Research notes

Wang Chao

January 19, 2015

Contents

1 Basic theory

Definition 1.1. *Given two measurable spaces $(S, \mathcal{S})$ and $(T, \mathcal{T})$, a mapping $\mu : S \times \mathcal{T} \to \mathbb{R}_+$ is called a (probability) kernel from S to T if the function $\mu_s(B) = \mu(s, B)$ is $\mathcal{S}-$measurable in $s \in S$ for fixed $B \in \mathcal{T}$ and a (probability) measure in $B \in \mathcal{T}$ for fixed $s \in S$.*

Any kernel μ determines an associated operator that maps suitable functions $f : T \to \mathbb{R}$ to their integrals $\mu f(s) = \int \mu(s, dt) f(t)$. Kernels play an important role in probability theory where they appear in guises of random measures, conditional distributions, Markov transition functions, and potentials.

The following characterizations of the kernel property are often useful. For simplicity we restrict our attention to probability kernels.

Lemma 1.2. *Fix two measurable spaces $(S, \mathcal{S})$ and $(T, \mathcal{T})$, a π-system $\mathcal{C}$ with $\sigma(\mathcal{C}) = \mathcal{T}$, and a family $\mu = \{\mu_s; s \in S\}$ of probability measures on T. Then these conditions are equivalent.*

(1) μ is probability kernel from S to T;

(2) μ is a measurable mapping from S to $\mathcal{P}(T)$;

(3) $s \mapsto \mu_s(B)$ is a measurable mapping from S to $[0, 1]$ for every $B \in \mathcal{C}$.

Proof. (2) $\to$ (3): Note that $s \mapsto \mu_s(B) = s \mapsto \mu_s \mapsto \mu_s(B)$ is a two-step mapping. The first part is measurable by (2), the second part is measurable as $\pi_B : \mu \mapsto \mu(B)$ on $\mathcal{P}(T)$ for every $B \in T$.

(3) $\to$ (1): by a $\lambda-$system argument.

(1) $\to$ (2): let $\pi_B : \mathcal{P}(T) \to [0, 1]$ be defined as $\pi_B(\mu) = \mu(B)$, then the $\sigma-$algebra $\mathcal{B}(\mathcal{P}(T)) = \sigma\{\pi_B; B \in \mathcal{T}\}$. To show μ is $\mathcal{S}/\mathcal{B}(\mathcal{P}(T))-$measurable, it is equivalent to show that $\pi_B \circ \mu$ is $\mathcal{S}/\mathcal{B}([0, 1])-$measurable. Since $s \mapsto \pi_B \circ \mu(s) = \mu_s(B)$ is measurable by the definition of probability kernel, $\pi_B \circ \mu$ is $\mathcal{S}/\mathcal{B}([0, 1])-$measurable. □

Let us now introduce a third measurable space $(U, \mathcal{U})$, and consider two kernels μ and ν, one from S to T and the other from $S \times T$ to U. Imitating the construction of product measures, we may attempt to combine μ and ν into a kernel from S to $T \times U$ given by

$$(\mu \otimes \nu)(s, B) = \int \mu(s, dt) \int \nu(s, t, du) 1_B(t, u), B \in \mathcal{T} \otimes \mathcal{U}.$$

The following lemma justifies the formula and provides some further useful information.

Lemma 1.3. *(kernels and functions) Fix three measurable spaces $(S,\mathcal{S})$, $(T,\mathcal{T})$, and $(U,\mathcal{U})$. Let μ and ν be probability kernels from S to T and $S\times T$ to U, respectively, and consider two measurable functions $f : S\times T\to\mathbb{R}_+$ and $g : S\times T\to U$. Then*

(1) $\mu_s f(s,\cdot)$ is a measurable function of $s\in S$;

(2) $\mu_s\circ(g(s,\cdot))^{-1}$ is a kernel from S to U;

(3) $\mu\otimes\nu$ is a kernel from S to $T\times U$.

Proof. (1): It is true if f is the indicator function of a set $A=B\times C$ with $B\in S$ and $C\in T$, i.e., $\mu_s 1_{B\times C}(s,t)=\mu_s(C)1_B(s)$. From here, we may extend to general $A\in\mathcal{S}\otimes\mathcal{T}$ by a monotone class argument and then to arbitrary f by linearity and monotone convergence.

