9/15/2011

Pathwise diff & canonical gradient (est eff. only if IC = canon grad.)

Thm: Suppose $O \sim P \in \mathcal{M}$

$\rightarrow P = Q(P)\,g(P)$, where Q, g are $\underbrace{\text{variation indep.}}$

$\Psi: \mathcal{M} \rightarrow \mathbb{R}$, $\Psi(P) = \Psi(Q(P))$

$\rightarrow \mathcal{M} = \{Q \cdot g : Q \in \mathcal{Q}, g \in G\}$

* We have to put modeling efforts separately on each parm.

Class of paths: $\left.\begin{matrix} \{Q(\varepsilon):\varepsilon\} \subset \mathcal{Q} \\ \{g(\varepsilon):\varepsilon\} \subset G \end{matrix}\right\}$ i.e. submodels

Scores for Q: $\frac{\partial}{\partial\varepsilon}\log[Q(\varepsilon)g]\big|_{\varepsilon=0} \xrightarrow{\text{generates}} T_Q$ = tangent space for Q

Scores for g: $\frac{\partial}{\partial\varepsilon}\log[g(\varepsilon)Q]\big|_{\varepsilon=0} \longrightarrow T_g$ = " " " g

$T_Q \perp T_g$ in $L_0^2(P)$, $\langle h_1, h_2\rangle = E_P[h_1(O)h_2(O)]$

$$s_1 = \frac{\frac{\partial}{\partial\varepsilon}Q(\varepsilon)\big|_{\varepsilon=0}}{Q} \in T_Q$$

$$s_2 = \frac{\frac{\partial}{\partial\varepsilon}g(\varepsilon)\big|_{\varepsilon=0}}{g} \in T_g$$

$$E_P[s_1(O)s_2(O)] = \int \frac{\frac{\partial Q(\varepsilon)}{\partial\varepsilon}\big|_{\varepsilon=0}}{\cancel{Q}} \frac{\frac{\partial g(\varepsilon)}{\partial\varepsilon}\big|_{\varepsilon=0}}{\cancel{g}} \cancel{Q(\varepsilon)g(\varepsilon)}\, d\mu(O)$$

$$= \int \frac{\partial}{\partial\varepsilon}Q(\varepsilon)\big|_{\varepsilon=0} \frac{\partial}{\partial\varepsilon}g(\varepsilon)\big|_{\varepsilon=0}\, \partial\mu(O)$$

$$\int \frac{\partial}{\partial\varepsilon}Q(\varepsilon)\big|_{\varepsilon=0}\, g(\varepsilon_1)\,\partial\mathcal{M}(O) = \frac{\partial}{\partial\varepsilon}\underbrace{\int Q(\varepsilon)\,g(\varepsilon_1)\,\partial\mathcal{M}(O)}_{\text{pdf} \Rightarrow 1}\Big|_{\varepsilon=0}$$

$$= \frac{\partial}{\partial\varepsilon} 1 \;\cancel{\partial\mathcal{M}(O)}$$

$$= 0$$

Likewise, $\frac{\partial}{\partial\varepsilon_1}\int \frac{\partial}{\partial\varepsilon}Q(\varepsilon)\big|_{\varepsilon=0}\, g(\varepsilon_1)\,\partial\mathcal{M}(O)\big|_{\varepsilon_1=0} = 0$

$$= \int \frac{\partial}{\partial\varepsilon}Q(\varepsilon)\big|_{\varepsilon=0} \frac{\partial}{\partial\varepsilon_1}g(\varepsilon_1)\big|_{\varepsilon_1=0}\,\partial\mathcal{M}(O)$$

should just call this g.

If $D^*(P)$ is a canonical gradient of the parm $\Psi: \mathcal{M}(g_0) \rightarrow \mathbb{R}$, g known
i.e. $\{Q g_0 : Q \in \mathcal{Q}\}$

then $D^*(P)$ is also the canonical gradient of $\Psi: \mathcal{M} \rightarrow \mathbb{R}$ at P

i.e. knowing g means Ψ is same as if whole model is known

Proof: $\frac{\partial}{\partial\varepsilon}\Psi(Q(\varepsilon))\big|_{\varepsilon=0} = E_P[D^*(P)(O)\,s_1(O)]$, $s_1 = \frac{\frac{\partial}{\partial\varepsilon}Q(\varepsilon)|_{\varepsilon=0}}{Q}$

