Empirical Processes - Introduction

April 8, 2005

Notation and Basic Setup

We will assume that $O_1, ..., O_n \sim P$ i.i.d.

P_n is the *empirical distribution*.

Pf means $E_p f(O) = \int f dP$, likewise $P_n f = \int f dP_n$.

$\mathcal{F}$ will represent a set of real-valued functions f whose domain is the space of O_i.

$l^\infty(\mathcal{F})$ is the normed space of mappings from $\mathcal{F}$ to $\mathcal{R}$. If $G \in l^\infty(\mathcal{F})$, the norm is defined by $\|G\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \mid G(f) \mid < \infty$.

If $P \mid f \mid < \infty$ for all $f \in \mathcal{F}$, then $G_n \equiv \sqrt{n}(P_n - P)$ is a random member of $l^\infty(\mathcal{F})$, so it is a random mapping from $\mathcal{F}$ to $\mathcal{R}$. It is defined by $G_n(f) = \sqrt{n}(P_n f - Pf)$. Note that it is the randomness of the empirical distribution P_n that makes G_n a random element of $l^\infty(\mathcal{F})$.

G is said to be a P-Brownian Bridge if it is a random element of $l^\infty(\mathcal{F})$ that is continuous with probability one, such that $(G(f_1), ..., G(f_k))$ is multivariate Normal with mean zero and $Cov(G(f_i), G(f_j)) = Cov_P(f_i(O), f_j(O))$ for any k members of $\mathcal{F}$. Note that continuity is with respect to the norm of l^∞ defined earlier, so G is continuous at $f_0 \in \mathcal{F}$ if for every $\epsilon > 0$ there exists $\delta > 0$ such that if $f \in \mathcal{F}, \|f - f_0\|_\infty < \delta$, then $\|G(f) - G(f_0)\| < \epsilon$.

Glivenko-Cantelli Classes

When $\mathcal{F}$ is a finite class of functions in $L^1(P)$, it is clear from the law of large numbers that $\|P_n - P\|_{\mathcal{F}} \to 0$ almost surely. Whenever this property holds for a class $\mathcal{F}$, we call $\mathcal{F}$ a *P-Glivenko-Cantelli class* . However, not all classes of functions are Glivenko-Cantelli. Consider $\mathcal{F}$ being the class of all real-valued functions bounded between zero and one. Then for each n, there exists a (random) $f_n \in \mathcal{F}$ such that $f_n = 1(O_1, ..., O_n)$ so that $P_n f_n = 1$, but we could have $P f_n = 0$ if P is a continuous distribution. Thus

$\|P_n - P\|_{\mathcal{F}} \geq 1$ for all n, so $\mathcal{F}$ is not Glivenko-Cantelli. Basically, the larger the class $\mathcal{F}$, the harder it is for $\mathcal{F}$ to be Glivenko-Cantelli.

This leads to the *first major goal of empirical process theory*: Determine sufficient conditions for $\mathcal{F}$ to be Glivenko-Cantelli that are as easy to check as possible, but apply to as large a class $\mathcal{F}$ as possible.

Convergence in Distribution

Note that the Glivenko-Cantelli property can be thought of as a uniform law of large numbers over $\mathcal{F}$. Once we have established a uniform law of large numbers, we might also wonder if we can establish a uniform version of the Central Limit Theorem, but first we have to talk about what that could even mean.

For a single random variable $f(O) \in L^2(P)$, the Central Limit Theorem teaches us that $\sqrt{n}(\frac{1}{n}\sum f(O_i) - E_P f) \Longrightarrow N(0, var_P(f))$, but what does it mean for this to hold uniformly over $\mathcal{F}$. In empirical process theory, we are interested in showing that $G_n \Longrightarrow G$ in $(l^\infty(\mathcal{F}), \|\cdot\|_{\mathcal{F}})$, which says that the empirical process converges to the P-Brownian Bridge, when both are viewed as random functions from $\mathcal{F}$ to $\mathcal{R}$.

A naive way of defining convergence in distribution would be to say that $G_n \Longrightarrow G$ if the finite-dimensional distributions $(G_n(f_1), ..., G_n(f_k))$ converged in distribution to $(G(f_1), ..., G(f_k))$ for all $(f_1, ..., f_k) \in \mathcal{F}$. However, there are many properties of a random function that are not determined by its finite dimensional distributions, so this naive definition is insufficient. Instead, we define convergence in distribution as follows, in a way that generalizes the usual definition for real-valued random variables.

Definition: If X_n, X are random elements of a normed space $(D, \|\cdot\|)$, we say that X_n converges in distribution to X (denoted $X_n \Longrightarrow X$) if for every bounded continuous function g from $(D, \|\cdot\|)$ to $\mathcal{R}$, $E[g(X_n)] \to E[g(X)]$.

Donsker Classes

If the empirical process G_n converges in distribution in $(l^\infty(\mathcal{F}), \|\cdot\|_{\mathcal{F}})$ to the P-Brownian Bridge G, we say that $\mathcal{F}$ is a *P-Donsker class.* Because the finite dimensional distributions of G are multivariate Normal, saying that $\mathcal{F}$ is Donsker is saying that the CLT holds uniformly over $\mathcal{F}$. Note that all Donsker classes are Glivenko-Cantelli classes, but the converse is not true (consider the function class to be a single function $f \in L^1(P)$ but $f \notin L^2(P)$). As with Glivenko-Cantelli classes, $\mathcal{F}$ can only be Donsker if the class is not too big. This leads to the *second major goal of empirical process theory*: Determine sufficient conditions for $\mathcal{F}$ to be Donsker that are as easy to check as possible, but apply to as large a class $\mathcal{F}$ as possible.

In general, we will not be very concerned with establishing that classes of functions are Glivenko-Cantelli or Donsker. For us, empirical process theory will be a means to an end, but we should mention that the class of indicator functions $\mathcal{F} \equiv \{1((-\infty, t]) : t \in \mathcal{R}\}$ is a Donsker class.

Measure Theory Detail

When working with real-valued random variables, measurability issues can often be completely ignored, as the construction of subsets of $\mathcal{R}$ that are non-Borel measurable usually requires cavilling with the Axiom of Choice. Unfortunately, this is not the case when working with random functions.

If $G_n \Longrightarrow G$, we need from the definition of weak convergence that $E[g(G_n)]$ converges to $E[g(G)]$ for bounded continuous g mapping $(l^\infty(\mathcal{F}), \|\cdot\|_{\mathcal{F}})$ to $\mathcal{R}$. But to even talk about $E[g(G_n)]$ according to the usual definition of expectation, we need for $g(G_n)$ to be a Borel measurable function. If $O_1, ..., O_n$ are defined on a probability space $(\Omega, \mathcal{B}, P)$, where $\mathcal{B}$ denotes the Borel subsets, it is possible to find $B \in \mathcal{B}$ and bounded continuous g such that $\{\omega : g(G_n)(\omega) \in B\} \notin \mathcal{B}$, so that $g(G_n)$ is non- Borel measurable.

We can circumvent this technical difficulty by defining the outer expectation as $E^\star g = \inf\{Ef : f \geq g \text{ is measurable}\}$. So technically, $X_n \in (D, \|\cdot\|)$ converges in distribution to $X \in (D, \|\cdot\|)$ if for all bounded continuous g mapping $(D, \|\cdot\|)$ to $\mathcal{R}$, $E^\star[g(X_n)]$ converges to $E[g(X)]$. However, this and other measurability details will be ignored in subsequent lectures.

An application of empirical process results to simultaneous confidence bands.

Result 0.1. *Let $G_{n,P} \in \ell^\infty(\mathcal{F})$ be an empirical process indexed by a class of functions $\mathcal{F}$. Suppose that $\mathcal{F}$ is a Donsker class: that is, $G_{n,P} \overset{D}{\Longrightarrow} G_P$ in $\ell^\infty(\mathcal{F})$, where G_P is the Gaussian process defined by its finite dimensional distributions being multivariate normal with covariance implied by pairwise covariances $COV(G_P(f_1), G_P(f_2)) = COV_P(f_1(O), f_2(O))$. Let $q_{0.95,P}$ be the 0.95-quantile of $\|G_P\|_{\mathcal{F}} \equiv \sup_{f\in\mathcal{F}} \mid G(f) \mid$. Then*

$$Pr(Pf \in P_n f \pm q_{0.95,P}/\sqrt{n} \text{ for all } f \in \mathcal{F}) \to 0.95. \tag{1}$$

To obtain the nicest type simultaneous confidence band (that is, a band which is wide were the estimator $P_n f$ is highly variable, and small at f where the estimator $P_n f$ is precise) one would choose $\mathcal{F}$ so that $\mathrm{VAR}_P(f(O)) = \sigma^2$ does not depend on the choice $f \in \mathcal{F}$. For example, given a class $\mathcal{F}_0$, one would define

$$\mathcal{F} \equiv \{f/\sigma(f) : f \in \mathcal{F}_0\},$$

where $\sigma^2(f) \equiv \mathrm{VAR}_P f(O)$. An interesting question for empirical process theory is if the fact that $\mathcal{F}_0$ is Donsker, implies that $\mathcal{F}$ is Donsker. Clearly, if $\inf_{f\in\mathcal{F}_0} \sigma^2(f) > 0$, then the answer is yes, but, if $\sigma(f)$ can approximate zero arbitrarily close so that functions $f/\sigma(f)$ can become unbounded (but finite variance), then we will probably need some condition,.

Proof: Consider the function $g : \ell^\infty(\mathcal{F}) \rightarrow \mathbb{R}$ defined by $g(G) \equiv \|G\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \mid G(f) \mid$. This function is continuous. Since $G_{n,P} \overset{D}{\Longrightarrow} G_P$ in $\ell^\infty(\mathcal{F})$, the continuous mapping theorem teaches us that $g(G_{n,P}) \overset{D}{\Longrightarrow} g(G_P)$. Since $q_{0.95,P}$ is a continuity point of the limit distribution $g(G_P)$, weak convergence implies that

$$\Pr(\sup_{f\in\mathcal{F}} \mid G_{n,P}(f) \mid \leq q_{0.95,P}) \rightarrow \Pr(\sup_{f\in\mathcal{F}} \mid G_P(f) \mid \leq q_{0.95,P}) = 0.95.$$

Since the left-hand side equals (1), this completes the proof.

Result 0.2. *Let $G_{n,P_n} \in \ell^\infty(\mathcal{F})$ be the empirical process indexed by a class of functions $\mathcal{F}$ corresponding with sampling from the empirical distribution P_n. Suppose that $\mathcal{F}$ is a Donsker class so that $G_{n,P_n} \overset{D}{\Longrightarrow} G_P$, conditional on almost every data realization $(P_n : n \geq 1)$ (van der Vaart, Wellner, 1996). Let $q_{n,0.95}$ be the 0.95-quantile of $\|G_{n,P_n}\|_{\mathcal{F}}$. Then $q_{n,0.95} \rightarrow q_{0.95,P}$, and thus (by the previous result)*

$$Pr(Pf \in P_n f \pm q_{n,0.95}/\sqrt{n} \text{ for all } f \in \mathcal{F}) \rightarrow 0.95. \tag{2}$$

Proof. We have

$$0.95 = \Pr(\|G_{n,P_n}\|_{\mathcal{F}} \leq q_{n,0.95}). \tag{3}$$

By the continuous mapping theorem, we have $\|G_{n,P_n}\|_{\mathcal{F}} \overset{D}{\Longrightarrow} \|G_P\|_{\mathcal{F}}$ a.e. Since pointwise convergence of cumulative distribution functions to a continuous cumulative distribution function implies uniform convergence, this implies that the cumulative distribution function of $\|G_{n,P_n}\|_{\mathcal{F}}$ converges uniformly to the cumulative distribution function of $\|G_P\|_{\mathcal{F}}$. Thus,

$$\Pr(\|G_{n,P_n}\|_{\mathcal{F}} \leq q_{n,0.95}) - \Pr(\|G_P\|_{\mathcal{F}} \leq q_{n,0.95}) \rightarrow 0,$$

for $n \rightarrow \infty$. Combining this with (3) yields

$$\Pr(\|G_P\|_{\mathcal{F}} \leq q_{n,0.95}) \rightarrow 0.95.$$

Finally, taking the inverse of the cdf of $\|G_P\|_{\mathcal{F}}$ on both sides yields the wished convergence of $q_{n,0.95}$ to $q_{0.95,P}$ a.e. This completes the proof.

References

See *Weak Convergence and Empirical Processes* by van der Vaart and Wellner. Older empirical process references include books by Pollard and Billingsley.

Almost Sure Representation Theorem

Theorem 0.1. *Suppose that X_n converges to X in distribution in a normed space $(D, \|\cdot\|)$, where X is Borel measurable and separable with probability one. Then there exist $Y_n, Y \in (D, \|\cdot\|)$ such that X_n equals Y_n in distribution (so $Eg(X_n) = Eg(Y_n)$ for any bounded continuous real-valued g), X equals Y in distribution and $Y_n \to Y$ almost surely (meaning $\|Y_n - Y\| \to 0$ almost surely).*

Here the separability of X means that with probability one, there exists a countable subset $\{d_1, d_2, ..\}$ of D such that for any Borel measurable compact subset D_0 of D, $P(X \in D_0) = P(X \in D_0 \cap \{d_1, d_2, ...\})$. When $(D, \|\cdot\|) = (l^\infty(\mathcal{F}), \|\cdot\|_{\mathcal{F}})$ and X is the P-Brownian Bridge G, this condition is satisfied because G is defined to be continuous. Here, the Borel sigma field just denotes the sigma field generated by the open subsets of $(D, \|\cdot\|)$.

Continuous Mapping Theorem

Theorem 0.2. *Suppose that X_n converges to X in distribution in a normed space $(D, \|\cdot\|)$, where X is Borel measurable and separable with probability one, and H is a continuous mapping from $(D, \|\cdot\|)$ to another normed space $(E, \|\cdot\|)$. Then $H(X_n)$ converges in distribution to $H(X)$.*

proof: Let g be a bounded continuous mapping from $(E, \|\cdot\|)$ to $\mathcal{R}$. By the a.s. representation theorem, there exists $Y_n =_d X_n$, $Y =_d X$ such that $Y_n \to Y$ a.s. Hence, $H(Y_n) \to H(Y)$ almost surely because H is continuous. Hence, $g(H(Y_n)) \to g(H(Y))$ almost surely because g is continuous. As g is bounded, the bounded convergence theorem tells us that $E[g(H(X_n))] = E[g(H(Y_n))] \to E[g(H(Y))] = E[g(H(X))]$, and the result follows from the definition of convergence in distribution. □

Note that the method used in this proof shows that almost sure convergence always implies convergence in distribution. See the class notes for a more powerful version of this theorem, called the extended continuous mapping theorem.

The continuous mapping theorem has many important applications. In particular, it can be invoked to show convergence in distribution for real-valued continuous functionals of the empirical process G_n. Usually we only care such functionals, rather than the behavior of the entire random function G_n in $l^\infty(\mathcal{F})$, and empirical process theory provides an elegant tool for answering questions of statistical interest.

For example, consider the class of indicator functions $\mathcal{F} = \{1((-\infty, t]) : t \in \mathcal{R}\}$, and define $H : l^\infty(\mathcal{F}) \to \mathcal{R}$ by $H(X) = \|X\|_{\mathcal{F}}$. Clearly H is continuous on $(l^\infty(\mathcal{F}), \|\cdot\|_{\mathcal{F}})$, so the continuous mapping theorem tells us that $H(G_n)$ converges in distribution to $H(G)$, for G the P-Brownian Bridge. Kolmogorov found the distribution of $H(G)$ analytically, and the quantiles of this distribution can be used to form asymptotically valid confidence bands for the cumulative distribution function of P.

The Ordinary Delta Method

Recall that if X_n is a real valued random variable and μ a constant such that $\sqrt{n}(X_n - \mu) \implies N(0, \sigma^2)$, and ϕ is a real-valued function with a continuous derivative, that $\sqrt{n}(\phi(X_n) - \phi(\mu)) \implies N(0, [\phi'(\mu)]^2\sigma^2)$. This idea can be generalized to find the limiting distribution of $\sqrt{n}(\phi(P_n) - \phi(P))$, when ϕ is a functional of probability distributions, so that $\phi(P_n)$ is an estimator of a parameter $\phi(P)$. But in order to state the formal result, we first have to discuss differentiation in general spaces.

Functional Derivatives

If $\phi : \mathcal{R} \to \mathcal{R}$, it is easy to define what is meant by ϕ being differentiable at a point. If $\phi : (l^\infty, \|\cdot\|_{\mathcal{F}}) \to \mathcal{R}$, or more generally, $\phi : (D, \|\cdot\|) \to (E, \|\cdot\|)$, obviously the usual definition of differentiability does not apply.

A heuristic is that if $\phi : (D, \|\cdot\|) \to (E, \|\cdot\|)$ is differentiable at $P \in D$, then the derivative of ϕ at P (denoted $d\phi_P$) is a continuous linear map $d\phi_P : (D, \|\cdot\|) \to (E, \|\cdot\|)$ that is a linear approximation to ϕ at P.

Unfortunately, there is ambiguity about how to define a derivative map between general spaces. The three most important types of derivatives are listed below.