(2): For given $U\in\mathcal{U}$, $\mu_s\circ g(s,\cdot)^{-1}(U)=\int_T\mu_s(dt)1\{g(s,t)\in U\}$. Let $f(s,t)=1\{t(s,t)\in U\}$, then $f : S\times T\to\mathbb{R}_+$ is measurable. Then $\mu_s\circ g(s,t)^{-1}(U)=\mu_s f(s,\cdot)$ is measurable by (1). On the other hand, $\mu_s\circ g(s,\cdot)^{(-1)}$ is a probability measure.

(3): Firstly, it's easy to see that $\mu\otimes\nu(s,\cdot)$ is a probability measure on $(T\times U)$ for fixed $s\in S$. For fixed $B\in\mathcal{T}\otimes\mathcal{U}$, assume $B=A\times C$ for $A\in\mathcal{T}$ and $C\in\mathcal{U}$, then

$$\begin{aligned}
&s\mapsto(\mu\otimes\nu)(s,A\times C)\\
&=\int\mu(s,dt)\int\nu(s,t,du)1_{A\times C}(t,u)\\
&=\int\mu(s,dt)\int\nu(s,t,C)1_A(t)\\
&=\int\mu(s,dt)\nu(s,t,C)1_A(t),
\end{aligned}$$

where $\nu(s,t,C)1_A(t)$ is a measurable function from $S\times T$ to $\mathbb{R}_+$, by (1), $s\mapsto(\mu\otimes\nu)(s,A\times C)$ is measurable, then by Lemma 1.2(3), it is measurable for all $B\in\mathcal{T}\times\mathcal{U}$. □

For any measurable function $f\geq 0$ on $T\times U$, we get

$$(\mu\otimes\nu)_s f=\int\mu(s,dt)\int\nu(s,t,du)f(t,u),\forall s\in S,$$

or simply

$$(\mu\otimes\nu)f=\mu(\nu f).$$

By iteration we may combine any kernels μ_k from $S_0 \times \cdots \times S_{k-1}$ to S_k, $k = 1, \ldots, n$, into a kernel $\mu_1 \otimes \cdots \otimes \mu_n$ from S_0 to $S_1 \times \cdots S_n$, given by

$$(\mu_1 \otimes \cdots \mu_n) f = \mu_1(\mu_2(\cdots \mu_n(f) \cdots))$$

for any measurable function $f \geq 0$ on $S_1 \times \cdots \times S_n$.

In applications we often encounter kernels μ_k from S_{k-1} to S_k, $k = 1, \ldots, n$, in which case the composition $\mu_1 \cdots \mu_n$ is defined as a kernel from S_0 to S_n for measurable $B \subset S_n$ by

$$\begin{aligned}(\mu_1 \cdots \mu_n)_s B &= (\mu_1 \otimes \cdots \otimes \mu_n)(S_1 \times \cdots \times S_{n-1} \times B) \\ &= \int \mu_1(s, ds_1) \int \mu_2(s_1, ds_2) \cdots \int \mu_{n-1}(s_{n-2}, ds_{n-1}) \mu_n(s_{n-1}, B)\end{aligned}$$

For any two measures μ and ν on $(\Omega, \mathcal{A})$, we say that ν is absolutely continuous w.r.t. μ and write $\nu \ll \mu$ if $\mu A = 0$ implies $\nu A = 0$ for all $A \in \mathcal{A}$. The following result gives a fundamental decomposition of a measure into an absolutely continuous and a singular component; at the same time it provides a basic representation of the former part.

Theorem 1.4. *(Lebesgue decomposition, Radon-Nikodym theorem) For any σ-finite measures μ and ν on Ω, there exist some unique measures $\nu_a \ll \mu$ and $\nu_s \perp\!\!\!\perp \mu$ s.t. $\nu = \nu_a + \nu_s$. Furthermore, $\nu_a = f \cdot \mu$ for some μ-a.e. measurable function $f \geq 0$ on Ω.*

Fix a probability space $(\Omega, \mathcal{A}, P)$ and an arbitrary sub-σ-field $\mathcal{F} \subset \mathcal{A}$, in $L^2 = L^2(\mathcal{A})$, then $M = \{\eta \in L^2; \eta = \tilde{\eta} \text{ a.s. for some } \tilde{\eta} \in L^2(\mathcal{F})\}$ is a closed subspace. The for all $\xi \in L^2$, there exists a.s. unique $\eta \in M$ s.t. $\xi - \eta \perp\!\!\!\perp M$. Define $\mathrm{E}^{\mathcal{F}} \xi = \mathrm{E}[\xi | \mathcal{F}]$ as an arbitrary $\mathcal{F}$-measurable version of η.

