Overlap Weighting Methods for Comparative Effectiveness Research

Fan Li 1,2,3  Laine E. Thomas 2,3

1 Department of Statistical Science, Duke University
2 Department of Biostatistics and Bioinformatics, Duke University
3 Duke Clinical Research Institute (DCRI)

January 6, 2020

2020 International Conference on Health Policy Statistics, San Diego
Workshop objectives

Aim to answer the following questions

- How does real world data alter the paradigm for causal inference?
- What are limitations of IPW, and trimming methods?
- What is overlap weighting, and what are its advantages?
- How do we apply overlap weights in practice?
- How does this extend to complex settings?
Website

All the lecture slides, papers, tutorials, computer code and package are posted and continuously updated on the webpage on overlap weights

https://www2.stat.duke.edu/~fl35/OW.html
Section 1.

Background

- Motivating Example: Framingham Heart Study
- Standard methods
- Problem
Introduction

- Observational treatment comparisons may be confounded

- Goal: Balance covariate distributions across groups to remove confounding

- One common approach is weighting based on a propensity score

- Main idea: weight the treatment and control groups to create a pseudo-population (the target population) where the covariate distributions are balanced

- Inverse probability of treatment weighting (IPW) is the dominant weighting approach
Example: Framingham Heart Study
(Thomas et al. 2020)

- **Goal:** evaluate the effect of statins on health outcomes

- **Patients:** cross-sectional population from the offspring cohort with a visit 6 (1995-1998)

- **Treatment:** statin use at visit 6 vs. no statin use

- **Outcomes:** CV death, myocardial infarction (MI), stroke

- **Confounders:** sex, age, body mass index, diabetes, history of MI, history of PAD, history of stroke...

- Significant imbalance between treatment and control groups in covariates motivates IPW (or some form of propensity score adjustment)
Unadjusted differences (Statins vs. Control)

Variable Names:
- Age
- BMI
- Chol
- DBP
- Diabetes
- Female
- FRS
- Glucose
- HDL
- MI hx
- PAD hx
- SBP
- SBP med
- Smoking
- Stroke hx
- Triglycerides

Standardized Mean Differences

Method:
- △ Unadjusted
Standard Setup

- Data: a random sample of $n$ units from a population

- $Z_i \in \{0, 1\}$ treatment indicator, $X_i = (X_{i1}, \ldots, X_{ip})^T$ pre-treatment covariates

- For each unit, two potential outcomes $(Y_i(1), Y_i(0))$, only observe $Y_i(Z_i)$

- **Estimand**: Average Treatment Effect (ATE)

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

- Assuming conditional exchangeability (unconfoundedness) and positivity (overlap), can identify ATE from data
**Inverse Probability Weighting (IPW)**

- The propensity score (PS) \( e(X) = Pr(Z = 1|X) \)

- Inverse probability of treatment weights:

\[
\begin{align*}
    w(X_i) &= \frac{1}{e(X_i)} \quad \text{for } Z_i = 1, \\
    w(X_i) &= \frac{1}{1-e(X_i)} \quad \text{for } Z_i = 0
\end{align*}
\]

- ATE estimated by weighted outcome mean difference between groups

\[
\hat{\tau} = \frac{\sum_{i=1}^{n} Z_iY_iw_i(X_i)}{\sum_{i=1}^{n} Z_iw_i(X_i)} - \frac{\sum_{i=1}^{n} (1-Z_i)Y_iw_i(X_i)}{\sum_{i=1}^{n} (1-Z_i)w_i(X_i)}
\]

- PS often estimated by a logistic model \( e(X_i; \hat{\beta}) = 1/(1 + \exp(-X_i^T\hat{\beta})) \)
Challenge: balance after IPW isn’t great
Challenge: Poor overlap
Example: Poor Overlap (Brennen et al. 2016)
Example: Poor Overlap (Nicholson et al. 2019)
Trends in Real World Data

- People with propensity 0 and 1 have no treatment uncertainty - nearly always get treated or untreated

- “Big data” including Medicare Claims, large registries or EHR make it possible to define the target population in increasingly broad terms

- We don’t start with a target population; we start with large data from which many target populations could be defined

- There is pressure to include “everyone”
  - “Inclusivity” is a perceived advantage of RWD
  - Limiting the population is not simple (not uni-dimensional)
  - Avoid the appearance of “picking and choosing”
  - Largest “N” is perceived to optimize precision
Causal Paradigm: Survey vs. RWD

- IPW originates from the Horvitz-Thompson estimator in survey literature

- Starting point of survey: **Design** – target population is defined *a priori*, extreme weights are rare

- Starting point of observational studies: **Data** - target population is usually NOT defined *a priori*, extreme weights are common

- Using IPW (and ATE) implicitly assumes that the sample is representative of a well-defined target population, as in the survey framework

- In RWD, IPW (and ATE) may correspond to the effect of an infeasible intervention
Challenge: Poor overlap
IPW Operational Challenges

- Propensity values near 0 and 1 yield extreme weights (after taking the inverse)

- Adverse finite-sample consequences – Basu’s elephant: severe bias and variance

- Normalization of weights helps, but not a lot

- Core problem: lack of overlap in the tail of the propensity distribution – causal comparisons of these units are highly uncertain
Other adjustment methods

How do poor overlap and extreme propensity scores impact other adjustment methods?