$\frac{\partial}{\partial\varepsilon}\Psi(P(\varepsilon))\big|_{\varepsilon=0} = \frac{\partial}{\partial\varepsilon}\Psi(Q(\varepsilon))\big|_{\varepsilon=0}$ (b/c variation of g doesn't matter)

$= E_P[D^*(P)(O)\,s_1(O)]$, $s_2 = \frac{\partial/\partial\varepsilon\, g(\varepsilon)|_{\varepsilon=0}}{g}$

element of T_Q

* Want to write func. as sum of $s_1 + s_2$

$= E_P[D^*(P)(O)(s_1+s_2)(O)]$, b/c inner prod of D^* and $s_2 = 0$

$\therefore D^*(P)$ is a gradient $\hookrightarrow$ cont.

cont.

Tangent space of model is $T = T_Q \oplus T_g$

Also, $D^*(P) \in T = T_Q \oplus T_g$

$\therefore$ Done; have shown its a canonical gradient & element of tangent space.

Application in ex

$O = (W, A, Y) \sim P = P_W \underbrace{P_{A|W}}_{\text{aka } g_A} P_{Y|A,W} = \underbrace{(P_W P_{Y|A,W})}_{Q} \cdot g$

$\Psi : \mathcal{M} \to \mathbb{R}$.

$\Psi(P) = \sum_w \sum_y y P_Y(y|A=1,w) \cdot P_W(w) = \Psi(Q)$

Say $\underbrace{(P_W, P_Y) \text{ are NP and } g \in G}_{\text{defines our model } \mathcal{M}}$

$\Rightarrow$ By the prev theorem, $T_Q \perp T_g$

$T_Q = T_W \oplus T_Y$, where $T_W = \{s_W(w) : E_P(s_W(w)) = 0\}$

$T_Y = \{S_Y(y|A,w) : \underbrace{E_P(S(y|A,w)|A,}_{=0}$

Define $D(P)(O) = Y \frac{I(A=1)}{g(A|W)} - \Psi(P)$ → is a gradient in our model $\mathcal{M}(g)$

↳ then prev. theorem says,

$D^* = \underbrace{\pi[D|T_Q]}_{\text{project } D \text{ on } T_Q}$ is the canonical gradient of $\Psi : \mathcal{M} \to \mathbb{R}$

Question: Why is D a gradient of $\Psi : \mathcal{M}(g) \to \mathbb{R}$?

answer ↳ $\frac{\partial}{\partial \varepsilon} \Psi(P(\varepsilon))|_{\varepsilon=0} = \frac{\partial}{\partial \varepsilon} \sum_W \left(\sum_y y (1 + \varepsilon s_Y) P(y|A=1,w) \right) (1 + \varepsilon s_W(w)) P_W(w)_{\varepsilon=0}$

$= \sum_w \sum_y y s_Y(y|1,w) P(y|A=1,w) P_W(w) + \sum_w \sum_y y P(Y|A=1,w) s_W(w) P_W(w)$

Goal: $E[D(s_W + s_Y)]$

$\stackrel{?}{=} \sum_w \sum_a \sum_y y \frac{I(a=1)}{g(A|w)} s_Y(y|a,w) P_Y(y|a,w) P_W(w)\, g(a|w)$

$+ \sum_w \sum_a \sum_y y \frac{I(a=1)}{g(a|w)} s_W(w) P(Y|A=1,w) P_W(w)\, g(a|w)$

$= E\left[Y \frac{I(a=1)}{g(A|W)} S_Y(Y|A,W)\right] + E\left[Y \frac{I(A=1)}{g(A,W)} S_W(W)\right]$

$= E\left[Y \frac{I(A=1)}{g(1,W)} [S_Y + S_W]\right]$

↳ now need to center

subtracting Ψ does nothing to the expectation

$\Rightarrow D(O) = \underbrace{Y \frac{I(A=1)}{g(a|w)} - \Psi(P)}_{\text{is a gradient}}$

