STA 640 — Causal Inference

Chapter 6.2: Post-treatment confounding: Principal Stratification

Fan Li

Department of Statistical Science
Duke University

Post-Treatment Confounding

- So far most of the problems discussed adjust for pre-treatment confounding, i.e. covariates
- Confounding occurs after treatment (but before the final outcome) poses different challenges to causal inference
- Post-treatment confounding: a post-treatment intermediate variable D lies in the causal pathway between Z and Y:

$$Z \longrightarrow D \longrightarrow Y.$$

- Known as "endogenous" selection problems in economics
- Rosenbaum (1984) show: adjusting post-treatment variables D in the same way as pre-treatment covariates X leads to biased causal effects
- Include a wide range of (seemingly different) problems

Example: Noncompliance in Randomized Experiments

- Noncompliance in RCT is a special case of post-treatment/assignment confounding
- (Randomly) assigned treatment Z; actual trt D; outcome Y
- Noncompliance: Z usually strongly affects D, but still $D \neq Z$ for some units
- Post-assignment confounding: units with $Z = 1, D = d$ are usually not the same as units with $Z = 0, D = d$, and thus a direct comparison leads to biased causal estimate

Example: Censoring (or Truncation) by Death

(Zhang and Rubin, 2003, JEBS)

- Goal: randomized study of a drug's effect on Quality Of Life (QOL) two years after treatment
  - Treatment Z: randomized to trt (0) and control (1)
  - Outcome Y: QOL two years post-randomization
  - Intermediate outcome D_i: Indicator of two-year survival
- Complication: Some subjects will die before completion of the study; QOL for these subjects is not well defined
- Such outcomes are called "censored" or "truncated" by death
- Statistical challenge: QOL is only defined on survived units ($D = 1$). If the treatment has a non-zero effect on survival, then the survived trt units $Z = 1, D = 1$ are different from survived con units $Z = 1, D = 0$

Principal Stratification

(Frangakis and Rubin, 2002, Biometrics)

- Frangakis and Rubin (2002) generalized the IV approach to noncompliance (Angrist et al. 1996) to principal stratification, applicable to all post-treatment confounding
- Assuming a binary D: units can be classified into four groups according to the joint potential values of D, $S_i = (D_i(0), D_i(1))$:

$$
\begin{aligned}
00 &= \{i : D_i(0) = 0, D_i(1) = 0\} \\
10 &= \{i : D_i(0) = 1, D_i(1) = 0\} \\
01 &= \{i : D_i(0) = 0, D_i(1) = 1\} \\
11 &= \{i : D_i(0) = 1, D_i(1) = 1\}
\end{aligned}
$$

- This cross-classification of units is the principal stratification with respect to the (binary) post-treatment variable D.

Properties of Principal Stratification

- Key property: Principal stratum membership S_i is not affected by treatment assignment
- Principal stratum membership only reflects subject's characteristics: it can be viewed as a pre-treatment variable
- Therefore comparison of potential outcomes $Y_i(0)$ and $Y_i(1)$ is a well-defined causal effect, because it is defined on a common set of units (the same principal stratum)
- Principal Causal Effect (PCE):

$$\tau^{PCE} = \mathbb{E}[Y_i(1) - Y_i(0)|S_i = (d_0, d_1)] \quad d_0, d_1 \in \{0, 1\}$$

PS Example 1: Treatment Noncompliance

Angrist et al., 1996

- $D_i(z) =$ Treatment received given assignment z for $z = 0, 1$

$$D_i(z) = \begin{cases} 0, & \text{if subject } i \text{ received control given assignment } z; \\ 1, & \text{if subject } i \text{ received active trt given assignment } z. \end{cases}$$

$$\begin{aligned} 00 &= \{i : D_i(0) = 0, D_i(1) = 0\} = \text{ Never Takers} \\ 10 &= \{i : D_i(0) = 1, D_i(1) = 0\} = \text{ Defiers} \\ 01 &= \{i : D_i(0) = 0, D_i(1) = 1\} = \text{ Compliers} \\ 11 &= \{i : D_i(0) = 1, D_i(1) = 1\} = \text{ Always Takers} \end{aligned}$$

