

# 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: Selection Bias in Cluster Randomized Trials

Li et al. (2022, Clinical Trials)

- ▶ Cluster randomized trials (CRT): treatment assigned at cluster (not individual) level
- ▶ CRT often first assigns treatment and then recruits study units, open-labeled
- ▶ Subject to post-assignment selection bias, if there is obvious or perceived advantage of one treatment over another
- ▶ Assignment:  $Z$ ; recruitment indicator:  $D$
- ▶ The problem: the recruited sample consists of units with  $D = 1$ , but the recruited trt group ( $Z = 1, D = 1$ ) is different from the recruited control group  $Z = 0, D = 1$
- ▶ This is due to the post-assignment self selection of recruitment status, breaking the initial randomization.

# 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  $D$ ,  $S_i = (D_i(0), D_i(1))$ :

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

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

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

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

- ▶ 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  $z = 0, 1$

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

00 =  $\{i : D_i(0) = 0, D_i(1) = 0\}$  = Never Takers

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

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

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

- ▶ 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  $z$ ,  $z = 0, 1$

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

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

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

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

## Censoring by Death: Additional Examples

- ▶ Evaluating the causal effects of job training programs on wages (Zhang et al., 2008, 2009)

$D(z)$  = Indicator of employment given assignment  $z$

- ▶ Evaluating the causal effects of a special educational intervention on final test scores (Zhang & Rubin, 2003)

$D(z)$  = Graduation indicator given assignment  $z$

- ▶ Evaluating the causal effect of Breast Self-Examination (BSE) teaching courses on quality of execution of BSE (Mattei & Mealli, 2007)

$D(z)$  = Indicator of BSE practice given assignment  $z$

- ▶ Evaluating the effectiveness of degree programs on employment status of their graduates (Grilli & Mealli, 2008)

$D(z)$  = Graduation indicator given assignment  $z$

## PS Example 3: Selection Bias in CRT

Li et al., 2022

- ▶  $D_i(z)$ : Recruitment status given assignment  $z$ ,  $D_i(z) = 0$  un-recruited;  $D_i(z) = 1$  recruited

- ▶ Principal strata

00 =  $\{i : D_i(0) = 0, D_i(1) = 0\}$  = Never recruited

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

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

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

- ▶ The recruited units are all with  $D = 1$ ; we do not observe unrecruited units  $D = 0$
- ▶ What are the causal estimands: (i)  $\tau^{overall} = E[Y_i(1) - Y_i(0)]$  for the overall population; (ii)  $\tau^{recruited} = E[Y_i(1) - Y_i(0) \mid D_i = 1]$
- ▶  $\tau^{overall} \neq \tau^{recruited}$ , unless under homogeneous treatment

# A hypothetical example: selection bias in CRT

Li et al., 2022

| Principal Stratum<br>$S$ | Full Data                         |        | Potential outcome |        | Obs Data from a Randomized Study<br>Average of $(R, Y)$ given assignment |           |
|--------------------------|-----------------------------------|--------|-------------------|--------|--|-----------|
|                          | Post-random recruitment<br>$R(1)$ | $R(0)$ | $Y(1)$            | $Y(0)$ | $Z = 1$  | $Z = 0$   |
| Always                   | 1                                 | 1      | 30                | 10     | $(1, 27.5)$  | $(1, 10)$ |
| Compliant                | 1                                 | 0      | 25                | 10     |  | $(0, ?)$  |
| Never                    | 0                                 | 0      | 20                | 10     | $(0, ?)$   |           |

- ▶ Assume (i) monotonicity, i.e. no defiant-recruited, and (ii) equal proportions of each stratum in the overall population.
- ▶ True causal effects:  $\tau^a = 20$ ,  $\tau^c = 15$ ,  $\tau^n = 10$ ,  $\tau^{overall} = 15$
- ▶ A naïve ITT estimate of the recruited data:  $\hat{\tau} = 27.5 - 10 = 17.5$ , biased