↳

cont. Now, will find canonical grad by proj. D on T_Q

Lemma: for $D(x)$, $\mathcal{H} = \{f(Q(x)) : f\}$

$$\pi[D \mid \mathcal{H}] = E[D(x) \mid Q(x)]$$

* we condition on to show orthogonality

Lemma: for $L_0^2(x)$, $\mathcal{H} = \{h(Q_1 x) \mid Q_2(x))$

$$E[h(Q_1(x)\, Q_2(x)) \mid Q_2(x)] = 0$$

$$\pi[D \mid \mathcal{H}] = E(D \mid Q_1(x)\, Q_2(x)) - E(D \mid Q_2(x))$$

$$\begin{aligned}
\pi[D \mid T_Q] &= \pi[D \mid T_W \oplus T_Y] \\
&= \pi[D \mid T_W] + \pi[D \mid T_Y] \\
&= E[D \mid W] + [\underbrace{E[D \mid W, A, Y]}_{=D} - E[D \mid W, A]] \\
&= \underbrace{E\left[Y \frac{I(A=1)}{g(1 \mid W)} - \psi \mid W\right]}_{= \cancel{A}} + \underbrace{\left[Y \frac{I(A=1)}{g(1 \mid W)} - E(Y \mid A, W) \frac{I(A=1)}{g(1 \mid W)}\right]}_{\frac{I(A=1)}{g(A \mid W)}(Y - E(Y \mid A, W))}
\end{aligned}$$

$$\begin{aligned}
E\left[Y \frac{I(A=1)}{g(1 \mid W)} \mid W\right] &= E\left[E[Y \mid A, W] \frac{I(A=1)}{g(1 \mid W)} \mid W\right] \\
&= \frac{E[Y \mid A=1, W]}{g(1 \mid W)} \underbrace{E[I(A=1) \mid W]}_{= P(A=1 \mid W) = \cancel{g(1 \mid W)}}
\end{aligned}$$

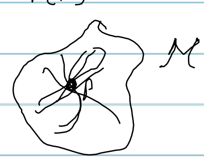
$$\boxed{= E[Y \mid A=1, W] - \Psi(P) + \frac{I(A=1)}{g(A \mid W)}(Y - E(Y \mid A, W))}$$ answer

Question: Why do we care about the efficient influence curve?
↳ Answered in appendix A4.

Thm A1: We have a rv $O \sim P \in \mathcal{M}$, $\Psi : \mathcal{M} \to \mathbb{R}$

We have class of submodels through P at $\varepsilon = 0 \Rightarrow \{P(\varepsilon), \varepsilon\}$

Let $T(P)$ be the tangent space of P



$D^*(P) \in T(P)$

$\partial/\partial\varepsilon \log P(\varepsilon)|_{\varepsilon=0}$

→ Assume Ψ is pathwise differentiable: $\frac{\partial}{\partial \varepsilon} \Psi(P(\varepsilon))\big|_{\varepsilon=0} = E_P[D^*(P)(O)\, S(O)]$

$$\hat{\Psi}(P_n) - \hat{\Psi}(P) = \frac{1}{n} \sum_{i=1}^{n} IC(P)(O_i) + R_n$$

$P(\varepsilon)_n$ = empirical dist for sample from $P(\varepsilon)$, ie. $O_i \overset{iid}{\sim} P(\varepsilon)$

Note: $P(\varepsilon_n)_n$ is sample from perturbed dist.

cont. $P(\varepsilon)_n - P = P(\varepsilon_n)_n - P(\varepsilon_n) + P(\varepsilon_n) - P$

$\rightarrow$ choose $\varepsilon_n = \frac{1}{\sqrt{n}}$

So, $\hat{\Psi}(P(\varepsilon_n)_n) - \hat{\Psi}(P) = P(\varepsilon_n)_n IC(P) + R(P(\varepsilon_n)_n, P)$,

where $\dfrac{E_{P(\varepsilon_n)} R(P(\varepsilon_n)_n, P)}{\varepsilon_n} \xrightarrow[n\to\infty]{} 0$

We want $\dfrac{E_{P(\varepsilon_n)} \hat{\Psi}(P(\varepsilon_n)_n) - \Psi(P(\varepsilon_n))}{\varepsilon_n} \xrightarrow[n\to\infty]{} 0$

$\hookrightarrow$ then $\pi[IC(P) \mid T(P)] = D^*(P)$

IC

D

* So, an estimator is asymptotically linear only if: ______.

$\hookrightarrow$ best estimator is asymptotically linear w/ IC = can. gradient.