- Principal causal effect: Complier Average Causal Effects (CACE)

$$\tau^{CACE} = E[Y_i(1) - Y_i(0) \mid D_i(0) = 0, D_i(1) = 1]$$

PS Example 2: Censoring (or Truncation) by Death

$D_i(z) =$ Indicator for two-year survival given assignment z, $\quad z = 0, 1$

$$D_i(z) = \begin{cases} 0, & \text{if subject } i \text{ dies given assignment } z; \\ 1, & \text{if subject } i \text{ lives given assignment } z. \end{cases}$$

- Never Survivals: Subjects who will die no matter how treated

$$00 = \{i : D_i(0) = 0, D_i(1) = 0\}$$

- Defiant Survivals: Subjects who will die if treated but live otherwise

$$10 = \{i : D_i(0) = 1, D_i(1) = 0\}$$

- Compliant Survivals: Subjects who will live if treated but die otherwise

$$01 = \{i : D_i(0) = 0, D_i(1) = 1\}$$

- Always Survivals: Subjects who will live no matter how treated

$$11 = \{i : D_i(0) = 1, D_i(1) = 1\}$$

Censoring (or Truncation) by Death

- A well defined causal effect of the active treatment versus the control treatment on QOL exists only for the always-survivors $11 = \{i : D_i(0) = 1, D_i(1) = 1\}$:

$$\tau^{SACE} = E\Big[Y_i(1) - Y_i(0) \mid D_i(0) = 1, D_i(1) = 1\Big]$$

  where $SACE$ stands for Survival Average Causal Effect

- For the $10 = \{i : D_i(0) = 1, D_i(1) = 0\}$ and $01 = \{i : D_i(0) = 0, D_i(1) = 1\}$ groups, the average causal effect on QOL involves to assume we know how to trade off a particular QOL and being dead (and out of misery)

- For the $00 = \{i : D_i(0) = 0, D_i(1) = 0\}$ group there is no QOL to compare

Principal Stratification: Central Challenge

- The above three examples differ in settings and goal, but all share the same fundamental feature: post-treatment confounding bias
- All can be formulated via the principal stratification framework; additional examples of PS are given at the end of the slides
- Central challenge in inference of Principal Stratification: individual principal strata memberships are not observed, because of the fundamental problem of causal inference.
- Additional assumptions are required for identifying PCE
- Different PS settings share the same estimation and inferential procedure, but differ in estimands of interest and specific identification assumptions

PS Estimation and Inference: Assumptions

- Assumption 1: Unconfoundedness of treatment assignment

$$\{Y_i(0), Y_i(1), D_i(0), D_i(1)\} \perp Z_i \mid X_i$$

- Ignorability implies
  - Principal stratum membership S_i has the same distribution between the treatment arms (within cells defined by X)

$$D_i(0), D_i(1) \perp Z_i \mid X_i$$

  - Latent Ignorability ("latent" because conditioning on a latent variable: PS):

$$\{Y_i(0), Y_i(1)\} \perp Z_i \mid D_i(0), D_i(1), X_i$$

- Assumption 2: Monotonicity $D_i(1) \geq D_i(0)$ for all i. No defiers.

Principal Stratification: Basic structure of identification

- Focusing on the case of binary Z and D
- The observed (Z, D) cells consist of mixtures of principal strata:

Z	D	S
0	0	[C, NT]
0	1	[AT, D]
1	0	[NT, D]
1	1	[C, AT]

- Monotonicity assumptions reduce some of these mixtures to one component, but generally not enough to identify all principal causal effects
- Estimation of PS inherently involves latent mixture models: disentangle the latent mixtures (i.e. principal strata) from observed data

PS: Mixture Model Approach

- Six quantities are associated with each unit: $Y_i(1), Y_i(0), D_i(1), D_i(0), \mathbf{X}_i, Z_i$,
- Four are observed, $Y_i^{obs} = Y_i(Z_i), D_i^{obs} = D_i(Z_i), Z_i, \mathbf{X}_i$, and the rest two are unobserved $Y_i^{mis} = Y_i(1 - Z_i), D_i^{mis} = D_i(1 - Z_i)$;
- Consequently the principal strata membership $S_i = (D_i(0), D_i(1))$—the label of components in mixture model— is unobserved
- Two ways to handling the latent mixture label: (i) integrate out (expectation) the label; (ii) impute the label.
- Key questions in the latent mixture model approach: (i) What models do we need to specify? (ii) What is the likelihood?

