Lecture 16

In this and the next couple of lecture we will go back to our initial story of estimating "smooth" functionals in infinite dimensional models.

The General Set Up: $X_1, \ldots, X_n$ are a random sample on some sample space $(\Omega, \mathcal{A})$ having distribution P. We assume $P \in \mathcal{P}$ ← a model and want to estimate $\Psi(P): \mathcal{P} \to \mathbb{R}$. We saw that in many examples $\mathcal{P} = \{P_\eta : \eta \in H\}$ where H is a subset of a normed linear space, and $\Psi(P_\eta) = \chi(\eta)$. We have considered the analysis of quadratic functional of a density as our working example so far. The analysis was initiated by a Von-Mises type expansion of the functionals, which requires certain kind of differentiability of the functionals and the distributions under study. We shall now try to make it more concrete. We refer to Aad Van der Vaart! Asymptotic Statistics (Chapter 25) for more details.

We first begin by defining tangent space and influence functions when we have $X_1, \ldots, X_n$ iid $P \in \mathcal{P}$ and we want to estimate $\Psi(P)$.

Definition (Differentiable Path) A differentiable path is a map $t \to P_t$ from a neighbourhood of 0, $(0, \varepsilon) \subset (0, +\infty)$ to $\mathcal{P}$ such that for some measurable fn. $g: \Omega \to \mathbb{R}$,

$$\int \left(\frac{dP_t^{1/2} - dP^{1/2}}{t} - \frac{1}{2} g \, dP^{1/2} \right)^2 \to 0 \quad \text{as } t \downarrow 0$$

The fn g is called the score fn of the submodel $\{P_t : 0 \le t < \varepsilon\}$ at $t=0$ with $P_0 = P$ the truth.

Indeed the preceding definition is not very enlightening. First of all we have not made clear what the objects $dP_t^{1/2}$, $dP^{1/2}$ are. They can be formalized by introducing Hilbert space of square root of probability measures. (Le Cam 1960's). A more simple form is

$$\int \left(\frac{p_{tt}^{1/2} - p_t^{1/2}}{t} - \frac{1}{2} g p_t^{1/2} \right)^2 d\mu_t \to 0 \quad \text{as } t \downarrow 0$$

where for each t, μ_t is an arbitrary measure relative to which both P and P_t have densities p_t and p_{tt} respectively. ~~It shall~~ The choice of μ_t is not important. We shall not pay too much attention to these subtleties and instead try to understand the implications of the definitions. To this end note that

$$\int \left(\frac{dP_t^{1/2} - dP^{1/2}}{t} - \frac{1}{2} g \, dP^{1/2} \right)^2$$

$$= \int \left(\frac{dP_t^{1/2} - dP^{1/2}}{t \, dP^{1/2}} - \frac{1}{2} g \right)^2 dP$$

$\therefore$ The definition $\Rightarrow \lim_{t \to 0} \mathbb{E}_P \left\{ \left(\frac{dP_t^{1/2} - dP^{1/2}}{t \, dP^{1/2}} - \frac{g}{2} \right)^2 \right\} = 0$

$\Rightarrow$ We have some kind of differentiability in quadratic means of square root of our densities. This differentiability requires convergence in $L_2(P)$ sense (or $L_2(\mu_t)$ sense), and not pointwise.

$\{P_t : 0 \le t < \varepsilon\}$ that is diff. at quadratic mean at $t=0$ is called Smooth Parametric Submodel of $\mathcal{P}$.

Definition (Tangent set) Letting $t \to P_t$ range over a collection of smooth and "regular" submodels we obtain a collection of score functions, which we call a Tangent set of the model $\mathcal{P}$ at P, denoted by $\dot{\mathcal{P}}_P$.

By regular we mean that the Fisher information at $t=0$ of $\{P_t : 0 \le t < \varepsilon\}$ ~~defined as~~ $\left(\mathbb{E}(g^2)\right)$ is non zero.

When we consider all possible differentiable paths $t \to P_t$, we obtain the maximal collection of score functions $\to$ Maximal Tangent Set.

The following Lemmas gives fundamental properties of ~~the~~ score functions.

Lemma 1: Every score function satisfies $\mathbb{E}_P(g) = 0$ and $\mathbb{E}_P g^2 < \infty$.

Proof: For given but arbitrary $(t_n)_{n \ge 1}$ such that $t_n \to 0$ as $n \to \infty$, let p_n and p be densities of P_{t_n} and P w.r.t a dominating measure μ. (For example a convex combination of the countably many measures $P_{t_n} + P$).
By definition $(\sqrt{p_n} - \sqrt{p})/t_n$ converges in $L_2(\mu)$ to $\frac{1}{2} g \sqrt{p}$.
$$\Rightarrow \left\| \frac{\sqrt{p_n} - \sqrt{p}}{t_n} - \frac{1}{2} g \sqrt{p} \right\|^2_{L_2(\mu)} \to 0.$$

Now $(\sqrt{p_n} - \sqrt{p})/t_n \in L_2(\mu) \ \forall n \ge 1$. $\Rightarrow \frac{g\sqrt{p}}{2}$ is a limit of $L_2(\mu)$ functions in $L_2(\mu) \Rightarrow g\sqrt{p}/2 \in L_2(\mu) \Rightarrow g \in L_2(P)$.

Now, $\lim_{t_n \to 0} \| \sqrt{p_n} - \sqrt{p} - \frac{1}{2} g \sqrt{p}\, t_n \|_{L_2(\mu)} = \lim_{t_n \to 0} \left\| \frac{\sqrt{p_n} - \sqrt{p}}{t_n} - \frac{1}{2} g \sqrt{p} \right\|_{L_2(\mu)} t_n$

$= 0$

But $\| \sqrt{p_n} - \sqrt{p} - \frac{1}{2} g \sqrt{p}\, t_n \|_{L_2(\mu)}$

$\geq \| \sqrt{p_n} - \sqrt{p} \|_{L_2(\mu)} - \frac{1}{2} t_n \| g \sqrt{p} \|_{L_2(\mu)}$

$\Rightarrow 0 \geq \lim_{n \to \infty} \left| \| p_n^{\frac{1}{2}} - p^{\frac{1}{2}} \|_{L_2(\mu)} - \frac{1}{2} t_n \| g \sqrt{p} \|_{L_2(\mu)} \right|$

$\Rightarrow \| p_n^{\frac{1}{2}} - p^{\frac{1}{2}} \|_{L_2(\mu)} \to 0 \quad$ as $\quad n \to \infty$

$\Rightarrow p_n^{\frac{1}{2}} \longrightarrow p^{\frac{1}{2}}$ in $L_2(\mu)$

$\Rightarrow E_p(g) = \int g \, dp = \int g p \, d\mu = \int \frac{1}{2} g \sqrt{p} \times 2\sqrt{p} \, d\mu$

$= \left\langle \frac{1}{2} g \sqrt{p}, 2\sqrt{p} \right\rangle_{L_2(\mu)}$

$= \lim_{n \to \infty} \int \frac{(\sqrt{p_n} - \sqrt{p})}{t_n} \times (\sqrt{p_n} + \sqrt{p}) \, d\mu$ (continuity of inner product)

$= \lim_{n \to \infty} \int \frac{(p_n - p)}{t_n} \, d\mu = 0$ ▨

It follows that a tangent set can be identified with a subset of $L_2(P)$ (upto equivalence classes). Tangent sets are often linear spaces, and referred to tangent spaces in such cases. Geometrically, we may visualize the model $\mathcal{P}$, or rather the corresponding "square root of measures" $dP^{1/2}$, as a subset of the unit ball of a Hilbert space & the set $\mathcal{P}_p$ (or rather the objects $\pm g p^{1/2}$) as its tangent set.