Definition 1.5. *(conditional expectation, Kolmogorov) Fix a probability space $(\Omega, \mathcal{A}, P)$, for any $\sigma-$field $\mathcal{F} \in \mathcal{A}$ there exists an a.s. unique linear operator $E^{\mathcal{F}} : L^1 \to L^1(\mathcal{F})$ s.t.*

(1) (averaging property) $E\left[E^{\mathcal{F}} \xi; A\right] = E[\xi; A], \forall \xi \in L^1, A \in \mathcal{F}$.

The following additional properties hold whenever the corresponding expression exist for the absolute values:

(2) (positivity) $\xi \geq 0$ implies $E^{\mathcal{F}} \xi \geq 0$ a.s.;

(3) (L^1-contractivity) $E|E^{\mathcal{F}} \xi| \leq E|\xi|$;

(4) (monotone convergence) $0 \leq \xi_n \uparrow \xi$ implies $E^{\mathcal{F}} \xi_n \uparrow E^{\mathcal{F}} \xi$ a.s.;

(5) (pull-out) $E^{\mathcal{F}}\xi\eta = \xi E^{\mathcal{F}}\eta$ *if* ξ *is* $\mathcal{F}-$*measurable;*

(6) (self-adjoint) $E\left[\xi E^{\mathcal{F}}\eta\right] = E\left[\eta E^{\mathcal{F}}\xi\right] = E\left[E^{\mathcal{F}}\eta E^{\mathcal{F}}\xi\right]$*;*

(7) (chain rule) $E^{\mathcal{F}}E^{\mathcal{G}}\xi = E^{\mathcal{F}}\xi$ *a.s. for all* $\mathcal{F} \subset \mathcal{G}$*.*

Note that $\mathrm{E}^{\mathcal{F}}\xi$ is a $\mathcal{F}$-measurable random variable. In particular, we note that $\mathrm{E}^{\mathcal{F}}\xi = \xi$ a.s. iff ξ has an $\mathcal{F}-$measurable version and that $\mathrm{E}^{\mathcal{F}}\xi = \mathrm{E}\xi$ a.s. iff $\xi \perp\!\!\!\perp \mathcal{F}$.

How to find condtional expectation in practice?

The next result shows that the conditional expectation $\mathrm{E}^{\mathcal{F}}\xi$ is *local* in both ξ and $\mathcal{F}$, an observation that simplifies many proofs. Given two σ-fields $\mathcal{F}$ and $\mathcal{G}$, we say that $\mathcal{F} = \mathcal{G}$ on A if $A \in \mathcal{F} \cap \mathcal{G}$ and $A \cap \mathcal{F} = A \cap \mathcal{G}$.

Lemma 1.6. *(local property) Let the* σ*-fields* $\mathcal{F}, \mathcal{G} \subset \mathcal{A}$ *and functions* $\xi, \eta \in L^1$ *be such that* $\mathcal{F} = \mathcal{G}$ *and* $\xi = \eta$ *a.s. on some set* $A \in \mathcal{F} \cap \mathcal{G}$*. Then* $E^{\mathcal{F}}\xi = E^{\mathcal{G}}\eta$ *a.s. on* A*.*

Proof. To prove the assertion, it's equivalent to prove $1_A\mathrm{E}^{\mathcal{F}}\xi = 1_A\mathrm{E}^{\mathcal{G}}\eta$ a.s..

