Nonparametric causal effects based on incremental propensity score interventions

Edward Kennedy

Department of Statistics & Data Science
Carnegie Mellon University

ICHPS, 11 Jan 2018

Take away

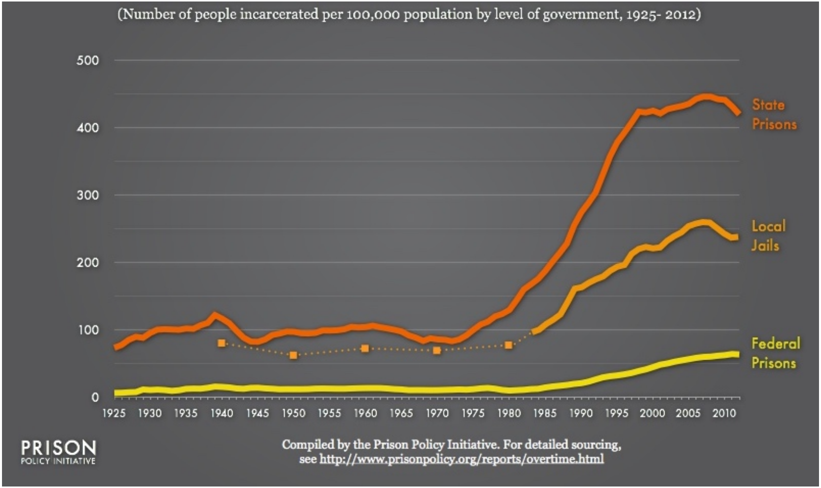
Standard causal methods require **strong statistical assumptions**

- e.g., all must have non-zero chance of treatment and control
- need parametric models if more than a few timepoints

We propose incremental propensity score interventions instead

- e.g., what would happen if we shifted everyone's PS?
- these completely avoid positivity and parametric assumptions

Motivating example



Motivating example

Source: http://www.prisonpolicy.org/global/

Motivating example

Incarceration is a colossal industry in the US

- currently **2.3 million** confined in correctional facilities
- another **4.6 million** on probation/parole

Important to study unintended consequences of mass incarceration

- e.g., effects on employment, health, psychology, social ties...

We will consider effects on entry into marriage

- impacts family/social support, children's outcomes, recidivism

Data & setup

We use data from National Longitudinal Survey of Youth 1997.

Observe iid sample $(\mathbf{Z}_1, ..., \mathbf{Z}_n)$ for

$$\mathbf{Z} = (\mathbf{X}_1, A_1, \mathbf{X}_2, A_2, ..., \mathbf{X}_T, A_T, Y) = (\overline{\mathbf{X}}_T, \overline{A}_T, Y)$$

where $T = 10$ years (2001-2010), $n = 4781$ subjects, and

- $\mathbf{X}_t$ = covariates at time t *(demographics, delinquency indicators, employment, earnings...)*
- A_t = exposure at time t *(whether incarcerated at year t)*
- Y = outcome *(whether married in 2010)*

Standard approaches

Let $Y^{\overline{a}_T}$ denote potential outcome that would have been observed under exposure sequence $\overline{a}_T = (a_1, \dots, a_T)$

- let $\mathbf{H}_t = (\overline{\mathbf{X}}_t, \overline{A}_{t-1})$ denotes past covariate/exposure history

Standard causal methods target deterministic intervention effects

$$\mathbb{E}(Y^{\overline{a}_T}) = m(\overline{a}_T; \beta) \qquad \text{(MSM)}$$

$$\mathbb{E}(Y^{\overline{a}_t,0} - Y^{\overline{a}_{t-1},0} \mid \mathbf{H}_t, A_t) = \gamma_t(\mathbf{h}_t, a_t; \theta) \qquad \text{(SNM)}$$

or similar related quantities (Robins 1986, 1994, 2000)

Issue 1: Parametric modeling

MSMs/SNMs have curse of dimensionality in T. Even in RCT:

- for $T = 10$, if $n < 5k$ then $> 99\%$ chance of non-empty cell, need $n \approx 12k$ to guarantee $< 1\%$ chance of empty cell

Parametric models reduce variance but can give extreme bias

- lots of parameters $\implies$ hard to interpret/visualize
- fewer parameters $\implies$ probably severely wrong

Let's be honest:
We use parametric models because they make life easier

Issue 2: Positivity

Let $\pi_t(\mathbf{h}_t) = \mathbb{P}(A_t = 1 \mid \mathbf{H}_t = \mathbf{h}_t)$ denote propensity score at t.