Usually we construct submodels $t \to P_t$ such that for every x, $g(x) = \frac{\partial}{\partial t} \log dP_t(x)|_{t=0}$. Such pointwise differentiability is not required by our definitions. We therefore often make use of the next lemma.

Lemma 2: If p_t is a density w.r.t fixed measure μ and $t \to p_t(x)$ is continuously diff. in a neighbourhood of 0 and $t \to \int \dot{p}_t^2/p_t \, d\mu$ is finite and continuous in this neighborhood, then $t \to P_t$ is a diff. path.

It also turns out that our definition of differentiability is a reasonable one because it ensures a desired form of local asymptotic normality.

Lemma 3: If $t \to P_t$ is a differentiable path, then

$$\log \prod_{i=1}^{n} \frac{dP_{1/\sqrt{n}}}{dP}(X_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} g(X_i) - \frac{1}{2} E_P(g^2) + o_P(1)$$

Now come to estimation of suitably differentiable functionals. $(P \to \Psi(P))$

Differentiable Functionals:

For defining "information" for estimating $\Psi(P)$, only those submodels $t \to P_t$ along which $t \to \Psi(P_t)$ is differentiable is of interest. A minimal requirement is differentiability at $t=0$, but we need more. We follow definition introduced by Levit (1978). Let $\Psi: \mathcal{P} \to \mathbb{B}$ where $\mathbb{B}$ is a Banach space.

Definition: A map $\Psi: \mathcal{P} \to \mathbb{B}$ is differentiable at P relative to a given tangent space $\dot{\mathcal{P}}_P$ if there exists a continuous linear map $\dot{\Psi}_P : L_2(P) \to \mathbb{B}$ such that $\forall g \in \dot{\mathcal{P}}_P$ and corresponding submodel $t \to P_t$,

$$\frac{\Psi(P_t) - \Psi(P)}{t} \xrightarrow{\mathbb{B}} \dot{\Psi}_P(g) \quad \text{i.e} \quad \left\| \frac{\Psi(P_t) - \Psi(P)}{t} - \dot{\Psi}_P(g) \right\|_{\mathbb{B}} \to 0$$

It requires that at $t \to \Psi(P_t)$ is differentiable in usual sense with an extra requirement for it to have a special representation. The map $\dot{\Psi}_P$ is much like a Hadamard derivative of Ψ viewed as a map on the space of square root of measures.

Implications: In the case $\mathbb{B} = \mathbb{R}^k$ for some $k \in \mathbb{N}$, we have by Reisz Representation theorem, that $\exists\, \tilde{\Psi}_P : \Omega \to \mathbb{R}^k$ s.t. $\dot{\Psi}_P(g) = \langle \tilde{\Psi}_P, g \rangle_{L_2(P)} = \int \tilde{\Psi}_P\, g\, dP \quad (\tilde{\Psi}_P \in L_2(P))$. The function $\tilde{\Psi}_P$ is defined uniquely only upto inner products with elements of the tangent set (which is not necessarily all of $L_2(P)$). However it is possible to find $\tilde{\Psi}_P$ whose coordinates $\in \overline{\text{lin}}\, \dot{\mathcal{P}}_P$ → this function is unique and is called the efficient influence function. Any other $\tilde{\Psi}_P$ is simply called an influence function of $\Psi(P)$ at P w.r.t $\dot{\mathcal{P}}_P$. In particular one takes as an influence fn. any m'ble $\dot{\Psi}_P : \Omega \to \mathbb{R}^k$ whose projection onto $\overline{\text{lin}}\, \dot{\mathcal{P}}_P$ is the efficient influence fn.

Remark: It can be shown that the maximal tangent set is a cone, which we will assume from now on.

Examples:

① Parametric Model: Consider a parametric model with parameter $\theta \in \Theta \subseteq \mathbb{R}^d$ and Θ is an open subset of $\mathbb{R}^d$. Assume $\{P_\theta\}_{\theta \in \Theta}$ to be the class of probability distributions and p_θ be density of P_θ w.r.t some common dominating measure μ.

Suppose $\exists$ vector valued measurable map $\dot{l}_\theta$ such that

$$\int \left(p_{\theta+h}^{1/2} - p_\theta^{1/2} - \frac{1}{2} h^T \dot{l}_\theta p_\theta^{1/2} \right)^2 d\mu = o(\|h\|^2) \quad \text{as } h \to 0 \text{ in } \mathbb{R}^d.$$

Then a tangent set at P_θ is given by linear space $\{h^T \dot{l}_\theta : h \in \mathbb{R}^m\} \to$ span of score functions for the coordinates of θ.

If the Fisher information matrix $I_\theta = P_\theta(\dot{l}_\theta \dot{l}_\theta^T)$ is invertible then every map $\chi : \Theta \to \mathbb{R}^k$ that is differentiable in an ordinary sense is differentiable as a map $\psi(P_\theta) = \chi(\theta)$ on the model relative to the tangent space. To see this, note that ~~this~~ the submodel $t \to \theta + th$ has score $h^T \dot{l}_\theta$ at P_θ and

$$\frac{\partial}{\partial t} \chi(\theta + th)\Big|_{t=0} = \dot{\chi}_\theta h = P_\theta (\dot{\chi}_\theta I_\theta^{-1} \dot{l}_\theta \dot{l}_\theta^T h)$$
$$= P_\theta (\underbrace{\dot{\chi}_\theta I_\theta^{-1} \dot{l}_\theta}_{\dot{\psi}_{P_\theta}(g)} \underbrace{h^T \dot{l}_\theta}_{g})$$

(2) Nonparametric Model: Suppose $\mathcal{P}$ consists of all probability laws on a given sample space. Then the tangent set at $P = \{g : \int g\,dP = 0,\ \int g^2 dP < \infty\} = L_2^0(P)$. This is indeed the maximal tangent set. To see that this is indeed the tangent set we exhibit suitable one dimensional submodels of $\mathcal{P}$ at P s.t. we can get any $g \in L_2^0(P)$ as a score.