- Regression adjustment: Increases model sensitivity
- Matching: Many patients get excluded because they don’t have a match
- Stratification: Adds bias due to residual imbalance within strata

Better to fix IPW than to abandon weighting altogether.
Symmetric trimming (Crump et al., 2009)

- exclude patients whose estimated PS is outside \([\alpha, 1 - \alpha]\)
- rule of thumb \(\alpha = 0.1\)

Asymmetric trimming (Sturmer et al., 2010)

- exclude patients with PS outside of the common PS range formed by the treated and control patients
- among the treated units, further exclude those whose PS is below the \(q\) quantile of the treated units
- among control units, exclude those whose PS is above the \((1 - q)\) quantile of the control units
## Trimming thresholds applied to Framingham

<table>
<thead>
<tr>
<th>Method</th>
<th>Left Excluded</th>
<th>Right Excluded</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symmetric $\alpha = 0.025$</td>
<td>1047</td>
<td>1</td>
<td>1048 (31%)</td>
</tr>
<tr>
<td>Symmetric $\alpha = 0.05$</td>
<td>1634</td>
<td>1</td>
<td>1635 (48%)</td>
</tr>
<tr>
<td>Symmetric $\alpha = 0.10$</td>
<td>2269</td>
<td>1</td>
<td>2270 (68%)</td>
</tr>
<tr>
<td>Asymmetric $q = 0.025$</td>
<td>1179</td>
<td>134</td>
<td>1313 (39%)</td>
</tr>
<tr>
<td>Asymmetric $q = 0.05$</td>
<td>1554</td>
<td>257</td>
<td>1811 (54%)</td>
</tr>
<tr>
<td>Asymmetric $q = 0.10$</td>
<td>1811</td>
<td>468</td>
<td>2279 (68%)</td>
</tr>
</tbody>
</table>
Symmetric Trimming $\alpha = 0.10$
Propensity Score Trimming - Cont’d

Reduce the impact of extreme PS and improve finite-sample property of IPW

Choice of threshold $\alpha$, $q$ may be arbitrary

- conceptual challenge: ambiguous target population/interpretation
- operational challenge: causal estimates sensitive to trimming threshold
- operational challenge: bias-variance tradeoff
- operational challenge: refitting PS after trimming (a hidden message)
Section 2.

New Weighting Methods

- Class of balancing weights
- Overlap weights
- R package PSweight
Weighting Beyond IPW

Two contributions

1. Provide a unified framework—the balancing weights—to allow different user-specified target populations

2. Propose a new weighting scheme—the overlap weighting

- Statistical optimality and conceptual advantages
- Generalized to many settings: multiple treatments, subgroups, time-varying treatments
Assume sample drawn from density \( f(X) \), can represent the density of target population by \( g(X) \propto f(X)h(X) \), where \( h(\cdot) \) is called a tilting function.

Denote \( \mu_1(X) = \mathbb{E}[Y(1)|X] \), \( \mu_0(X) = \mathbb{E}[Y(0)|X] \).

A class of weighted average treatment effect (WATE): ATE over the target population \( g \)

\[
\tau^h = \tau_1^h - \tau_0^h = \frac{\mathbb{E}[h(X)\{\mu_1(X) - \mu_0(X)\}]}{\mathbb{E}[h(X)]} = \mathbb{E}_g[Y(1) - Y(0)]
\]
Let \( f_z(x) = \Pr(X = x | Z = z) \), we have

\[
f_1(x) \propto f(x)e(x), \quad f_0(x) \propto f(x)(1 - e(x))
\]

For a given \( h(x) \), to estimate \( \tau_h \), we can weight \( f_z(x) \) to the target population using weights

\[
\begin{align*}
  w_1(x) &\propto \frac{f(x)h(x)}{f_1(x)} = \frac{f(x)h(x)}{f(x)e(x)} = \frac{h(x)}{e(x)}, \\
  w_0(x) &\propto \frac{f(x)h(x)}{f_0(x)} = \frac{f(x)h(x)}{f(x)(1 - e(x))} = \frac{h(x)}{1 - e(x)}.
\end{align*}
\]

We call the class of weights \((w_0, w_1)\) balancing weights:

\[
f_1(x)w_1(x) = f_0(x)w_0(x) = f(x)h(x).
\]

They balance the distributions of the weighted covariates between comparison groups.
Balancing Weights: Examples

- Choice of $h(x)$ determines the target population, estimand, weights.
- Statistical, scientific and policy considerations all come into play in specifying $h(x)$.

<table>
<thead>
<tr>
<th>target population</th>
<th>$h(x)$</th>
<th>estimand</th>
<th>weight $(w_1, w_0)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>combined</td>
<td>1</td>
<td>ATE</td>
<td>$\left( \frac{1}{e(x)}, \frac{1}{1-e(x)} \right)$ [IPW]</td>
</tr>
<tr>
<td>treated</td>
<td>$e(x)$</td>
<td>ATT</td>
<td>$\left( 1, \frac{e(x)}{1-e(x)} \right)$</td>
</tr>
<tr>
<td>control</td>
<td>$1 - e(x)$</td>
<td>ATC</td>
<td>$\left( \frac{1-e(x)}{e(x)}, 1 \right)$</td>
</tr>
<tr>
<td>overlap</td>
<td>$e(x)(1 - e(x))$</td>
<td>ATO</td>
<td>$(1 - e(x), e(x))$</td>
</tr>
<tr>
<td>trimming</td>
<td>$1(\alpha &lt; e(x) &lt; 1 - \alpha)$</td>
<td></td>
<td>$\left( \frac{1(\alpha &lt; e(x) &lt; 1 - \alpha)}{e(x)}, \frac{1(\alpha &lt; e(x) &lt; 1 - \alpha)}{1-e(x)} \right)$</td>
</tr>
<tr>
<td>matching</td>
<td>$\min{e(x), 1 - e(x)}$</td>
<td></td>
<td>$\left( \frac{\min{e(x), 1-e(x)}}{e(x)}, \frac{\min{e(x), 1-e(x)}}{1-e(x)} \right)$</td>
</tr>
</tbody>
</table>
Overlap Weighting (OW)