Standard MSMs/SNMs require positivity assumptions of the form

$$\mathbb{P}\{0 < \pi_t(\mathbf{H}_t) < 1\} = 1$$

i.e., everyone has to have chance at treatment/control. But:

- very sick may always take trt, very healthy may never
- multi-year incarceration, many have $\pi_t(\mathbf{h}_t) \approx 0$

Even near-violations can wreak havoc for finite n! (even if $T = 1$)

Related work

Restrictive modeling/positivity assumptions can be weakened by shifting focus to effects of other types of interventions
$\rightarrow$ Lots of recent interest in dynamic & stochastic interventions:

- $T = 1$: Pearl (00), Tian (08), Diaz & van der Laan (12, 13), Moore et al (12), Haneuse & Rotnitzky (13)
- $T > 1$: Murphy et al (01), Robins et al (04), vdL & Petersen (07), Taubman et al (09), Cain et al (10), Young et al (11, 14)

But none of these approaches simultaneously

- are completely nonparametric, even when T is large
- avoid positivity conditions entirely

Our proposal

We propose incremental propensity score intervention effects and corresponding estimators

Advantages:

- completely nonparametric even with large T
- no positivity required
- estimators can converge at fast parametric $\sqrt{n}$ rates, even if constructed via machine learning / high-dimensional regression
- can be used in general longitudinal studies
- yields neat Fisher-type test of no longitudinal trt effect

Incremental PS interventions

Incremental PS interventions shift π_t values instead of setting A_t

Let $Y^{\mathbf{Q}(\delta)}$ be potential outcome under the fluctuated trt process

$$q_t(\mathbf{h}_t; \delta, \pi_t) = \frac{\delta \pi_t(\mathbf{h}_t)}{\delta \pi_t(\mathbf{h}_t) + 1 - \pi_t(\mathbf{h}_t)}$$

where $\delta \in (0, \infty)$ is an increment parameter

- $q_t = \pi_t$ if $\delta = 1$, $\quad q_t \to 1$ as $\delta \to \infty$, $\quad q_t \to 0$ as $\delta \to 0$

The increment δ is just an OR

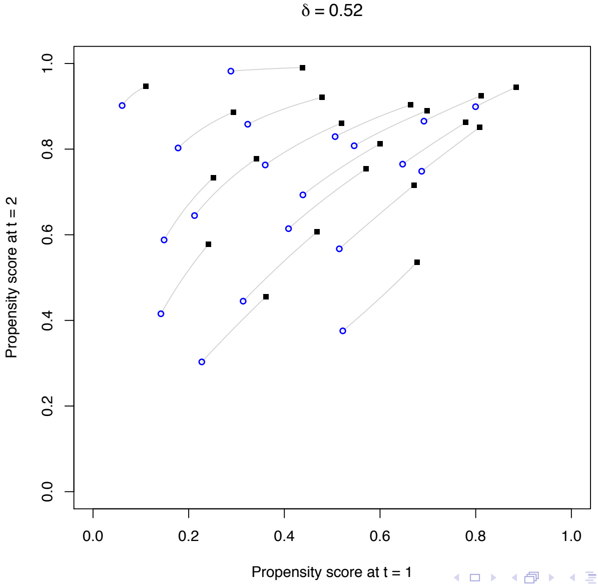
The increment parameter is easy to interpret if we notice