PS: Complete data likelihood

- The probability density function of all random variables as

$$
\begin{aligned}
& \prod_i \Pr(Y_i(0), Y_i(1), D_i(0), D_i(1), Z_i, \mathbf{X}_i, \boldsymbol{\theta}) \\
= & \prod_i \Pr(Z_i | Y_i(0), Y_i(1), S_i, \mathbf{X}_i, \boldsymbol{\theta}) \Pr(Y_i(0), Y_i(1) | S_i, \mathbf{X}_i, \boldsymbol{\theta}) \Pr(S_i | \mathbf{X}_i, \boldsymbol{\theta}) \Pr(\mathbf{X}_i | \boldsymbol{\theta}) \\
\propto & \prod_i \Pr\left(Y_i(0) \mid S_i, \mathbf{X}_i; \boldsymbol{\theta}\right)^{(1-Z_i)} \Pr\left(Y_i(1) \mid S_i, \mathbf{X}_i; \boldsymbol{\theta}\right)^{Z_i} \Pr\left(S_i \mid \mathbf{X}_i; \boldsymbol{\theta}\right)
\end{aligned}
$$

  where $\boldsymbol{\theta}$ is the global parameter with prior distribution $p(\boldsymbol{\theta})$

- Second equality/proportional to sign is due to (1) unconfoundedness and (2) we condition on X
- So one needs to specify two models:
  - The principal strata model (S-model): $\Pr(S_i \mid \mathbf{X}_i, \boldsymbol{\theta})$
  - The outcome model given stratum (Y-model): $\Pr(Y_i(z) \mid S_i, \mathbf{X}_i, \boldsymbol{\theta})$

Complete intermediate data likelihood

- Complete intermediate data likelihood:

$$\prod_i \Pr\left(Y_i(0) \mid S_i, \mathbf{X}_i; \boldsymbol{\theta}\right)^{(1-Z_i)} \Pr\left(Y_i(1) \mid S_i, \mathbf{X}_i; \boldsymbol{\theta}\right)^{Z_i} \Pr\left(S_i \mid \mathbf{X}_i; \boldsymbol{\theta}\right).$$

- Without any constraints, the complete intermediate data likelihood is a product of four components, each corresponding to an observed cell of Z, D and being a mixture of two principal strata:

$$\begin{aligned} Lik \propto & \prod_{i:Z_i=0,D_i=0} \left(\pi_{i,c} f_{i,c0} + \pi_{i,n0} f_{i,n0}\right) \times \prod_{i:Z_i=0,D_i=1} \left(\pi_{i,a} f_{i,a0} + \pi_{i,d} f_{i,d0}\right) \\ & \times \prod_{i:Z_i=1,D_i=0} \left(\pi_{i,n} f_{i,n1} + \pi_{i,d} f_{i,d1}\right) \times \prod_{i:Z_i=1,D_i=1} \left(\pi_{i,a} f_{i,a1} + \pi_{i,c} f_{i,c1}\right), \end{aligned}$$

where $f_{i,sz} = \Pr(Y_i(z) | S_i = s, \mathbf{X}_i; \boldsymbol{\theta})$ and $\pi_{i,s} = \Pr(S_i = s | \mathbf{X}_i; \boldsymbol{\theta})$