- Consistent estimator for $\tau^h$ with any $h$:
  \[
  \hat{\tau}^h = \hat{\tau}_1 - \hat{\tau}_0 = \frac{\sum_{i=1}^{n} Z_i Y_i w_i(X_i)}{\sum_{i=1}^{n} Z_i w_i(X_i)} - \frac{\sum_{i=1}^{n} (1 - Z_i) Y_i w_i(X_i)}{\sum_{i=1}^{n} (1 - Z_i) w_i(X_i)}
  \]

  with balancing weights

  \[
  \begin{cases}
  w(X_i) \propto \frac{h(X_i)}{e(X_i)} & \text{for } Z_i = 1, \\
  w(X_i) \propto \frac{h(X_i)}{1 - e(X_i)} & \text{for } Z_i = 0
  \end{cases}
  \]

- Overlap weights are defined by choosing $h(X)$ that minimizes the asymptotic variance of $\hat{\tau}^h \Rightarrow h(X) = e(X)(1 - e(X))$

- Overlap weights

  \[
  \begin{cases}
  w(X_i) \propto 1 - e(X_i) & \text{for } Z_i = 1, \\
  w(X_i) \propto e(X_i) & \text{for } Z_i = 0
  \end{cases}
  \]

- First conceived by Alan Zaslavsky; one of the earliest use: Schneider et al. (2001, JAMA)
Overlap Weighting (OW) - Cont’d

Conceptual Advantages

- Target population $g(X) \propto f(X)e(X)(1 - e(X))$ emphasizes units at clinical equipoise, i.e., with substantial probability of receiving both treatments (substantial overlap in covariates)

- Addresses the RWD question: **How inclusive can we be without compromising validity?**

- Exemplifies the principle of “observational studies analyzed like randomized trials”

- $\tau = \mathbb{E}_g [Y(1) - Y(0)]$ – average treatment effect among the overlap population (ATO)
Overlap Weighting (OW) - Cont’d

Statistical Advantages

- **Minimum variance** of the nonparametric estimator among all balancing weights
- Weights are **bounded** (unlike IPW)
- **Continuously down-weights units in the tails**, avoids *ad hoc* trimming
- **Exact balance** for means of included covariates in logistic propensity score model (next page)
Framingham Heart Study - Results

All-cause Mortality

Method

- Unadjusted
- IPTW
- OW
- Sym 0.10
- Asym 0.10

Hazard Ratio

0.5 0.7 1.0 1.5 2.0
Overlap Weighting: Exact Balance

**Theorem.** When the propensity scores are estimated by maximum likelihood under a logistic regression model, 

\[
\text{logit}\{e(x_i)\} = \beta_0 + x'_i \beta,
\]

the overlap weights lead to exact balance in the means of any included covariate between treatment and control groups:

\[
\frac{\sum_i x_{ij} Z_i (1 - \hat{e}_i)}{\sum_i Z_i (1 - \hat{e}_i)} = \frac{\sum_i x_{ij} (1 - Z_i) \hat{e}_i}{\sum_i (1 - Z_i) \hat{e}_i}, \quad \text{for } j = 1, \ldots, p, \tag{2}
\]

where \(\hat{e}_i = \{1 + \exp[-(\hat{\beta}_0 + x'_i \hat{\beta})]\}^{-1}\) and \(\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_j)\) is the MLE for the regression coefficients.

- **Remark:** the exact balance property applies to any included covariate or derived covariate, including high order terms and interaction terms of the covariates.
Framingham Heart Study - Balance diagnostics

Variable Names

- Age
- BMI
- DBP
- Diabetes
- Female
- Fram. risk
- Glucose
- HDL Chol
- MI hx
- PAD hx
- BP
- SBP med
- Smoking
- Stroke hx
- Total Chol
- Triglycerides

Standardized Mean Differences

Method

- IPTW
- Trimmed
- OW
Balancing Weights: Augmented Estimator
Mao et al. 2019

- We can augment the nonparametric weighting estimator with an outcome regression $\hat{\mu}_z(X_i) = \hat{E}[Y(z)|X]$, for $z = 0, 1$.

- The augmented estimator of any balancing weight $w$ (equivalently, tilting function $h$):
  
  $$
  \hat{\tau}^h_{\text{aug}} = \hat{\tau}^h_1 - \hat{\tau}^h_0 - \left( \frac{\sum_{i=1}^n (1 - e(X_i))W_i\hat{\mu}_1(X_i)}{\sum_{i=1}^n h(X_i)} - \frac{\sum_{i=1}^n e(X_i)W_i\hat{\mu}_0(X_i)}{\sum_{i=1}^n h(X_i)} \right)
  $$

- Same form as the augmented IPW (double-robust) estimator
  - $\hat{\mu}^h_{\text{aug}}$ is semiparametric efficient for estimating $\tau^h$ when both the PS model and the outcome regression model are correctly specified
  - But not double-robust because the estimand depends on propensity score; no big deal in practice, still more efficient and less bias than the nonparametric estimator in most cases
We are developing a R package **PSweight**, which incorporates:

- Overlap weighting
- Inverse probability weighting, with trimming
- Binary treatment and multiple treatments
- Nonparametric estimator and augmented estimator
- Continuous, binary, count, survival outcome
- Diagnostic tables and graphics
PSweight demo

- PSweight R package available on OW website
  https://www2.stat.duke.edu/~fl35/OW.html

- R function `OW(ps.formula, yname, data)`
  - `ps.formula`: the propensity model (ie. \( Z \sim X \))
  - `yname`: character name of the outcome variable
  - `data`: the input dataframe (including \( X, Y, Z \))
    - \( X \): the \( n \times p \) matrix of covariates for the PS model (w/out intercept term)
    - \( Y \): the \( n \times 1 \) vector of outcomes
    - \( Z \): the \( n \times 1 \) vector of binary treatment status
Short Demonstration