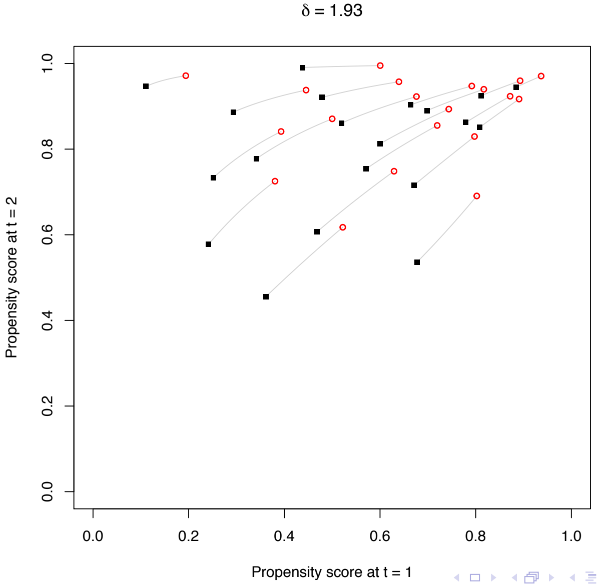
$$\delta = \frac{q_t(\mathbf{h}_t)/\{1 - q_t(\mathbf{h}_t)\}}{\pi_t(\mathbf{h}_t)/\{1 - \pi_t(\mathbf{h}_t)\}} = \frac{\text{odds}_{\mathbb{Q}}(A_t = 1 \mid \mathbf{H}_t = \mathbf{h}_t)}{\text{odds}_{\mathbb{P}}(A_t = 1 \mid \mathbf{H}_t = \mathbf{h}_t)}$$

when $0 < \pi_t < 1$ (else $q_t = \pi_t$) $\implies$ δ is simply an odds ratio

Example: Suppose $\delta = 1.5$, so odds of treatment increase by 50%

- if $\pi_t = 50\%$ then $q_t = 60\%$
- if $\pi_t = 25\%$ then $q_t \approx 33\%$
- if $\pi_t = 5\%$ then $q_t \approx 7.3\%$


δ = 0.52
Propensity score at t = 2
Propensity score at t = 1
0.0
0.2
0.4
0.6
0.8
1.0


δ = 1.93
Propensity score at t = 2
Propensity score at t = 1
0.0
0.2
0.4
0.6
0.8
1.0

Identification

We focus on estimating mean $\psi(\delta) = \mathbb{E}(Y^{\mathbf{Q}(\delta)})$

- *mean outcome if odds of treatment were multiplied by* δ

Assume: 1. Consistency: $Y = Y^{\overline{A}_T}$

2. Exchangeability: $A_t \perp\!\!\!\perp Y^{\overline{a}_T} \mid \mathbf{H}_t$

Identification follows from Robins' extended g-formula:

$$\psi(\delta) = \sum_{\overline{a}_T} \int_{\mathcal{X}} \mu(\mathbf{h}_t, a_t) \prod_{t=1}^{T} q_t(a_t \mid \mathbf{h}_t)\, d\mathbb{P}(\mathbf{x}_t \mid \mathbf{h}_{t-1}, a_{t-1})$$

$\rightarrow$ no positivity needed! since $q_t = \pi_t$ if $\pi_t = 0, 1$ for $0 < \delta < \infty$

Efficiency theory

Understanding the efficient influence function (EIF) is crucial

- variance gives us efficiency bound $\rightarrow$ estimation benchmark
- recipe for constructing estimators that are efficient yet robust
- clarifies regularity conditions needed for efficient estimation

Uncentered EIF for $T = 1$ case:

$$\frac{\delta\pi(\mathbf{X})\phi_1(\mathbf{Z}) + \{1-\pi(\mathbf{X})\}\phi_0(\mathbf{Z})}{\delta\pi(\mathbf{X}) + \{1-\pi(\mathbf{X})\}} + \frac{\delta\{\mu(\mathbf{X},1)-\mu(\mathbf{X},0)\}\{A-\pi(\mathbf{X})\}}{\{\delta\pi(\mathbf{X})+1-\pi(\mathbf{X})\}^2}$$

for $\phi_a = \frac{\mathbb{1}(A=a)}{\pi(a|\mathbf{X})}\{Y-\mu(\mathbf{X},A)\} + \mu(\mathbf{X},a)$ EIF for $\mathbb{E}\{\mu(\mathbf{X},a)\}$

Estimation

It is easy to construct an IPW estimator of $\psi(\delta)$:

$$\hat{\psi}^*_{ipw}(\delta) = \mathbb{P}_n \left\{ \prod_{t=1}^{T} \frac{(\delta A_t + 1 - A_t) Y}{\delta \hat{\pi}_t(\mathbf{H}_t) + 1 - \hat{\pi}_t(\mathbf{H}_t)} \right\}$$