- This is the latent mixture model

Parameter Estimation: EM algorithm

Zhang and Rubin, 2003, JBES

- The complete intermediate data likelihood is not directly observable because of the missing principal strata membership S
- With monotonicity and ER for noncompliers, the complete intermediate data likelihood reduces to:

$$Lik \propto \prod_{i:Z_i=0,D_i=0} \left(\pi_{i,c} f_{i,c0} + \pi_{i,n0} f_{i,n0}\right) \times \prod_{i:Z_i=1,D_i=0} \left(\pi_{i,n} f_{i,n1}\right) \times \prod_{i:Z_i=1,D_i=1} \left(\pi_{i,c} f_{i,c1}\right),$$

- Using the EM algorithm to estimate the parameters
  - E-step: the unobserved principal strata are replaced by their expectations given the data and the current estimates of the parameters
  - M-step: the likelihood conditional on the expected principal strata is maximized

Parameter Estimation: Bayesian Approach

Imbens and Rubin (1997, AOS)

- Similar to the likelihood approach: adding priors, substitute EM with posterior simulation via MCMC
- *Bayesian inference considers the observed values of the six quantities to be realizations of random variables, and the unobserved values to be unobserved random variables* (Rubin, 1978, AOS)
- Goal: to get the posterior predictive distributions of the missing data $Y_i^{mis} = Y_i(1 - Z_i), D^{mis} = D_i(1 - Z_i)$

Outline of Bayesian Inference

- The posterior predictive distribution of the missing potential outcomes is:

$$\begin{aligned} & \Pr(\mathbf{Y}^{mis}, \mathbf{D}^{mis} | \mathbf{Y}^{obs}, \mathbf{D}^{obs}, \mathbf{Z}, \mathbf{X}, \boldsymbol{\theta}) \qquad (1) \\ \propto & \prod_i \Pr(Y_i(0), Y_i(1) | S_i, \mathbf{X}_i, \boldsymbol{\theta}) \Pr(S_i | \mathbf{X}_i, \boldsymbol{\theta}) p(\boldsymbol{\theta}). \end{aligned}$$

  where $\boldsymbol{\theta}$ is the global parameter with prior distribution $p(\boldsymbol{\theta})$

- We need to specify (i) the outcome model $\Pr(Y_i(0), Y_i(1) \mid S_i, \mathbf{X}_i, \boldsymbol{\theta})$; (ii) the principal strata model: $\Pr(S_i \mid \mathbf{X}_i, \boldsymbol{\theta})$; and (iii) a prior distribution for $\boldsymbol{\theta}$: $p(\boldsymbol{\theta})$

- The above implicitly assumes: the parameters for each component in the second row are *a priori* distinct and independent

Outline of Bayesian Inference

- The posterior distribution of $\boldsymbol{\theta}$ is generally not tractable.
- One can use a Gibbs sampler to simulate from the joint posterior distribution $\Pr(\boldsymbol{\theta}, \mathbf{D}^{mis} \mid \mathbf{Y}^{obs}, \mathbf{D}^{obs}, \mathbf{Z}, \mathbf{X})$ by iteratively drawing between
  - $\Pr(\mathbf{D}^{mis} \mid \mathbf{Y}^{obs}, \mathbf{D}^{obs}, \mathbf{Z}, \mathbf{X}, \boldsymbol{\theta})$ (Parallel to the E-step)
  - $\Pr(\boldsymbol{\theta} \mid \mathbf{Y}^{obs}, \mathbf{D}^{obs}, \mathbf{D}^{mis}, \mathbf{Z}, \mathbf{X})$ (Parallel to the M-step)
- $\Pr(\boldsymbol{\theta} \mid \mathbf{Y}^{obs}, \mathbf{D}^{obs}, \mathbf{D}^{mis}, \mathbf{Z}, \mathbf{X})$ is proportional to the complete intermediate data likelihood:

$$\Pr(\boldsymbol{\theta} \mid \mathbf{Y}^{obs}, \mathbf{D}^{obs}, \mathbf{D}^{mis}, \mathbf{Z}, \mathbf{X})$$
$$\propto \quad p(\boldsymbol{\theta}) \prod_i \Pr\left(Y_i(0) \mid S_i, \mathbf{X}_i\right)^{(1-Z_i)} \Pr\left(Y_i(1) \mid S_i, \mathbf{X}_i\right)^{Z_i} \Pr\left(S_i \mid \mathbf{X}_i\right)$$