For a bounded fn $g \in L_2^0(P)$, consider

$$p_t(x) = c(t)\exp(tg(x))\,p(x)$$

p being density of P w.r.t some fixed measure μ.

$$\Rightarrow g(x) = \frac{\partial}{\partial t}\log p_t(x)\Big|_{t=0}$$

One can also show that g is a score in an L_2-sense (by direct calculation using boundedness of g)

For unbounded functions g these submodels might not be well defined. However one can always use

$$p_t(x) = c(t)\,f(tg(x))\,p(x) \quad \text{with}$$

$$f(0) = f'(0) = 1$$

e.g: $f(x) = 2(1+e^{-2x})^{-1}$ works.

For differentiability of $\Psi(P)$ at P, we need for any differentiable submodel $\{P_t\}$

$$\frac{\Psi(P_t) - \Psi(P)}{t} \longrightarrow \int \dot{\Psi}_P\, g\, dP \quad \text{for some } \dot{\Psi}_P.$$

e.g: When P is on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $\Psi(P) = \int x\,dP$, $\dot{\Psi}_P(x) = x$ (check by direct calculation)

Score and Information Operators

Consider first the situation that $\mathcal{P} = \{P_\eta : \eta \in H\}$ and η is itself a probability measure on some measurable space. We are interested in estimation of $\psi(P_\eta) = \chi(\eta)$ for $\chi : H \to \mathbb{R}^d$ on the model H. We will work $k=1$ here.

The model H gives rise to a tangent set $\dot{H}_\eta$ at η. If the map $\eta \to P_\eta$ is differentiable in an appropriate sense then the corresponding derivative will map every score $b \in \dot{H}_\eta$ into a score g for the model $\mathcal{P}$ at P_η. Let's make this precise.

Assume that a smooth parametric submodel $t \to \eta_t$ at η induces a smooth parametric submodel $t \to P_{\eta_t}$ at P_η where $\eta = \eta_0$, and that the score functions b of the submodel $t \to \eta_t$ and g of $t \to P_{\eta_t}$ are related by

$$g = A_\eta b$$

for some operator $A_\eta : \dot{H}_\eta \subset L_2(\eta) \to L_2(P_\eta)$.

We assume that

$$A_\eta : \operatorname{lin} \dot{H}_\eta \subset L_2(\eta) \to L_2(P_\eta) \quad \text{is continuous \& linear.}$$

Next assume that the function $\eta \to \chi(\eta)$ is differentiable with influence function $\tilde{\chi}_\eta$ w.r.t. $\dot{H}_\eta$.

Thus by definition the function $\psi(P_\eta) = \chi(\eta)$ is pathwise differentiable w.r.t the tangent set $\dot{\mathcal{P}}_{P_\eta} = A_\eta \dot{\mathcal{H}}_\eta$ iff $\exists$ a function $\tilde{\psi}_{P_\eta}$ such that

$$\langle \tilde{\psi}_{P_\eta}, A_\eta b \rangle_{P_\eta} = \frac{\partial}{\partial t} \psi(P_{\eta_t})\Big|_{t=0} \frac{\partial}{\partial t} \chi(\eta_t)\Big|_{t=0} = \langle \tilde{\chi}_\eta, b \rangle_\eta$$
$$\forall\, b \in \dot{\mathcal{H}}_\eta$$

This equation can be written in terms of the adjoint score operator $A_\eta^* : L_2(P_\eta) \to \overline{\text{lin}}\, \dot{\mathcal{H}}_\eta$. By definition of the adjoint

$$\langle h, A_\eta b \rangle_{P_\eta} = \langle A_\eta^* h, b \rangle_\eta \quad \text{for } \forall\, h \in L_2(P_\eta),\ b \in \dot{\mathcal{H}}_\eta$$

Note that we define A_η^* to have range $\overline{\text{lin}}\, \dot{\mathcal{H}}_\eta$ so that it is the adjoint of $A_\eta : \dot{\mathcal{H}}_\eta \to L_2(P_\eta)$. This is the adjoint of an extension $A_\eta : L_2(\eta) \to L_2(P_\eta)$ followed by an orthogonal projection onto $\overline{\text{lin}}\, \dot{\mathcal{H}}_\eta$.

Some Results from Functional Analysis: Every continuous linear map $A : \mathcal{H}_1 \to \mathcal{H}_2$ between Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ has an adjoint map $A^* : \mathcal{H}_2 \to \mathcal{H}_1$ which is a continuous linear map & uniquely determined by

$$\langle A^* h_2, h_1 \rangle_1 = \langle h_2, A h_1 \rangle_2 \quad \text{for all } h_1 \in \mathcal{H}_1,\ h_2 \in \mathcal{H}_2.$$

If A is considered to be the restriction to $\mathcal{H}_1 \subset \tilde{\mathcal{H}}_1$ of a continuous linear map $\tilde{A} : \tilde{\mathcal{H}}_1 \to \mathcal{H}_2$ where $\tilde{\mathcal{H}}_1$ an Hilbert that contains $\mathcal{H}_1$ isometrically, then $A^* = \Pi \tilde{A}^*$ where $\Pi : \tilde{\mathcal{H}}_1 \to \mathcal{H}_1$ the orthogonal projection on $\mathcal{H}_1$.

Going back to our problem, we have

$$\langle h, A_\eta b\rangle_{P_\eta} = \langle A_\eta^* h, b\rangle_\eta \quad \forall\, h \in L_2(P_\eta),\ b \in \dot{H}_\eta$$

But pathwise differentiability requires

$$\langle \tilde{\psi}_{P_\eta}, A_\eta b\rangle_{P_\eta} = \langle \tilde{\chi}_\eta, b\rangle_\eta \quad \forall\, b \in \dot{H}_\eta$$

But

$$\Rightarrow \langle \tilde{\psi}_{P_\eta}, A_\eta b\rangle_{P_\eta} = \langle A_\eta^* \tilde{\psi}_{P_\eta}, b\rangle_\eta \quad \forall\, b \in \dot{H}_\eta$$

$$\Rightarrow A_\eta^* \tilde{\psi}_{P_\eta} = \tilde{\chi}_\eta \quad \ldots\ldots (*)$$

$\Rightarrow \psi(P_\eta) = \chi(\eta)$ is differentiable w.r.t. to the tangent set $\dot{\mathcal{P}}_{P_\eta} = A_\eta \dot{H}_\eta$ iff $(*)$ can be solved for $\tilde{\psi}_{P_\eta}$. Equivalently, this happens iff $\tilde{\chi}_\eta \in \text{Range}(A_\eta^*)$. Since A_η^* is not necessarily onto $\overline{\text{lin}}\,\dot{H}_\eta$, this is indeed a condition!