But for general $\hat{\pi}_t$ this won't be $\sqrt{n}$-consistent & asymp. normal
$\rightarrow$ *only if $\hat{\pi}_t$ constructed with correct parametric models*

Or can solve EIF estimating equation $\hat{\psi}^*(\delta) = \mathbb{P}_n\{\varphi(\mathbf{Z}; \hat{\boldsymbol{\eta}}, \delta)\}$

- can be $\sqrt{n}$ CAN even if $\hat{\boldsymbol{\eta}} = (\hat{\pi}_t, \hat{m}_t)$ converge at slower rates
- but must restrict complexity of $\hat{\boldsymbol{\eta}}$ (random forests, boosting?)

Sample-splitting estimator

Can exploit K-fold sample splitting to use arbitrary ML methods:

$$\hat{\psi}(\delta) = \mathbb{P}_n\{\varphi(\mathbf{Z}; \hat{\boldsymbol{\eta}}_{\text{-}S}, \delta)\}$$

where $S \in \{1, ..., K\}$ is splitting rv, $\hat{\boldsymbol{\eta}}_{\text{-}s}$ is fit *excluding* fold s

- still need faster than $n^{-1/4}$ rate for $\hat{\boldsymbol{\eta}} = (\hat{\pi}_t, \hat{m}_t)$ for CAN, as with estimating equation estimator

Large-sample properties

Suppose $\mathcal{D} = [\delta_\ell, \delta_u]$ is bounded with $0 < \delta_\ell \leq \delta_u < \infty$, and:

- $\Big(\sup_\delta \|\hat{m}_{t,\delta} - m_{t,\delta}\| + \|\hat{\pi}_t - \pi_t\| \Big) \|\hat{\pi}_s - \pi_s\| = o_{\mathbb{P}}(1/\sqrt{n})$ for $s \leq t$

Then normalized $\hat{\psi}(\cdot)$ converges to mean-zero Gaussian process:

$$\frac{\hat{\psi}(\delta) - \psi(\delta)}{\hat{\sigma}(\delta)/\sqrt{n}} \rightsquigarrow \mathbb{G}(\delta) \quad \text{in } \ell^\infty(\mathcal{D})$$

where $\hat{\sigma}^2(\delta) = \mathbb{P}_n[\{\varphi(\mathbf{Z}; \hat{\boldsymbol{\eta}}_{\text{-}S}, \delta) - \hat{\psi}(\delta)\}^2]$

- for *pointwise CIs*: empirical variance of estimated IF
- for uniform CIs can use multiplier bootstrap (Chernozhukov etc)

$\rightarrow$ very easy to compute (don't need to do any refitting!)

Testing no effect

Given a uniform CI, we can invert to test no effect hypothesis

$$H_0 : \psi(\delta) = \mathbb{E}(Y) \ \text{ for all } \delta \in \mathcal{D} \cup \{1\}$$

$\rightarrow$ note: this null is somewhere *in between Fisher and Neyman*

Specifically, for lower/upper uniform limits $\hat{\psi}_{\ell/u,\alpha}$

$$\hat{p} = \sup \left\{ \alpha : \inf_{\delta \in \mathcal{D}} \hat{\psi}_{u,\alpha}(\delta) \geq \sup_{\delta \in \mathcal{D}} \hat{\psi}_{\ell,\alpha}(\delta) \right\}$$

is a valid p-value for testing H_0.

- this is just biggest α giving CI that contains straight line

Back to NLSY application

Recall we have data across $T = 10$ years for $n = 4781$ individuals

- goal: learn about effects of incarceration on marriage

We estimated nuisance functions (π_t, m_t) with random forests

- used $K = 5$ fold sample splitting
- need to do $T + 1 = 11$ fits for each δ value (and split)
- but the `ranger` package in R is very fast