Since $1_A\mathrm{E}^{\mathcal{F}}\xi$ and $1_A\mathrm{E}^{\mathcal{G}}\eta$ are $\mathcal{F}\cap\mathcal{G}$-measurable, we get $B = A\cap\left\{\mathrm{E}^{\mathcal{F}}\xi > \mathrm{E}^{\mathcal{F}}\eta\right\} \in \mathcal{F} \cap \mathcal{G}$, and the average property yields

$$\begin{aligned} &\mathrm{E}\left[\mathrm{E}^{\mathcal{F}}\xi; B\right] = \mathrm{E}[\xi; B] = \mathrm{E}\left[\eta; B\right] = \mathrm{E}\left[\mathrm{E}^{\mathcal{G}}\eta; B\right] \\ \Longrightarrow\ &\mathrm{E}\left[\mathrm{E}^{\mathcal{F}}\xi - \mathrm{E}^{G}\eta; B\right] = 0 \end{aligned}$$

however, $1_B(\mathrm{E}^{\mathcal{F}}\xi - \mathrm{E}^{G}\eta) \geq 0$, so we have $1_B(\mathrm{E}^{\mathcal{F}}\xi - \mathrm{E}^{G}\eta) = 0$ a.s. which means $\mathrm{E}^{\mathcal{F}}\xi \leq \mathrm{E}^{G}\eta$ a.s. on A. The opposite inequality is obtained by interchanging the roles of $(\mathcal{F}, \xi)$ and $(\mathcal{G}, \eta)$. □

The conditional probability of an event $A \in \mathcal{A}$, given $\mathcal{F}$, is defined as

$$P^{\mathcal{F}}A = \mathrm{E}^{\mathcal{F}}1_A.$$

Thus, $P^{\mathcal{F}}A$ is the a.s. unique r.v. in $L^1(\mathcal{F})$ s.t.

$$\mathrm{E}\left[P^{\mathcal{F}}A; B\right] = P(A \cap B), \forall B \in \mathcal{F}.$$

Note that

- $P^F A = PA$ a.s. iff $A \perp\!\!\!\perp \mathcal{F}$.
- $\mathcal{P}^F A = 1_A$ a.s. iff A agrees a.s. with a set in $\mathcal{F}$.
- $0 \leq P^F A \leq 1$ by positivity of $\mathrm{E}^{\mathcal{F}}$.

- $P^{\mathcal{F}} \cup_n A_n = \sum_n P^{\mathcal{F}} A_n$ a.s., if $\{A_n\}$ are disjoint (by monotone convergence)

If η is a r.e. in some measurable space $(S, \mathcal{S})$, we define conditional on η as conditional w.r.t. the induced σ-field $\sigma(\eta)$. Thus,

$$\mathrm{E}^{\eta}\xi = \mathrm{E}^{\sigma(\eta)}\xi,\ P^{\eta}A = P^{\sigma(\eta)}A.$$

By Lemma (??), the η-measurable function $\mathrm{E}^{\eta}\xi$ may be represented in the form $f(\eta)$, where f is a measurable function on S, determined a.e. by $\mathcal{L}(\eta)$ by the averaging property

$$\mathrm{E}\left[f(\eta); \eta \in B\right] = \mathrm{E}\left[\xi; \eta \in B\right], \forall B \in \mathcal{S}.$$

In particular, the function f depends only on the distribution of (ξ, η). Conditional w.r.t. a σ-field $\mathcal{F}$ is the special case when η is the identity map from $(\Omega, \mathcal{A})$ to $(\Omega, \mathcal{F})$.

We proceed to examine the existence of measure-valued version of $P^{\mathcal{F}}$ and P^{η}. Note that kernels on the basic probability space Ω is called *random measurs*.

Now fix a σ-field $\mathcal{F} \subset \mathcal{A}$ and a r.e. ξ in some measurable space $(S, \mathcal{S})$. By a *regular conditional distribution of ξ, given $\mathcal{F}$*, we mean a version of $P\left[\xi \in \cdot | \mathcal{F}\right]$ on $\Omega \times \mathcal{S}$ which is a probability kernel from $(\Omega, \mathcal{F})$ to $(S, \mathcal{S})$, hence an $\mathcal{F}$-measurable random probability measure on S. More generally, if η is another r.e. in some measurable space $(T, \mathcal{T})$, a regular conditional distribution of ξ, given η, is defined as a random measure of the form

$$\mu(\eta, B) = P[\xi \in B | \eta]\ a.s., \forall B \in \mathcal{S},$$

where μ is a probablity kernel from T to $\mathcal{S}$. Special cases:

- If ξ is $\mathcal{F}$-measurable or independent of $\mathcal{F}$, then $P[\xi \in B | \mathcal{F}]$ has regular version $1\{\xi \in B\}$ or $P\{\xi \in B\}$, respectively.