Generate data similar to Framingham

```r
> X1 = matrix(rbinom(6*1000,1, .25), nrow=1000, ncol=6)
> X2 = rmvnorm(1000, mean = rep(0, 10), diag(rep(1,10)))
> X = cbind(X1, X2)

> coef = matrix(c(1,.5,.5,.25,0,0,1,.5,.5,.5,.25,.25,.25,0,0,0))
> linear = X %*% coef
> propensity = exp(-3 + linear)/(1 + exp(-3 + linear))
> Z = rbinom(1000, 1, propensity)

> Y = rnorm(1000, 130 + X %*% coef*4 - Z*10, 30)

> test_data = data.frame(Z=Z, data.frame(X), Y=Y)
```
Generated data

> head(Y)
[1] 182.00957 158.66518 107.36351 113.06511 49.31825 110.11113

> head(Z)
[1] 0 1 0 0 0 0

> X[1,]
[1] 0.0000000 0.0000000 0.0000000 1.0000000 0.0000000
[6] 1.0000000 0.5714657 -1.2378502 -0.6383928 -0.2067148
[11] 0.6171518 -0.9211108 -0.3725148 -0.4424818 0.8781408
[16] -1.3936681
Distribution of the generated propensity score

![Graph showing the distribution of propensity scores for untreated and on statins patients.](image-url)
**OW Analysis of Continuous Outcome**

```r
> mean(Y[Z==1]) - mean(Y[Z==0])
[1] -2.73

> form.ps <- "Z ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + X11 + X12 + X23 + X14 + X15 + X16"

> res <- OW(ps.formula = form.ps, yname = 'Y', data = test_data)

> summary(res)

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std.Error</th>
<th>Lower.CL</th>
<th>Upper.CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>-9.98</td>
<td>2.96</td>
<td>-15.79</td>
<td>-4.18</td>
</tr>
</tbody>
</table>
```
Outcome model augmentation

```r
> form.out <- "Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + X10 + + X11 + X12 + X23 + X14 + X15 + X16"

> res.aug <- OW(ps.formula = form.ps, yname = 'Y', data = test_data, + augmentation = TRUE, out.formula = form.out, family = "gaussian")

> summary(res.aug)

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std.Error</th>
<th>Lower.CL</th>
<th>Upper.CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-10.22</td>
<td>-16.04</td>
<td>-4.40</td>
</tr>
</tbody>
</table>

> summary(res)

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std.Error</th>
<th>Lower.CL</th>
<th>Upper.CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-9.98</td>
<td>-15.79</td>
<td>-4.18</td>
</tr>
</tbody>
</table>
```
Comparison to IPW with/out augmentation

```r
> res.ipw <- IPW(ps.fomula = form.ps, yname = 'Y', data = test_data)

> res.ipw.aug <- IPW(ps.fomula = form.ps, yname = 'Y', data = test_data,
  + augmentation= TRUE, formula2=form.out, family="gaussian")

> summary(res.ipw)

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std.Error</th>
<th>Lower.CL</th>
<th>Upper.CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 -5.50</td>
<td>3.00</td>
<td>-11.40</td>
<td>0.392</td>
</tr>
</tbody>
</table>

> summary(res.ipw.aug)

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Std.Error</th>
<th>Lower.CL</th>
<th>Upper.CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 -9.56</td>
<td>3.93</td>
<td>-17.27</td>
<td>-1.86</td>
</tr>
</tbody>
</table>
```
Result Summary

Estimated treatment effect

Method

Difference in mean Y

-15 -10 -5 0
Corresponding balance diagnostics

> res$smdplot
> res.ipw$smdplot
## Target Population

- res$demo
- res.ipw$demo

<table>
<thead>
<tr>
<th></th>
<th>Unadjusted</th>
<th>Adjusted</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Statin users N=348</td>
<td>No statins N=3008</td>
<td>Overall Unadjusted N=3356</td>
<td>IPTW Weighted N=3356</td>
<td>IPTW Sym Trim (0.10) N=1086</td>
<td>Overlap Weighted N=3356</td>
<td></td>
</tr>
<tr>
<td>Age – yr*</td>
<td>64 (57, 70)</td>
<td>57 (50, 65)</td>
<td>58 (51, 66)</td>
<td>58 (51, 66)</td>
<td>64 (58, 70)</td>
<td>63 (57, 69)</td>
<td></td>
</tr>
<tr>
<td>Female sex</td>
<td>44.3</td>
<td>54.3</td>
<td>53.2</td>
<td>52.1</td>
<td>41.8</td>
<td>44.2</td>
<td></td>
</tr>
<tr>
<td>Systolic Blood Pressure – mmHg</td>
<td>131 (119, 144)</td>
<td>125 (114, 138)</td>
<td>126 (114, 138)</td>
<td>126 (115, 139)</td>
<td>132 (120, 145)</td>
<td>131 (119, 144)</td>
<td></td>
</tr>
<tr>
<td>Diastolic Blood Pressure – mmHg</td>
<td>76 (69, 82)</td>
<td>75 (68, 81)</td>
<td>75 (69, 81)</td>
<td>75 (69, 81)</td>
<td>76 (69, 81)</td>
<td>76 (69, 81)</td>
<td></td>
</tr>
<tr>
<td>BMI – kg/m²</td>
<td>29 (26, 32)</td>
<td>27 (25, 31)</td>
<td>28 (25, 31)</td>
<td>28 (25, 31)</td>
<td>29 (26, 32)</td>
<td>29 (26, 32)</td>
<td></td>
</tr>
<tr>
<td>History of CVD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Myocardial Infarction</td>
<td>16.4</td>
<td>3.0</td>
<td>4.4</td>
<td>5.9</td>
<td>13.6</td>
<td>12.7</td>
<td></td>
</tr>
<tr>
<td>Stroke</td>
<td>5.5</td>
<td>1.8</td>
<td>2.2</td>
<td>2.1</td>
<td>5.2</td>
<td>4.5</td>
<td></td>
</tr>
<tr>
<td>Peripheral Artery Dis.</td>
<td>7.5</td>
<td>2.2</td>
<td>2.7</td>
<td>2.7</td>
<td>6.9</td>
<td>5.9</td>
<td></td>
</tr>
<tr>
<td>Any ASCVD</td>
<td>35.3</td>
<td>8.9</td>
<td>11.6</td>
<td>13.4</td>
<td>32.8</td>
<td>29.4</td>
<td></td>
</tr>
</tbody>
</table>
Recall $\tau_1 = Pr_g[Y(1) = 1] = \mathbb{E}_g[Y(1)] = \frac{\mathbb{E}_f[h(X)\mu_1(X)]}{\mathbb{E}_f(h(X))}$, 
$\tau_0 = Pr_g[Y(0) = 1] = \mathbb{E}_g[Y(0)] = \frac{\mathbb{E}_f[h(X)\mu_0(X)]}{\mathbb{E}_f(h(X))}$