Implemented our proposed methods, also standard MSM analysis

Standard MSM analyses

Model: $\mathbb{E}(Y^{\bar{a}_T}) = \beta_0 + \beta_1 \sum_t a_t$

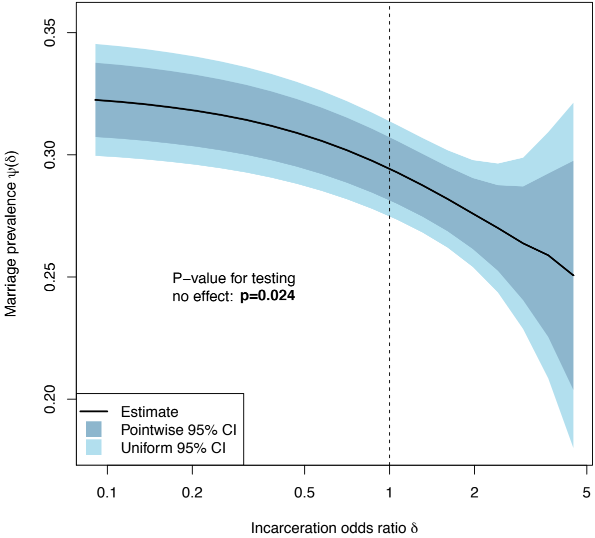
	Estimate	Robust.SE	z.val	p.val
(Intercept)	-2.72e+15	8.15e+14	-3.34	0.001
totincarc	-1.12e+13	1.25e+14	-0.09	0.928

After stabilization:

	Estimate	Robust.SE	z.val	p.val
(Intercept)	-0.8592	0.033315	-25.79	0.000
totincarc	-0.3241	0.112994	-2.87	0.004

Model: $\mathbb{E}(Y^{\bar{a}_T}) = \beta_0 + \sum_t \beta_t a_t$

```
Error in solve.default... system is computationally singular...
```


Marriage prevalence ψ(δ)
0.35
0.30
0.25
0.20
P−value for testing
no effect: p=0.024
Estimate
Pointwise 95% CI
Uniform 95% CI
0.1
0.2
0.5
1
2
5
Incarceration odds ratio δ

Summary

Available causal methods require positivity/parametrics/both

- especially in longitudinal studies with e.g., 5+ timepoints

We propose incremental propensity score interventions

- no parametric assumptions or positivity required
- efficient estimators that can incorporate machine learning
- uniform inference $\rightarrow$ novel tests of no effect

The paper is in press at JASA and on arxiv:
arxiv.org/abs/1704.00211

You can implement the method with the R package “npcausal”
http://www.ehkennedy.com/code.html

Feel free to email with any questions
or if you want to collaborate in applying these methods:
edward@stat.cmu.edu

Thank you!

Taxonomy of intervention types

Restrictive modeling/positivity assumptions can be weakened by shifting focus to effects of other types of interventions

1. Deterministic
   a. static: $A_t^* = a_t$
   b. dynamic: $A_t^* = d_t(\mathbf{H}_t)$ for some $d_t : \mathcal{H}_t \mapsto \mathcal{A}$

2. Stochastic
   a. static: $A_t^* \sim \text{Bern}(q_t)$
   b. dynamic: $A_t^* \sim \text{Bern}\{q_t(\mathbf{H}_t)\}$ for some $q_t : \mathcal{H}_t \mapsto [0, 1]$

Identification when $T = 1$

When $T = 1$ the identifying expression for $\psi(\delta)$ simplifies:

$$\psi(\delta) = \mathbb{E}\left[\frac{\delta\pi(\mathbf{X})\mu(\mathbf{X}, 1) + \{1 - \pi(\mathbf{X})\}\mu(\mathbf{X}, 0)}{\delta\pi(\mathbf{X}) + 1 - \pi(\mathbf{X})}\right]$$

where $\mu(\mathbf{X}, A) = \mathbb{E}(Y \mid \mathbf{X}, A)$ is regression function