The general case requires some regularity condition on S.

Theorem 1.7. *(conditional distribution) For any Borel space S and measurable space T, let ξ and η be r.e. in S and T, respectively. Then there exists a probablity kernel μ from T to S s.t. $P[\xi \in \cdot | \eta] = \mu(\eta, \cdot)$ a.s., and μ is unique a.e. $\mathcal{L}(\eta)$.*

Proof. We may assume that $S \in \mathcal{B}(\mathbb{R})$. For every rational number r we may choose some measurable function $f_r = f(\cdot, r) : T \to [0, 1]$ s.t.

$$f(\eta, r) = P[\xi \leq r | \eta]\ a.s., \forall r \in \mathcal{Q}.$$

Let A be the set of all $t \in T$ s.t. $f(t,r)$ is non-decreasing in $r \in \mathcal{Q}$ with limits 1 and 0 at $\pm\infty$. Since A is specified by countably many measurable conditions, each of which holds a.s. at η, we have $A \in \mathcal{T}$ and $\eta \in \mathcal{A}$ $a.s.$ Define

$$F(t,x) = 1_A \inf_{r>x} f(t,r) + 1_{A^c} 1\{x \geq 0\}, \ x \in \mathbb{R}, t \in T,$$

note that $F(t,\cdot)$ is a distribution function on $\mathbb{R}$ for every $t \in T$ □

¡++¿

Definition 1.8. *A set K is totally bounded iff $\forall \varepsilon > 0$, it can be covered by finitely many balls of radius ε.*

Definition 1.9. *(Uniform tightness) A set of random vectors $\{X_\alpha : \alpha \in A\}$ is uniformly tight if $\forall \varepsilon > 0, \exists M$ s.t. $\sup_\alpha \mathrm{P}(\|X_\alpha\| > M) < \varepsilon$.*

Uniformly tight = bounded in probability.

Definition 1.10. *(Uniform integrability) (a notion for expected values) A family of r.v. $\{\xi_t, t \in T\}$ are said to be uniformly integrable if*

$$\lim_{r\to\infty} \sup_{t\in T} E\left[|\xi_t|; |\xi_t| > r\right] = 0.$$

Examples:

(1) For sequences $\{\xi_n\}$ in L^1, it is equivalent to

$$\lim_{r\to\infty} \limsup_n \mathrm{E}\left[|\xi_n|; |\xi_n| > r\right] = 0.$$

(2) If ξ_t are L^p-bounded for some $p > 1$, in the sense that $\sup_t \mathrm{E}|\xi|^p < \infty$. As $\mathrm{E}\left[|\xi|; |\xi| > r\right] \le \mathrm{E}\left[|\xi| \left(\frac{|\xi|}{r}\right)^{p-1}; |\xi| > r\right] \le \frac{\mathrm{E}|\xi|^p}{r^{p-1}}$.

¡++¿

Lemma 1.11. *The r.v. $\xi_t, t \in T$ are said to be uniformly integrable iff*

(1) $\sup_t E|\xi_t| < \infty$,

(2) $\lim_{\mathrm{P}\, A\to 0} \sup_{t\in T} E[|\xi_t|; A] = 0$

Theorem 1.12. *(Prohorov's theorem) Let X_n be vectors,*

(1) If $X_n \xrightarrow{\mathcal{L}} X$ for some X, then $\{X_n\}$ is uniformly tight.

(2) If X_n is uniformly tight, then $\exists$ a subsequence with $X_{n_j} \xrightarrow{\mathcal{L}} X$ as $j \to \infty$.