$\tau^{ATO} = \tau_1 - \tau_0$ interpreted as the causal risk difference

We define the causal risk ratio and odds ratio among the overlap population as

$\tau_{RR} = \frac{\tau_1}{\tau_0}$, $\tau_{OR} = \frac{\tau_1/(1 - \tau_1)}{\tau_0/(1 - \tau_0)}$

Estimate PS, and use

$\hat{\tau}_1 = \frac{\sum_{i=1}^{n} Z_i Y_i (1 - \hat{e}_i)}{\sum_{i=1}^{n} Z_i (1 - \hat{e}_i)}$, $\hat{\tau}_0 = \frac{\sum_{i=1}^{n} (1 - Z_i) Y_i \hat{e}_i}{\sum_{i=1}^{n} (1 - Z_i) \hat{e}_i}$
Section 3.

Multiple (or multi-valued) treatments

- Motivating Example: Racial Disparity in Health Expenditure
- Standard methods
- Balancing weights
- Generalized overlap weights
Example: Racial Disparity in Medical Expenditure

- **Goal:** estimate racial disparity in medical expenditure after balancing covariates (Le Cook et al, 2009)

- **Data:** 2009 Medical Expenditure Panel Survey (MEPS): 9830 Whites, 4020 Blacks, 1446 Asians, 5150 Hispanics

- **“Treatments”:** 4 racial groups; race not manipulable, *unconfounded descriptive comparison*

- **Outcome:** total medical expenditure

- **Confounders:** patient health status reflecting clinical appropriateness and need
Standard Setup

- Data: a random sample of \( n \) units drawn from a population
- Treatments: \( Z_i \in \{1, \ldots, J\} \) with \( J \geq 3 \)
- For each unit \( i \), a set of potential outcomes \( \{Y_i(1), \ldots, Y_i(J)\} \), only \( Y_i(Z_i) \) observed
- Observed data: pre-treatment variables (covariates) \( X_i \), treatment status \( Z_i \) and \( Y_i = Y_i(Z_i) \)
- **Estimand**: for unordered nominal treatments, pairwise average treatment effect (pATE)
  \[
  \tau_{j,j'}^{\text{pATE}} = \mathbb{E}[Y(j) - Y(j')] , \quad j \neq j'
  \]
- The equivalent estimand is \( \mu_j = \mathbb{E}[Y(j)] \) for all \( j \), where \( \mathbb{E}[\cdot] \) is over the combined population
Generalized Propensity Score (GPS)
(Imbens, 2000)

Definition: Generalized Propensity Score (GPS) – the conditional probability of being assigned to a treatment group given the covariates:

\[ e_j(X) \equiv \Pr(Z = j|X) \]

- Each unit has \( J \) GPSs: \( e = \{e_1, \ldots, e_J\} \), and \( \sum_{j=1}^{J} e_j(X) = 1 \)
- Example: \( J = 3 \), three units with \( e = (.3, .6, .1) \), \( (.3, .25, .45) \), \( (.3, .1, .6) \)
- GPS is usually estimated by a multinomial logistic regression
- Individual matching less suited to multiple treatments (Imbens, 2000)
Example: Racial Disparity in Medical Expenditure

- Imbalance
- Propensity to Asian
- Propensity to Black

Unweighted IPW
0.0 0.5 1.0 1.5
Imbalance
Propensity to Asian
Estimated GPS
Density
0.0 0.2 0.4 0.6 0.8 1.0
White Group
Asian Group
Black Group
Hispanic Group

Density
Estimated GPS
0.0 0.2 0.4 0.6 0.8 1.0
White Group
Asian Group
Black Group
Hispanic Group
Causal Assumptions

(Imbens, 2000)

- **(Weak Unconfoundedness)** The assignment is weakly unconfounded if for all $j$:

  $$Y(j) \perp \mathbb{1}\{Z = j\}|X$$

  $$\Rightarrow Y(j) \perp X|e_j(X)$$ for all $j$

- **(Overlap)** The probability of being assigned to any treatment group $e_j(X) = \Pr(Z = j|X) > 0$ for all $X$ and $j$
Previous Methods

- Matching (Lechner 2002)

- Subclassification (Zanutto et al. 2005)

- Vector matching (Lopez and Gutman, 2017): match on the vector of GPS, infeasible with even moderate number of treatments

- Inverse probability weighting (Feng et al. 2012; McCaffrey et al. 2013)

- Trimming: *ad hoc* choice of threshold

- Optimal trimming (Yang et al. 2016): data-dependent, results in ambiguous target population

- Matching weights (Yoshida, et al. 2017): special case of balancing weights, but not optimal and non-smooth
Inverse Probability Weighting