## Example: Selection bias in CRT

- ▶ Core reason of selection bias in CRT:
  - ▶ The recruited trt group  $Z = 1, D = 1$  consists of always-recruited and compliant-recruited
  - ▶ The recruited control group  $Z = 1, D = 1$  consists of only always-recruited
- ▶ Thus the two groups are not comparable, breaking randomization.
- ▶ “Randomization protects against confounding, but not against selection bias when the selection occurs after the randomization” (Hernan and Robins, p103)
- ▶ Also known as *recruitment bias* or *identification bias* in literature
- ▶ Consequences
  - ▶ With recruited sample alone, we can only estimate  $\tau^{recruited}$
  - ▶ To estimate the overall effect  $\tau^{overall}$ , we need additional data on unrecruited units; standard covariate adjustment is not adequate

## 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):

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



## 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 Estimation and Inference: Moment-based

- ▶ In the simplest case of binary treatment, a single binary intermediate variable and no covariates, under monotonicity and exclusion restriction (ER), moment estimate (2SLS in Angrist et al. 2002) of CACE is available
- ▶ Large sample variance is available
- ▶ Without ER and/or monotonicity, one can instead find nonparametric bounds for CACE (e.g. Grilli and Mealli, 2007), but the bounds can be often too wide to be informative
- ▶ Limitations: (i) does not utilize the mixture structure, inefficient; (ii) difficult to extend to general setting, e.g. covariates, nonbinary treatments, truncation by death...

## 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

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

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):  $\Pr(S_i | \mathbf{X}_i, \boldsymbol{\theta})$
  - ▶ The outcome model given stratum (Y-model):  $\Pr(Y_i(z) | S_i, \mathbf{X}_i, \boldsymbol{\theta})$

## Complete intermediate data likelihood

- ▶ Complete **intermediate** data likelihood:

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

- ▶ 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:

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

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: Likelihood

- ▶ The complete intermediate data likelihood is not directly observable because of the missing principal strata membership  $S$
- ▶ How do assumptions simplify the computation?
  - ▶ Monotonicity:  $\pi_{i,d} = 0$
  - ▶ ER:  $f_{i,n0} = f_{i,n1} = f_{i,n}$ ;  $f_{i,a0} = f_{i,a1} = f_{i,a}$
- ▶ With monotonicity and ER, the complete intermediate data likelihood reduces to:

$$\begin{aligned} Lik \propto & \prod_{i:Z_i=0, D_i=0} (\pi_{i,c} f_{i,c0} + \pi_{i,n} f_{i,n}) \times \prod_{i:Z_i=0, D_i=1} (\pi_{i,a} f_{i,a}) \\ & \times \prod_{i:Z_i=1, D_i=0} (\pi_{i,n} f_{i,n}) \times \prod_{i:Z_i=1, D_i=1} (\pi_{i,a} f_{i,a} + \pi_{i,c} f_{i,c1}), \end{aligned}$$

# Parameter Estimation: EM algorithm

- ▶ 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
  
- ▶ Case study: Zhang and Rubin (2003, JBES)

# 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  $Y_i^{mis} = Y_i(1 - Z_i), D_i^{mis} = D_i(1 - Z_i)$



# Outline of Bayesian Inference of PS

- ▶ Assuming unconfoundedness
- ▶ The *posterior predictive distribution* of the missing potential outcomes is proportional to the complete data likelihood (as before):

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

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

- ▶ Besides the S-model and Y-model, we need specify a prior distribution for  $\boldsymbol{\theta}$ :  $p(\boldsymbol{\theta})$
- ▶ The above implicitly assumes: the parameters for each component in the second row of (1) are *a priori distinct and independent*

# Outline of Bayesian Inference of PS

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

$$\Pr(\theta \mid \mathbf{Y}^{obs}, \mathbf{D}^{obs}, \mathbf{D}^{mis}, \mathbf{Z}, \mathbf{X})$$

$$\propto p(\theta) \prod_i \Pr(Y_i(0) \mid S_i, \mathbf{X}_i)^{(1-Z_i)} \Pr(Y_i(1) \mid S_i, \mathbf{X}_i)^{Z_i} \Pr(S_i \mid \mathbf{X}_i)$$

## 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 (in Lindley's sense)
- ▶ 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, AOS)