Definition 1.13. *(Nze and Doukhan, 2004, Section 3.1) (Mixing) Let* $(\Omega, \mathcal{A}, \mathrm{P})$ *be a probability space and let* $\mathcal{A}_1, \mathcal{A}_2$ *be two sub* σ*-algebra of* $\mathcal{A}$*. the coefficients*

$$\begin{aligned}
\alpha(\mathcal{A}_1, \mathcal{A}_2) &= \sup\{|\,\mathrm{P}(a \cap b) - \mathrm{P}(a)\,\mathrm{P}(b)|; a \in \mathcal{A}_1, b \in \mathcal{A}_2\} \\
\beta(\mathcal{A}_1, \mathcal{A}_2) &= E\sup\{|\,\mathrm{P}(b|\mathcal{A}_1) - \mathrm{P}(a)|; b \in \mathcal{A}_2\} \\
\phi(\mathcal{A}_1, \mathcal{A}_2) &= \sup\{|\,\mathrm{P}(b|a) - \mathrm{P}(a)|; a \in \mathcal{A}_1, b \in \mathcal{A}_2\} \\
\rho(\mathcal{A}_1, \mathcal{A}_2) &= \sup\left\{|corr(a, b)|; a \in L^2(\mathcal{A}_1), b \in L^2(\mathcal{A}_2)\right\}
\end{aligned}$$

are, respectively, the strong mixing coefficient α*, absolute regularity coefficient* β*, uniform mixing coefficient* ϕ*, and maximal correlation coefficient* ρ*.*

Lemma 1.14. *(Slutsky) Let* X_n, X, Y_n *be r.v. If* $X_n \xrightarrow{\mathcal{L}} X$ *and* $Y_n \xrightarrow{\mathcal{L}} c$ *for a const* c*, then*

(1) $X_n + Y_n \xrightarrow{\mathcal{L}} X + c$*,*

(2) $X_n Y_n \xrightarrow{\mathcal{L}} cX$*,*

(3) $X_n / Y_n \xrightarrow{\mathcal{L}} X/c$ *if* $c \neq 0$*.*

Definition 1.15. *A stochastic process* $X = \{X_t : t \in T\}$ *is a collection of random variables* $X_t : (\Omega, \mathcal{F}, \mathrm{P}) \mapsto R$*, indexed by an arbitrary set* T*.*

For each fixed $\omega \in \Omega$*, the map* $t \mapsto X_t(\omega)$ *is called a sample path. If every sample path is a bounded function,* X *an be viewed as a map* $X :\mapsto \ell^\infty(T)$*.*

Definition 1.16. *(Weak convergence of stochastic processes, van der Vaart (2000, page 261, thm 18.4)).*

A sequence of arbitrary maps $X_n : \Omega_n \mapsto \ell^\infty(T)$ *converges weakly to a right random element iff both of the following conditions hold:*

(1) $\forall t_1, \ldots, t_k \in T$*, the sequence* $(X_{n,t_1}, \ldots, X_{n,t_k})$ *converges in distribution.*

(2) $\forall \varepsilon, \eta > 0, \exists$ *a partition of* T *into finitely many sets* $T_1, \ldots, T_k$ *s.t.*

$$\limsup_{n\to\infty} \overset{*}{\mathrm{P}}\left(\sup_i \sup_{s,t\in T_i} |X_{n,s} - X_{n,t}| \geq \varepsilon\right) \leq \eta.$$

Definition 1.17. *A time series is weak stationary, if it has finite second-order moment, invariant unconditional mean and well-defined auto-covariance function.*

2 M- and Z- estimators

Theorem 2.1. *(van der Vaart, 2000, page 45) Let M_n be random functions and let M be a fixed function of θ s.t.* [1]

(1) $\sup_{\theta\in\Theta} |M_n(\theta) - M(\theta)| \xrightarrow{P} 0$,

(2) $\forall\varepsilon, \sup_{\theta:d(\theta,\theta_0)>\varepsilon} M(\theta) < M(\theta_0)$,

(3) $\hat{\theta}_n$ *is a sequence of estimators s.t.* $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1)$.

then $\hat{\theta}_n \xrightarrow{P} \theta_0$.

Proof. $\forall\varepsilon$, consider the event $A = \left\{|\hat{\theta}_n - \theta_0| > \varepsilon\right\}$. Let $\eta = \sup_{\theta:|\theta-\theta_0|>\varepsilon} |M_{\theta_0} - M_\theta|$, then $\eta > 0$.

Let $B = \{\sup_\theta |M_n(\theta) - M(\theta)| \leq \eta/4\}$. Condition (1) implies $\mathrm{P}(B^c) \to 0$.