- Foundation: similar to binary treatments, we can use weighting to identify

\[ \mu_j = \mathbb{E}[Y(j)] = \mathbb{E}_X \left\{ \mathbb{E} \left[ \frac{1\{Z_i = j\}Y_j}{e_j(X)} \right] \right\} \]

- Inverse probability weights (IPW):

\( \{w_1(X_i), \ldots, w_J(X_i)\} = \left\{ \frac{1}{e_1(X_i)}, \ldots, \frac{1}{e_J(X_i)} \right\} \)

- A consistent nonparametric estimator of pairwise ATE is

\[ \hat{\tau}_{j,j'}^{\text{pATE}} = \frac{\sum_{i=1}^{n} 1\{Z_i = j\}Y_i/e_j(X_i)}{\sum_{i=1}^{n} 1\{Z_i = j\}/e_j(X_i)} - \frac{\sum_{i=1}^{n} 1\{Z_i = j'\}Y_i/e_{j'}(X_i)}{\sum_{i=1}^{n} 1\{Z_i = j'\}/e_{j'}(X_i)} \]

- IPW balances the weighted distribution of pre-treatment covariates across multiple groups relative to the combined population
IPW and Trimming: Challenges

- **Target population of IPW:** Among the combined patients receiving any of $J$ treatments
- **Relevant question:** Among patients who could reasonably receive any of these $J$ treatments, what are the pairwise treatment effects?
- **Same challenges as binary treatment case, but exaggerated**
- **More opportunities for extreme propensities to arise; variance inflation and bias**
- **Solutions like trimming are hard to apply**
  - May lose a lot of patients
  - Rule of thumb for cut points is hard to specify
  - Order of trimming matters
Assume sample drawn from density $f(X)$, can represent the density of target population by $g(X) = f(X)h(X)$, where $h(\cdot)$ is a *tilting function*.

Define $\mu_j(X) = \mathbb{E}[Y(j)|X]$ as the regression function.

Average potential outcome in the target population $g(x)$

$$
\mu^h_j = \frac{\mathbb{E}[\mu_j(X)h(X)]}{\mathbb{E}[h(X)]} = \mathbb{E}_g[Y(j)].
$$
Balancing Weights

- Recall for all $j$

$$f_j(X) = f(X|Z = j) \propto f(X)e_j(X)$$

- For a target population $g(X) = f(X)h(X)$, to estimate its average potential outcome of group $j$, $\mu^h_j$, we use the following weights to reweigh $f_i(X)$ to $g(X)$

$$w_j(X) \propto \frac{f(X)h(X)}{f(X)e_j(X)} = \frac{h(X)}{e_j(X)}$$

- The class of weights $\{h(X)/e_1(X), \ldots, h(X)/e_J(X)\}$ is the balancing weights for multiple treatments:

$$f_j(X)w_j(X) = f(X)h(X) = g(X)$$

- Balancing weights balance the weighted distributions of covariates across $J$ comparison groups

- IPW is obtained by setting $h(X) = 1$
Let \( a = (a_1, \ldots, a_J) \) be a vector,

**Estimand:** \( \tau_h(a) = \sum_{j=1}^{J} a_j \mu^h_j \)

Choice of coefficient \( a \) determines the causal estimand.

For nominal treatments, we focus on the \( a \) corresponding to pairwise comparisons, e.g. \((1, -1, 0, \ldots, 0)\)

Choice \( h \) determines the target population and weights.

Setting \( a \) to the pairwise set and \( h(X) = 1 \), we obtain pairwise ATE.
## Balancing Weights: Examples

<table>
<thead>
<tr>
<th>Target population</th>
<th>Tilting function $h(X)$</th>
<th>Weights ${w_j(X)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined</td>
<td>1</td>
<td>${1/e_j(X)}$</td>
</tr>
<tr>
<td>Treated ($j'$th group)</td>
<td>$e_{j'}(X)$</td>
<td>${e_{j'}(X)/e_j(X)}$</td>
</tr>
<tr>
<td>Trimming</td>
<td>$\mathbb{1}{X \in \mathbb{C}}$</td>
<td>${\mathbb{1}{X \in \mathbb{C}}/e_j(X)}$</td>
</tr>
<tr>
<td>Matching</td>
<td>$\min_{1 \leq l \leq J}{e_l(X)}$</td>
<td>${\min_l{e_l(X)}/e_j(X)}$</td>
</tr>
<tr>
<td>Overlap</td>
<td>$\frac{1}{\sum_{l=1}^{J} 1/e_l(X)}$</td>
<td>$\left{ \frac{1/e_j(X)}{\sum_{l=1}^{K} 1/e_l(X)} \right}$</td>
</tr>
</tbody>
</table>
Weighting Estimator

- General principle: estimating the average of the potential outcomes separately for each treatment level with the balancing weights, 
  \( w_j(X) = h(X)/e_j(X) \)

- Consistent estimator

\[
\hat{\mu}_j^h = \mathbb{E}_h[Y(j)] = \frac{\sum_{i=1}^{n} 1(Z_i = j)Y_i w_j(X_i)}{\sum_{i=1}^{n} 1(Z_i = j)w_j(X_i)}
\]

\[
\hat{\tau}^h(a) = \sum_{j=1}^{J} a_j \hat{\mu}_j^h
\]
For nominal treatments, the following $h$ minimizes the total asymptotic variance of the weighting estimators for all pairwise comparisons (Proposition 3, Li and Li 2019)

$$\tilde{h}(X) = \frac{1}{\sum_{l=1}^{J} 1/e_l(X)}$$

Consequently, the (generalized) overlap weights for multiple treatments:

for treatment $j$

$$w_j(X) = \frac{1/e_j(X)}{\sum_{l=1}^{J} 1/e_l(X)}$$
Generalized Overlap Weights
(Li and Li, 2019)