# Case study: Effect of debit card on cash holding

Mercatanti and Li, 2017, JRSSC

- ▶ Goal: study the effects of using debit cards on the cash holding behaviour of consumers
- ▶ As debit cards can be viewed as a substitute of cash in some ways, they may reduce the cash demand, which is important for central banks as cash demand is an important factor in monetary policy
- ▶ Complication: some people who possess debit cards do not use them
- ▶ The interest is in the effect of **using** cards rather than **possessing** cards
- ▶ Key: cards are issued by banks (treatment assigned) but the choice of use depends on individual (treatment received), which can be viewed a post-treatment variable

## Effect of debit card on cash holding

- ▶  $Z_i$ : indicator of unit  $i$  possessing a debit card
- ▶  $D_i$ : indicator of unit  $i$  using a debit card
- ▶  $Y_i$ : average amount of cash held by unit  $i$
- ▶  $X_i$ : pre-treatment covariates
- ▶ Potential outcomes of  $Y$  and  $D$ .
- ▶ Principal strata of card usage:  $S_i = (D_i(0), D_i(1))$
- ▶ Causal estimands: CACE

$$\tau^{CATE} \equiv \mathbb{E}[Y_i(1) - Y_i(0) \mid S_i = c]$$

and CATT (average causal effect of the treated compliers)

$$\tau^{CATT} \equiv \mathbb{E}[Y_i(1) - Y_i(0) \mid S_i = c, Z_i = 1]$$

- ▶ In randomized experiments, CATT equals CACE.

## Assumptions

- ▶ Monotonicity (no defiers or always-takers):  $D_i(0) = 0$ . Satisfied automatically.
- ▶ Overlap:  $0 < \Pr(Z_i = 1|\mathbf{X}_i) < 1$ , for all  $i$ .
- ▶ Unconfoundedness:  $\{Y_i(1), Y_i(0), D_i(1), D_i(0)\} \perp Z_i|\mathbf{X}_i$ .
- ▶ Exclusion Restriction for never-takers:

$$\mathbb{E}[Y_i(1)|\mathbf{X}_i, S_i = n] = \mathbb{E}[Y_i(0)|\mathbf{X}_i, S_i = n].$$

- ▶ Exclusion Restriction for compliers: For all units with  $S_i = (0, 1)$ , the effect of card possession is only through using the card.

In observational studies, these assumptions must be conditional on covariates, and CATT and CATE are usually different

## Parametric Models

- ▶ A logistic regression model for principal stratum membership  $G$ :

$$\text{logit}(\Pr(S_i = n | X_i = x)) = \alpha_0 + x \cdot \alpha.$$

- ▶ A linear regression model for continuous potential outcomes, with different intercepts and slopes for different strata:

$$\begin{aligned} \Pr(Y_i(z) | S_i, X_i = x) = & \mathbf{1}_{S_i=c} \cdot (\beta_{c0} + z \cdot \theta_c + x \cdot \beta_{c1}) \\ & + \mathbf{1}_{S_i=n} \cdot (\beta_{n0} + x \cdot \beta_{n1}) + \epsilon_i, \end{aligned}$$

where  $\epsilon_i \sim N(0, \sigma^2)$  and  $\mathbf{1}_{S_i=s}$  is an indicator function

- ▶ Here  $\theta_c$  equals CATE, and the CATT can be subsequently estimated by averaging the differences between the observed outcomes for treated compliers and their estimated counterfactuals:

$$\widehat{\tau}^{CATT} = \frac{\sum_i D_i \cdot Z_i \cdot [Y_i - (\hat{\beta}_{c0} + X_i \cdot \hat{\beta}_{c1})]}{\sum_i D_i \cdot Z_i}.$$

## Extending Monotonicity

- ▶ Monotonicity plays a crucial role in identification and estimation of principal stratification
- ▶ To extend PS to complex settings, e.g. multi-valued or multiple intermediate variables  $D$ , it is crucial to **specify case-specific monotonicity assumptions to reduce no. of strata**
- ▶ One example: both treatment switching and discontinuation
- ▶ A case study (Liu, Wruck, Li, 2024)
  - ▶ Treatment: aspirin low (81mg daily) vs. high dose (325mg daily)
  - ▶ Both treatment switching and discontinuation (patients stopped taking aspirin entirely)
  - ▶ Consider actual treatment levels  $D$  are three: discontinue (\*), low (0), high (1)
  - ▶ Without restrictions, nine ( $3 \times 3$ ) possible strata  $S = (D(0), D(1))$