On $A \cap B$

$$\begin{aligned} M_n(\theta_0) - M_n(\hat{\theta}_n) \geq & M(\theta_0) - \eta/4 - (M(\theta) + \eta/4) \\ = & M(\theta_0) - M(\theta) - \eta/2 \\ \geq & \eta - \eta/2 \\ = & \eta/2 \end{aligned}$$

which has prob. $\to 0$ by condition (3).

As $A = (A \cap B) \cup (A \cap B^c)$, $\mathrm{P}(A) = \mathrm{P}(A \cap B) + \mathrm{P}(A \cap B^c) \leq \mathrm{P}(A \cap B) + \mathrm{P}(B) \xrightarrow{P} 0$. □

Theorem 2.2. *Let Ψ_n be random vector-valued functions and Ψ a fixed vector-valued function of θ s.t.*

(1) $\sup_{\theta\in\Theta} \|\Psi_n(\theta) - \Psi(\theta)\| \xrightarrow{P} 0$,

(2) $\forall\varepsilon > 0, \inf_{\theta:d(\theta,\theta_0)\geq\varepsilon} \|\Psi(\theta)\| \geq 0 = \|\Psi_(\theta_0)\}$,

(3) $\hat{\theta}_n$ *s.t.* $\Psi_n(\hat{\theta}_n) = o_P(1)$.

then $\hat{\theta}_n \xrightarrow{P} \theta_0$.

Proof. It follows by applying Theorem 2.1 to the functions $M_n(\theta) = -\|\Psi_n\|$ and $M = -\|\Psi\|$. □

[1] Some expression may not be measurable, in which case, they shall be understood outer measure

The uniform convergence of functions is equivalent to the set of functions $\{\psi(\theta), \theta \in \Theta\} < ++ >$ being Glivenko-Cantelli.

Theorem 2.3. *(van der Vaart, 2000, page 266, thm 19.1) (Glivenko-Cantelli) If X_i are iid r.v. with dist. func. F, then*

$$\|\mathbb{F}_n - F\| \xrightarrow{a.s.} 0.$$

Proof. (Based on compact range and monotonicity of cdf function)

By SLLN, $\forall t$, $\mathbb{F}_n(t) \xrightarrow{a.s.} F(t)$ and $\mathbb{F}_n(t-) \xrightarrow{a.s.} F(t-)$.

$\forall \varepsilon$, $\exists$ partition $-\infty = t_0 < t_1 \cdots < t_k = \infty$ s.t. $\forall i, F(t_i-) - F(t_{i-1}) < \varepsilon$. Then $\forall t \in [t_{i-1}, t_i)$,

$$\mathbb{F}_n(t) - F(t) \leq \mathbb{F}_n(t_i-) - F(t_i-) + \varepsilon,$$
$$\mathbb{F}_n(t) - F(t) \geq \mathbb{F}_n(t_{i-1}) - F(t_{i-1}) - \varepsilon,$$

Also, $\mathbb{F}_n \xrightarrow{a.s.} F$ is uniformly for any finite set $\{t_1, \ldots, t_{k-1}\}$. That implies $\limsup \|\mathbb{F}_n - F\|_\infty \leq \varepsilon, a.s.$ Since ε is arbitrary, the lim sup is zero. □

Definition 2.4. *(P-Glivenko-Cantelli & P-Donsker) For a class $\mathscr{F}$ of measurable functions $f : \mathbf{X} \mapsto \mathbb{R}$,*

(1) it is called P-Glivenko-Cantelli if

$$\|\mathbb{P}_n f - \mathrm{P} f\|_{\mathscr{F}} = \sup_{f \in \mathscr{F}} \|\mathbb{P}_n f - \mathrm{P} f\| \xrightarrow{a.s.} 0.$$

(2) it is P-Donsker if the sequence of empirical process $\{\mathbb{G}_n f : f \in \mathscr{F}\}$ converges in distribution to a tight limit process in the space $\ell^\infty(\mathscr{F})$,where the empirical process evaluated at f is defined as $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n f - \mathrm{P} f)$.

