Assessing External Validity Over Worst-case Subpopulations

Hongseok Namkoong

Decision, Risk, and Operations Division
Columbia Business School
namkoong@gsb.columbia.edu

Based on a joint work with Sookyo Jeong
Potential outcomes

- A feature vector $X \in \mathbb{R}^k$
- A treatment assignment $Z \in \{0,1\}$
- Potential outcomes: $Y(1), Y(0)$
- **Observe** $Y := Y(Z)$, **never** $Y(1 - Z)$

Average Treatment Effect (ATE)

$$ATE = \mathbb{E}[Y(1) - Y(0)]$$

$$= \mathbb{E}_{X \sim P_X} \left[ \mathbb{E}[Y(1) \mid X] - \mathbb{E}[Y(0) \mid X] \right]$$

$$= \mathbb{E}_{X \sim P_X} \left[ \mu_1^*(X) - \mu_0^*(X) \right] =: \mathbb{E}_{X \sim P_X} \left[ \mu^*(X) \right]$$

- $P_X$ is the data generating distribution for $X$
What if $P_X$ changes?

- Demographic compositions shift over time
What if $P_X$ changes?

- Even for carefully designed randomized trials, “statistics” starts only at treatment assignment, with big biases in selection into study.

Distribution of log-district size in studies versus total population

[Tipton et al. 2019] The convenience of large urban school districts: a study of recruitment practices in 37 randomized trials
What if $P_X$ changes?

- “Clinical trials for new drugs skew heavily white”
  - Out of 10,000+ cancer trials, less than 2% focused on racial minorities, and less than 5% of participants were non-white

- Especially problematic when treatment effect is heterogeneous
  - [Leigh et al. ‘16, Imai et al. ‘13, Gijsberts et al. ‘15, Basu et al. ‘17, Baum et al. ‘17, Duan et al. ‘19]

- Recently, two large trials with $n = 5K-10K$ had opposite findings on a treatment to lower blood pressure on cardiovascular disease
  - [ACCORD ‘10, SPRINT ‘15]
Potential solution?

• Directly estimate conditional average treatment effect (CATE) using ML methods?

[Leigh et al. '16, Imai et al. '13, Gijsberts et al. '15, Basu et al. '17, Baum et al. '17, Duan et al. '19, Nie and Wager '20]

• ML models perform very poorly on underrepresented groups

• ML estimates are unstable and resulting inference is underpowered

• Predefined subgroup analysis difficult due to intersectionality

Effect of Medicaid enrollment on doctor’s office utilization

<table>
<thead>
<tr>
<th>ethnic group</th>
<th>effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>hispanic</td>
<td>V</td>
</tr>
<tr>
<td>US citizen</td>
<td>V</td>
</tr>
<tr>
<td>college+</td>
<td>X</td>
</tr>
</tbody>
</table>
Automatically find **worst-off subpopulations** and measure **treatment effect** on them

\[ Q_X \text{ is a subpopulation} \quad \iff \exists \text{ proportion } a \in (0, 1], \text{ prob. } Q'_X \]

\[ \text{ s.t. } P_X(\cdot) = aQ_X + (1-a)Q'_X \]
Subpopulations

Automatically find **worst-off subpopulations** and measure **treatment effect** on them

\[ Q_X \text{ is a subpopulation} \quad \text{s.t.} \quad P_X(\cdot) = aQ_X + (1 - a)Q'_X \]
Subpopulations

Automatically find **worst-off subpopulations** and measure **treatment effect** on them

\[ Q_X \text{ is a subpopulation} \quad \leftrightarrow \quad \exists \text{proportion } a \in (0, 1], \text{ prob. } Q'_X \]

s.t. \[ P_X(\cdot) = aQ_X + (1 - a)Q'_X \]
Subpopulations

Automatically find worst-off subpopulations and measure treatment effect on them

\[ Q_X \text{ is a subpopulation } \iff \exists \text{ proportion } a \in (0, 1], \text{ prob. } Q'_X \text{ s.t. } P_X(\cdot) = aQ_X + (1 - a)Q'_X \]
Subpopulations

Automatically find worst-off subpopulations and measure treatment effect on them

$Q_X$ is a subpopulation $\iff \exists$ proportion $a \in (0, 1]$, prob. $Q'_X$

s.t. $P_X(\cdot) = aQ_X + (1 - a)Q'_X$
Worst-case subpopulation

Notation

\( Q_X \geq \alpha \) \iff \exists \text{probability } Q'_X, \text{ and } a \geq \alpha

\text{s.t. } P_X = aQ_X + (1 - a)Q'_X

subpopulation with proportion larger than \( \alpha \in (0, 1] \)

worst-case treatment over subpopulation larger than \( \alpha \in (0, 1] \)

\[
WTE_\alpha := \sup_{Q_X \geq \alpha} \mathbb{E}_{Q_X} [\mu^*(X)]
\]

where \( \mu^*(X) := \mathbb{E}[Y(1) - Y(0) | X] \) is the conditional average treatment effect (CATE).

