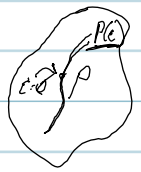
9/8/2011 * We want estimators w/ bias $< \frac{1}{\sqrt{n}}$

For a r.v. $O \sim P_0 \in \mathcal{M}$, we wish to learn about $\Psi: \mathcal{M} \to \mathbb{R}$

Ψ_0 = true value of parm of interest

parm is pathwise differentiable at P =>



Section A.4 Cramer-Rao

↓

we want it differentiable at every P

↳ needs to be within model

For class of 1-dim submodel (generates rich enough ~~space~~ class)

$$\left\{ \frac{P(\varepsilon) : \varepsilon \in (-\delta, \delta)}{P(\varepsilon = 0) = P} \right\} \subset \mathcal{M}$$

* Want to choose as rich a class of submodels as possible

Pathwise derivative: $\frac{\partial}{\partial \varepsilon} \Psi(P(\varepsilon))\Big|_{\varepsilon=0} = \lim_{\varepsilon \to 0} \frac{\Psi(P(\varepsilon)) - \Psi(P)}{\varepsilon}$ (we want this to exist for every path)

$S(O)$ = Score of parm model = $\frac{\partial / \partial \varepsilon P(\varepsilon)}{P}\Big|_{\varepsilon=0}$

↓ if exists

$= E_P[D(P)(O)\, S(O)]$ for some $D(P)$

Hilbert space: $L_0^2(P), \langle h_1, h_2 \rangle = E_0[h_1(O), h_2(O)]$

linear space $= \mathrm{cov}_P(h_1, h_2)$

we can define norm $\|h\| \equiv \sqrt{\langle h_1, h_2 \rangle}$

$\frac{\partial}{\partial \varepsilon} \Psi(P(\varepsilon))$ can be thought of as mapping $L_0^2(P)$ to real space $\mathbb{R}$

$A: \underbrace{L_0^2(P)}_{\text{Hilbert space}} \to \mathbb{R}$

A is linear bounded $(A(h_1, h_2) = A(h_1) + A(h_2))$

$\Rightarrow A(h) = \langle O, h \rangle = E_P[D(O), h(O)]$

aka Riesz-Fréchet theorem

Ex. Say $\mathcal{M}$ = NP model

O = real valued

$\Psi(P) = \int O\, P(O)\, \partial M(O)$

$P(\varepsilon) = (1 + \varepsilon h) P$, $\begin{cases} E_P h(O) = 0 \\ \|h\|_\infty < \infty \end{cases}$

↑ is submodel

↗ Cont.

cont. $\frac{\partial}{\partial \varepsilon}\Psi(P(\varepsilon)) = \frac{\partial}{\partial \varepsilon}\int O(1+\varepsilon h(O))P(O)\,\partial\mu(O)$

$= \int O \cdot h(O) P(O)\,\mu(O)$

$= \langle f, h\rangle$ ← needs to have mean 0

$= \langle \underbrace{f - E_P(f)}_{D(P)}, h\rangle$

$\Rightarrow$ inner product of $D(P)$ and h (the score)

Note: $D(P) \in L_0^2(P)$

↳ Hilbert space

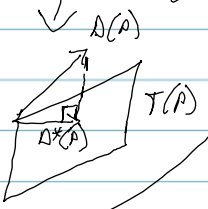
Let $T(P)$ be the tangent space of this class of submodels through P

= closure of the linear span of all scores

$\frac{\partial}{\partial \varepsilon}\Psi(P(\varepsilon))\big|_{\varepsilon=0} = E_P(D^*(P)(O)S(O))$ for a unique gradient $D^* \in T(P)$

Note: $\forall$ gradient D, $\pi[D \mid T(P)] = D^*$

i.e. projection of D onto a tangent space.



called the canonical gradient

↓

also the efficient influence curve

Def^n Projection: $\pi[D \mid T(P)]$ is defined to be ① $\pi(D \mid T(P)) \in T(P)$

② $D - \pi(D \mid T(P)) \perp T(P)$

i.e. $\langle D - \pi(D \mid T(P)), h\rangle = 0 \quad \forall\, h \in T(P)$

Ex on $O = (W, A, Y)$ projection

$E[h(W, a, Y) \mid W] = \pi(h \mid T_W)$

$\{$"$h(W)\cdot h \in L_0^2(P)\}$

$\langle h(W,a,Y) - E(h \mid W), h(W)\rangle$

$= E[(h(W,a,Y) - E(h \mid W), h(W)] = 0$

$= E[h(W,a,Y)h_1(W)) - E(E(h \mid W)h_1(W)] = 0$

* the efficient influence curve is not influenced in any way by placing restrictions on global space.

Generalized Cramer-Rao lower bound

Assume $O \sim P(\varepsilon_0) \in \{P(\varepsilon) : \varepsilon\}$ (Note: $\varepsilon_0 = 0$, don't know this yet)

Target parm: $\Psi(P(\varepsilon_0))$

Cr-Rao: $\frac{\left(\frac{\partial}{\partial \varepsilon_0}\Psi(P(\varepsilon_0))\right)^2}{I(\varepsilon_0) = E[S^2(O)]} = \frac{[E_P(D^*(P)(O)S(O))]^2}{\underbrace{E_P[S^2(O)]}_{\text{Var score}}}$

Generalized Cr-R lower bound: $\sup_{S \in T(P)} \frac{(E_P D^*(P) S(O))^2}{E_P S^2(O)} \leq \sup_{S \in T(P)} \frac{E(D^{*2}) \cdot E(S^2)}{E(S^2)}$

↳ least favorable submodel

$= E(D^*(P)^2(O)) = \boxed{\mathrm{Var}(D^*)}$

Cauchy-Schwarz inequality: $E_P[D^* S] \leq \sqrt{E(D^{*2})}\sqrt{E(S^2)}$

ie. $\langle D^*, S \rangle \leq \|D\| \cdot \|S\|$

↳ inner product.

Note: Can choose score $S(O) = D^*(P)$

ie. equal to canonical gradient

Example:

$O = (W, A, Y) \sim P_0 \in \mathcal{M} = NP$ model

$\Psi(P) = E_P E_P[Y|A=1, W] = \sum_w \sum_y y\, P(Y|A=1, W) P(W)$

Recall: $P(w, a, y) = P(w) P(a|w) P(y|a, w)$

For pathwise deriv.: need: fluctuations

$P_\varepsilon(w) = (1 + \varepsilon S_W(w)) P(w)$, $S_W(w)$ has mean 0

$\sum_w (S_W(w)) P(w) = 0$

$P_\varepsilon(a|w) = (1 + \varepsilon S_W(a|w)) P(a|w)$, $E(S_A(A|W)) = 0$

$P_\varepsilon(y|a, w) = (1 + \varepsilon S_Y(y|a, w)) P(y|a, w)$, $E(S_Y(Y|A, W)|A, W) = 0$

$\frac{\partial}{\partial \varepsilon} \Psi($