Now two solutions of $(*)$ can only differ by an element of the null space $\mathcal{N}(A_\eta^*) = \mathcal{R}(A_\eta)^\perp$ where $\mathcal{R}(A_\eta)$ is the range $A_\eta : \text{lin}\,\dot{H}_\eta \to L_2(P_\eta)$. Thus there is at most one solution $\tilde{\psi}_{P_\eta}$ that is contained in $\overline{\mathcal{R}(A_\eta)} = \overline{\text{lin}}\, A_\eta \dot{H}_\eta$.

If $\tilde{\chi}_\eta$ is contained in ~~in~~ the smaller range of $A_\eta^* A_\eta$ then equation $(*)$ can be solved & "$\tilde{\psi}_{P_\eta} = A_\eta (A_\eta^* A_\eta)^- \tilde{\chi}_\eta$"

$A_\eta^* A_\eta$ is called an information operator & $(A_\eta^* A_\eta)^-$ is a "generalized inverse". The following lemma shows that this attractive solution is available for any functional χ if the range of the score operator is closed, a situation which fails often (unfortunately).

Results from Functional Analysis

Lemma 1: Let $A: H_1 \to H_2$ be a continuous linear map between Hilbert spaces H_1 and H_2. Then the following are equivalent.

(i) $\mathcal{R}(A)$ is closed.

(ii) $\mathcal{R}(A^*)$ is closed.

(iii) $\mathcal{R}(A^*A)$ is closed.

(iv) $\mathcal{R}(A^*A) = \mathcal{R}(A^*)$.

Lemma 2: Let $A: H_1 \to H_2$ be a continuous linear map between two Hilbert spaces H_1 and H_2. Then

(i) $\mathcal{N}(A) = \mathcal{R}(A^*)^\perp$

(ii) $\mathcal{N}(A^*) = \mathcal{R}(A)^\perp$

Also, $A^*A: H_1 \to H_1$ is one-to-one and onto and has a continuous inverse iff A is ~~not~~ one-to-one and onto and $R(A)$ is closed iff A^*A is one-to-one and onto.

More generally: So far we have assumed that η is a probability distribution. However a close inspection of the arguments reveal that this is not necessary. More generally let $\mathcal{P} = \{P_\eta : \eta \in H\}$ where H is an arbitrary set. Let H_η be a subset of a Hilbert spaces that indexes "directions" b in which η can approximated within H. Suppose $\exists$ continuous linear map/operators $A_\eta: \text{lin } H_\eta \to L_2(P_\eta)$ and $\dot{\chi}_\eta: \text{lin } H_\eta \to \mathbb{R}$ & $\forall\, b \in H_\eta$ $\exists$ a path $t \to \eta_t$ s.t. as $t \downarrow 0$

$$\int \left[\frac{dP_{\eta_t}^{1/2} - dP_\eta^{1/2}}{t} - \frac{1}{2} A_\eta b \, dP_\eta^{\frac{1}{2}} \right]^2 \to 0$$

$$\Leftarrow \frac{\chi(\eta_t) - \chi(\eta)}{t} \to \dot{\chi}_\eta b$$

then by Reisz representation theorem, $\exists \tilde{\chi}_\eta \in \overline{\text{lin}}\, \dot{H}_\eta$ s.t. $\dot{\chi}_\eta b = \langle \tilde{\chi}_\eta, b \rangle$. Thereafter proceeding as before, we can obtain the following theorem.

Theorem: The map $\psi: \mathcal{P} \to \mathbb{R}$ given by $\psi(P_\eta) = \chi(\eta)$ is differentiable at P_η relative to the tangent set $A_\eta \dot{H}_\eta$ iff $\tilde{\chi}_\eta \in \mathcal{R}(A_\eta^*)$ for $A_\eta^*: L_2(P_\eta) \to \overline{\text{lin}}\, \dot{H}_\eta$.

Proof: By assumption $A_\eta \dot{H}_\eta$ is tangent set. The rest are similar arguments as before.

There are two reasons why one might fail to have a solution to (*)

$$A_\eta^* \tilde{\psi}_{P_\eta} = \tilde{\chi}_\eta \quad , \quad \tilde{\chi}_\eta \in \overline{\text{lin}}\, \dot{H}_\eta$$

$$A_\eta: \dot{H}_\eta \to L_2(P_\eta)$$

$$A_\eta^*: L_2(P_\eta) \to \overline{\text{lin}}\, \dot{H}_\eta$$

Firstly $\mathcal{R}(A^*) \subsetneq \overline{\text{lin}}\, \dot{H}_\eta$. Since $b \perp \mathcal{R}(A_\eta^*)$ iff $b \in \mathcal{N}(A_\eta)$, this can only happen when A_η is not $1\text{-}1$. This means two directions b may lead to same score for $A_\eta b$ $\Rightarrow$ Information matrix corresponding to two dimensional model is singular. An intuitive explanation is that the parameter is not locally identifiable & hence we have a problem.

Secondly, $\mathcal{R}(A^*)$ may be dense but not closed. Then for any $\tilde{\chi}_\eta$ $\exists$ elements in $\mathcal{R}(A_\eta^*)$ that are arbitrarily close to $\tilde{\chi}_\eta$ but (*) cannot be solved. Such situations can give rise to problematic situations.

Theorem: If $\tilde{\chi}_\eta \notin \mathcal{R}(A_\eta^*)$ then

(i) $\nexists$ estimators for $\chi(\eta)$ that are "regular" at P_η.

(ii) $\sup_{b \in \dot{H}_\eta} \frac{\langle \tilde{\chi}_\eta, b \rangle_\eta^2}{\| A_\eta b \|_{P_\eta}^2} = +\infty$ (Has implications in $\sqrt{n}$-consistent estimation).

Example: (Information loss Model)

Suppose that a typical observation is distributed as a measurable transformation $X = m(Y)$ of where Y is not observed. Suppose now that $Y \sim \eta$, $\eta \in H$. This yields a natural parametrization of the distribution of $X \sim P_\eta$. Assume the form of m is known.

