_{ni}) < \infty$, then, $\sum_{i=1}^n x_{ni} \xrightarrow{d} \mathcal{N}(0, 1)$.

⁹ Leibniz's rule:

$$\frac{d}{d\theta} \int_{a(\theta)}^{b(\theta)} f(x, \theta) dx = f(b(\theta), \theta) \frac{db(\theta)}{d\theta} - f(a(\theta), \theta) \frac{da(\theta)}{d\theta} + \int_{a(\theta)}^{b(\theta)} \frac{\partial}{\partial \theta} f(x, \theta) dx.$$

which implies after some algebra that

$$n\mathbb{E} \left(\left| e_i - x_i' h_1 n^{-1/2} \right| - |e_i| \right) \left(\left| e_i - x_i' h_2 n^{-1/2} \right| - |e_i| \right) \rightarrow h_1' \mathbb{E} (x_i x_i') h_2.$$

References

- HANSEN, B. E. (2000): “Sample splitting and threshold estimation,” *Econometrica*, 68, 575–603.
- KIM, J., AND D. POLLARD (1990): “Cube root asymptotics,” *The Annals of Statistics*, 18(1), 191–219.
- KOENKER, R., AND J. BASSETT, GILBERT (1978): “Regression Quantiles,” *Econometrica*, 46, pp. 33–50.
- MANSKI, C. F. (1975): “Maximum score estimation of the stochastic utility model of choice,” *Journal of Econometrics*, 3(3), 205–228.
- (1985): “Semiparametric analysis of discrete response. Asymptotic properties of the maximum score estimator,” *Journal of Econometrics*, 27(3), 313–333.
- (1988): “Identification of binary response models,” *Journal of the American Statistical Association*, 83(403), 729–738.