Wether a class is Glivenko-Cantelli or Donsker depends on the "size" of the class. A relative simple way to measure the size of a class $\mathscr{F}$ is in terms of entropy. Consider the bracketing entropy relative to the $L^p(P)$-norm

$$\|f\|_{P,r} = (\mathrm{P}\,[|f|^r])^{1/r}.$$

Given two functions l and u, the bracket $[l, u]$ is the set of all functions f s.t. $l \leq f \leq u$. An ε-bracket in $L^r(P)$ is a bracket $[l, u]$ with $\|u - l\|_{P,r} \leq \varepsilon$. The bracketing number $N_{[]}(\varepsilon, \mathscr{F}, L^r(P))$ is the minimum number of ε-brackets needed to cover $\mathscr{F}$, where the bracketing function l and u must have finite $L^r(P)$-norm but need not belong to $\mathscr{F}$. The entropy with bracketing is the log of the bracketing number.

Theorem 2.5. *(Glivenko-Cantelli, thm 19.4 (van der Vaart, 2000)) Every class $\mathscr{F}$ of measurable functions s.t. $N_{[]}(\varepsilon, \mathscr{F}, L^1(P)) < \infty$ for every $\varepsilon > 0$ is P-Glivenko-Cantelli.*

For most class of interest, the bracketing number grow to ininfty as $\varepsilon \to 0$. A sufficient condition for a class to be Donsker is that they do not grow too fast. The speed can be measured in terms of the bracketing integral

$$J_{[]}\left(\delta, \mathscr{F}, L^2(P)\right) = \int_0^\delta \sqrt{\log N_{[]}\left(\varepsilon, \mathscr{F}, L^2(P)\right)} d\varepsilon.$$

If this integral is finite-valued, then the class $\mathscr{F}$ is P-Donsker.

Theorem 2.6. *Every class $\mathscr{F}$ of measurable functions with $J_{[]}\left(\delta, \mathscr{F}, L^2(P)\right) < \infty$ is P-Donsker.*

Theorem 2.7. *(Arcones and Yu, 1994) Given a sequence of strictly stationary sequence $\{X_i\}_{j=1}^{\infty}$ in a measurable space $(S, \mathcal{S})$. Suppose that*

(1) $\mathscr{F}$ is a measurable V-C subgraph class of functions, s.t.

(2) $PF^p < \infty$ for some $p \in (2, \infty)$ and $F(x) = \sup_{f \in \mathscr{F}} |f(x)|$.

(3) If the β-mixing coefficient of the stationary sequence satisfies

$$k^{p/(p-2)} (\log k)^{2(p-1)/(p-2)} \beta_k \to 0,$$

then the empirical process

$$\left\{ n^{-1/2} \sum_{t=1}^{n} \left(f(X_i) - \mathrm{P}\, f\right) : f \in \mathscr{F} \right\}$$

converges in law to a Gaussian process $\{G(f) : f \in \mathscr{F}\}$ which has a version with uniformly bounded and uniformly continuous paths w.r.t. the L^2 norm.

¡++¿

Definition 2.8. *(V-C class)*

¡++¿
The Reinsch form Let

$$S_\lambda = N(N^\intercal N + \lambda \Omega_N)^{-1} N^\intercal$$

be the spline smoother and $N = UDV^\intercal$ the singular value decomposition of N. Since N is $n \times n$, U is orthogonal hence invertible with $U^{-1} = U^\intercal$, and D is invertible since N has full rank n. Then

$$\begin{aligned} S_\lambda &= N(N^\intercal N + \lambda\Omega_N)^{-1}N^\intercal \\ &= UDV^\intercal(VD^2V^\intercal + \lambda\Omega_N)^{-1}VDU^\intercal \\ &\dots \\ &= (I + \lambda U^\intercal D^{-1}V^\intercal\Omega_N VD^{-1}U)^{-1} \\ &= (I + \lambda K)^{-1} \end{aligned}$$

3 Linear Algebra

Eigen-decomposition: A matrix M can be written as $M = P^{-1}DP$.

Schur complement:

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

References

Arcones, M. A. and Yu, B. (1994) Central limit theorems for empirical anduprocesses of stationary mixing sequences. *Journal of Theoretical Probability* **7**(1), 47–71.

Nze, P. A. and Doukhan, P. (2004) Weak dependence: models and applications to econometrics. *Econometric Theory* **20**(06), 995–1045.

van der Vaart, A. W. (2000) *Asymptotic Statistics*. Cambridge: Cambridge University Press.