There is cool property that this type of censoring mechanisms entail: if $t \to \eta_t$ is a differentiable submodel of H, then the induced submodel $t \to P_{\eta_t}$ is a differentiable of $\{P_\eta : \eta \in H\}$. Moreover, the score $g = A_\eta b$ for induced model $t \to P_{\eta_t}$ can be obtained from score b of the model $t \to \eta_t$ by taking conditional expectations.

$$A_\eta b(x) = \mathbb{E}_\eta(b(Y) \mid X = x).$$

Lemma: Suppose that $\{\eta_t : 0 < t < 1\}$ is a collection of probability measures on a measurable space $(\mathcal{Y}, \mathcal{B})$ s.t. for some measurable fn. $b : \mathcal{Y} \to \mathbb{R}$,

$$\int \left(\frac{d\eta_t^{1/2} - d\eta^{\frac{1}{2}}}{t} - \frac{1}{2} b \, d\eta^{\frac{1}{2}} \right)^2 \to 0$$

For a measurable map $m : \mathcal{Y} \to \mathcal{X}$ $(\mathcal{X}, \mathcal{A})$ let P_η be the distribution of $m(Y)$ if $Y \sim \eta$ & let $A_\eta b(x) = \mathbb{E}(b(Y) \mid m(Y) = x)$

Then

$$\int \left(\frac{dP_{\eta_t}^{\frac{1}{2}} - dP_\eta^{\frac{1}{2}}}{t} - \frac{1}{2} A_\eta b \, dP_\eta^{\frac{1}{2}} \right)^2 \to 0$$

Proof: Assume for the sake of notational simplicity that η_t and η have densities h_t and h w.r.t. some dominating prob. measure ν. (If this is not the case chose $\nu = \nu_t = \frac{1}{2}(\eta + \eta_t)$ and add an extra t-subscript throughout the following arguments).

Furthermore assume b is uniformly bdd by M (o.w. truncate b at $M_t \to \infty$ & add t^2's in the following chain of arguments accordingly).

Then $\int u_t^2 \nu \to 0$ where

$$u_t = \left(\frac{h_t^{\frac{1}{2}} - h^{\frac{1}{2}}}{t} - \frac{1}{2} b h^{\frac{1}{2}} \right)$$

Define μ to be the law of $X = m(Y)$ where Y has ν as its distribution. Then

$$p_t(x) = \mathbb{E}_\nu \left(h_t(Y) \mid X = x \right) \text{ & } p(x) = \mathbb{E}_\nu \left(h(Y) \mid X = x \right)$$

are densities of P_{η_t} and P_η w.r.t. μ. For P_η this follows because

$$\begin{aligned} P_\eta(A) = \int 1_A(m(y))\, d\eta(y) &= \mathbb{E}_\nu \left(1_A(m(Y))\, h(Y) \right) \\ &= \mathbb{E}_\nu \left(1_A(X)\, \mathbb{E}_\nu \left(h(Y) \mid X \right) \right) \\ &= \int_A p(x)\, d\mu(x) . \end{aligned}$$

By a similar argument we have a.s. under P_η

$$A_\eta b(X) p(X) = \mathbb{E}_\nu \left(b(Y) h(Y) \mid X \right)$$

Now, from definition of u_t,

$$h_t = h + t b h + t^2 u_t^2 + t \left(t u_t b h^{\frac{1}{2}} + 2 u_t h^{\frac{1}{2}} + \frac{1}{4} t b^2 h \right)$$

Evaluating these at Y and taking conditional expectations with respect to X,

$$p_t = p + t A_\eta b p + c_t + d_t$$

$$c_t(x) = t^2 \mathbb{E}_\nu \left(u_t^2(Y) \mid X \right)$$

$$d_t(x) = t\, \mathbb{E}_\nu \left(\left(t u_t b h^{\frac{1}{2}} + 2 u_t h^{\frac{1}{2}} + \frac{1}{4} t b^2 h \right)(Y) \mid X \right)$$

$$|d_t(x)|^2 = t^2 \mathbb{E}_\nu \left(\left(t u_t b h^{\frac{1}{2}} + 2 u_t h^{\frac{1}{2}} + \frac{1}{4} t b^2 h \right)(Y) | x \right)^2$$

$$\lesssim t^2 \mathbb{E}_\nu \left(\left(u_t h^{\frac{1}{2}} (tM+1) + t M^2 h \right)(Y) | x \right)$$

$$\lesssim t^2 \left(\mathbb{E}_\nu \left(u_t^2(Y) | x \right) (tM+1)^2 + t^2 M^4 p(x) \right) p(x) \quad \text{(by Cauchy Schwarz Inequality)}$$

Now using lemma 3 below we have

$$\left[\frac{P_t^{\frac{1}{2}} - P^{\frac{1}{2}}}{t} - \frac{1}{2} (A_\eta b) p^{\frac{1}{2}} \right]^2 \lesssim \mathbb{E}_\nu \left(u_t^2(Y) | x \right) (tM+1)^2 + t^2 M^4 p(x)$$
$$+ \mathbb{E}_\nu \left(u_t^2(Y) | x \right) + \left| \frac{1}{\sqrt{1-Mt}} - 1 \right|^2 M^2 p(x)$$

on the set $A = \{p > 0\}$.

The integral over the set A of the above for relative to ν converges to 0 as $t \to 0$.

Finally $\eta(m^{-1}(A^c)) = P_\eta(A^c) = 0 \Rightarrow P_{\eta_t}(A^c) = \eta_t(m^{-1}(A^c)) = o(t^2)$ because $\nu(t u_t^{-1} B) = \eta_t(B)$ if $\eta(B) = 0$. Thus the integral over A^c converges to 0 as well.

lemma 3: For $a, b, c, d \in \mathbb{R}$ with $a > 0$, $b/a \le \varepsilon < 1$, $c \ge 0$ and $a + b + c + d \ge 0$,

$$\left| \sqrt{a+b+c+d} - \sqrt{a} - \frac{1}{2} \frac{b}{\sqrt{a}} \right| \le \frac{3d^2}{a(1-\varepsilon)} + 3c + \left| \frac{1}{\sqrt{1-\varepsilon}} - 1 \right|^2 \frac{b^2}{a}.$$

□

Now view $A_\eta : L_2(\eta) \to L_2(P_\eta)$. To get its adjoint note

$$\langle g, A_\eta b \rangle_{L_2(P_\eta)} = \mathbb{E}\left(g(X) A_\eta b(X) \right)$$
$$= \mathbb{E}\left(g(X) \mathbb{E}_\eta (b(Y)|X) \right)$$
$$= \mathbb{E}\left(g(X) b(Y) \right)$$
$$= \mathbb{E}\left(\mathbb{E}_\eta (g(X)|Y)\, b(Y) \right)$$
$$= \langle A_\eta^* g, b \rangle_{L_2(\eta)}$$

$A_\eta^* g = \mathbb{E}_\eta (g(X) | Y = y) : L_2(P_\eta) \to L_2(\eta)$

Although the score operator is defined for all $L_2(\eta)$ we need its restriction to $\overline{\text{lin}}\, H_\eta$. Hence forth we need to project the conditional expectation $\mathbb{E}_\eta(g(X)|Y=y)$ onto $\overline{\text{lin}}\, H_\eta$.

(This ends our discussion on existence of first order influence functions)

Morally Correct Calculations of Influence Functions

We work with an example, that of estimation of a treatment effect functional.