Weak Identifiability

- From the Bayesian perspective, PCEs are always identified under ignorability because with proper prior distributions of the parameters, posterior distributions of the causal estimands are always proper
- But some estimands are weakly identified, with substantial regions of flatness in their posterior distributions
- This is different from the none-or-all point identification under the frequentist paradigm
- Some weakly identifiable parameters are still informative about the causal effects
- Bayesian inference for causal estimands can be sharpened by additional assumptions such as monotonicity and ER
- Case study: Imbens and Rubin (1997, Ann Stat)

PS: pros and cons of flexibility

- Principal Stratification: a key strength is its flexibility in formulating a wide range of seemingly different settings
- However, different settings target at different principal strata and require different assumptions (besides unconfoundedness and monotonicity):

setting	target strata	assumptions
noncompliance	compliers (01)	ER for noncompliers
censoring by death	always-survivors (11)	case-dependent
selection bias in CRT	(11) & (01)	need additional data

- The flexibility brings challenges in providing a generally applicable algorithm, case-dependent implementation

Mixture models: pros and cons

- Conceptually, inference of all the PS settings can be handled by mixture models, straightforward to extend to more complex settings (e.g. clustering, covariates, non-binary IV)
- Challenges of mixture model: (i) requires substantial stat and programming expertise; (ii) often not stable
- Difficulty in implementation prevents the wide adoption of principal stratification
- An alternative approach is the weighting via *principal score* (Jo and Stuart, 2009)

“PStrata” R Package

- Estimation and inference of PS is challenging: mixture models are hard; Bayesian inference is hard (for most practitioners)
- R package “PStrata" (Liu and Li, 2022): implement most common PS settings via Bayesian mixture model: noncompliance, truncation by death; continuous/binary outcomes, survival outcomes
  - R: interface, main function, standardized inputs (S-model, Y-model, priors, setting, assumptions (e.g. monotonicity, ER))
  - Stan: backstage posterior sampling given a likelihood
  - C++: connecting R and Stan – take in R input and output text Stan code of the data likelihood
- Also implement the principal score approach
- `https://github.com/LauBok/PStrata`
- For illustration, see HW4.

References

Ding, P. and Lu, J. (2017). Principal stratification analysis using principal scores. JRSSB, 79, 757-777

Frangakis, C. E., Rubin, D. B. (2002). Principal stratification in causal inference. Biometrics, 58(1), 21-29.

Imbens, G. W., Rubin, D. B. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. The annals of statistics, 305-327.

Jiang, Z., Yang, S. and Ding, P. (2021+). Multiply robust estimation of causal effects under principal ignorability. JRSSB (forthcoming) https://arxiv.org/abs/2012.01615

Jin, H., Rubin, D. B. (2008). Principal stratification for causal inference with extended partial compliance. Journal of the American Statistical Association, 103(481), 101-111.

Jo, B., Stuart, E. A. (2009). On the use of propensity scores in principal causal effect estimation. Statistics in medicine, 28(23), 2857-2875.

References

Li F, Tian Z, Bobb J, Papadogeorgou G, Li F. (2021). Clarifying selection bias in cluster randomized trials. *Clinical Trials*. 19(1), 33-41.

Mercatanti A, Li F. (2017). Do debit cards decrease cash demands?: Causal inference and sensitivity analysis using Principal Stratification. Journal of Royal Statistical Society - Series C (Applied Statistics). 66(4), 759-776.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. Journal of the Royal Statistical Society: Series A (General), 147(5), 656-666.

Zhang, J. L., Rubin, D. B. (2003). Estimation of causal effects via principal stratification when some outcomes are truncated by“death”. Journal of Educational and Behavioral Statistics, 28(4), 353-368.

Zhang, J. L., Rubin, D. B., Mealli, F. (2009). Likelihood-based analysis of causal effects of job-training programs using principal stratification. Journal of the American Statistical Association, 104(485), 166-176.