- Maximum $h$ is attained when $e_j(X) = 1/J$ for all $j$ – substantial probability to receive each treatment

- **Target population**: subpopulation with the most overlap in covariates among all groups

- **Target estimand**: pairwise average treatment effect among the overlap population (pATO)
Optimal Tilting Function: Ternary Plot

- For $J = 3$, visualize $h(e_1(X), e_2(X), e_3(X))$ over a two-dimensional probability simplex

![Ternary Plot Image]
Generalized Overlap Weights: Statistical Advantages

- **Maximum total efficiency** for pairwise comparisons among all balancing weights

- Weights are by construction **bounded** and **robust** to extreme propensities (prevalence with multiple treatments)

- Avoid *ad hoc* trimming decisions: **continuously down-weighting the units along the “edges”**

- Simulations confirmed that causal comparisons enabled by generalized overlap weights are consistently more efficient than *ad hoc* trimming methods
Balance Check

- Nominal treatments: GPS estimated by multinomial model
- Adequacy of GPS model informed by overlap-weighted covariate balance
- Recall
  \[ f_j(X)w_j(X) = f(X)h(X) = f_{j'}(X)w_{j'}(X), \quad j \neq j' \]
- Population standardized difference (PSD)
  \[ \text{PSD}_j = \frac{|\bar{X}_j - \bar{X}_p|}{S_X}; \quad \max_j \{\text{PSD}_j\} \]
- Absolute standardized differences (ASD)
  \[ \text{ASD}_{j,j'} = \frac{|\bar{X}_j - \bar{X}_{j'}|}{S_X}; \quad \max_{j \neq j'} \{\text{ASD}_{j,j'}\} \]

where \( \bar{X}_j \) is the weighted covariate mean, \( \bar{X}_p \) is the covariate mean in the target population, \( S_X \) is the pooled standard deviation
Example: Racial Disparity in Medical Expenditure

Figure: Boxplots of population standardized difference (PSD) and absolute standardized difference (ASD) for all covariates
Example: Racial Disparity in Medical Expenditure

- Estimates (CIs) for difference in medical expenditure ($)

<table>
<thead>
<tr>
<th></th>
<th>White-Asian</th>
<th>White-Black</th>
<th>White-Hispanic</th>
</tr>
</thead>
<tbody>
<tr>
<td>IPW</td>
<td>2402</td>
<td>908</td>
<td>719</td>
</tr>
<tr>
<td></td>
<td>(530, 4274)</td>
<td>(505, 1311)</td>
<td>(129, 1309)</td>
</tr>
<tr>
<td>Trimming</td>
<td>1335</td>
<td>1148</td>
<td>1257</td>
</tr>
<tr>
<td></td>
<td>(671, 1999)</td>
<td>(781, 1515)</td>
<td>(804, 1711)</td>
</tr>
<tr>
<td>Overlap</td>
<td>1160</td>
<td>886</td>
<td>1221</td>
</tr>
<tr>
<td></td>
<td>(660, 1661)</td>
<td>(518, 1253)</td>
<td>(849, 1593)</td>
</tr>
</tbody>
</table>

- One Asian subject has over 30% of the weight (out of 1446 Asians)

- Optimal trimming excludes 2125 Whites, 44 Asians, 1001 Blacks and 603 Hispanics
Simulated Example

- Consider $J = 3$ groups with total sample size $n = 1500$
- Generate $Z|X$ from a multinomial logistic model
- Specify response function $Y(j)|X$ for all $j$
- Consider adequate overlap and lack of overlap (↓)
## Simulated Example

<table>
<thead>
<tr>
<th></th>
<th>Absolute Bias</th>
<th></th>
<th></th>
<th>RMSE</th>
<th></th>
<th></th>
<th>95% Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\tau_{1,2}$</td>
<td>$\tau_{1,3}$</td>
<td>$\tau_{2,3}$</td>
<td>$\tau_{1,2}$</td>
<td>$\tau_{1,3}$</td>
<td>$\tau_{2,3}$</td>
<td>$\tau_{1,2}$</td>
</tr>
<tr>
<td>IPW</td>
<td>0.19</td>
<td>0.02</td>
<td>0.17</td>
<td>1.04</td>
<td>0.61</td>
<td>1.16</td>
<td>0.79</td>
</tr>
<tr>
<td>Opt Trim</td>
<td>0.03</td>
<td>0.01</td>
<td>0.01</td>
<td>0.38</td>
<td>0.28</td>
<td>0.47</td>
<td>0.93</td>
</tr>
<tr>
<td>Overlap</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.28</td>
<td>0.23</td>
<td>0.35</td>
<td>0.95</td>
</tr>
</tbody>
</table>

- Generalized overlap weights: smallest bias, largest efficiency and nominal coverage
- Similar findings with $J = 4$ and $J = 6$
Augmented Estimator

- For each balancing weight $w$, we can augment the nonparametric weighting estimator with an outcome regression $\hat{\mu}_j(X_i) = \hat{E}[Y(j)|X]$

- The augmented estimator:

$$\hat{\mu}^{h,\text{aug}}_j = \hat{\mu}_j - \frac{\sum_{i=1}^{n}(D_{ij} - e_j(X_i))w_j(X_i)\hat{\mu}_j(X_i)}{\sum_{i=1}^{n} h(X_i)},$$

where $D_{ij} = 1\{Z_i = j\}$

- $\hat{\mu}^{h,\text{aug}}_j$ is semiparametric efficient for estimating $\mu^h_j$ when both the PS model and the outcome regression model are correctly specified

- Not double-robust because the estimand depends on propensity score; nonetheless, more efficient than the nonparametric estimator in most cases
Ordinal Treatments

- Target estimands: require different choice of $a$
- Example: the quadratic contrasts between unit increases in the treatment level
  \[ \tau^h = (\mu^h_{j+1} - \mu^h_j) - (\mu^h_j - \mu^h_{j-1}) \]
- Example: weighted average of unit increase in the treatment level
  \[ \tau^h = \sum_{j=1}^{J-1} \pi_j (\mu^h_{j+1} - \mu^h_j) \]
- Example: the accumulative effect of the maximum treatment,
  \[ \tau^h = \mu^h_J - \mu^h_1 \]
- The general framework of balancing weights still applies
Section 4.