Suppose your typical observation consists of (Y, A, X) where $A \in \{0,1\}$ corresponds to a treatment, $Y \in \mathbb{R}$ is an outcome and $X \in [0,1]$ is a covariate. We want to study the treatment effect of A on Y. One common way of quantifying this through the "variance weighted treatment effect"

$$\tau = \frac{\mathbb{E}(\text{var}(A|X)\, c(X))}{\mathbb{E}(\text{var}(A|X))}$$

where $c(X) = \mathbb{E}(Y|A=1,X) - \mathbb{E}(Y|A=0,X)$ is treatment effect in the stratum of X under no unmeasured confounding. While considering fixed treatment effect $c(x) = \varphi^*$ for all x, $\tau = \varphi^*$, justifying its value as a marker for treatment effect. τ also φ^* under a commonly studied model in this area: $\mathbb{E}(Y|A,X) = \varphi^* A + b(X)$.

How to estimate τ?

Note that $\tau = \dfrac{\mathbb{E}(\text{var}(A|X)\, c(X))}{\mathbb{E}(\text{var}(A|X))} = \dfrac{\mathbb{E}\,\text{cov}(Y,A|X)}{\mathbb{E}\,\text{var}(A|X)}$

To see this note that

$$\begin{aligned}\text{cov}(Y,A|X) &= \mathbb{E}(AY|X) - \mathbb{E}(A|X)\,\mathbb{E}(Y|X) \\ &= \left(\mathbb{E}(Y|A=1,X) - \mathbb{E}(Y|X)\right)\mathbb{E}(A|X)\end{aligned}$$

Similarly,

$$\begin{aligned}\text{cov}(Y,1-A|X) &= \left(\mathbb{E}(Y|A=0,X) - \mathbb{E}(Y|X)\right)\mathbb{E}(1-A|X) \\ &= \left(\mathbb{E}(Y|A=1,X) - c(X) - \mathbb{E}(Y|X)\right)\left(1 - \mathbb{E}(A|X)\right)\end{aligned}$$

$$= \frac{\operatorname{cov}(Y, A|X)}{\mathbb{E}(A|X)}\left(1-\mathbb{E}(A|X)\right) - c(X)\left(1-\mathbb{E}(A|X)\right)$$

$$= \frac{\operatorname{cov}(Y, A|X)\left(1-\mathbb{E}(A|X)\right) - c(X)\operatorname{var}(A|X)}{\mathbb{E}(A|X)}$$

$$\Rightarrow \operatorname{cov}(Y, 1-A|X)\,\mathbb{E}(A|X) = \operatorname{cov}(Y, A|X)\left(1-\mathbb{E}(A|X)\right) - c(X)\operatorname{var}(A|X)$$

$$\Rightarrow -\operatorname{cov}(Y, A|X)\,\mathbb{E}(A|X) = \operatorname{cov}(Y, A|X)\left(1-\mathbb{E}(A|X)\right) - c(X)\operatorname{var}(A|X)$$

$$\Rightarrow \operatorname{cov}(Y, A|X) = c(X)\operatorname{var}(A|X).$$

Therefore its enough to understand estimation of $\mathbb{E}\operatorname{cov}(Y, A|X)$. (the estimation of $\mathbb{E}\operatorname{var}(A|X)$ is similar with "$Y = A$").

But $\mathbb{E}\operatorname{cov}(Y, A|X) = \mathbb{E}\left(\mathbb{E}(AY|X) - \mathbb{E}(A|X)\,\mathbb{E}(Y|X)\right)$

$$= \mathbb{E}(AY) - \mathbb{E}\left(\mathbb{E}(A|X)\,\mathbb{E}(Y|X)\right)$$

now $\mathbb{E}(AY)$ can be efficiently estimated by $\frac{1}{n}\sum_{i=1}^{n} A_i Y_i$.

We therefore focus on the estimation of

$$\mathbb{E}\left(\mathbb{E}(A|X)\,\mathbb{E}(Y|X)\right) = \int a(x)\,b(x)\,g(x)\,dx$$

where $a(x) = \mathbb{E}(A|X=x)$

$b(x) = \mathbb{E}(Y|X=x)$

$g(x) \leftarrow$ prob. density of X on $[0,1]$ w.r.t λ.

An influence function for $\Psi(P_\eta) = \chi(\eta)$ where $P_\eta = P_{(a,b,c,g)}$, $\eta = (a,b,c,g)$ and $\chi(\eta) = \int abg$ will operationalize our Von-Mises calculus. In particular the likelihood of the model indexed by (a,b,c,g) is given by (at a single obs. (A,Y,Z)) if $Y \in \{0,1\}$

$$a(X)^A\,(1-a(X))^{1-A}\,(c(1-a)+b)(X)^{AY}\,(1-c(1-a)-b)(X)^{(1-Y)A} \times (-ca+b)(X)^{Y(1-A)}\,(1+ca-b)(X)^{(1-Y)(1-A)} = p_\eta(Y,A,X) \quad \text{(say)}$$

Then an influence function of $\Psi(P_\eta) = \chi(\eta)$ at P_η is given by $(Y - b(X))(A - a(X)) - \chi(\eta)$, w.r.t. the tangent space given by perturbations in given directions. Namely, we consider submodels (one-dimensional) in given directions $a', b', c',$ and g' as follows:

$$a_t = a + t\,a'$$
$$b_t = b + t\,b'$$
$$c_t = c + t\,c'$$
$$g_t = g(1 + t\,g'), \quad \text{※ } \int g'g = 0$$

Then one can calculate the corresponding likelihood $p_{\eta_t} = p_{(a_t, b_t, c_t, g_t)}$ and the corresponding score in the given direction (a', b', c', g') is calculated by $\frac{d}{dt} \log p_{\eta_t}\big|_{t=0} := s_\eta(a', b', c', g')$ (say). For any such score one can verify by tedious but straightforward calculations that

$$\frac{d}{dt}\chi(\eta_t)\Big|_{t=0} = \frac{d}{dt}\int a_t b_t g_t \Big|_{t=0}$$
$$= \mathbb{E}_\eta\left(\left\{(Y - b(X))(A - a(X)) - \chi(\eta)\right\} s_\eta(a', b', c', g')\right)$$

This then implies that $(Y - b(X))(A - a(X)) - \chi(\eta)$ is an influence function at P_η w.r.t the tangent space obtained above.

Remark: Indeed to justify that s_η is a score and $(Y - b(X))(A - a(X)) - \chi(\eta)$ is a valid influence function one needs to be more rigorous. However, we will see that this heuristic and morally correct calculation suffices for us since all we care about is obtaining an estimator of $\chi(\eta)$ and correct its bias suitably. However for systematic handling of the procedure one definitely needs to pin down the gory details.

Now one can check by tedious but straightforward calculations that

$$\frac{d}{dt}\chi(\eta_t)\Big|_{t=0} = \frac{d}{dt}\chi\left((a_t, b_t, c_t, g_t)\right)\Big|_{t=0}$$

$$= \frac{d}{dt}\int a_t b_t g_t\Big|_{t=0} = \mathbb{E}_\eta\left(\{(Y-b(X))(A-a(X))\}B_\eta(a', b', c', \varphi)\right)$$

$-\chi(\eta)$ ← influence function.