Definitions: If there is a continuous linear map $d\phi_P : (D, \|\cdot\|) \to (E, \|\cdot\|)$ such that $Rem(h) = \|\phi(P+h) - \phi(P) - d\phi_P(h)\|$ for $h \in D$, then ϕ is (*Gateaux*, *Hadamard*, *Frechet*) differentiable at P if $Rem(\epsilon h)/\epsilon \to 0$ uniformly over h for h in any (singleton of D, compact subset of D, bounded subset of D) respectively.

Note that Frechet differentiability is stronger than Hadamard differentiability, which is stronger than Gateaux differentiability. If the Frechet derivative exists, it is equal to the Hadamard derivative, and if the Hadamard derivative exists, it is equal to the Gateaux derivative.

Note that for $\phi : \mathcal{R}^k \to \mathcal{R}$, the Gateaux derivative corresponds to the directional derivative, and the Frechet and Hadamard conicide and are equal to the total derivative. When $k = 1$, all three derivatives are the same, and coincide with the ordinary definition of a derivative. This is because the Gateaux derivative at x in the direction h is given by $\frac{d}{d\epsilon}\phi(x + \epsilon h) \mid_{\epsilon=0} = \phi'(x)h$. Thus, $d\phi_x$ is the linear map map $h \to \phi'(x)h$. But when $k > 1$, it is possible for the directional derivative to be defined in every direction, but for the total derivative to not exist.

Note that the derivative depends on what norm is chosen for both spaces D and E. Generally a strong norm in D and a weak norm in E makes it easier for the derivative to exist.

The Hadamard derivative is thought to be the most useful in empirical process theory, as Frechet differentiability is too hard to establish in many cases, but Gateaux differentiability is not strong enough to imply desired results. Because the Hadamard derivative is so useful, it is convenient to rephrase Hadamard differentiability in another form.

Theorem 0.3. *A map $\phi : (D, \| \cdot \|) \rightarrow (E, \| \cdot \|)$ is Hadamard differentiable at $P \in D$ with derivative $d\phi_P : (D, \| \cdot \|) \rightarrow (E, \| \cdot \|)$ if $d\phi_P$ is a continuous linear map such that $\frac{\phi(P+t_n h_n)-\phi(P)}{t_n} \rightarrow d\phi_P(h)$ for all scalar sequences $t_n \rightarrow 0$ and $h_n \in D \rightarrow h \in D$.*

Often we are interested in functionals defined on a space of probability distributions, so that $\phi(P)$ is an unknown parameter and $\phi(P_n)$ is an estimator. But notice that the definition of the derivative requires specifying a normed space. So what normed space should we choose to include all probability distributions? Should we think of P and P_n as members of $(l^\infty(\mathcal{F}), \| \cdot \|_{\mathcal{F}})$? If so, for what $\mathcal{F}$? One common choice is to think of P and P_n as distribution functions (which are monotone right-continuous functions with left limits) and define the domain D of ϕ to be the space of *cadlag* functions. Cadlag is a french acronym for *continuous from the right with limits from the left*, and as the name implies, it is the space of all real-valued right-continuous functions with left limits. The most common norm used for the space of cadlag functions is the supremum norm.

Once differentiability has been established (any of the three kinds), actually finding the derivative is easy, because $d\phi_P(h) = \frac{d}{d\epsilon}\phi(P + \epsilon h) \mid_{\epsilon=0}$. The right side is just the derivative of a real-valued function of ϵ, evaluated at zero, so it can often be found using high school calculus.

Example: Consider ϕ defined on the space of cadlag functions with supremum norm, by $\phi(F) = \int_0^t \frac{1}{1-F(s_-)} dF(s)$. This is an important functional in survival analysis, represents the *cumulative hazard* when $O \sim P$, and F is the cumulative distribution function of P. Evaluating $\frac{d}{d\epsilon}\phi(P + \epsilon h) \mid_{\epsilon=0}$ we see that the Gateaux derivative at F in the direction h is given by $d\phi_F(h) = \int_0^t \frac{1}{1-F(s_-)} dh(s) + \int_0^t \frac{h(s_-)}{(1-F(s_-))^2} dF(s)$. The Hadamard derivative of the cumulative hazard function will be important later when we analyze the well-known Kaplan-Meier estimator.

Finally, there is a generalization of the well-known chain rule for Hadamard differentiation.

Theorem 0.4. *Suppose $\phi : D \rightarrow E$ has Hadamard derivative $d\phi_P$ at $P \in D$ and that $\psi : E \rightarrow F$ has Hadamard derivative $d\psi_{\phi(P)}$ at $\phi(P) \in E$. Then the composition $\psi(\phi) : D \rightarrow F$ has Hadamard derivative $d\psi_{\phi(P)}(d\psi_P)$ at P.*

proof: Consider a scalar sequence $t_n \rightarrow 0$. For $h_n \in D \rightarrow h$, $k_n \equiv \frac{\phi(P+t_n h_n)-\phi(P)}{t_n} \rightarrow d\phi_P(h)$, by the Hadamard differentiability of ϕ at P. Hence, $\frac{\psi(\phi(P+t_n h_n))-\psi(\phi(P))}{t_n} = \frac{\psi(\phi(P)+t_n k_n)-\psi(\phi(P))}{t_n} \rightarrow d\psi_{\phi(P)}(d\phi_P(h))$ by the Hadamard differentiability of ψ at $\phi(P)$, thus proving the desired result. □

We now state the main result, which is called the *functional delta method.*

Theorem 0.5. *For D and E normed spaces, suppose $\phi : D \to E$ has Hadamard derivative $d\phi_P$ at $P \in D$, and that for $X_n, X \in D$, $\sqrt{n}(X_n - P) \implies X$. If X is Borel measurable and separable, then $\sqrt{n}(\phi(X_n) - \phi(P)) \implies d\phi_P(X)$. Further, $\|\sqrt{n}(\phi(X_n) - \phi(P)) - d\phi_P(\sqrt{n}(X_n - P))\|$ converges in distribution (so also in outer probability) to zero.*

proof: Taken from page 374 of van der Vaart and Wellner. By the almost sure representation theorem, there exist $Y_n =_d \sqrt{n}(X_n - P)$, $Y =_d X$ such that $\|Y_n - Y\| \to 0$ almost surely. Thus $\sqrt{n}(\phi(X_n) - \phi(P)) =_d \frac{\phi(P+Y_n/\sqrt{n})-\phi(P)}{1/\sqrt{n}} \to_{a.s.} d\phi_P(Y) =_d d\phi(X)$, by the Hadamard differentiability of ϕ, using $1/\sqrt{n}$ in place of t_n. This proves the first part. The second part follows by considering the map $\psi : D \to (E, E)$ defined by $\psi(d) = (\phi(d), d\phi_P(d))$, with Hadamard derivative $(d\phi_P, d\phi_P)$. From this Hadamard differentiability, it follows that $(\sqrt{n}(\phi(X_n) - \phi(P)), \sqrt{n}(d\phi_P(X_n) - d\phi_P(P)) \implies (d\phi_P(X), d\phi_P(X))$, and then apply the continuous mapping theorem to the difference of the two coordinates. □

Basics of convergence in Distribution

Review of probabilistic big-oh, little-oh notation: Recall that for random variables A_n, B_n taking values in a normed space, $A_n = o_P(B_n)$ means that $\|A_n\|/\|B_n\|$ converges to zero in probability under P. So $A_n = o_P(n^{-1/2})$ means that $\sqrt{n}\|A_n\|$ converges to zero in probability under P. A_n is said to be *bounded in probability* (denoted $A_n = O_P(1)$) if for all $\epsilon > 0$ there exists $M < \infty$ such that $P(\|A_n\| > M) \leq \epsilon$. $A_n = O_P(B_n)$ means that $\|A_n\|/\|B_n\| = O_P(1)$. Some helpful rules to keep in mind are that $O_P(1)o_P(1) = o_P(1)$ and that $O_P(1) + o_p(1) = O_P(1)$.

One important result for proving convergence in distribution is *Slutsky's Theorem.* Suppose that a scalar sequence a_n converges in probability to a, a sequence b_n in a metric space $(D, \|\cdot\|)$ converges to a constant value in $(D, \|\cdot\|)$ and $X_n \in (D, \|\cdot\|)$ converges in distribution to separable $X \in (D, \|\cdot\|)$. Slutsky's Theorem states that $a_n X_n + b_n$ converges in distribution to $aX + b$.

Influence Curves

Definition: For $O_1, ..., O_n \sim P$ i.i.d., $\phi_n = \phi_n(O_1, ..., O_n)$ is said to an *asymptotically linear* estimator of $\phi(P) \in \mathcal{R}^k$ with *influence curve* $IC(O|P) \in \mathcal{R}^k$ if $E_P IC(O|P) = 0$, $E_P\|IC(O|P)\|_2^2 < \infty$, and $\phi_n = \phi(P) + \frac{1}{n}\sum_{i=1}^n IC(O_i|P) + o_P(n^{-1/2})$.

Note: The influence curve depends on the unknown P, so it is not a statistic.

If ϕ_n is asymptotically linear for $\phi(P)$, the CLT tells us that $\sqrt{n}(\phi_n - \phi(P) \implies N(0, \Sigma_P)$

under P, where $\Sigma_P = E_P[IC(O|P)IC(O|P)^T]$. If we can consistently estimate Σ_P from the data (often with the empirical estimator $\frac{1}{n}\sum_{i=1}^n IC(O_i \mid P_n)IC(O_i \mid P_n)^T$), then we can form asymptotically valid confidence regions for $\phi(P)$. This is why asymptotic linearity is considered a stronger and more desirable property that $\sqrt{n}$-consistency ($\|\phi_n - \phi(P)\| = O_P(n^{-1/2})$).

Examples

If $O_i \in \mathcal{R}$, and $E_P|O_i| < \infty$, then the sample mean is asymptotically linear for $E_P(O)$ with influence curve $IC(O|P) = O - E_P(O)$. Likewise, the empirical distribution at a fixed point t ($\frac{1}{n}\sum_{i=1}^n 1(O_i \leq t)$) is asymptotically linear for $P(O \leq t)$ with influence curve $IC(O|P) = 1(O \leq t) - P(O \leq t)$. In both of these examples, there is no remainder (the $o_P(n^{-1/2}$ term). The heuristic of asymptotic linearity is that up to an $o_P(n^{-1/2})$ term, the estimator behaves like a sample mean.

Sample Median

This example is slightly harder, because although the sample median is an asymptotically linear estimator of the median under regularity conditions, there is a remainder.

Theorem 0.6. *Suppose real-valued $O_1, ...O_n \sim P$ i.i.d. has cumulative distribution function F, and F_n denotes the empirical c.d.f. Suppose that F has a density f that is positive and continuous in a neighborhood of the unique median θ, where $F(\theta) = 1/2$. If θ_n is the sample median $F^{-1}(1/2) = \inf\{x : F_n(x) \geq 1/2\}$, then θ_n is an asymptotically linear estimator of θ with influence curve $IC(O_i|P) = -\frac{1}{f(\theta)}(1(O \leq \theta) - 1/2)$.*

proof: We first establish the consistency of θ_n. As the median is given to be unique, $F(\theta + \epsilon) > 1/2$ and $F(\theta - \epsilon) < 1/2$ for any $\epsilon > 0$. Hence, $P(\mid \theta_n - \theta| > \epsilon) = P(F_n(\theta - \epsilon) \geq 1/2) + P(F_n(\theta + \epsilon) < 1/2) \rightarrow_P 0$ because $F_n(t) \rightarrow_P F(t)$ for any t by the law of large numbers. Thus $\theta_n \rightarrow_P \theta$.

Let G_n denote the empirical process $\sqrt{n}(F_n - F)$ (technically this is $\sqrt{n}(P_n - P) \in (l^\infty(\mathcal{F}), \|\cdot\|_\mathcal{F})$ where $\mathcal{F} = \{1((-\infty, t]) : t \in \mathcal{R}\}$). It can be shown via empirical process theory that $G_n \Longrightarrow G$, for G the P-Brownian Bridge. So as $(G_n, \theta_n) \Longrightarrow (G, \theta)$ and G is continuous, the continuous mapping theorem yields $G_n(\theta) - G_n(\theta_n) = o_P(1)$.

As f is continuous in a neighborhood of θ, Taylor expanding $F(\theta_n)$ about θ gives $F(\theta_n) = 1/2 + (\theta_n - \theta)(f(\theta) + o_P(1))$. Clearly $F_n(\theta_n) = 1/2 + o_P(n^{-1/2})$, so $G_n(\theta_n) = \sqrt{n}(F_n(\theta_n) - F(\theta_n)) = \sqrt{n}(1/2 + o_P(n^{-1/2}) - 1/2 - (\theta_n - \theta)(f(\theta) + o_P(1)) = -\sqrt{n}(\theta_n - \theta)(f(\theta) + o_P(1))$.

Rearranging terms gives $\sqrt{n}(\theta_n - \theta) = -\frac{G_n(\theta_n)}{f(\theta)} + o_P(1) = -\frac{G_n(\theta)}{f(\theta)} + \frac{G_n(\theta) - G_n(\theta_n)}{f(\theta)} + o_P(1) = -\frac{G_n(\theta)}{f(\theta)} + o_P(1)$ from our comments above. From the definition of G_n, dividing both sides by $\sqrt{n}$ gives that $\theta_n = \theta - \frac{1}{n}\sum_{i=1}^n \frac{1}{f(\theta}(1(O_i \leq \theta) - 1/2) + o_P(n^{-1/2})$, proving the desired result. □

Bootstrapping

Note that in the case of the median under the conditions of the previous theorem, the influence curve teaches us that $\sqrt{n}(\theta_n - \theta) \Longrightarrow N(0, \frac{1}{4f^2(\theta)})$. So how can we give asymptotically valid confidence intervals for the median θ if we have to estimate an asymptotic variance that involves a density function. Densities can be very poorly behaved and hard to estimate. The easiest way out is to just use the bootstrap, which provides automatic confidence intervals. The bootstrap has been shown to work for the median, and under very general conditions for asymptotically linear estimators. In fact, the sample median was one of the examples discussed in Efron's 1979 paper where he invented the bootstrap method.

The Influence Curve and the Functional Delta Method

Often the easiest way to compute the influence curve of an estimator $\phi(P_n)$ of a parameter $\phi(P)$ is to apply the functional delta method. The major difficulty is often showing that ϕ is a Hadamard differentiable map. We summarize the result in the following theorem.

Theorem 0.7. *Suppose that $O_1, ..., O_n \sim P$ i.i.d., and that $G_n \in (l^\infty(\mathcal{F}), \|\cdot\|_\mathcal{F})$ is the empirical process, so $G_n(f) = \sqrt{n}(P_n(f) - P(f))$ for $f \in \mathcal{F}$. Let $G_{1,O} \in (l^\infty(\mathcal{F}), \|\cdot\|_\mathcal{F})$ be defined by $G_{1,O}(f) = f(O) - E_P(f(O))$. Suppose $\mathcal{F}$ is a Donsker class (so $G_n \Longrightarrow G$ for G the P-Brownian Bridge). If $\phi : (l^\infty(\mathcal{F}), \|\cdot\|_\mathcal{F}) \to \mathcal{R}^k$ has Hadamard derivative $d\phi_P$ at P, then $\phi(P_n)$ is an asymptotically linear estimator for $\phi(P)$ with influence curve $IC(O|P) = d\phi_P(G_{1,O})$.*

proof: As $d\phi_P$ is a linear map and $G_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n G_{1,O_i}$, $d\phi_P(G_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^n d\phi_P(G_{1,O_i}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n IC(O_i|P)$. But by the functional delta method (the second statement in the theorem), $\sqrt{n}(\phi(P_n) - \phi(P)) = d\phi_P(G_n) + o_P(1) = \frac{1}{\sqrt{n}} \sum_{i=1}^n IC(O_i|P) + o_P(1)$, so dividing both sides by $\sqrt{n}$ gives that $\phi(P_n) = \phi(P) + \frac{1}{n} \sum_{i=1}^n IC(O_i|P) + o_P(n^{-1/2})$, thus proving the desired result. □

Terminology for Normed Spaces

Definition: A (real) *linear space* D is a set with addition and scalar multiplication defined such that if $d_1, d_2, d_3 \in D$ and $c_1, c_2 \in \mathcal{R}$:

$c_1 d_1 + c_2 d_2 \in D$ (closed under linear combination).
$d_1 + d_2 = d_2 + d_1$ (commutative). $d_1 + (d_2 + d_3) = (d_1 + d_2) + d_3$ (associative addition).

There exists $0 \in D$ such that $d_1 + 0 = d_1$ and $d_1 + (-d_1) = 0$.
$1d_1 = d_1$.
$c_1(c_2 d_1) = (c_1 c_2) d_1$ (associative multiplication).
$c_1(d_1 + d_2) = c_1 d_1 + c_1 d_2$ and $(c_1 + c_2) d_1 = c_1 d_1 + c_2 c_1$ (distributive).

Definition: A (real) *normed space* (also called a normed linear space) $(D, \|\cdot\|)$ is a real linear space D and a function $\|\cdot\| : D \to \mathcal{R}$ such that for $d_1, d_2 \in D$ and $c \in \mathcal{R}$:

$\|d_1\| \geq 0$ and $\|d_1\| = 0$ if and only if $d_1 = 0$.
$\|cd_1\| =\mid c \mid \|d_1\|$.
$\|d_1 + d_2\| \leq \|d_1\| + \|d_2\|$ (triangle inequality).