EIF for $T > 1$

EIF (again uncentered) in longitudinal studies is more complicated:

$$\varphi = \sum_{t=1}^{T} \left[\frac{A_t\{1-\pi_t(\mathbf{H}_t)\} - (1-A_t)\delta\pi_t(\mathbf{H}_t)}{\delta/(\delta-1)} \right] \left\{ \sum_{a=0}^{1} m_t(\mathbf{H}_t, a) q_t(a \mid \mathbf{H}_t) \right\}$$
$$\times \left\{ \prod_{s=1}^{t} \frac{(\delta A_s + 1 - A_s)}{\delta\pi_s(\mathbf{H}_s) + 1 - \pi_s(\mathbf{H}_s)} \right\} + \prod_{t=1}^{T} \frac{(\delta A_t + 1 - A_t) Y}{\delta\pi_t(\mathbf{H}_t) + 1 - \pi_t(\mathbf{H}_t)}$$

where for $m_{T+1} = Y$ we recursively define

$$m_t(\mathbf{H}_t, A_t) = \sum_{a=0}^{1} \mathbb{E}\left\{ m_{t+1}(\mathbf{H}_{t+1}, a) q_{t+1}(a \mid \mathbf{H}_{t+1}) \,\middle|\, \mathbf{H}_t, A_t \right\}$$

Estimation algorithm

$\forall \delta, k$, with $\mathbf{D}_0$ / $\mathbf{D}_1$ train/test data, resp., with $\mathbf{D} = \mathbf{D}_0 \cup \mathbf{D}_1$:

1. Regress $A_t \sim \mathbf{H}_t$ in $\mathbf{D}_0$, obtain preds $\hat{\pi}_t(\mathbf{H}_t)$ in $\mathbf{D}$.
2. Compute weights $W_t = \frac{\delta A_t + 1 - A_t}{\delta \hat{\pi}_t(\mathbf{H}_t) + 1 - \hat{\pi}_t(\mathbf{H}_t)}$ in $\mathbf{D}_1$.
3. Compute cumulative product weight $\widetilde{W}_t = \prod_{s=1}^{t} W_s$ in $\mathbf{D}_1$.
4. For each time $t = T, T-1, ..., 1$ (starting with $R_{T+1} = Y$):
   (a) Regress $R_{t+1} \sim (\mathbf{H}_t, A_t)$ in $\mathbf{D}_0$, obtain preds $\hat{m}_t(\mathbf{H}_t, a)$ in $\mathbf{D}$.
   (b) Construct pseudo-outcome $R_t = \sum_a \hat{m}_t(\mathbf{H}_t, a) q_t(a \mid \mathbf{H}_t)$ in $\mathbf{D}$.
5. Compute weights $V_t = \frac{A_t\{1 - \hat{\pi}_t(\mathbf{H}_t)\} - (1 - A_t)\delta \hat{\pi}_t(\mathbf{H}_t)}{\delta/(\delta - 1)}$ in $\mathbf{D}_1$.
6. Set $\hat{\psi}_k(\delta)$ as average of $\varphi = \widetilde{W}_T Y + \sum_t \widetilde{W}_t V_t R_t$ vals in $\mathbf{D}_1$.

$\rightarrow$ Set $\hat{\psi}(\delta)$ as average of K estimators $\hat{\psi}_k(\delta)$, $k = 1, ..., K$.

Uniform inference

Easy to get *pointwise CIs*: empirical variance of estimated IF

- for uniform CIs can use multiplier bootstrap (Chernozhukov etc)

i.e., to find critical value $\hat{c}_\alpha$ such that

$$\mathbb{P}\left\{\hat{\psi}(\delta) - \frac{\hat{c}_\alpha \hat{\sigma}(\delta)}{\sqrt{n}} \leq \psi(\delta) \leq \hat{\psi}(\delta) + \frac{\hat{c}_\alpha \hat{\sigma}(\delta)}{\sqrt{n}}, \forall \delta \in \mathcal{D}\right\} = 1 - \alpha + o(1)$$

we can generate $\xi_i \sim N(0,1)$ and solve

$$\mathbb{P}\left(\sup_{\delta \in \mathcal{D}} \left| \sqrt{n}\, \mathbb{P}_n \left[\xi \left\{ \frac{\varphi(\mathbf{Z}; \hat{\boldsymbol{\eta}}_{\text{-}S}, \delta) - \hat{\psi}(\delta)}{\hat{\sigma}(\delta)} \right\} \right] \right| \geq \hat{c}_\alpha \,\middle|\, \mathbf{Z}_1, \ldots, \mathbf{Z}_n \right) = \alpha$$

$\rightarrow$ very easy to compute (don't need to do any refitting!)