Therefore $(Y-b(X))(A-a(X))-\chi(\eta)$ is indeed a ~~some~~ w.r.t the tangent space generated by these submodels.

Therefore our estimator in first order is

$$\hat{\chi}_{n,1} = \chi(\hat{\eta}) + \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{b}(X_i))(A_i - \hat{a}(X_i)) - \chi(\hat{\eta})$$

where $\chi(\hat{\eta}) = \int \hat{a}\hat{b}\hat{g}$ and $\hat{a}, \hat{b}, \hat{g}$ are estimated from $X_{n+1}, \ldots, X_{2n}$.

Therefore $\hat{\chi}_{n,1} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{b}(X_i))(A_i - \hat{a}(X_i))$

Let's understand the bias of this estimator. By direct calculations,

$$\mathbb{E}_{\eta,1}(\hat{\chi}_{n,1}) - \chi(\eta) = \int(\hat{a}-a)(\hat{b}-b)g$$

where $\mathbb{E}_{\eta,1}$ denotes expectation with second half of the sample kept fixed.

Our second order Von-Mises expansion aims to correct for this bias. Indeed if $\exists$ a kernel $K(x,y)$ s.t. $\int h(x)K(x,y)\,dy = h(y)$ a.e. g and all $h \in L_2[0,1]$ we can correct this bias by considering

$$-\frac{1}{n(n-1)}\sum_{i_1 \neq i_2}\frac{(Y_{i_1} - \hat{b}(X_{i_1}))}{\sqrt{\hat{g}(X_{i_1})}}K(X_{i_1}, X_{i_2})\frac{(A_{i_2} - \hat{a}(X_{i_2}))}{\sqrt{\hat{g}(X_{i_2})}}$$

because this has expectation $-\int(b-\hat{b})(a-\hat{a})g$.

But $\nexists$ such a kernel ~~with~~ as an element of $L_2([0,1]\times[0,1])$. So we work with partial representation.

That is we correct our bias using projection kernels onto finite but growing with n dimensional spaces. In particular, recall our definition of $V_j = \{\phi_{J_0 k}, \psi_{\ell k} : \ell \geq J_0, k \in \mathbb{Z}_\ell\}$ to be the linear span of the first few basis functions from an MRA generalized by (ϕ, ψ). Then our bias corrected second order estimator becomes:

$$\hat{\chi}_{n,2} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{b}(X_i))(A_i - \hat{a}(X_i)) - \frac{1}{n(n-1)}\sum_{i_1 \neq i_2} \frac{(Y_{i_1} - \hat{b}(X_{i_1}))}{\hat{g}(X_{i_1})^{\frac{1}{2}}} K_{V_j}(X_{i_1}, X_{i_2}) \frac{(A_{i_2} - \hat{a}(X_{i_2}))}{\hat{g}(X_{i_2})^{\frac{1}{2}}}$$

We next show that $\hat{\chi}_{n,2}$ has desired bias and variance properties.

Bias: $\left| \mathbb{E}_\eta(\hat{\chi}_{n,2}) - \chi(\eta) \right|$

$$= \left| \mathbb{E}_{\eta,2} \int (b - \hat{b})(x)(a - \hat{a})(x) g(x) - \mathbb{E}_\eta \left\{ \frac{(Y_1 - \hat{b}(X_1))}{\hat{g}(X_1)^{\frac{1}{2}}} K_{V_j}(X_1, X_2) \frac{(A_2 - \hat{a}(X_2))}{\hat{g}(X_2)^{\frac{1}{2}}} \right\} \right|$$

For notational brevity, put $\delta b = b - \hat{b}$, $\delta a = a - \hat{a}$, $\delta g = g - \hat{g}$.

$$\Rightarrow \mathbb{E}_\eta\left(\frac{(Y_1 - \hat{b}(X_1))}{\hat{g}(X_1)^{\frac{1}{2}}} K_{V_j}(X_1, X_2) \frac{(A_2 - \hat{a}(X_2))}{\hat{g}(X_2)^{\frac{1}{2}}} \right)$$

$$= \mathbb{E}_{\eta,2}\left\{ \iint \frac{\delta b(x_1) g(x_1)}{\sqrt{\hat{g}(x_1)}} K_{V_j}(x_1, x_2) \frac{\delta a(x_2) g(x_2)}{\sqrt{\hat{g}(x_2)}} dx_1 dx_2 \right\}$$

$$= \mathbb{E}_{\eta,2} \int \frac{\delta b(x_1) g(x_1)}{\sqrt{\hat{g}(x_1)}} \Pi_{V_j}\left(\frac{\delta a\, g}{\sqrt{\hat{g}}} \right)(x_1)\, dx_1$$

$$= \mathbb{E}_{\eta,2} \int \frac{\delta b(x_1) \delta a(x_2) g(x_1) g(x_2)}{\sqrt{\hat{g}(x_1)}\sqrt{\hat{g}(x_2)}} dx_1 - \mathbb{E}_{\eta,2} \int \frac{\delta b(x_1) g(x_1)}{\sqrt{\hat{g}(x_1)}} \Pi_{V_j}^{-1}\left(\frac{\delta a\, g}{\sqrt{\hat{g}}} \right)(x_1) dx_1$$

$$= \mathbb{E}_{\eta,2} \int \delta b(x_1) \delta a(x_2) g(x_1)\, dx_1 + \mathbb{E}_{\eta,2} \int \delta b(x_1) \delta a(x_2) g^2(x_1) \left(\frac{1}{\hat{g}(x_1)} - \frac{1}{g(x_1)} \right) dx_1$$

$$- \mathbb{E}_{\eta,2} \int \frac{\delta b(x_1) g(x_1)}{\sqrt{\hat{g}(x_1)}} \Pi_{V_j}^{-1}\left(\frac{\delta a\, g}{\sqrt{\hat{g}}} \right)(x_1) dx_1$$

$$= \mathbb{E}_{\eta,2}\int \delta b\,\delta a\, g \;+\; \mathbb{E}_{\eta,2}\int \delta b\,\delta a\,\delta g\;\frac{g}{\hat g}$$

$$-\mathbb{E}_{\eta,2}\int \Pi_{V_j^\perp}\left(\frac{\delta b\, g}{\sqrt{\hat g}}\right)\Pi_{V_j^\perp}\left(\frac{\delta a\, g}{\sqrt{\hat g}}\right)$$

$\therefore$ Bias: $\left|\mathbb{E}_{\eta,2}\int \delta b\,\delta a\,\delta g\,\frac{g}{\hat g} - \mathbb{E}_{\eta,2}\int \Pi_{V_j^\perp}\left(\frac{\delta b\, g}{\sqrt{\hat g}}\right)\Pi_{V_j^\perp}\left(\frac{\delta a\, g}{\sqrt{\hat g}}\right)\right|$