Basic example: $\mathcal{R}^k$ is normed space. For $p \geq 1$, $\|(x_1, ..., x_n)\| = (\sum_{i=1}^n |x_i|^p)^{1/p}$ defines a norm. $p = 2$ corresponds to the Euclidean norm.

More complicated examples: The space of cadlag functions $(D(a,b), \|\cdot\|)$, which is the space of all functions defined on (a,b) that are right continuous with left limits, such that $\|d\| = \sup_{a \leq t \leq b} \mid d(t) \mid$. Also, $(l^\infty(\mathcal{F}), \|\cdot\|_\mathcal{F})$, defined previously, is a normed space. Sets of random variables can also live in normed spaces. $L_0^P \equiv \{X : E_P X = 0, E_P|X|^p < \infty\}$ is a normed space for $p \geq 1$ with $\|X\| = (E|X|^p)^{1/p}$.

Definition: If $d_1, d_2, ...$ is a sequence in a normed space $(D, \|\cdot\|)$, the sequence is said to converge to $d \in D$ if for all $\epsilon > 0$ there exists a positive integer N such that $n \geq N$ implies $\|d_n - d\| \leq \epsilon$.

Definition: If $d_1, d_2, ...$ is a sequence in a normed space $(D, \|\cdot\|)$, we say the sequence is *Cauchy* if for all $\epsilon > 0$ there exists a positive integer N such that $n \geq N$ implies $sup_{m,n \geq N} \|d_m - d_n\| \leq \epsilon$.

Definition: We say that a normed space $(D, \|\cdot\|)$ is a *Banach space* if every Cauchy sequence in $(D, \|\cdot\|)$ converges to some $d \in D$.

Definition: If $f : (D, \|\cdot\|) \to (E, \|\cdot\|)$ is a mapping from one normed space to another, we say that f is *continuous* at $d \in D$ if for every sequence $d_1, d_2, ...$ in D converging to d we have that $f(x_1), f(x_2), ...$ converges to $f(d)$ in $(E, \|\cdot\|)$. If f is continuous at all $d \in D$, we say that f is a continuous function.

Definition: If $(D, \|\cdot\|)$ is a normed space, then for any $\epsilon > 0$ and $d \in D$, the *open ball* of radius ϵ at d is defined by $B_\epsilon(d) = \{d' \in D : \|d' - d\| < \epsilon\}$.

Definition: If $D_0 \subset D$ for $(D, \|\cdot\|)$ a normed space, we say that D_0 is an *open set* if for every $d_0 \in D_0$ there exists $\epsilon(d_0)$ such that $B_{\epsilon(d_0)}(d_0) \subset D_0$. A subset of D is said to be *closed* if its complement is open.

Definition: If $D_0 \subset D$ for $(D, \|\cdot\|)$ a normed space, we say that D_0 is *bounded* if for each $d_0 \in D_0$ there exists $r(d_0) > 0$ such that $D_0 \subset B_{r(d_0)}(d_0)$.

Definition: If $D_0 \subset D$ for $(D, \|\cdot\|)$ a normed space, a collection of sets is a *cover* if D_0 is a subset of the union of sets in the collection. If each set in the collection is open, the collection is said to be an *open cover*. If the union of sets in a subcollection of the collection still contains D_0, the subcollection is said to be a *subcover* of D_0. If every open cover of D_0 contains a finite subcover, D_0 is said to be *compact*. A set D_0 is compact if and only if every sequence $d_1, d_2, ... \in D_0$ contains a subsequence converging to an element of D_0.

Definition: If $(D, \|\cdot\|)$ is a normed space, the normed space is said to be *separable* if there exists a countable *dense* subset $\{d_1, d_2, ...\}$ of D such that for any $d \in D$ and any $\epsilon > 0$ there exists d_n in the countable subset such that $\|d - d_n\| \leq \epsilon$. That is, the normed space can be approximated arbitrarily well by a countable set.

Note on Integration Theory

Definition: Let $D[0, b]$ denote the space of cadlag functions on $[0, b]$. $BV_M \subset D[0, b]$ is the set of cadlag functions on $[0, b]$ of *bounded variation*, indexed by some $M > 0$. $A \in BV_M$ if $\sum_{i=1}^{n} |A(t_i) - A(t_{i-1})| \leq M$ for all $0 \leq t_1 \leq ... \leq t_n \leq b$.

Suppose a is a Borel measurable function on $[0, b]$. If A is a monotone cadlag function on $[0, b]$ such that $-\infty < A(0) \leq A(b) < \infty$, recall that $\int_0^b adA$ is defined as $\int_0^b ad\mu$. Here μ is the measure on $[0, b]$ (with respect to the Borel sigma-field) uniquely defined (by Caratheodory's Extension Theorem) by $\mu((0, b]) = A(b) - A(0)$. It can be shown that if $A \in BV_M \subset D[0, b]$ then A can be uniquely written as $A_1 - A_2$ where A_1, A_2 are monotone cadlag functions such that $-\infty < A_j(0) \leq A_j(b) < \infty$. In this case $\int_0^b adA$ can be defined as $\int_0^b adA_1 - \int_0^b adA_2$. If A is not necessarily of bounded variation, but a is of bounded variation, then $\int_0^b adA$ can be defined by *integration by parts* as $a(A(b)) - a(A(0)) - \int_0^b A(t_-)da(t)$.

Product Integrals

Our treatment is based on section 3.9 of van der Vaart and Wellner. For an excellent overview, see *www.math.uu.nl/people/gill/Preprints/prod_int_0.pdf*

Definition: Let $D[0, b]$ denote the space of cadlag functions on $(0, b]$. The *product integral* of $A \in BV_M \subset D[0, b]$ is a function $\phi(A) \in D[0, b]$ denoted by $\phi(A)(t) = \Pi_{0<s\leq t}(1+dA(s))$. It is defined by $\lim_{\max_i |t_i - t_{i-1}| \to 0} \Pi_i(1 + A(t_i) - A(t_{i-1}))$, where the limit is taken over partitions $0 = t_0 < t_1 < ... < t_n = t$ with $\sup_{0\leq s\leq b} \min_{1\leq i\leq n} |t_i - s| \to 0$. It can be shown that the limit exists and is unique, and doesn't depend on which sequence of partitions $(t_0, t_1, ..., t_n)$ is chosen, so $\phi(A)$ is well-defined. For $s < t$, $\phi(A)(s,t]$ is notation for $\frac{\phi(A)(t)}{\phi(A)(s)}$.

Another way to represent the product integral is as follows.

Theorem 0.8. *For $A \in BV_M \subset D(0,b]$, the product integral $\phi(A) \in D[0,b]$ is equal to the unique solution of the Volterra equation $\phi(A)(t) = 1 + \int_0^t \phi(A)(s_-)dA(s)$, for $0 \leq t \leq b$.*

Suppose that $a_1, ..., a_n, b_1, ..., b_n$ are real numbers. It is easy to check by induction that $\Pi_{i=1}^n a_i - \Pi_{i=1}^n b_i = \sum_{i=1}^n (\Pi_{j=1}^{i-1} a_j)(a_i - b_i)(\Pi_{k=i+1}^n b_k)$. For $n \leq 2$, it is just algebra to check that $a_1 a_2 - b_1 b_2 = (a_1 - b_1)b_2 + a_1(a_2 - b_2)$. For $n > 2$, let $\tilde{a}_2 = \Pi_{i=2}^n a_i$ and $\tilde{b}_2 = \prod_{i=2}^n b_i$. Then $\Pi_{i=1}^n a_i - \Pi_{i=1}^n b_i = a_1\tilde{a}_2 - b_1\tilde{b}_2 = (a_1 - b_1)\tilde{b}_2 + a_1(\tilde{a}_2 - \tilde{b}_2) = (a_1 - b_1)\prod_{i=2}^n b_i + a_1 \sum_{i=2}^n (\Pi_{j=2}^{i-1} a_j)(a_i - b_i)(\Pi_{k=i+1}^n b_k) = \sum_{i=1}^n (\Pi_{j=1}^{i-1} a_j)(a_i - b_i)(\Pi_{k=i+1}^n b_k)$. This *telescoping trick* for representing differences of products can be generalized to differences of product integrals with the following result, known as the *Duhamel equation*.

Theorem 0.9. *Suppose that $A, B \in BV_M \subset D[0,b]$. If ϕ denotes the product integral, then $\phi(B)(t) - \phi(A)(t) = \int_0^t \phi(A)(u)\phi(B)(u,t]d(B-A)(u)$.*

We will need a further result before we can give the Hadamard derivative of the product integral. This can be proven by integration by parts. See problem 3.9.8 of van der Vaart and Wellner.

Theorem 0.10. *Suppose that $A, B \in BV_M \subset D[0,b]$. Recall that $\phi(A), \phi(B) \in D[0,b]$. For $d \in D[0,b]$, $\|d\|_\infty$ denotes $\sup_{0 \leq t \leq b} |d(t)|$. If ϕ denotes the product integral, then $\|\phi(B) - \phi(A)\|_\infty \leq C(M)\|B - A\|_\infty$ for a constant $C(M)$ depending on M. Thus, product integration is uniformly continuous.*

We are now ready to provide the main result on product integration.

Theorem 0.11. *$\phi : (BV_M, \|\cdot\|_\infty) \subset (D[0,b], \|\cdot\|_\infty) \to (D[0,b], \|\cdot\|_\infty)$ is Hadamard differentiable at $A \in BV_M \subset D[0,b]$, with Hadamard derivative $d\phi_A(\alpha)(t) = \int_0^t \phi(A)(u)\phi(A)(u,t]d\alpha(u)$, where ϕ denotes the product integral.*

sketch of proof: Suppose a scalar sequence $t_n \to 0$ and $\alpha_n \in D[0,b] \to \alpha \in D[0,b]$.
$\|\frac{\phi(A+t_n\alpha_n)-\phi(A)}{t_n} - \int_0^\cdot \phi(A)(u)\phi(A)(u,\cdot]d\alpha(u)\|_\infty$
$= \|\frac{1}{t_n}\int_0^\cdot \phi(A)(u)\phi(A+t_n\alpha_n)(u,\cdot]d(A - A + t_n\alpha_n) - \int_0^\cdot \phi(A)(u)\phi(A)(u,t]d\alpha(u)\|_\infty$
$= \| \int_0^\cdot \phi(A)(u)\phi(A_n)(u,\cdot]d\alpha_n(u) - \int_0^\cdot \phi(A)(u)\phi(A)(u,t]d\alpha(u)\|_\infty$

If α_n or α is replaced with $\tilde{\alpha}$, the error in each integral is bounded by a constant times $\|\alpha_n - \tilde{\alpha}\|_\infty$ or $\|\alpha - \tilde{\alpha}\|_\infty$ respectively, which can be shown with integration by parts. By choosing $\tilde{\alpha}$ of bounded variation close to α (and thus α_n for sufficiently large n) it suffices to show that $\| \int_0^\cdot \phi(A)(u)\phi(A_n)(u,t]d\tilde{\alpha}(u) - \int_0^\cdot \phi(A)(u)\phi(A)(u,t]d\tilde{\alpha}(u)\|_\infty \to 0$. This follows because $\phi(A_n)$ converges uniformly to $\phi(A)$ by the previous theorem. □

Product Integrals and Cumulative Hazards

We first prove the Hadamard differentiability of a very simple map.

Theorem 0.12. *Let $D^\star[a,b]$ denote the subset of functions in $D[0,b]$ that are bounded below by constant $\epsilon > 0$, where ϵ is fixed. Then $\phi : (D[0,b], \|\cdot\|_\infty) \to (D[0,b], \|\cdot\|_\infty)$ is Hadamard differentiable at $B \in D^\star[a,b]$ with Hadamard derivative $d\phi_B(\beta) = -\frac{\beta}{B^2}$.*

proof: Consider a scalar sequence $t_n \to 0$, and $\beta_n \in D[0,b] \to \beta \in D[0,b]$.

Note that for $|b(s) - B(s)| \leq t_n|\beta_n(s)|$ that $|b(s) - B(s)| \leq t_n|\beta_n(s)| \leq t_n\|\beta_n\|_\infty \leq t_n\|\beta_n - \beta\|_\infty + t_n\|\beta\|_\infty \to 0$. For the real-valued function $f(x) = 1/x$, note that $f'(x) = -1/x^2$ is uniformly continuous for $0 < \epsilon \leq x \leq \|B\|_\infty$, so as $B(s)$ is bounded away from $\epsilon > 0$, $f'(b(s)) \to -\frac{1}{B(s)}$ uniformly for $0 \leq s \leq b$.

Then by a first order Taylor expansion of $f(x) = 1/x$ about $B(s)$, $\frac{\phi(B+t_n\beta_n)(s)-\phi(B)}{t_n} + \frac{b}{B^2}(s) = \frac{1}{t_n}(\frac{1}{B(s)+t_n\beta_n(s)} - \frac{1}{B(s)}) + \frac{b(s)}{B^2(s)} = \frac{1}{t_n}(t_n\beta_n(s))f'(b(s)) + \frac{\beta(s)}{B^2(s)} = \beta_n(s)f'(b(s)) + \frac{\beta(s)}{B^2(s)} \to 0$ uniformly for $0 \leq s \leq b$, by above, and the fact that $\|\beta_n - \beta\|_\infty \to 0$. □

We can now prove the Hadamard differentiability of the so-called cumulative hazard.

Theorem 0.13. *Let E denote the space $\{(A,B) : A,B \in D[0,b],\ A \in BV_M,\ B \geq \epsilon > 0\}$. The map $\phi : D[0,b]^2 \to D[0,b]$ given by $\phi(A,B) = \int_0^\cdot \frac{1}{B}dA$ is Hadamard differentiable (using the supremum norm for the domain and range spaces) for $(A,B) \in E$ with Hadamard derivative $d\phi_{(A,B)}(\alpha,\beta) = \int_0^\cdot (1/B)d\alpha - \int_0^\cdot (\beta/B^2)dA$.*

proof: We write the map as $(A,B) \to (A, 1/B) \to \int_0^\cdot \frac{1}{B}dA$. From the previous theorem, the map $(A,B) \to (A,1/B)$ has derivative map at (A,B) given by $(\alpha,\beta) \to (\alpha, -\beta/B^2)$. From the Homework 1 question on the Wilcoxin statistic, $(A,1/B) \to \int_0^\cdot \frac{1}{B}dA$ has derivative map at $(A,1/B)$ given by $(\alpha, -\beta/B^2) \to \int_0^\cdot \frac{1}{B}d\alpha - \int_0^\cdot \frac{\beta}{B^2}dA$. The result now follows by the chain rule. □

We can now consider composing comulative hazard functions and product integrals.

Theorem 0.14. *Let E denote the space $\{(A,B) : A,B \in D[0,b],\ A \in BV_M,\ B \geq \epsilon > 0\}$. Consider the map $\phi : D[0,b]^2 \to D[0,b]$ defined by $(P_1,P_2) \to (\Lambda = \int_0^\cdot \frac{1}{P_2}dP_1) \to \Pi_{0<s\leq\cdot}(1 - d\Lambda(s))$. Then if $\psi(\Lambda)(t) \equiv \Pi_{0<s\leq t}(1 + d\Lambda(s))$ for $\Lambda(t) = \int_0^t \frac{1}{P_2}dP_1)$, ϕ is Hadamard differentiable (using the supremum norm for the domain and range spaces) at $(A,B) \in E$ with Hadamard derivative given by:*
$(\alpha,\beta) \to \int_0^\cdot \psi(\Lambda)(u)\psi(\Lambda)(u,\cdot](\frac{1}{B(u)}d\alpha(u) - \frac{\beta(u)}{B(u)^2}dA(u))$.

proof: This is just the chain rule, applied to the previously given Hadamard derivatives of the cumulative hazard and product integral maps. □

Survival Analysis

The most basic setup in a field of statistics known as *survival analysis* is as follows. We are interested a random variable $0 \leq T$ with cumulative distribution function F and survival function $S = 1 - F$. Here T is often called a *survival time* or *failure time.* We would like to estimate S, but aren't able to observe n i.i.d. copies of T. Instead we observe n i.i.d. copies of $O = (\tilde{T}, \Delta) \sim P$, where $\tilde{T} = \min(T, C)$ and $\Delta = 1(T \leq C)$ for $0 \leq C \sim G \perp F$ a *censoring time.* As usual, P_n denotes the empirical distribution. For simplicity, we will here assume that F and G represent continuous distributions. The most common application is where the survival time T measures the time untill death or recurrence of illness in a medical study, and the censoring time C is the time at which the subject drops out of the study or is unavailable for follow-up. Another common application is in *reliability analysis* where T measures the length of time a product such as a car lasts before breaking down. The assumption that $T \perp C$ (meaning T and C are independent) is fairly strong, and there are many examples where it is violated, but it can be slightly weakened to a so called *coarsening at random* assumption discussed in van der Laan and Robins. But without coarsening at random, essentially nothing can be said about the distribution of T.

Identifiability and Estimation in Survival Analysis

We first note that if censoring occurs before some time b, then it should be very difficult to estimate the survival function after b, because we have no data on T conditional on $T > b$. We therefore consider some b such that $P(\tilde{T} > b) > 0$, and will consider estimating the surival function S on $[0, b]$.