Recap

- Covariates: \( X \)
- Treatment assignment: \( Z \)
- Potential outcome: \( Y(0), Y(1) \)
- Response \( Y := Y(Z) \)
Sensitivity analysis

• Posit a set of “plausible” changes to $P_X$, and take worst-case over them

• If effects are still valid under plausible violations, we can certify robustness

• Sensitivity of a finding: magnitude of violation when endpoint crosses a threshold

• Today: Worst-case bounds on the Doubly Robust / AIPW estimator

\[ \text{WTE}_\alpha := \sup_{Q_X \geq \alpha} \mathbb{E}_{Q_X} \left[ \mu^*(X) \right] \]
Sensitivity analysis

• Posit a set of “plausible” changes to $P_X$, and take worst-case over them

• If effects are still valid under plausible violations, we can certify robustness

• Sensitivity of a finding: magnitude of violation when endpoint crosses a threshold

• Today: Worst-case bounds on the Doubly Robust / AIPW estimator

Is this a “sensible” amount of distribution shift / violation?
Sensitivity analysis

• Does not assume a fixed target; often appropriate for operational decisions

• Heuristically, set $\alpha$ small if the collected data is not diverse

• Conservative but can still be useful; future work needed on this

• Need to be accompanied by a design-based perspective to maximizing diversity in $P_X$
Effect of Medicaid on doctor visits over time

- Evaluate effect of Medicaid enrollment on doctors’ office utilization
- Medicaid costs $553 billion/yr; need to ensure valid effects through time
- Outcome: visit to doctors in the two-weeks prior to a random survey date
- Control for demographics, medical history, employment, earnings, insurance, government assistance etc (d = 396)
- Take the viewpoint of an analyst in 2009 (n = 82,993)
Effect of Medicaid on doctor visits over time

• Evaluate effect of Medicaid enrollment on doctors’ office utilization in 2009
Effect of Medicaid on doctor visits over time

- Evaluate effect of Medicaid enrollment on doctors’ office utilization
Effect of Medicaid on doctor visits over time

- Evaluate effect of Medicaid enrollment on doctors’ office utilization
Welfare attitudes experiment

- Evaluate effect of wording on survey results ("welfare" vs "assistance to the poor")
- WTE guarantees positive findings even for small subpopulations
- WTE is stable across model classes used, similar to ATE, unlike CATE

(a) ATE and WTE_α
(b) CATE by years of education
(c) CATE by age
WTE = Tail-average

Recap
- Covariates: X
- Treatment assignment: Z
- Potential outcome: $Y(0), Y(1)$
- CATE $\mu^*(X) = \mathbb{E}[Y(1) - Y(0) \mid X]$

**Lemma (Shapiro et al. '09)**

\[
\sup_{Q_x \geq \alpha} \mathbb{E}_{Q_x} [\mu^*(X)] = \mathbb{E}[\mu^*(X)h^*(X)]
\]

where

\[
h^*(x) := \frac{1}{\alpha} \mathbb{1}\{\mu^*(x) \geq P_{1-\alpha}(\mu^*)\}
\]
Estimation Approach

• Use ML methods to fit nuisance parameters
  \[ \mu_z^*(X) = \mathbb{E}[Y(z) \mid X = x], \quad z \in \{0, 1\} \]
  \[ e^*(X) = \mathbb{P}(Z = 1 \mid X) \quad h^*(X) = \frac{1}{\alpha} 1 \{ \mu^*(X) \geq \frac{1}{P_{1-\alpha}}(\mu^*) \} \]

• Today: Construct a WTE estimator insensitive to error in nuisance estimates

• Design an mean zero augmentation term that includes nuisance parameters
  \[ WTE_\alpha + \mathbb{E} \left[ h^*(X) \left( \frac{Z}{e^*(X)}(Y - \mu_1^*(X)) - \frac{1 - Z}{1 - e^*(X)}(Y - \mu_0^*(X)) \right) \right] \]

Recap

‣ Covariates: X
‣ Treatment assignment: Z
‣ Potential outcome: Y(0), Y(1)

Neyman orthogonal: Directional derivative w.r.t. nuisance parameters, taken at the true nuisance value \( (\mu_1^*, \mu_0^*, e^*, h^*) \) is zero. [Neyman '59, Chernozhukov et al. '18]
Assumptions

Standard; required for identification and estimation of ATE

- No unobserved confounding: $Y(0), Y(1) \perp Z \mid X$
- Overlap: $\exists c > 0 \text{ s.t. } \Pr(e^*(X) \in [c, 1-c]) = 1$
- SUTVA: single version of treatment, no interference between units

Recap
- Covariate $X$, Treatment $Z$
- Potential outcome: $Y(0), Y(1)$
- Propensity score $e^*(X) = \Pr(Z = 1 \mid X)$
Main Results

Theorem (Jeong & N.'20)

1. Under slower-than-parametric rates of convergence on the nuisance parameters, $\sqrt{n}(\hat{w}_\alpha - WTE_\alpha) \Rightarrow N(0, \sigma_\alpha^2)$

2. $\sigma_\alpha^2$ is the optimal asymptotic variance

- Central limit rates even when nuisance estimates converge more slowly
- Augmented estimator is semiparametrically efficient for both randomized and observational studies
Summary

- Worst-case bounds on the Doubly Robust / AIPW estimator under distribution shift
- Allow flexible use of ML methods to estimate nuisance parameters
- Central limit results even when nuisance parameters converge slower
- Our procedures are *optimal; semiparametrically efficient*