$\le$  $\le \mathbb{E}_{\eta,2}\left(\left\|\frac{g}{\hat g}\right\|_\infty \|\delta g\|_2 \|\delta a\|_4 \|\delta b\|_4\right)$

$$-\mathbb{E}_{\eta,2}\left\|\Pi_{V_j^\perp}\left(\frac{\delta b\, g}{\sqrt{\hat g}}\right)\right\|_2 \left\|\Pi_{V_j^\perp}\left(\frac{\delta a\, g}{\sqrt{\hat g}}\right)\right\|_2$$

Now its time to assume some regularity properties on a, b, g. In particular assume $a \in H(\alpha, M)$, $b \in H(\beta, M)$, and $g \in H(\nu, M)$ on $[0,1]$ where $\nu > 2\max\{\alpha, \beta\}$. Also assume that $\|a\|_\infty \vee \|b\|_\infty \vee \|g\|_\infty \le B_0$ and $g(x) \ge B_0 > 0$ for all x. Then we have estimators $\hat a, \hat b, \hat g$ of a, b, g s.t. the following hold:

(i) $\mathbb{E}_\eta \|\delta a\|_\infty \le c^*\left(\frac{n}{\log n}\right)^{-\frac{\alpha}{2\alpha+1}}$, $\mathbb{E}_\eta \|\delta b\|_\infty \le c^*\left(\frac{n}{\log n}\right)^{-\frac{\beta}{2\beta+1}}$,

$\mathbb{E}_\eta\left(\|\delta g\|_\infty\right) \le c^*\left(\frac{n}{\log n}\right)^{-\frac{\nu}{2\nu+1}}$ $\forall\, \eta \in \mathcal{P}(\alpha,\beta,\nu)$ for $n \ge n^*$

(ii) $\frac{B_0}{2} \le \hat g(x) \le 2B_0$ for all x. $\forall\, \eta \in \mathcal{P}(\alpha,\beta,\nu)$, for $n \ge n^*$

(iii) $\sup_{\eta \in \mathcal{P}(\alpha,\beta,\nu)} \mathbb{P}_\eta\left(\hat a \in H(\alpha, c^{**})\right) \ge 1 - \frac{1}{n^3}$

$\sup_{\eta \in \mathcal{P}(\alpha,\beta,\nu)} \mathbb{P}_\eta\left(\hat b \in H(\beta, c^{**})\right) \ge 1 - \frac{1}{n^3}$ $\quad \forall\, n \ge n^*$

$\sup_{\eta \in \mathcal{P}(\alpha,\beta,\nu)} \mathbb{P}_\eta\left(\hat g \in H(\nu, c^{**})\right) \ge 1 - \frac{1}{n^3}$

where c^*, c^{**}, n^* depends on B_0 and M, and

$$\mathcal{P}(\alpha,\beta,\nu) = \left\{(a,b,c,g) : \|a\|_\infty \vee \|b\|_\infty \vee \|g\|_\infty \le B_0,\ g(x) \ge B_0\ \forall x \in [0,1],\ a \in H(\alpha,M),\ b \in H(\beta,M),\ g \in H(\nu,M),\ \int g = 1\right\}$$

This implies that,

$$|\text{Bias}| \le \underset{\sim}{C}(B_U, M)\left(\frac{n}{\log n}\right)^{-\left(\frac{\alpha}{2\alpha+1}+\frac{\beta}{2\beta+1}+\frac{\nu}{2\nu+1}\right)} + C(B_U, B_L, M)\left\{2^{-2j\alpha}+\frac{1}{n^3}\right\}^{\frac{1}{2}}\left\{2^{-2j\beta}+\frac{1}{n^3}\right\}^{\frac{1}{2}}$$

$$\le C(B_U, B_L, M)\left\{n^{-\frac{2\alpha+2\beta}{2\alpha+2\beta+1}} + \left(2^{-2j\alpha}+\frac{1}{n^3}\right)^{\frac{1}{2}}\left(2^{-2j\beta}+\frac{1}{n^3}\right)^{\frac{1}{2}}\right\} \quad (\text{since } \nu > 2(\alpha\vee\beta))$$

$\blacklozenge$

Now we go the variance calculation.

We additionally assumed $\|a\|_\infty \vee \|b\|_\infty \le B_U$. (provided we have (Φ, Ψ) with sufficient regularity & compact support)

This implies that we can add an extra property to our list (i) - (iii) on page (21) and obtain estimates $\hat{a}, \hat{b}$ which satisfy $\sup_{\eta \in \mathcal{P}(\alpha,\beta,\nu)} \mathbb{P}_\eta\left(\max(\|\hat{a}\|_\alpha, \|\hat{b}\|_\infty) \le 2B_U\right) \ge 1 - \frac{1}{n^3} \quad \forall\, n \ge n^*$.

If finally we assume $|Y| \le B_U$ a.s. (which implies $\|b\|_{\mathcal{B}} \le B_U$ a.s.)

we can do our variance calculation by standard ~~bias variance~~ Hoeffding's decomposition and get

$$\text{Var}(\hat{\chi}_{n,2}) \le C(B_U)\left\{\frac{1}{n} + \frac{2^j}{n^2}\right\} \quad (\text{check!})$$

The right bias variance is then obtained by $2^j \sim n$ if $\frac{\alpha+\beta}{2} \ge \frac{1}{4}$ and $2^j \sim n^{\frac{2}{2\alpha+2\beta+1}}$ if $\frac{\alpha+\beta}{2} < \frac{1}{4}$

This gives us the following result.

Theorem: Assume $|Y| \le B_U$ a.s. and ~~$|X| \le 1$~~ $|A| \le B_U$ a.s.

Then $\exists$ an estimator $\hat{\chi}_n$ of $\Psi(P_\eta) = \mathbb{E}(W(A, Y|X))$

s.t. if $\nu > 2\max\{\alpha, \beta\}$ and $n \ge n^*$ (depending on M, B_U, B_L)

~~sup~~ $\mathbb{E}_\eta\left(\hat{\chi}_n - \Psi(P_\eta)\right)^2 \le C(B_U, B_L, M)\, n^{-\frac{4\alpha+4\beta}{2\alpha+2\beta+1}}$

$\eta \in \mathcal{P}(\alpha,\beta,\nu)$

$$\left\{= C(B_U, B_L, M)\, n^{-\frac{8\delta_{av}}{4\delta_{av}+1}}\right\}$$

where $\delta_{av} = \frac{\alpha+\beta}{2} \le \frac{1}{4}$

$$= C(B_U, B_L, M)\frac{1}{n} \quad \delta_{av} \ge \frac{1}{4}.$$