Let $\mathcal{F}_1 = \{f(O) = 1(\tilde{T} \leq u, \Delta = 1) : 0 \leq u \leq b\}$.
Let $\mathcal{F}_2 = \{f(O) = 1(\tilde{T} \geq u) : 0 \leq u \leq b)\}$.
Let $\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2$.

Then for $P_1(t) = P(\tilde{T} \leq t, \Delta = 1)$ and $P_2(t) = P(\tilde{T} \geq t)$, the map $P \rightarrow (P_1, P_2)$ maps $l^\infty(\mathcal{F})$ (with the norm $\| \cdot \|_{\mathcal{F}}$) to the bivariate space on cadlag functions $D[0, b]^2$ (with the supreumum norm). Because this map is linear and continuous, it is equal to its own Hadamard derivative.

Because we've assumed that $P(\tilde{T} > b) > 0$, P_2 is bounded away from O on $[0, b]$. From our previous notes, this implies the map $(P_1, P_2) \rightarrow \int_0^\cdot \frac{1}{P_2(u)} dP_1(u)$ from $D[0, b]^2$ to $D[0, b]$ is Hadamard differentiable (using the supremum norm in the domain and range spaces) at (P_1, P_2) with derivative map $(\alpha, \beta) \rightarrow \int_0^\cdot \frac{1}{B} d\alpha - \int_0^\cdot \frac{\beta}{B^2} dP_1$. However, the independence of T and C implies that $\int_0^\cdot \frac{1}{P_2(u)} dP_1(u) = \int_0^\cdot \frac{1}{1 - F(u_-)} dF(u) \equiv \Lambda$, the *cumulative hazard function.* For $\phi : P \rightarrow (P_1, P_2) \rightarrow \int_0^\cdot \frac{1}{P_2(u)} dP_1(u)$ (mapping $l^\infty(\mathcal{F})$ to $D[0, b]$), this suggests estimating $\Lambda(\cdot) = \phi(P)(\cdot)$ with $\Lambda_n(\cdot) \equiv \phi(P_n)(\cdot)$, and this is called the *Nelson-Aalen estimator.*

For n observations, let $t_1, ..., t_m$ denote the ordered times at which failures occur (so $(\tilde{T} = t_i, \Delta = 1)$, and let n_i denote the number of observations still at risk of failing at time t_i (so $n_i = \sum_{j=1}^{n} 1(\tilde{T} \geq t_i)$. Then the Nelson-Aalen estimator can be written as $\Lambda_n(t) = \sum_{\{i:t_i \leq t\}} \frac{d_i}{n_i}$.

Theorem 0.15. *The Nelson-Aalen estimator $\Lambda_n(t)$ is an asymptotically linear estimator of $\Lambda(t)$ for $0 \leq t \leq b$ with influence curve given by:*
$IC(O|P) = \int_0^t \frac{1}{P_2(u)} d(1(\tilde{T} \leq u, \Delta = 1) - P_1(u)) - \int_0^t \frac{1(\tilde{T} \geq u) - P_2(u))}{P_2(u)^2} dP_1(u)$.

proof: This follows from applying the functional delta method to $\phi : P \rightarrow (P_1, P_2) \rightarrow \int_0^t \frac{1}{P_2(u)} dP_1(u)$ (mapping $(l^\infty(\mathcal{F}), \| \cdot \|_{\mathcal{F}})$ to $\mathcal{R}$), as it can be shown that $\mathcal{F}$ is a Donsker class. Technically we use the chain rule, but the mapping $P \rightarrow (P_1, P_2)$ from $l^\infty(\mathcal{F})$ to $D[0, b]^2$ is equal to its own derivative. Consider $G \in l^\infty(\mathcal{F})$, and the functions $f_{1,u} = 1(\tilde{T} \leq t, \Delta = 1) \in \mathcal{F}_1$ and $f_{2,u} = 1(\tilde{T} \geq u) \in \mathcal{F}_2$. The previously given Hadamard differentiability of the cumulative hazard map tells us that the Hadamard derivative of ϕ at P, denoted by $d\phi_P : l^\infty(\mathcal{F}) \rightarrow \mathcal{R}$, is given by $d\phi_P(G) = \int_0^t \frac{1}{P_2(u))} dG(f_{1,u}) - \int_0^t \frac{G(f_{2,u})}{B^2(u)} dP_1(u)$. The desired result follows from the functional delta method, recalling the influence curve is $IC(O|P) = d\phi_P(G_{1,O})$, where $G_{1,0}$ is the empirical process for the single observation O, so $G_{1,O}(f_{1,u}) = 1(\tilde{T} \leq u, \Delta = 1) - P_1(u)$ and $G_{1,O}(f_{2,u}) = 1(\tilde{T} \geq u) - P_2(u)$. $\square$

From the definition of Λ, $1 - \int_0^{\cdot} \frac{1}{1 - F(u_-)} d\Lambda(u) = S(\cdot)$. From the representation of the product integral in the previous notes as the unique solution of the Volterra equation, this implies that $S(\cdot) = \Pi_{0 \leq u \leq \cdot}(1 - d\Lambda(u))$, where Π here denotes the product integral. As $\Lambda \in D[0, b]$ was shown to be identifiable from P through (P_1, P_2), this shows that $S \in D[0, b]$ is also identifiable from P. So for $\psi : P \rightarrow (P_1, P_2) \rightarrow\rightarrow \Pi_{0 \leq u \leq \cdot}(1 - d\Lambda(u))$ (mapping $l^\infty(\mathcal{F})$ to $D[0, b]$), $S(\cdot) = \psi(P)(\cdot)$. This suggests estimating $S(\cdot)$ with $S_n(\cdot) \equiv \psi(P_n)(\cdot)$, called the *Kaplan-Meier estimator*. Using the previous notation, the Kapalan-Meier estimator can be written as $S_n(t) = \Pi_{\{i:t_i \leq t\}}(1 - \frac{d_i}{n_i})$.

Theorem 0.16. *The Kaplan-Meier estimator $S_n(t)$ is an asymptotically linear estimator of $S(t)$ for $0 \leq t \leq b$ with influence curve given by:*
$IC(O|P) = -S(t)[\int_0^t (\frac{1}{P_2(s)} d(1(\tilde{T} \leq s, \Delta = 1) - P_1(s)) - \int_0^t \frac{1(\tilde{T} \geq s)}{P_2^2(s)} dP_1(s)]$.
This simplfies to $IC(O|P) = -S(t)[\frac{1(\tilde{T} \leq t, \Delta = 1)}{P_2(\tilde{T})} - \int_0^{\min(\tilde{T}, t)} \frac{1}{P_2^2(s)} dP_1(s)]$.

proof: We again apply the functional delta method to the map $\psi : P \rightarrow (P_1, P_2) \rightarrow \Lambda \rightarrow \Pi_{0 \leq s \leq t}(1 - d\Lambda(s)) = S(t)$ from $(l^\infty(\mathcal{F}), \| \cdot \|_{\mathcal{F}})$ to $\mathcal{R}$, as it can be shown that $\mathcal{F}$ is a Donsker class. In our use of the chain rule, the mapping $P \rightarrow (P_1, P_2)$ from $l^\infty(\mathcal{F})$ to $D[0, b]^2$ is equal to its own derivative. Consider $G \in l^\infty(\mathcal{F})$, and the functions $f_{1,s} = 1(\tilde{T} \leq s, \Delta = 1) \in \mathcal{F}_1$ and $f_{2,s} = 1(\tilde{T} \geq s) \in \mathcal{F}_2$. The previously given Hadamard differentiability of the composition of the cumulative hazard function and the product integral gives that ψ is Hadamard differentiable at P, with Hadamard derivative denoted by $d\psi_P : l^\infty(\mathcal{F}) \rightarrow \mathcal{R}$, given by:

$d\psi_P(G) = -\int_0^t \Pi_{0 \leq u < s}(1 - d\Lambda(u)) \Pi_{s < t}(1 - d\Lambda(u)) (\frac{1}{P_2(s)} dG(f_{1,s}) - \frac{G(f_{2,s})}{P_2^2(s)} dP_1(s))$

$= -S(t)\int_0^t(\frac{1}{P_2(s)}dG(f_{1,s}) - \frac{G(f_{2,s})}{P_2^2(s)}dP_1(s))$.

Note that the $-$ enters trivially from the chain rule because we are applying the product integral map to $-\Lambda$. The $S(t)$ comes from the fact that $\Pi_{0\leq u<s}(1-d\Lambda(u))\Pi_{s<t}(1-d\Lambda(u)) = \Pi_{0\leq u\leq t}(1-d\Lambda(u)) = S(t)$ because we have assumed that S is a continuous distribution. The desired result follows from the functional delta method, recalling the influence curve is $IC(O|P) = d\phi_P(G_{1,O})$, where $G_{1,0}$ is the empirical process for the single observation O, so $G_{1,O}(f_{1,s}) = 1(\tilde{T} \leq s, \Delta = 1) - P_1(s)$ and $G_{1,O}(f_{2,s}) = 1(\tilde{T} \geq s) - P_2(s)$. $\Box$

Lecture of February 22, 2005

Functional δ-method for analyzing Z-estimators

The general methodology for analyzing Z-estimators discussed in the last lecture requires that the map $\theta \longrightarrow U(\theta, P)$ be Frechét differentiable. This requirement is easily met if θ is finite-dimensional, but may be difficult to establish in the infinite-dimensional case. We will now discuss another approach for analyzing the asympotic behavior of Z-estimators that is based on the functional δ-method and that does not require the map $\theta \longrightarrow U(\theta, P)$ to be Frechét differentiable.

Suppose we observe n i.i.d. copies $O_1, ..., O_n$ of $O \sim P$. Consider the parametef of interest $\theta(P) \in (D_1, \| \|_1)$. Let $P \in (D_2, \| \|_2)$, e.g. $(D_2, \| \|_2) = (l^\infty(\mathcal{F}), \| \|_\mathcal{F})$ for a sufficiently rich class of functions $\mathcal{F}$. Suppose there exists a mapping $U : (D_1, \| \|_1) \times (D_2, \| \|_2) \longrightarrow (D_3, \| \|_3)$ such that $\theta(P)$ can be defined as the solution of $U(\theta, P) = 0$. Typically, we will have that $(D_3, \| \|_3) = (D_1, \| \|_1)$, although this is not required. Let $\varphi : (D_2, \| \|_2) \longrightarrow (D_1, \| \|_1)$ be the mapping that maps P into the solution θ of the equation $U(\theta, P) = 0$. Now let the estimator θ_n be defined as $\theta_n = \varphi(P_n)$, i.e. let θ_n be the solution of $U(\theta_n, P_n) = 0$.

As before, we want to prove that $\sqrt{n}(\theta_n - \theta) = \sqrt{n}(\varphi(P_n) - \varphi(P)) \overset{D}{\Longrightarrow} \mathbb{Z}$ in $(D_1, \| \|_1, \mathcal{B})$ as $n \rightarrow \infty$. To apply the functional δ-method, we first use empirical process theory to verify that $G_n = \sqrt{n}(P_n - P) \overset{D}{\Longrightarrow} G$ in $(D_2, \| \|_2)$. Next we need to establish that the map φ is Hadamard differentiable tangentially to a subspace $\mathbb{D}_2^* = (D_2^*, \| \|_2) \subseteq (D_2, \| \|_2)$ such that $G \in \mathbb{D}_2^*$, i.e. we need to show that for any sequences $\{h_n = \sqrt{n}(P_n' - P')\}_{n\geq 1}$ with $h_n \rightarrow h$ for some $h \in \mathbb{D}_2^*$ we have that $\| \sqrt{n}(\varphi(P_n') - \varphi(P')) - d\varphi(P')(h) \|_1 \rightarrow 0$. We will lay out a roadmap of five steps, A1-A5, to establish the required differentiability.

Consider the same setup as for the proof from last lecture:

$$\sqrt{n}\,[U(\theta_n', P') - U(\theta', P')] = -\sqrt{n}\,[U(\theta_n', P_n') - U(\theta_n', P')] \qquad (4)$$

A1: Show that $P_n' \rightarrow P' \Longrightarrow \varphi(P_n') \rightarrow \varphi(P')$.

To establish this continuity condition on φ see for example the consistency proof from last lecture.

A2: Show that $\| \sqrt{n}(U(\theta_n', P_n') - U(\theta_n', P')) - \frac{d}{dP'}U(\theta', P')(\sqrt{n}(P_n' - P)) \|_3 \longrightarrow 0$.

Here you can use that, by A1, $\theta_n' \to \theta' = \varphi(P')$. Now A2 implies that the right-hand side and hence also the left-hand side in (1) converges to $-\frac{d}{dP'}U(\theta', P')(h)$.

A3: Show that there exists a linear mapping $df_n : (D_1, \| \|_1) \longrightarrow (D_3, \| \|_3)$, possibly depending on θ_n', θ', and P, such that $U(\theta_n', P') - U(\theta', P') = df_n(\theta_n - \theta)$.

Write $U(\theta_n', P') - U(\theta', P') = f(\theta_n') - f(\theta')$. Consider the example

$$f(\theta_n') = \int \frac{1}{\theta_n'} h(P)dP, \qquad f(\theta') = \int \frac{1}{\theta'} h(P)dP$$

Then

$$f(\theta_n') - f(\theta') = \int \left(\frac{1}{\theta_n'} - \frac{1}{\theta'} \right) h(P)dP = \int \frac{\theta_n' - \theta'}{\theta_n'\theta'} h(P)dP = df_n(\theta_n' - \theta')$$

where $df_n(\theta_n' - \theta')$ depends on θ_n', θ', and P, but is linear in $\theta_n' - \theta'$.

Note that this step does not require any Frechét differentiability of the map $\theta \longrightarrow U(\theta, P)$. Now A3 implies that $df_n(\sqrt{n}(\theta_n' - \theta')) \to -\frac{d}{dP'}U(\theta', P')(h)$, or equivalently that $df_n(\sqrt{n}(\theta_n' - \theta')) + \frac{d}{dP'}U(\theta', P')(h) \to 0$

A4: Show that df_n is 1-1 and onto for all n and that $\limsup_n \| df_n^{-1} \| < \infty$.

This type of continuous differentiability of df_n and the linearity of df_n^{-1} imply that

$$df_n^{-1}\left(df_n(\sqrt{n}(\theta_n' - \theta')) + \frac{d}{dP'}U(\theta', P')(h) \right) = \sqrt{n}(\theta_n' - \theta') + df_n^{-1}\left(\frac{d}{dP'}U(\theta', P')(h) \right) \longrightarrow 0$$

A5: Show that, for any g, $\theta_n' \to \theta' \implies df_n^{-1} - df^{-1}(g) \to 0$, where $df^{-1}(g) = \left[\frac{d}{d\theta}U(\theta', P')\right]^{-1}(g)$.

Now A5 implies that

$$\sqrt{n}(\varphi(P_n)-\varphi(P)) \overset{D}{\Longrightarrow} -\left[\frac{d}{d\theta}U(\theta,P)\right]^{-1}\left(\frac{d}{dP}U(\theta,P)(\sqrt{n}(P_n-P))\right) \equiv d\varphi(P)(\sqrt{n}(P_n-P))$$

■

Homework 3 Solution, prepared by Dan Rubin

The Exponential with Censoring

We can write the likelihood for one observation as:
$L(\lambda) = [f(1-G)(\tilde{T})]^{\Delta}[SdG(\tilde{T}]^{1-\Delta} = [\lambda\exp(-\lambda\tilde{T})(1-G(\tilde{T}))]^{\Delta}[\exp(-\lambda\tilde{T})dG(\tilde{T})]^{1-\Delta}$
$l(\lambda) = \log L(\lambda) = \Delta[\log(\lambda) - \lambda\tilde{T}] - (1-\Delta)\lambda\tilde{T} + C$, where C does not depend on λ.
$U(\lambda) = \frac{d}{d\lambda}l(\lambda) = \Delta/\lambda - \tilde{T}$.
Since scores have mean zero, this implies $\lambda = E\Delta/E\tilde{T}$
Setting the score for n observations to zero, we see the mle is $\lambda_n = \overline{\Delta}_n/\overline{\tilde{T}}_n$, where $\overline{\Delta}_n = \frac{1}{n}\sum_{i=1}^n \Delta_i$, $\overline{\tilde{T}}_n = \frac{1}{n}\sum_{i=1}^n \tilde{T}_i$, and it is the mle because U is strictly decreasing in λ, so the log-likelihood is strictly concave.

So for $f(x,y) = x/y$, with continuous gradient $[1/y, -x/y^2]$, a first-order Taylor expansion about $(E\Delta, E\tilde{T})$, and the fact that $\|(\overline{\Delta}_n, \overline{\tilde{T}}_n) - (E\Delta, E\tilde{T})\| \to 0$ in probability by LLN, $\sqrt{n}(\lambda_n - \lambda) = \sqrt{n}(f(\overline{\Delta}_n, \overline{\tilde{T}}_n) - f(E\Delta, E\tilde{T}))$
$= \sqrt{n}([\overline{\Delta}_n - E\Delta](1/E\tilde{T} + o_p(1)) + [\overline{\tilde{T}}_n - E\tilde{T}](-E\Delta/(E\tilde{T})^2 + o_p(1)))$.
As $\sqrt{n}(\overline{\Delta}_n - E\Delta)$, $\sqrt{n}(\overline{\tilde{T}}_n - E\tilde{T}) = O_p(1)$ by the CLT, this linearlization gives the influence curve $IC(O|P) = (\Delta - E\Delta)/E\tilde{T} - (E\Delta)(\tilde{T} - E\tilde{T})/(E\tilde{T})^2$.