**Extensions**

- Heterogeneity of treatment (HTE)
- Subgroup analysis (SGA)
- Time-varying treatments
Heterogeneity of treatments: Setup

- For simplicity, back to the case of binary treatments
- Subgroup analysis as a special case of HTE (common one)
- Subgroup: $S$, defined by indicator functions of covariates $X$
- Examples of $S$
  - one block covariate: sex, age, etc
  - intersection of multiple covariates: interaction terms – most common in clinical research, focus here

- Target estimand:

$$\tau(S) = \mathbb{E}[Y(1) - Y(0)|x \in S]$$

- Existing methods: Subgroup balancing propensity score (Dong et al. 2019, SMMR), focus on IPW.
Example: COMPARE-UF

- **Goal:** Compare myomectomy to hysterectomy for treatment of uterine fibroids

- **Patients:** COMPARE-UF Registry with 557 and 721 eligible patients, respectively

- **Outcomes:** Quality of life, recovery time, treatment failure

- **Confounders:** Age, race, baseline quality of life, bleeding symptoms, bulk symptoms, duration of symptoms, insurance status, prior procedures, uterine volume...

- **Subgroups:** Age, race, bleeding symptoms (specified a-priori)...
Poor Balance within Subgroups

Even after IPW women younger than 40 have poor balance:
Overlap Weights: Exact Balance Revisited
(Li, Morgan, Zaslavsky, 2018)

**Theorem.** When the propensity scores are estimated by maximum likelihood under a logistic regression model,

\[
\text{logit}\{e(x_i)\} = \beta_0 + x_i \beta',
\]

the overlap weights lead to exact balance in the means of any included covariate between treatment and control groups:

\[
\frac{\sum_i x_{ij} Z_i (1 - \hat{e}_i)}{\sum_i Z_i (1 - \hat{e}_i)} = \frac{\sum_i x_{ij} (1 - Z_i) \hat{e}_i}{\sum_i (1 - Z_i) \hat{e}_i}, \quad \text{for } j = 1, \ldots, p,
\]  

(3)

where \( \hat{e}_i = \{1 + \exp[-(\hat{\beta}_0 + x_i \hat{\beta}')]\}^{-1} \) and \( \hat{\beta} = (\hat{\beta}_1, ..., \hat{\beta}_j) \) is the MLE for the regression coefficients.

- **Remark:** the exact balance property applies to any included covariate and derived covariate, including high order terms and interaction terms of the covariates
Overlap Weights: Exact Balance in Subgroups

Corollary. If the postulated propensity score model includes any interaction term with a binary variable, then the overlap weights lead to exact mean balance in the subgroups defined by that binary variable.

Remarks:

- Exact mean balance is achieved within subgroups by augmenting the propensity score model to include interactions between all adjustment variables and subgroups of interest.
Propensity Score Estimation: Bias-variance Tradeoff

- A richer PS model gives better balance and thus reduce bias, but at the cost of inflated variance of the weights

- Extreme case: a saturated PS model estimates PS to be exact 0 and 1 – variance goes to infinity

- Analytic judgment: where to stop in the bias-variance tradeoff?

- In practice: variable selection among all possible interactions between covariates and subgroups.
  - F Random forest
  - L LASSO
  - R Relaxed LASSO (using LASSO for selection, and maximum likelihood for estimation)
Comparison of Methods in COMPARE-UF: Age<40

**Figure:** Standardized mean differences (SMD) across alternative adjustment methods (inverse probability of treatment weighting, IPW, with main effects logistic regression, -M, LASSO, -L, for relaxed LASSO,-R, random forest -F; overlap weighting, OW, with the same suffixes).
Comparison of Methods in COMPARE-UF: All subgroups

Figure: Standardized mean differences (SMD) across alternative adjustment methods (inverse probability of treatment weighting, IPW, with main effects logistic regression, -M, LASSO, -L, for relaxed LASSO,-R, random forest -F; overlap weighting, OW, with the same suffixes).
Extensive simulations have shown: in the presence of treatment effect heterogeneity between subgroups

- Overlap weights give better subgroup balance than IPW, regardless of how the propensity score is estimated

- Among all propensity score models: LASSO performs the best, main effects model (i.e. no subgroups) the worst, boosting (twang package) in between

Practical guide for subgroup analysis:
1. Estimate propensity scores by using LASSO to select interaction terms in a logistic regression model
2. Use overlap weights based on the propensity score estimated above
Extension to Time-Varying Treatments

- Lack of overlap is even more severe in longitudinal (time-varying) treatments.

- Longitudinal treatments can be viewed as a special case of multiple treatments: Each observed treatment path is a group.

- Extension of balancing weights is much trickier – involve the counterfactual intermediate outcomes under all treatment paths.

- IPW bypass this problem at the cost of variance inflation.

- One possibility is to first impute all missing potential outcomes, and then re-weigh using balancing weights.

- But lose the simplicity appeal; infeasible with large $T$ (say $> 5$).

- Open question.
Acknowledgements

Collaborators

- Fan Li, PhD (Yale School of Public Health)
- Tianui Zhou (Duke)
- Guangyu Tong (Duke)
- Michael Pencina, PhD (Duke Biostatistics and Bioinformatics)
- Alan M Zaslavsky, PhD (Harvard)

- PCORI ME-2018C2-13289

- NSF SES 1424688
Key References