ML Consistency

$\int \log\frac{f_\theta}{f_{\theta_0}} f_{\theta_0}dx \le \log\int \frac{f_\theta}{f_{\theta_0}} f_{\theta_0}dx = \log(1) = 0$ with equality iff $f_\theta = f_{\theta_0}$ by Jensen's inequality, as $\log(\cdot)$ is strictly concave on $(0,\infty)$. By identifiability, this implies $KL(\cdot)$ is uniquely maximized at θ_0.

As θ_n is the mle, $P_n \log f_{\theta_n} \ge P_n \log f_{\theta_0}$.
We can rewrite this as $P\log f_{\theta_n} + \frac{1}{\sqrt{n}}G_n \log f_{\theta_n} \ge P\log f_{\theta_0} + G_n \log f_{\theta_0}$, for $G_n = \sqrt{n}(P_n - P)$.
$KL(\theta_0) = P\log f_{\theta_0} \le P\log f_{\theta_n} \le P\log f_{\theta_0} - \frac{2}{\sqrt{n}}\sup_{\theta\in\Theta}|G_n \log f_\theta| \to P\log f_{\theta_0} = KL(\theta_0)$ in probability, by the given Glivenko-Cantelli condition, implying that $KL(\theta_n)$ converges to $KL(\theta_0)$ in probability.

As $\{\theta \in \Theta : |\theta - \theta_0| \ge \epsilon\}$ is compact (check it is closed and bounded if Θ is), KL takes on its maximum on the set, which is less than $KL(\theta_0)$ by identifiability, so there is a $\delta(\epsilon) > 0$ such that $P(|\theta_n - \theta_0| \ge \epsilon) \le P(KL(\theta_0) - KL(\theta_n) > \delta(\epsilon)) \to 0$.

Bracketing and Covering Numbers

Definition: The *covering number* $N(\epsilon, \mathcal{F}, \|\cdot\|)$ of a class of functions $\mathcal{F}$ is the minimum number of balls $\{g : \|g - f\| \leq \epsilon\}$ such that the union of the balls contians $\mathcal{F}$. The *entropy* is defined as the logarithm of the covering number.

Note that the covering number increases as epsilon decreases, and that it depends on what norm is chosen.

Definition: Given two functions l, u, define $[l, u] = \{f : u \leq f \leq l\}$. The *bracketing number* $N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of brackets $[l, u]$ with $\|u - l\| \leq \epsilon$ needed such that the union contains $\mathcal{F}$. The logarithm of the bracketing number is called the *bracketing entropy.*

Theorem 0.17. *If the norm* $\|\cdot\|$ *is such that* $|f| \leq |f|$ *implies* $\|f\| \leq \|g\|$*, then* $N(\epsilon, \mathcal{F}, \|\cdot\|) \leq N_{[]}(2\epsilon, \mathcal{F}, \|\cdot\|)$.

proof: For such norms, if f is in the 2ϵ bracket $[l, u]$ then it is the ϵ-ball centered at $(l + u)/2$. □

Unfortunately, the above theorem has no converse allowing us to bound bracketing numbers from given covering numbers. This means that a good bracketing result is much stronger than a good covering number result.

Definition $F(o) \equiv \sup_{f \in \mathcal{F}} |f(o)|$ is the *envelope* for $\mathcal{F}$. In general, we will need $PF^2 < \infty$ to establish that $\mathcal{F}$ is a P-Donsker class. The $L_r(Q)$ norm is defined by $\|f\|_{Q,r} = (\int f^r dQ)^{1/r}$. the *uniform entropy number* (relative to $L_r(\cdot)$) is given by $\sup_Q N(\epsilon \|F\|_{Q,r}, \mathcal{F}, \|\cdot\|_{Q,r})$, where the supremum is taken over all possible probability distributions.

Theorem 0.18. $\mathcal{F}$ *is* P*-Glivenko-Cantelli if* $N_{[]}(\epsilon, \mathcal{F}, L_1(P)) < \infty$ *for every* $\epsilon > 0$.

Theorem 0.19. F *is* P*-Glivenko-Cantelli if* $PF < \infty$ *and* $\sup_{Q: QF^r < \infty} N(\epsilon \|F\|_{Q,1}, \mathcal{F}, L_1(Q)) < \infty$ *for every* $\epsilon > 0$.

Theorem 0.20. $\mathcal{F}$ *is* P*-Donsker if* $\int_0^\infty \sqrt{\log N_{[]}(\epsilon, \mathcal{F}, L_2(P)} d\epsilon < \infty$.

Theorem 0.21. *If* $\mathcal{F}$ *is such that* $\int_0^\infty \sup_Q \sqrt{\log N(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)} d\epsilon < \infty$*, then* $\mathcal{F}$ *is* P*-Donsker for every* P *such that* $PF^2 < \infty$*. This integral condition holds if* $sup_Q \log N(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \leq K(1/\epsilon)^{2-\delta}$*, for some* $\delta > 0$.

Note that in the last two theorems, we only have to worry about how the integral behaves for small ϵ, because the bracketing and covering numbers increase as ϵ decreases, and for Glivenko-Cantelli and Donsker classes these numbers will be one for sufficiently large ϵ. Also note that the bracketing number conditions are much weaker than the covering number conditions, and this is because it is harder to find a good bracketing number than a good covering number.

Some Examples of Donsker classes

The set of functions with uniformly bounded derivatives is a Donsker class. The class of all monotone functions $\{f : 0 \leq f \leq F\}$ is P-Donsker provided P-Donsker provided that $PF^2 < \infty$. The set of indicators of compact, convext subsets of a fixed bounded subset of $\mathcal{R}^d$ is Donsker for $d \geq 2$.

Permance Properties of Donsker classes

If $\mathcal{F}$ is Donsker and $\mathcal{G} \subset \mathcal{F}$, then $\mathcal{G}$ is Donsker. If $\mathcal{F}$ and $\mathcal{G}$ are Donsker, then so are $c_1\mathcal{F} + c_2\mathcal{G}$ (for scalars c_1, c_2), $\mathcal{F} \cup \mathcal{G}$, $\mathcal{F} \cap \mathcal{G}$, the closure of $\mathcal{F}$ (set of functions that are limit points both pointwise and in $L_2(P)$), and the set of convex combinations of functions in $\mathcal{F}$. Also, if $\mathcal{F}$ is Donsker with $PF < \infty$ and $1/f \geq \delta > 0$ for $f \in \mathcal{F}$, then $1/\mathcal{F} = \{1/f : f \in \mathcal{F}\}$ is Donsker. See section 2.10 of van der Vaart and Wellner for more examples.

1 Estimating Functions

In this section, we will first define the orthogonal complement of the nuisance tangent space and review efficiency theory. Subsequently, we link this orthogonal complement of the nuisance tangent space to the construction of estimating functions that are elements of this space when evaluated at the true parameter values and show that the expectation of such estimating functions has a derivative zero w.r.t. to fluctuations of a variation-independent nuisance parameter ρ. To avoid additional notation, in our presentation we use the introduced notation for the full data model, but obviously X now represents an observed data random variable so that our presentation applies, in particular, to our censored data model. At the end of this section we provide a representation of the orthogonal complement of the nuisance tangent space in the censored data model in terms of the orthogonal complement of the nuisance tangent space in the full data model.

1.1 Orthogonal complement of a nuisance tangent space

Consider a full data structure model $\mathcal{M}^F$ for the full data distribution F_X. Given F_X, for each g ranging over an index set, let $\epsilon \to F_{\epsilon,g}$ be a one-dimensional submodel of $\mathcal{M}^F$ with parameter $\epsilon \in (-\delta, \delta)$, for some small $\delta > 0$ (δ can depend on g), crossing $F_X = F_{0,g}$ at $\epsilon = 0$, and score $s(X) \in L_0^2(F_X)$, where $L_0^2(F_X)$ is the Hilbert space of functions of X with expectation zero and finite variance endowed with inner product $\langle h_1, h_2 \rangle_{F_X} = \int h_1(x)h_2(x)dF_X(x)$. Here, we define the score h as an $L_0^2(F_X)$ limit:

$$\lim_{\epsilon \to 0} \int \left\{ s(x) - \frac{1}{\epsilon} \frac{dF_{\epsilon,g} - dF_X}{dF_X}(x) \right\}^2 dF_X(x) = 0.$$

One can also define the score pointwise as

$$s(X) = s(g)(X) = \frac{d}{d\epsilon} \log \left(\frac{dF_{\epsilon,g}}{dF_X}(X) \right) \Big|_{\epsilon=0} \in L_0^2(F_X).$$

A typical choice of submodel is of the form $dF_{\epsilon,g}(x) = (1+\epsilon g(x))dF_X(x) + o(\epsilon)$ so that $s(g) = g$. Let $\mathcal{S} \subset L_0^2(F_X)$ be the set of scores corresponding with the class $\{F_{\cdot,g} : g\}$ of one-dimensional submodels. Let $T^F(F_X) \subset L_0^2(F_X)$ be the closure of the linear space spanned by $\mathcal{S}$. We refer to this Hilbert space $T^F(F_X)$ as the tangent space of the full data model. It is crucial that one chooses a rich class $\{F_{\cdot,g} : g\}$ of models that locally cover all possible score directions that the model $\mathcal{M}^F$ allows.

Example 1.1. For example, let $\mathcal{M} = \{f_{\mu,\sigma^2} : \mu, \sigma^2\}$ be the family of normal distributions. For each (δ_1, δ_2) we can define a one-dimensional submodel $\{f_{\mu+\epsilon\delta_1,\sigma^2+\epsilon\delta_2} : \epsilon\}$ that has score $\delta_1 S_1(X \mid \mu, \sigma^2) + \delta_2 S_2(X \mid \mu, \sigma^2)$, where S_1, S_2 are the scores for μ (i.e., $d/d\mu \log(f_{\mu,\sigma^2}(X))$) and σ^2, respectively. Thus, the tangent space is the two-dimensional space $\langle S_1, S_2 \rangle$ spanned by these two scores. □

Nuisance tangent space. Let $\mu = \mu(F_X) \in \mathbb{R}^k$ be a Euclidean parameter of interest. We will now define the so-called nuisance tangent space. Since only the score of $F_{\epsilon,g}$ is relevant for the definition of tangent spaces and the efficiency bound, from now on we will index the one-dimensional submodels by their score s, thereby making clear that two different one-dimensional submodels with the same score are only counted as one. In a full data model $\mathcal{M}^F = \{F_{\mu,\eta} : \mu, \eta\}$ with μ and η independently varying parameters over certain parameter spaces, one can directly determine the nuisance tangent space $T^F_{nuis}(F_X)$ as the space generated by all scores of one-dimensional submodels F_{μ,η_ϵ} just varying the nuisance parameter. In general, we define the nuisance tangent space as follows.

Definition 1.1. *Suppose that for each submodel $\{F_{\epsilon,s} : \epsilon\}$ with score s, $s \in \mathcal{S}$, $d/d\epsilon \mu(F_{\epsilon,s})|_{\epsilon=0}$ exists. The nuisance scores are given by the scores of the models $F_{\epsilon,s}$ for which μ does not locally vary:*

$$\left\{ s \in \mathcal{S} : \frac{d}{d\epsilon}\, \mu(F_{\epsilon,s})\big|_{\epsilon=0} = 0 \right\}.$$

The nuisance tangent space $T_{nuis}(F_X)$ is now the closure (in $L_0^2(F_X)$) of the linear space generated by these nuisance scores:

$$T^F_{nuis}(F_X) \equiv \overline{\left\{ s \in \mathcal{S} : \frac{d}{d\epsilon}\, \mu(F_{\epsilon,s})\big|_{\epsilon=0} = 0 \right\}}.$$

Example 1.2. Suppose that $X \sim F_X$ is a univariate real-valued variable and that the model for F_X is nonparametric. Then, we can choose as the class of one-dimensional submodels $dF_{\epsilon,s}(x) = (1+\epsilon s(x))dF_X(x)$ with $s \in \mathcal{S} = \{s \in L_0^2(F_X) : s \text{ uniformly bounded}\}$. It follows immediately that the tangent space is saturated: $T^F(F_X) = L_0^2(F_X)$.

Suppose that for a given t_0, $\mu = F(t_0)$ is the parameter of interest. We have

$$\mu(F_{\epsilon,s}) - \mu = \int \left\{ I_{(0,t_0]}(x) - \mu \right\} s(x) dF_X(x).$$

This shows that

$$T^F_{nuis}(F_X) = \{s \in L_0^2(F_X) : \langle s, D_{eff} \rangle_{F_X} = 0\},$$

where $D_{eff}(x) \equiv I_{(0,t_0]}(x) - \mu \in L_0^2(F_X)$. □

Pathwise derivative and gradients. Throughout this book, it is assumed that μ is pathwise differentiable along each of the one-dimensional submodels in the sense that for each $s \in \mathcal{S}$

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon} \left(\mu(F_{\epsilon,s}) - \mu(F_X) \right) = \langle \ell(\cdot \mid F_X), s \rangle_{F_X}$$

for an element $\ell(\cdot \mid F_X) \in L_0^2(F_X)^k$. Note that the right-hand side is a k-dimensional vector. For a k-dimensional function $\ell \in L_0^2(F_X)^k$ and $s \in L_0^2(F_X)$, we define the vector inner product $\langle \ell, s \rangle_{F_X}$ as the vector with jth component $\langle \ell_j, s \rangle_{F_X}$. Similarly, we will define a projection $\Pi(s \mid T^F(F_X)))$ of a k-dimensional function $s = (s_1, \ldots, s_k)$ onto a subspace (say) $T^F(F_X)$ of $L_0^2(F_X)$ componentwise as $\Pi(s_j \mid T^F(F_X))_{j=1}^k$. Any such element $\ell(\cdot \mid F_X) \in L_0^2(F_X)^k$ is called a gradient of the pathwise derivative, and the unique gradient $S_{eff}^{F*}(\cdot \mid F_X) \in T^F(F_X)^k$ (i.e., the gradient whose components are elements of the full data tangent space) is called the canonical gradient or efficient influence curve. Notice that $S^*_{eff,j} = \Pi(\ell_j \mid T^F(F_X))$ is the $T^F(F_X)$-component of any gradient component ℓ_j (i.e., the pathwise derivative w.r.t. a class of submodels with tangent space $T^F(F_X)$ only uniquely determines the $T^F(F_X)$-component of the gradient-components of the pathwise derivative). Thus, the set of all gradients is given by

$$\left\{ \ell \in L_0^2(F_X)^k : \langle \ell_j, s \rangle_{F_X} = \langle S_{eff,j}^{F*}(X \mid F_X, \mu), s \rangle_{F_X}, s \in T^F(F_X), \forall j \right\}, \tag{5}$$

where $j \in \{1, \ldots, k\}$. Note that if the full data model is nonparametric, then $T^F(F_X) = L_0^2(F_X)$, so $T_{nuis}^{F,\perp}(F_X) = \langle S_{eff}^{F*}(\cdot \mid F_X) \rangle$ is the k-dimensional space spanned by the components of the canonical gradient and the only gradient is the canonical gradient $S_{eff}^{F*}(\cdot \mid F_X)$. Note that for a vector function $a \in L_0^2(F_X)^k$, we define $< a >= \{c^\top a : c \in \mathbb{R}^k\}$ as the k-dimensional space spanned by the components of a. In general, the larger the model, the smaller the set of gradients.

Nuisance tangent space in terms of the canonical gradient. Under the pathwise differentiability condition, the nuisance tangent space is given by

$$T_{nuis}^F(F_X) = \{ s \in T^F(F_X) : s \perp S_{eff}^{F*}(\cdot \mid F_X) \}$$

and the tangent space equals

$$T^F(F_X) =< S_{eff}^{*F}(\cdot \mid F_X) > \oplus T_{nuis}^F(F_X). \tag{6}$$

Let $T_{nuis}^{F\perp}(F_X)$ be the orthogonal complement of the nuisance tangent space $T_{nuis}^F(F_X)$ in the Hilbert space $L_0^2(F_X)$. Let $\Pi(\cdot \mid T_{nuis}^F(F_X))$ be the projection operator onto the nuisance tangent space. We have $\Pi(s \mid T_{nuis}^{F,\perp}(F_X)) = s - \Pi(s \mid T_{nuis}^F(F_X))$ and

$$T_{nuis}^{F,\perp} = \{ s(X) - \pi(s \mid T_{nuis}^F) : s \in L_0^2(F_X) \}.$$

Alternatively, if $\Pi_{F_X} : L_0^2(F_X) \to T^F(F_X)$ is the projection operator onto $T^F(F_X)$, then it also follows that

$$T_{nuis}^{F,\perp}(F_X) = \{ D \in L_0^2(F_X) : \Pi_{F_X}(D \mid T^F(F_X)) \in < S_{eff}^{*F}(\cdot \mid F_X) > \}. \tag{7}$$

Example 1.3. Let us continue Example 1.2. Notice that μ is indeed pathwise differentiable with canonical gradient $S_{eff}^{F*}(X) = I_{(0,t_0]}(X) - \mu$. The orthogonal complement of the nuisance tangent space is thus $\langle I_{(0,t_0]}(X) - \mu \rangle$. □

Equivalence between gradients and orthogonal complement of nuisance tangent space. By (5), another characterization of a gradient is that each of its components is an element of $T^{F,\perp}_{nuis}$ whose projection onto $T^F(F_X) = T^F_{nuis} \oplus \langle S^{F*}_{eff} \rangle$ equals the corresponding component of S^{F*}_{eff}. Thus, gradients are orthogonal to T^F_{nuis} and need to be appropriately standardized. Since the projection of $D \in L^2_0(F_X)$ onto $\langle S^{F*}_{eff} \rangle$ is given by

$$\Pi(D \mid \langle S^{F*}_{eff} \rangle) = E(D S^{F*\top}_{eff}) E(S^{F*}_{eff} S^{F*\top}_{eff})^{-1} S^{F*}_{eff},$$

it follows that the set of gradients $T^{F,\perp,*}_{nuis}(F_X)$ is given by the following standardized versions of $T^{F,\perp}_{nuis}$:

$$\left\{ E(S^{F*}_{eff}(X) S^{F*\top}_{eff}(X)) \left\{ E(D(X) S^{F*\top}_{eff}(X \mid F_X)) \right\}^{-1} D : D \in T^{F,\perp}_{nuis}(F_X)^k \right\}.$$

This shows that the space spanned by the components of all of the gradients (5) equals the orthogonal complement $T^{F,\perp}_{nuis}$ of the nuisance tangent space.

If $D(X)$ plays the role of an estimating function, then the standardization matrix in front of D actually reduces to the much simpler derivative standardization provided in (**??**), as we will now show.

Lemma 1.1. *Suppose that there exists a mapping (i.e., estimating function) $(h, \mu, \rho) \rightarrow D_h(\cdot \mid \mu, \rho)$ on $\mathcal{H}^F \times \{(\mu(F_X), \rho(F_X)) : F_X \in \mathcal{M}^F\}$ into functions of X such that one can represent*

$$T^{F,\perp}_{nuis}(F_X) = \{D_h(\cdot \mid \mu(F_X), \rho(F_X)) : h \in \mathcal{H}^F(F_X)\} \qquad (8)$$

as the range of an index set $\mathcal{H}^F(F_X) \subset \mathcal{H}^F$ of $(h \rightarrow D_h(\cdot \mid \mu(F_X), \rho(F_X))$ for all $F_X \in \mathcal{M}^F$. In addition, assume that for all $h \in \mathcal{H}(F_X)$ and each one-dimensional submodel $F_{\epsilon,s}$, $s \in \mathcal{S}$, we have for $\epsilon \rightarrow 0$ $\| D_h(\cdot \mid \mu(F_{\epsilon,s}), \rho(F_{\epsilon,s})) - D_h(\cdot \mid \mu(F_X), \rho(F_X)) \|_{F_X} \rightarrow 0$. Assume that μ is a pathwise differentiable parameter at F_X with canonical gradient $S^{F*}_{eff}(\cdot \mid F_X)$ with $\langle S^{F*}_{eff} \rangle \subset \mathcal{S}$.

Let

$$f_h(s) \equiv \frac{d}{d\epsilon} E_{F_X} D_h(X \mid \mu(F_{\epsilon,s}), \rho(F_{\epsilon,s})) \Big|_{\epsilon=0}.$$

We have that an element $D = D_h(\cdot \mid \mu(F_X), \rho(F_X)) \in T^{F,\perp}_{nuis}(F_X)^k$ for $h \in \mathcal{H}(F_X)^k$ is a gradient if and only if

$$f_h(s) = \begin{cases} 0 & \textit{if } s \textit{ is nuisance score} \\ -\, d/d\epsilon \mu(F_{\epsilon,s})|_{\epsilon=0} & \textit{if } s \in \langle S^{F*}_{eff} \rangle \end{cases}.$$

Proof. Firstly, by assumption,

$$\begin{aligned} &\frac{1}{\epsilon} E_{F_X} \left\{ D_h(X \mid \mu(F_{\epsilon,s}), \rho(F_{\epsilon,h})) - D_h(X \mid \mu(F_X), \rho(F_X)) \right\} = \\ &\int D_h(x \mid \mu(F_{\epsilon,s}), \rho(F_{\epsilon,s})) \frac{1}{\epsilon} \frac{dF_X - dF_{\epsilon,s}}{dF_X}(x) dF_X(x) \\ &\rightarrow -\langle D_h(\cdot \mid \mu(F_X), \rho(F_X)), s \rangle_{F_X} \text{ if } \epsilon \rightarrow 0. \end{aligned}$$

By definition, D_h is a gradient if and only if for each $s \in T^F_{nuis}(F_X)$ the latter inner product equals zero and for $s \in \langle S^{F*}_{eff} \rangle$ the latter inner product equals $- \left. d/d\epsilon \mu(F_\epsilon)\right|_{\epsilon=0}$, which proves the lemma. □

Under further smoothness (in μ, ρ) conditions on $(\mu, \rho) \rightarrow D_h(\cdot \mid \mu, \rho)$ and under the assumption that μ, ρ are variation-independent parameters, one can now typically show that $D_h(\cdot \mid \mu(F_X), \rho(F_X))$ is a gradient if and only if the Gateaux derivative $ED_g(X \mid \mu(F_X), \rho)$ w.r.t. the nuisance parameter ρ at $\rho = \rho(F_X)$ (in every direction allowed by the model) equals zero and the derivative of $D_h(X \mid \mu, \rho(F_X))$ w.r.t. μ at $\mu(F_X)$ equals minus the identity matrix:

$$\frac{d}{d\mu} E_{F_X} D(X \mid \mu, \rho) \;=\; -I, \tag{9}$$

$$\frac{d}{d\rho} E_{F_X} D(X \mid \mu, \rho) \;=\; 0. \tag{10}$$

Formally, we have the following lemma.

Lemma 1.2. *Make the same assumptions as in the previous lemma. Assume that μ and ρ are variation independent parameters of F_X. Assume that for all $h \in \mathcal{H}^F(F_X)^k$, $\mu \rightarrow E_{F_X} D_h(\cdot \mid \mu, \rho(F_X))$ is differentiable at $\mu(F_X)$ with an invertible derivative matrix. Assume that $E(S^{F*}_{eff}(X) S^{F*\top}_{eff}(X))$ is invertible. If for $h \in \mathcal{H}(F_X)^k$, $D_h(X \mid \mu(F_X, \rho(F_X)) \in T^{F,\perp*}_{nuis}(F_X)$, then*

$$d/d\mu E_{F_X} D_h(X \mid \mu, \rho(F_X)) = -I,$$

where I denotes the $k \times k$ identity matrix. As a consequence, $T^{F,\perp,*}_{nuis}(F_X)$ is given by

$$\left\{ - \left\{ \frac{d}{d\mu} E_{F_X} D_h(X \mid \mu, \rho(F_X)) \Big|_{\mu=\mu(F_X)} \right\}^{-1} D_h : h \in \mathcal{H}^F(F_X)^k \right\}. \tag{11}$$

Proof. Let $s \in \langle S^{F*}_{eff} \rangle$. Note that $\epsilon \rightarrow E_{F_X} D(X \mid \mu(F_{\epsilon,s}), \rho(F_X))$ is a composition $h_1(h_2(\epsilon))$ with $h_2(\epsilon \mid s) = \mu(F_{\epsilon,s})$ and $h_1(\mu) = E_{F_X} D(X \mid \mu, \rho(F_X))$. As in the proof of the previous lemma, we show that

$$\left. d/d\epsilon h_1(h_2(\epsilon)) \right|_{\epsilon=0} = - \left. d/d\epsilon \mu(F_{\epsilon,s}) \right|_{\epsilon=0},$$

where the right-hand side actually equals $-h_2'(0 \mid s)$. By the chain rule, we have that the left-hand side equals $d/d\mu h_1(\mu) * h_2'(0 \mid s)$. Thus, $d/d\mu h_1(\mu) * h_2'(0 \mid s) = h_2'(0 \mid s)$ for all $s \in \langle S^{F,*}_{eff} \rangle$. We now need $h_2'(0 \mid s = S^{F*}_{eff,j})$, $j = 1, \ldots, k$, to be independent vectors. However, the latter vectors are given by $E(S^{F*}_{eff} S^{F*j}_{eff})$, $j = 1, \ldots, k$, so that this independence is a consequence of the assumed invertibility of $E(S^{F*}_{eff}(X) S^{F*\top}_{eff}(X))$. This proves that $d/d\mu h_1(\mu) = -I$. □

Example 1.4. (Parametric model) Consider a parametric model $X \sim f_{\theta,\eta}$, where $\mu = \theta \in \mathbb{R}^k$ is the parameter of interest and $\eta \in \mathbb{R}^m$ is the nuisance parameter. As the class of one-dimensional models, we choose the $k + m$ models just varying one of the parameters: let $a = (\theta, \eta)$. For every $\delta \in \mathbb{R}^{k+m}$, we let $\{f_{a+\epsilon\delta} : \epsilon\}$ be a one-dimensional submodel. The tangent space generated by these one-dimensional models is the $k+m$-dimensional subspace of $L_0^2(F_X)$ spanned by the score components of $h = (h_1, \ldots, h_k)$ of θ and $g = (g_1, \ldots, g_m)$ of η. We can find the orthogonal complement of the nuisance tangent space in two ways. Since the density of X is parametrized naturally with μ and nuisance parameter η, we can calculate the nuisance space directly: $T_{nuis}(F_X) = \langle g_1, \ldots, g_m \rangle$. Since $\Pi(\cdot \mid T_{nuis}(F_X)) : L_0^2(F_X) \to T_{nuis}(F_X)$ is given by

$$\Pi(D \mid T_{nuis}(F_X)) = E(D(X)g(X)^\top)E(g(X)g(X)^\top)^{-1}g(X),$$

we have that the orthogonal complement of the nuisance tangent space is given by

$$T_{nuis}^\perp(F_X) = \{D(X) - E(D(X)g(X)^\top)E(g(X)g(X)^\top)^{-1}g(X) : D \in L_0^2(F_X)\}.$$

Moreover, the efficient score $S_{eff}(X)$ for μ is given by

$$\Pi(h_j \mid T_{nuis}^\perp(F_X))_{j=1}^k = h(X) - E(h(X)g(X)^\top)E(g(X)g(X)^\top)^{-1}g(X), \qquad (12)$$

and the efficient influence curve/canonical gradient is given by the standardized version of the efficient score:

$$S_{eff}^*(X) = E(S_{eff}S_{eff}^\top(X))^{-1}S_{eff}(X).$$

Let us now find $T_{nuis}^\perp$ and the canonical gradient S_{eff}^* in terms of the pathwise derivative. We have

$$\left.\frac{d}{d\epsilon}\mu(f_{a+\epsilon\delta})\right|_{\epsilon=0} = (\delta_1, \ldots, \delta_k)^\top,$$

while the score of $f_{a+\epsilon\delta}$ at $\epsilon = 0$ equals the linear combination $\delta^\top(h,g)^\top$ of the scores and nuisance scores with coefficients given by δ. Thus, the gradients are all functions $\ell \in L_0^2(F_X)^k$ that satisfy

$$(\delta_1, \ldots, \delta_k)^\top = \langle \ell, \delta^\top(h,g)^\top \rangle_{F_X}.$$

To begin with, it follows that $\ell \perp \langle g \rangle$ is orthogonal to the nuisance scores $g_1, \ldots, g_m$. The canonical gradient is the only gradient (and thus orthogonal to $g_1, \ldots, g_m$), which is also an element of the tangent space. It follows that the canonical gradient equals S_{eff}^* defined above. □

1.2 Review of efficiency theory

The orthogonal complement of the nuisance tangent space forms the basis of the general estimating function approach presented in this book. It is also a space that is fundamental to efficiency theory. We will now review some of these fundamental results (Bickel, Klaassen, Ritov, and Wellner, 1993). Firstly, we need to recall that an estimator μ_n of μ is called regular relative to the given class of one-dimensional submodels $\{F_{\epsilon,s} : s \in \mathcal{S}\}$ if for $\epsilon = 1/\sqrt{n}$ the distribution of $\sqrt{n}(\mu_n - \mu(F_{\epsilon,s}))$, under $X_i \sim F_{\epsilon,s}$, $i = 1, \ldots, n$, converges to a limit distribution Z that does not depend on s.

Consider now a regular estimator relative to the class of one-dimensional submodels $\{F_{\epsilon,s} : s \in \mathcal{S}\}$ and assume that it is asymptotically linear at F_X with influence curve $IC(X \mid F_X, \mu)$:

$$\mu_n - \mu = \frac{1}{n}\sum_{i=1}^{n} IC(X_i \mid F_X, \mu) + o_P(1/\sqrt{n}).$$

Then

$$IC(X \mid F_X, \mu) \in T_{nuis}^{F,\perp,*}(F_X);$$

that is, the influence curve of a regular asymptotically linear estimator is a gradient. This shows that the orthogonal complement of the nuisance tangent space identifies asymptotically all regular asymptotically linear estimators of μ in the full data model $\mathcal{M}^F$. More important to us, as heavily exploited in this book, the orthogonal complement of the nuisance tangent space $T_{nuis}^{\perp}(F_X)$ can be used to define a class of estimating functions defining all estimators of interest.

The canonical gradient $S_{eff}^{F*}(X \mid F_X)$ is of great importance since the asymptotic variance of a regular asymptotically linear estimator of μ at F_X is bounded below by the variance of the canonical gradient, and a regular asymptotically linear (RAL) estimator is efficient at F_X if and only if it is asymptotically linear with influence curve equal to the canonical gradient (i.e., efficient influence curve) at F_X. The fact that the variance of the efficient influence curve provides a lower bound for the asymptotic variance of any RAL estimator can be understood as follows. For simplicity, consider the situation where μ is univariate. Consider the one-dimensional model $\{F_{\epsilon,s} : \epsilon\}$ with parameter ϵ, and note that the true parameter value is $\epsilon_0 = 0$. Note also that the score of ϵ at ϵ_0 equals $s(X)$. The parameter of interest in this model is the following function of ϵ: $\phi(\epsilon) \equiv \mu(F_{\epsilon,s})$. The Cramer–Rao lower bound for the variance of an unbiased estimator of $\phi(\epsilon) \in \mathbb{R}$ at $\epsilon = \epsilon_0 = 0$ equals

$$\frac{\left(\frac{d}{d\epsilon}\phi(\epsilon)\big|_{\epsilon=0}\right)^2}{E_{F_X} s^2(X)} = \frac{\langle S_{eff}^{F*}, s\rangle_{F_X}}{\langle s, s\rangle_{F_X}}, \tag{13}$$

where we just noted that $\phi'(0) = d/d\epsilon\mu(F_{\epsilon,s})\big|_{\epsilon=0}$. Of course, any regular asymptotically linear estimator for the model $\mathcal{M}$ should have an influence curve with variance larger than the Cramer–Rao lower bound (13) for the one-dimensional submodel $F_{\epsilon,s}$. This is true for every possible one-dimensional submodel. As a consequence, any regular asymptotically linear estimator for the model $\mathcal{M}^F$ should have an influence curve with variance larger than the supremum over all $s \in T^F(F_X)$ of the one-dimensional Cramer–Rao lower bounds (13) for the one-dimensional submodel $F_{\epsilon,s}$. By the Cauchy–Schwarz inequality, it follows immediately that this supremum is attained at $s = S_{eff}^{F*}$ with maximum $ES_{eff}^{F*}(X)^2$. This maximum is called the generalized Cramer–Rao lower bound.

Thus, one can view $S_{eff,j}^{F*}$ as the score (index) of the one-dimensional submodel that makes estimation of $\mu_j(F_X)$, $j = 1, \ldots, k$, most difficult, and its variance equals the generalized Cramer–Rao lower bound. An estimator that achieves this bound of a one-dimensional submodel must be efficient. We will actually view the canonical gradients of μ at $F_X \in \mathcal{M}^F$ as the basis for an optimal estimating function.

1.3 Estimating functions.

Let us now discuss estimating functions. Consider a class of k-dimensional estimating functions $\{D_h(X \mid \mu, \rho) : h \in \mathcal{H}^{F^k}\}$ indexed by an index h ranging over a set $\mathcal{H}^{F^k}$. An estimating function $D_h : \mathcal{X} \times \{\mu(F_X), \rho(F_X) : F_X \in \mathcal{M}^F\}$ is a function of X, the parameter of interest μ, and possibly a nuisance parameter $\rho = \rho(F_X)$. An estimating function is (uniformly) unbiased if

$$E_{F_X} D_h(X \mid \mu(F_X), \rho(F_X)) = 0 \text{ for all } F_X \in \mathcal{M}^F.$$

Suppose now that the estimating functions are an element of the orthogonal complement of the nuisance tangent space in the sense that for all $h \in \mathcal{H}^F$

$$D_h(\cdot \mid \mu(F_X), \rho(F_X)) \in T^{F\perp}_{nuis}(F_X) \text{ at all } F_X \in \mathcal{M}^F. \tag{14}$$

We showed in (10) that if ρ and μ are variation-independent parameters, then for any one-dimensional model $F_{\epsilon,s}$, $s \in T^F_{nuis}(F_X)$ (i.e., a one-dimensional model only fluctuating the nuisance parameter so that $\mu(F_{\epsilon,s}) - \mu(F_X) = o(\epsilon)$) we have

$$\frac{d}{d\epsilon} E_{F_X} D_h(X \mid \mu(F_X), \rho(F_{\epsilon,s}))\Big|_{\epsilon=0} = 0. \tag{15}$$

This is a very nice property of an estimating function since it shows that it either does not involve a nuisance parameter ρ or the derivative of the corresponding estimating equation w.r.t. ρ (fixing μ) is zero asymptotically. Consequently, under regularity conditions, for any decent *consistent* estimator ρ_n, the solution μ_n of the corresponding estimating equation

$$0 = \sum_{i=1}^{n} D_g(X_i \mid \mu, \rho_n)$$

will be asymptotically linear with influence curve

$$IC(X) = -\left\{ \frac{d}{d\mu} E_{F_X} D_g(X \mid \mu, \rho(F_X))\Big|_{\mu=\mu(F_X)} \right\}^{-1} D_g(X \mid \mu(F_X), \rho(F_X)). \tag{16}$$

In other words, it will have the same influence curve as in the case where $\rho_n = \rho(F_X)$ is known. This makes statistical inference straightforward, given that we already have the estimating function. Notice that by (11) the influence curve $IC(X)$ is indeed a gradient. Finding a class of estimating functions satisfying (14) requires computing $T^{F,\perp}_{nuis}(F_X)$ at all $F_X \in \mathcal{M}^F$.

Note that the definition of the orthogonal complement of the nuisance tangent space depends on the class of submodels one had at the start. It is of interest to note that if one chooses the class of one-dimensional submodels to be toosmall in the sense that we are excluding certain scores that should have been in the tangent space, then the orthogonal complement of the nuisance tangent space will not be truly orthogonal to all directions that the model allows. As a consequence, in that case, the corresponding estimating functions will not be orthogonal to all nuisance parameters in the sense that (15) is unequal to zero along certain one-dimensional submodels. There is nothing wrong with using a class of estimating functions that are only orthogonal to a subspace of the nuisance tangent space, but in that case one will not have the influence curve of the corresponding estimators equal the standardized estimating function (16).

Example 1.5. Consider the previous example and let us choose too small a class of one-dimensional submodels $F_{\epsilon,h}$ by also requiring that the score h satisfy $\langle h, f - E_{F_X}f(X)\rangle_{F_X} = 0$ for a given function f. The nuisance tangent space equals $\{h \in L_0^2(F_X) : h \perp \langle D^*, f - E_{F_X}f\rangle\}$ so that the orthogonal complement of the nuisance tangent space equals all linear combinations of D^* and $f - E_{F_X}f(X)$. Notice that the estimating function $D(X \mid \mu, \rho) \equiv D^*(X \mid \mu) + f(X) - \rho$ has a nuisance parameter $\rho(F_X) \equiv E_{F_X}f(X)$, and indeed its derivative w.r.t. ρ now is not zero: (15) now fails. In fact, the solution of the estimating equation $\sum_{i=1}^n D(X_i \mid \mu, \rho_n)$ with $\rho_n = \frac{1}{n}\sum_{i=1}^n f(X_i)$ (the only nonparametric way of estimating the nuisance parameter) equals the empirical cumulative distribution function at t_0 that solves $0 = \sum_{i=1}^n D^*(X_i \mid \mu)$. $\square$

1.4 Orthogonal complement of a nuisance tangent space in an observed data model

Consider the following class of parametric submodels through $G_{Y|X}$:

$$\{(1 + \epsilon V(y))dG(y|x) : V \in L_0^2(P_{F_X,G}), E(V(Y) \mid X) = 0\}.$$

The tangent space of G in the model $\mathcal{M}(\mathcal{G}_{CAR})$ generated by this class of submodels is given by

$$T_{CAR}(P_{F_X,G}) = \{v \in L_0^2(P_{F_X,G}) : E(v(Y) \mid X) = 0\}.$$

Let $A_{F_X} : L_0^2(F_X) \rightarrow L_0^2(P_{F_X,G})$ be the nonparametric score operator for F_X:

$$A_{F_X}(s)(Y) = E(s(X) \mid Y).$$

Let $\{\epsilon \rightarrow F_{\epsilon,s} : s \in \mathcal{S}^F\}$ be the class of one-dimensional submodels in the full data model with tangent space $T^F(F_X)$. The one-dimensional submodels $\epsilon \rightarrow F_{\epsilon,s}$ imply one-dimensional submodels $P_{F_{\epsilon,s},G}$ with scores

$$A_{F_X}(s)(Y) = E(s(X) \mid Y),$$

as proved in Gill (1989). Therefore, we have that the nuisance tangent space $T_{nuis}(P_{F_X,G})$ in model $\mathcal{M}(CAR)$ is given by

$$T_{nuis}(P_{F_X,G}) = \overline{\{A(s_{nuis}) : s_{nuis} \in T_{nuis}^F(F_X)\}} \oplus T_{CAR}(P_{F_X,G}).$$

The next theorem provides a representation of $T_{nuis}^{\perp}(P_{F_X,G})$. It will be useful to define the adjoint $A_G^{\top} : L_0^2(P_{F_X,G}) \rightarrow L_0^2(F_X)$ of A_{F_X}, which is given by

$$A_G^{\top}(V)(X) = E(V(Y) \mid X).$$

Let $\mathbf{I}_{F_X,G} : L_0^2(F_X) \rightarrow L_0^2(F_X)$ be defined by $\mathbf{I}_{F_X,G} = A_G^{\top}A_{F_X}$.

Theorem 1.1. *Since F_X, G is fixed in this theorem, we suppress possible dependence on F_X, G; In particular, U below can depend on both F_X and G.*

Suppose that $U : T_{nuis}^{F,\perp} \to L_0^2(P_{F_X,G})$ satisfies $E(U(D)(Y) \mid X) = D(X)$ for all $D \in T_{nuis}^{F,\perp}$. In the model $\mathcal{M}(CAR)$, we have

$$T_{nuis}^{\perp} = \{U(D) - \Pi(U(D) \mid T_{CAR}) : D \in T_{nuis}^{F,\perp}\}. \tag{17}$$

Specifically, for any $V \in T_{nuis}^{\perp}$, we have that $D_V \equiv A^{\top}(V) \in T_{nuis}^{F,\perp}$ and

$$V = U(D_V) - \Pi(U(D_V) \mid T_{CAR}). \tag{18}$$

We also note that in the model $\mathcal{M}(G)$ with G known, we have

$$T_{nuis}^{\perp} = \{U(D) + \Phi : D \in T_{nuis}^{F,\perp}, \Phi \in T_{CAR}\}. \tag{19}$$

Finally, we note that for a $D \in R(\mathbf{I})$ (range of linear operator $\mathbf{I} : L_0^2(F_X) \to L_0^2(P_{F_X,G})$)

$$U(D) - \Pi(U(D) \mid T_{CAR}) = A\mathbf{I}^{-1}(D). \tag{20}$$

How do the ideas from the course interact?

So far we have covered several different topics.

I: Empirical Process Theory

II: Derivatives in Function Space

III: The Functional Delta Method

IV: Estimating Equations

V: Efficiency Theory

The topics relate to each other as follows.

I $\longrightarrow$ III,IV: Empirical process theory is a tool used for showing that the functional delta method gives asymptotically linear estimates (along with Hadamard differentiability of the function), and a tool for showing solutions to estimating equations are asymptotically linear.

II $\longrightarrow$ III,V: Hadamard differentiability is needed to show the functional delta method gives asymptotically linear estimates. The smoothness condition needed to define *regular* parametric models (which form the foundation of efficiency theory) is Frechet differentiability of the mapping from a Euclidean parameter to the square-root density in $L_2(P)$.

III $\longrightarrow$ IV: The functional delta method can be used to show that estimating equations give asymptotically linear estimators, as discussed in class, although this is not a standard approach.

V $\longrightarrow$ III,IV: Efficiency theory lets us examine estimators (coming from either the functional delta method or estimating equations) to check whether they are efficient (the asymptotically best regular estimator). Efficiency theory also provides the class of all estimating functions of interest, through the orthogonal complement of the nuisance tangent space.

The general methodology of van der Laan and Robins

The general methodology of van der Laan and Robins for estimating regular Euclidean parameters with (or without) censored data can be summarized as follows.

Setup: $O_1, ..., O_n \sim P \in \mathcal{M}$ are n i.i.d. observations, the parameter of interest is the pathwise differentiable $\psi(P)$, for $\psi : \mathcal{M} \to \mathcal{R}^k$. The support for X and C in the model $\mathcal{M}$ are denoted $\mathcal{X}$ and $\mathcal{C}$. The full (unobserved) data is stored in X, and $O = \Phi(X, C)$ for C a censoring variable that satisfies *coarsening at random*, so for $C(o) = \{x \in \mathcal{X} : o = \Phi(x, c) \text{ for some } c \in \mathcal{C}\}$, $dP(o|X = x_1) = dP(o|X = x_2)$ for $x_1, x_2 \in C(o)$.

- i: Find $T^{\perp}_{nuis}(F_X)$ in the full-data world.
- ii: Map this into $T^{\perp}_{nuis}(P)$ in the observed-data world.
- iii: Estimate $\psi(P)$ by solving an estimating equation from $T^{\perp}_{nuis}(P)$
- iv: If possible, look for the efficient estimate, so use the efficient score as the estimating equation

Care needs to be taken in step (iii). Typically an estimating function can be written as $U(O|\psi(P), \eta(P))$, for $\eta(P)$ a nuisance parameter. We usually estimate this with η_n from the data, and solve $0 = \frac{1}{n}\sum_{i=1}^n U(O_i|\psi_n, \eta_n)$. In order to show that the resulting estimate ψ_n is asymptotically linear for $\psi(P)$, we must typically show that the nuisance parameter is estimated sufficiently quickly so that $\frac{1}{n}\sum_{i=1}^n U(O_i|\psi_n, \eta(P)) = o_P(n^{-1/2})$. Of course, we still need to check the usual regularity conditions of estimating equations to show asymptotic linearity of the resulting estimate. That is, we must show ψ_n is consistent, the map $\theta \to E_P[U(O|\theta, \eta(P)]$ from $\mathcal{R}^k$ to $\mathcal{R}^k$ should have an invertible derivative at $\theta = \psi(P)$, and there is also an empirical process condition to check for the functions $\{U(\cdot|\theta, \eta(P)) : \|\theta - \psi(P)\| < \epsilon\}$. The point is, obviously not every function in $T^{\perp}_{nuis}(P)$ can be used as an estimating equation to give an asymptotically linear estimate (eg. the zero vector is in this space), and we still have to verify the regularity conditions after deciding to use a particular estimating equation.

What's so great about $T_{nuis}^{\perp}(P)$?

There are many methodologies out there for estimating parameters.

A: *The plug-in principle.* If $\mathcal{M} \supset \{$all empirical distributions$\}$, take $\psi_n = \psi(P_n)$, for P_n the empirical distribution.

B: *Minimum distance estimates.* For $\Pi(P_n|\mathcal{M})$ the closest distribution in $\mathcal{M}$ (defined with some metric) to P_n, take $\psi_n = \psi(\Pi(P_n|\mathcal{M})$.

C: *The method of sieves.* Approximate $\mathcal{M}$ with an increasing nested sequence of regular parametric models, and use the efficient estimate of $\psi(P)$ in one of those submodels, where we increase the size of this submodel as we get more data.

D: *Empirical/Modified likelihood.* Use maximum likelihood to estimate $\psi(P)$ when working in the random model $\mathcal{M}_n = \{Q \in \mathcal{M} : Q(\{O_1, ..., O_n\}) = 1\}$.

Some of these methodologies will be revisited when we study estimation of irregular parameters. But for regular parameters, estimators formed in the ways listed above (or in any other way) can be characterized by $T_{nuis}^{\perp}(P)$. A general result is that if ψ_n is a regular asymptotically linear estimator of $\psi(P)$, then its influence curve is a gradient (so in $T_{nuis}^{\perp}(P)$). The estimator then corresponds to using an estimating function from $T_{nuis}^{\perp}(P)$ (the gradient premultiplied by the inverse of its covariance matrix). That is, all regular asymptotically linear estimators are asymptotically equivalent in first-order with estimators obtained from $T_{nuis}^{\perp}(P)$. It is for this reason that we consider $T_{nuis}^{\perp}(P)$ such a fundamental object.

Example 1.14, page 35, van der Laan, Robins, 2002, Unified Methods for Censored Longitudinal Data and Causality.

Example 1.6. (Repeated measures data with right-censoring; continuation of Example ??) In the previous coverage of this example, we explained that, in the full data world, globally efficient estimators are not practical but that attractive locally efficient estimators exist. The latter was shown by showing that the orthogonal complement of the nuisance tangent space in the full data model, which identifies all estimating functions, was given by functions $h(X^*)\epsilon(\alpha)$, which thus do not depend on nuisance parameters. In the observed data model $\mathcal{M}(CAR)$, defined by the restrictions (??) and (??), we claim that it is even impossible to construct any practical estimators at all.

In order to show this, we need to find the orthogonal complement of the nuisance tangent space in the model $\mathcal{M}(CAR)$ and show that each of the elements of this space depends on nuisance parameters that are very hard to estimate with normal sample sizes. Note that in the observed data model $\mathcal{M}(CAR)$, the nuisance parameter consists of the full data nuisance parameter η and G.

Let $\Delta = I(C \geq p)$ be the indicator that the subject does not drop out of the study before p. We have that

$$P(\Delta = 1 \mid X) = \bar{G}(p \mid X) = \prod_{j=0}^{p-1}(1 - \lambda_C(j \mid X)),$$

where $\bar{G}(t \mid X) \equiv P(C \geq t \mid X)$. To begin with, we consider the inverse of probability of censoring weighted estimating functions that are obtained by inverse weighting any full data structure estimating function $D_h(X \mid \alpha) \equiv h(X^*)\epsilon(\alpha)$:

$$\left\{ IC_0(Y \mid G, D) \equiv D_h(X \mid \alpha)\frac{\Delta}{\bar{G}(p \mid X)} : h \right\}. \tag{21}$$

Thus, given an estimator of the nuisance parameter $\bar{G}(p \mid X)$, one could use as estimating equation for α:

$$0 = \frac{1}{n}\sum_{i=1}^{n} h(X_i^*)\epsilon_i(\alpha)\frac{\Delta_i}{\bar{G}_n(p \mid X)}.$$

However, under the sole restriction CAR (**??**), estimation of $\bar{G}$ requires fitting nonparametrically a multinomial regression of very high dimension. As a consequence, these Horvitz–Thompson types of estimating functions do not result in practical estimators that are consistent and asymptotically normal (CAN) over the whole model $\mathcal{M}(CAR)$. We have that (21) is a subset of the orthogonal complement of the nuisance tangent space of η in the observed data model with $\bar{G}$ known. However, by CAR, the tangent space $T_{CAR}(P_{F_X,G})$ of G only assuming (**??**) (i.e, CAR) is also contained in the orthogonal complement of the nuisance tangent space of η in the observed data model with G known. In fact, our representation theorem (Theorem 1.1) shows that adding the tangent space $T_{CAR}(P_{F_X,G})$ to (21) yields the complete orthogonal complement of the nuisance tangent space of η in the model with G known. The tangent space $T_{CAR}(P_{F_X,G})$ of the conditional distribution G of C, given X, consists of all functions of Y with conditional mean zero, given X, w.r.t. G.

We will now derive a representation of the tangent space $T_{CAR}(P_{F_X,G})$ and determine the projection onto $T_{CAR}(P_{F_X,G})$. Our derivation yields an elegant (and easy-to-understand) proof of a very important fundamental result used throughout this book. We will do this in the general situation where we have observed data $(\tilde{T} = \min(T, C), \Delta = I(C \geq T), \bar{X}(\tilde{T}))$ and the full data structure $\bar{X}(T) = \{X(s) : s \leq T\}$, where T is possibly random: in the current example, we have $T = p$ fixed. We define $C = \infty$ if $C \geq T$ so that C is always observed. In the current example this implies that C can never take value p and $A(p)$ is a deterministic function of $A(p-1)$. Let $A(j) = I(C \leq j)$ so that $dA(j) = I(C = j)$, $j = 0, \ldots, p$. Let $\mathcal{F}(j) = (\bar{A}(j-1), \bar{X}(\min(j, C)))$ be the history observed up and including time j. Let $\alpha(j \mid \mathcal{F}(j)) = E(dA(j) \mid \mathcal{F}(j))$ be the probability that $C = j$, given the history $\mathcal{F}(j)$. Note that

$$\alpha(j \mid \mathcal{F}(j)) = I(\tilde{T} \geq j)P(C = j \mid \mathcal{F}(j), \tilde{T} \geq j) = I(\tilde{T} \geq j)\lambda_C(j \mid \bar{X}(j)),$$

where $\lambda_C(j \mid \bar{X}(j))$ is the hazard of C, given X at time j, which by CAR only depends on X through $\bar{X}(j)$. Under CAR, the G part of the likelihood of Y is given by

$$g(\bar{A}(p-1) \mid X) = \prod_{j=0}^{p-1} \alpha(j \mid \mathcal{F}(j))^{dA(j)} \{1 - \alpha(j \mid \mathcal{F}(j))\}^{1-dA(j)}. \tag{22}$$

Since $\alpha(j \mid \mathcal{F}(j))^{dA(j)}\{1 - \alpha(j \mid \mathcal{F}(j))\}^{1-dA(j)}$ is just a Bernoulli likelihood for the random variable $dA(j)$ with probability $\alpha(j \mid \mathcal{F}(j))$, it follows that the tangent space of $\alpha(j \mid \mathcal{F}(j))$ is the space of all functions of $(dA(j), \mathcal{F}(j))$ with conditional mean zero, given $\mathcal{F}(j)$. Straightforward algebra shows that any such function can be written as

$$\begin{aligned} V(dA(j), \mathcal{F}(j)) - E(V \mid \mathcal{F}(j)) &= \{V(1, \mathcal{F}(j)) - V(0, \mathcal{F}(j))\} \\ &\quad \times \{dA(j) - \alpha(j \mid \mathcal{F}(j))\}. \end{aligned} \tag{23}$$

Thus, the tangent space of the parameter $\alpha(j \mid \mathcal{F}(j))$ equals

$$T_{CAR,j} \equiv \{H(\mathcal{F}(j))\{dA(j) - \alpha(j \mid \mathcal{F}(j))\} : H\},$$

where H ranges over all functions of $\mathcal{F}(j)$ for which each element of $T_{CAR,j}$ has finite variance. By factorization of the likelihood (22), we have that

$$T_{CAR}(P_{F_X,G}) = T_{CAR,0} \oplus T_{CAR,1} \ldots \oplus T_{CAR,p-1}. \tag{24}$$

Equivalently,

$$T_{CAR}(P_{F_X,G}) = \left\{ \sum_{j=0}^{p-1} H(j, \mathcal{F}(j)) dM_G(j) : H \right\},$$

where

$$dM_G(j) = I(C = j) - \lambda_C(j \mid \bar{X}(j)) I(\tilde{T} \geq j).$$

Note that $I(C = j) = I(C = j, \Delta = 0)$ for $j < p$.

Thus, the complete orthogonal complement of the nuisance tangent space of η in the observed data model with G known is given by:

$$\left\{ h(X^*)\epsilon(\alpha) \frac{\Delta}{\bar{G}(p \mid X)} - \sum_{j=0}^{p-1} H(j, \mathcal{F}(j)) dM_G(j) : h, H \right\}. \tag{25}$$

This shows that in the observed data model with G known we have access to a rich class (25) of estimating functions for α *without* a nuisance parameter, namely any choice of h, H provides an estimating function for α.

The orthogonal complement of the nuisance parameter (η, G) in $\mathcal{M}(CAR)$ is the subspace of (25) consisting of the functions in (25) which are *also* orthogonal to G. Thus this space is given by:

$$\left\{ h(X^*)\epsilon(\alpha) \frac{\Delta}{\bar{G}(p \mid X)} - \sum_{j=0}^{p-1} H_{opt,h}(j, \mathcal{F}(j)) dM_G(j) : h, H \right\}, \tag{26}$$

where $H_{opt,h}$ is chosen so that $\sum_{j=0}^{p} H_{opt,h}(j, \bar{X}(j))dM_G(j)$ equals the projection of $h(X^*)\epsilon(\alpha)\frac{\Delta}{\bar{G}(p|X)}$ onto $T_{CAR}(P_{F_X,G})$ in the Hilbert space $L_0^2(P_{F_X,G})$.

We will now derive this projection. By representation (24) of T_{CAR}, we have that

$$\Pi(IC_0(Y \mid G, D) \mid T_{CAR}) = \sum_{j=0}^{p-1} \Pi(IC_0(Y \mid G, D) \mid T_{CAR,j}).$$

The projection onto $T_{CAR,j}$ is obtained by first projecting on all functions of $(dA(j), \mathcal{F}(j))$ and subsequently subtracting its conditional expectation, given $\mathcal{F}(j)$,

$$\begin{aligned} \Pi(IC_0(D) \mid T_{CAR,j}) &= E(IC_0(D) \mid dA(j), \mathcal{F}(j)) \\ &\quad -E(E(IC_0(D) \mid dA(j), \mathcal{F}(j)) \mid \mathcal{F}(j)), \end{aligned}$$

where we used short hand notation for $IC_0(Y \mid G, D)$. By (23), this can be written as

$$\{E(IC_0(D) \mid dA(j) = 1, \mathcal{F}(j)) - E(IC_0(D) \mid dA(j) = 0, \mathcal{F}(j))\}\, dM_G(j).$$

Finally, we note that $E(IC_0(D) \mid dA(j) = 1, \mathcal{F}(j)) = 0$ since $dA(j) = 1$ implies $\Delta = 0$ for $j \leq p - 1$. This proves that

$$\Pi(IC_0(D) \mid T_{CAR}) = -\sum_{j=0}^{p-1} \{E(IC_0(D) \mid dA(j) = 0, \mathcal{F}(j))\}\, dM_G(j).$$

This can be represented as $\sum_{j=0}^{p-1} H_{opt,D}(j, \mathcal{F}(j))dM_G(j)$ with

$$\begin{aligned} H_{opt,D}(j, \mathcal{F}(j)) &= -E(IC_0(Y \mid G, D) \mid C > j, \bar{X}(j)) \\ &= -\frac{1}{\bar{G}(j+1 \mid X)} E(D_h(X \mid \alpha) \mid \bar{X}(j), C > j), \end{aligned}$$

where, by definition, $\bar{G}(j + 1 \mid X) = P(C > j \mid X)$. We also note that by CAR $Q_{X,h} \equiv E(D_h(X \mid \alpha) \mid \bar{X}(j), C > j) = E(D_h(X \mid \alpha) \mid \bar{X}(j))$ which is thus a parameter of the full data distribution F_X.

We conclude that the orthogonal complement of the nuisance parameter (η, G) in the model $\mathcal{M}(CAR)$ is given by

$$\left\{ h(X^*)\epsilon(\alpha)\frac{\Delta}{\bar{G}(p \mid X)} - \sum_{j=0}^{p} Q_{X,h}(j, \bar{X}(j))\frac{dM_G(j)}{\bar{G}(j+1 \mid X)} : h \right\}. \qquad (27)$$

Each of the elements, indexed by h, in this orthogonal complement of the nuisance tangent space implies an estimating function for α (and a corresponding influence curve) with nuisance parameters being G and the full data parameter $Q_{X,h}$. Without additional assumptions on the full data model and censoring mechanism, neither of these two nuisance parameters can be reasonably well-estimated in practice. This shows that no practical estimators exist in model $\mathcal{M}(CAR)$.

Above, we formally proved the following fundamental results for the general right-censored data structure $(\tilde{T}, \Delta, \bar{X}(\tilde{T}))$ for the case where censoring is discrete:

Theorem 1.2. *Let $R(t) = I(T \leq t)$ for a time variable T. Let $X(t)$ be a time-dependent process including $R(t)$. Let $X = \bar{X}(T)$ be the full data. We have observed data $Y = (\tilde{T} = \min(C,T), \Delta = I(T \leq C), \bar{X}(\tilde{T}))$, where C is a univariate discrete variable with conditional distribution $G(\cdot \mid X)$, given X. Let $A(t) = I(C \leq t)$, where we define $C = \infty$ if $C \geq T$ so that C is always observed. Let $\mathcal{F}(t) = (\bar{A}(t-), \bar{X}(\min(t,C)))$ be the history observed up to time t.*

Assume CAR on G: $E(dA(t) \mid \bar{A}(t-), X) = E(dA(t) \mid \mathcal{F}(t))$ or equivalently, for $t \leq T$,

$$\lambda_{C|X}(t \mid X) \equiv P(C = t \mid C \geq t, X) = m(t, \bar{X}(t))$$

for some measurable function m. Then, the tangent space $T_{CAR}(P_{F_X,G})$ of G is given by

$$T_{CAR}(P_{F_X,G}) = \overline{\left\{ \int H(u, \mathcal{F}(u)) dM_G(u) : H \right\}} \cap L_0^2(P_{F_X,G}),$$

where $dM_G(u) = I(C \in du, \Delta = 0) - I(\tilde{T} \geq u)\lambda_{C|X}(du \mid X)$. For any function $V(Y)$, we have that $\Pi(V \mid T_{CAR})$ is given by

$$\int \{E(V(Y) \mid dA(u) = 1, \mathcal{F}(u)) - E(V(Y) \mid dA(u) = 0, \mathcal{F}(u))\} \, dM_G(u).$$

Remark: For the data structure Y of theorem 1.2 any variable $V(Y)$ can be written as $\Delta d_1(X) + (1-\Delta)V_2(\bar{X}(C), C)$ for some functions d_1 and V_2. It follows that $E[V(Y) \mid dA(u) = 1, F(u)] = V_2(\bar{X}(u), u)$ is actually a deterministic function of $V(Y)$.

If G is actually continuous, then the G part of the likelihood of Y is defined as the partial likelihood of $A(t)$, w.r.t. the left-continuous history $\mathcal{F}(t-)$ as in Andersen, Borgan, Gill and Keiding (1993), if in theorem 1.2 we replace $F(u)$ and $F(t)$ by $F(u-)$ and $F(t-)$.

The formulas for T_{CAR} and the projection onto T_{CAR} in Theorem 1.2 can be applied to the continuous case as well, but the proof of the representation of T_{CAR} involves calculating the scores from this partial likelihood, and the projection formula needs to be formally defined and proved, taking into account possible measurability conditions needed to define the conditional expectations. A formal treatment of the continuous case is given in van der Vaart (2001). Since the latter is beyond the scope and purpose of this book, we will avoid stating these continuous projection results as theorems, but still use them to define the corresponding estimating functions. □

2 Robustness of Estimating Functions

2.1 Robustness of estimating functions against misspecification of linear convex nuisance parameters.

We showed above that an estimating function that is an element of the orthogonal complement of the nuisance tangent space in the sense defined by (14) has a corresponding estimating equation with first derivative (directional) w.r.t. its nuisance parameter equal to zero. In fact, we will now prove that if the data-generating distribution is linear in the nuisance parameter ρ of the estimating function and the nuisance parameter space is convex, then, at the misspecified nuisance parameter, the estimating function remains unbiased *and orthogonal to the tangent space generated by this parameter*. Similar type results have been obtained by Bickel (1982), Bickel, Klaassen, Ritov, and Wellner (1993), Newey (1990), and Robins, Rotnitzky, van der Laan (2000). The following lemma is a slight modification (including now an orthogonality result) of the result in van der Laan, Yu (2001), but uses essentially the same proof.

Lemma 2.1. *Consider an estimating function $D(X \mid \mu, \rho)$ (i.e., a mapping $\mathcal{X} \times \{(\mu(F_X), \rho(F_X)) : F_X \in \mathcal{M}^F\} \rightarrow \mathbb{R}$) that satisfies (14). Assume that μ is pathwise differentiable at each $F \in \mathcal{M}^F$ along a F-specific class of one-dimensional models including nuisance score lines $F_\epsilon = \epsilon F_1 + (1-\epsilon)F \in \mathcal{M}^F$ indexed by a set of F_1's with 1) $\mu(F_1) = \mu(F)$, 2) $d/d\epsilon \mu(F_\epsilon)|_{\epsilon=0} = 0$, and 3) $dF_1/dF < \infty$ (i.e., being uniformly bounded). In addition, we assume that these classes of lines satisfy the following "connectivity" property: If we have a line $\epsilon F + (1-\epsilon)F_1$ in the F_1-specific class, and a line $\epsilon F_1^* + (1-\epsilon)F$ in the F-specific class (i.e., $\mu(F) = \mu(F_1) = \mu(F_1^*)$, $dF/dF_1 < \infty$, $dF_1^*/dF < \infty$, and the pathwise derivative of μ at F_1 and F, respectively, equals zero along these lines), then the line $\epsilon F_1^* + (1-\epsilon)F_1$ is an element of the F_1-specific class as well; in other words, beyond the properties 1 and 3 which follow directly it also satisfies property 2. Let $T_1(F_X)$ be the tangent space of this class of one-dimensional submodels at F_X.*

Let $F_1 \in \mathcal{M}^F$ be such that $dF/dF_1 < \infty$, $\mu(F_1) = \mu(F)$, and $d/d\epsilon \mu(\epsilon F + (1-\epsilon)F_1)|_{\epsilon=0} = 0$. Then

$$E_{F_X} D(X \mid \mu(F_X), \rho(F_1)) = 0.$$

In fact,

$$D(X \mid \mu(F_X), \rho(F_1)) \in T_1(F_X)^\perp.$$

We note that the connectivity assumption is a natural assumption on the classes of lines. An important corollary of this lemma is that, if the data-generating distribution F_X can be parametrized as $P_{\mu,\eta}$ with $\eta \rightarrow P_{\mu,\eta}$ linear, η ranging over a convex parameter space, and ρ a function of η, then $D(X \mid \mu(F_X), \rho_1) \in T_{nuis}^{F,\perp}(F_X)$ for all ρ_1. (Note that the connectivity assumption obviously holds for models with such variation independent parametrizations). This implies that one can treat ρ as the index h of the estimating function.

Note also that, if $\rho = (\rho_1, \rho_2)$ with ρ_1, ρ_2 being variation-independent (w.r.t. to each other and to μ) linear convex parameters (i.e., satisfying the assumptions of Lemma 2.1), then double application of Lemma 2.1 in the model first with ρ_1 known and then with ρ_2 known, respectively, yields the following double robustness property of the estimating function D:

$$E_{F_X} D(X \mid \mu(F_X), \rho_1, \rho_2) = 0 \text{ if either } \rho_1 = \rho_1(F_X) \text{ or } \rho_2 = \rho_2(F_X).$$

However, note that in this case $D(X \mid \mu(F_X), \rho_1, \rho_2(F_X))$ (or $D(X \mid \mu(F_X), \rho_1(F_X), \rho_2)$) is not necessarily still orthogonal to the tangent space corresponding with just varying ρ_2 (or just varying ρ_1).

Lemma 2.1 requires that $F_X \ll F_1$, but it follows straightforwardly that, if $\rho(F_1)$ can be approximated by a $\rho(F_{1m}$ with $F_X \ll F_{1m}$ so that application of this lemma yields $E_{F_X} D(X \mid \mu(F_X), \rho(F_{1m})) = 0$, and $D(X \mid \mu(F_X), \rho(F_{1m}))$ converges to $D(X \mid \mu, \rho(F_1))$ in $L^2(F_X)$, then we will have $E_{F_X} D(X \mid \mu(F_X), \rho(F_1)) = 0$ as well.

Proof of lemma. Let F_1, F be as in the lemma. Thus $\mu(F_1) = \mu(F)$. Then, $F_{1,\epsilon,s} = \epsilon F + (1-\epsilon)F_1$ is a one-dimensional submodel of $\mathcal{M}^F$ with score $s = d(F - F_1)/dF_1$ satisfying

$$0 = \frac{d}{d\epsilon}\mu(F_{1,\epsilon,s})\bigg|_{\epsilon=0}.$$

By the fact that μ is pathwise differentiable along $F_{1,\epsilon,s}$ at F_1, we have for any gradient $\ell(X \mid \mu(F_1), \rho(F_1))$

$$0 = \int \ell(x \mid \mu(F_1), \rho(F_1)) \frac{d(F - F_1)}{dF_1} dF_1 = \int \ell(x \mid \mu(F_1), \rho(F_1)) dF(x).$$

Since a standardized version of $D(\cdot \mid \mu(F_1), \rho(F_1))$ is a gradient, this implies also that for any such pair F_1, F

$$0 = \int D(x \mid \mu(F_1), \rho(F_1)) dF(x) = \int D(x \mid \mu(F), \rho(F_1)) dF(x),$$

which proves the protection of unbiasedness as stated in the lemma.

This protection will now also provide us with the claimed orthogonality. Any score in $T_1(F_X)$ is of the form $d(F_1^* - F)/dF$ for some F_1^* with $dF_1^*/dF < \infty$, $\mu(F_1^*) = \mu(F)$, and $d/d\epsilon\ \mu(\epsilon F_1^* + (1-\epsilon)F)|_{\epsilon=0} = 0$. We have

$$\begin{aligned}\int \ell(X \mid \mu, \rho(F_1)) \frac{d(F_1^* - F)}{dF} dF &= \int \ell(X \mid \mu, \rho(F_1)) d(F_1^* - F) \\ &= \int \ell(X \mid \mu, \rho(F_1)) dF_1^* - \int \ell(X \mid \mu, \rho(F_1)) dF.\end{aligned}$$

We just proved that the second term equals zero. By the same proof, the first term equals zero if $dF_1^*/dF_1 < \infty$, and $d/d\epsilon\ \mu(\epsilon F_1^* + (1-\epsilon)F_1)|_{\epsilon=0} = 0$. We have $dF_1^*/dF_1 = (dF_1^*/dF) * (dF/dF_1) < \infty$, which proves the first condition. The pathwise derivative conditions holds by our connectivity assumption on the class of lines. This completes the proof. □