## Extending Monotonicity: Example

- ▶ Specify more elaborate monotonicity assumptions according to the real context
- ▶ Liu et al. (2024) assumes:
  - i. If  $D_i(0) = *$ , then  $D_i(1) = *$ , namely, if a patient would discontinue under the 81 mg assignment, then he/she would also discontinue under the 325 mg assignment;
  - ii. if  $D_i(0) = 1$ , then  $D_i(1) = 1$ , namely, if a patient would take 325 mg under the 81 mg assignment, then he/she would also take 325 mg under the 325 mg assignment.
- ▶ Assumption ii implies the original monotonicity when there is no discontinuation
- ▶ With this monotonicity, the number of strata reduced from 9 to 5:  
 $S \in \{(*, *), (0, *), (0, 0), (0, 1), (1, 1)\}$

## 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)

## Bypassing mixture models: Principal Score

- ▶ Principal score (Jo and Stuart, 2009):  $e_s(X) = \Pr(S = s \mid X)$
- ▶ Key assumption: **principal ignorability (PI)**  $Y(z) \perp S \mid X$ , which implies the same conditional expectations across mixture components (strata)
- ▶ PI: a strong assumption in parallel with the standard ignorability assumption
- ▶ PI holds only when we have adequate background covariates
- ▶ Assuming unconfoundedness, monotonicity and PI, one can use weighting to estimate PCEs (Ding and, 2017; Jiang et al. 2022), bypassing mixture models
- ▶ PI assumption is the key: assuming strata membership is fully encoded in covariates

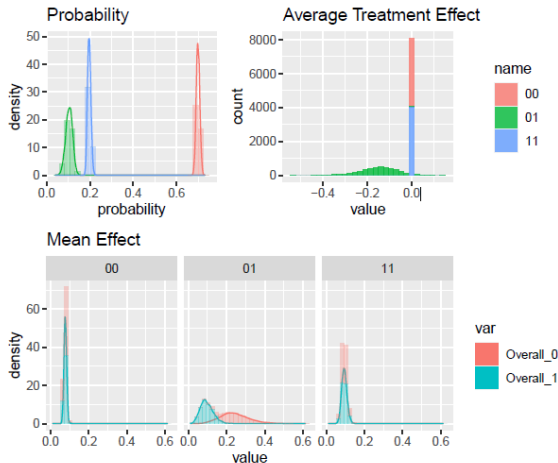
## “PStrata” R Package

- ▶ Estimation and inference of PS is challenging: mixture models are hard; Bayesian inference is hard (for most practitioners)
- ▶ R package “PStrata” (Liu and Li, 2023): 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://cran.r-project.org/web/packages/PStrata/index.html>
- ▶ For illustration, see HW4.

## Characterize and Visualize Principal Strata

- ▶ Principal strata, by definition, are latent. But it is helpful to characterize and visualize
- ▶ With mixture model, we can
  - ▶ calculate each unit's probability of being in each stratum,
  - ▶ based on which randomly draw each unit's principal stratum membership.
  - ▶ Then plot these individual probabilities or/and present a table of baseline characteristics of units in each stratum
- ▶ Good separation in probabilities between strata is desirable, implying stronger identification of the strata

# Characterize and Visualize Principal Strata: Example



The upper left plot shows the probability of each strata, and the upper right plot shows the average treatment effect within each strata. Point mass of 0 are observed in never takers and always takers due to ER. The lower plot shows the estimated mean effect of treatment group and control group, under each stratum.

## Multiple Intermediate Variables

- ▶ Often there are multiple intermediate variables in a study. Exponentially increases the number of principal strata. Additional assumptions are required to identify the causal effects
- ▶ The Faenza Randomized Experiment on Breast Self-Examination (BSE): conducted between Jan 1988 and Dec 1990 at the Oncologic Center of the Faenza Health District in Italy
- ▶ Two BSE teaching methods were compared:
  - ▶ a *standard treatment* of receiving mailed information only, and
  - ▶ a *new treatment* of additional attendance in a self-exam course.
- ▶ The question of interest is the effect of an enhanced training class on BSE practices and quality of self-exam execution.



# Complications

- ▶ The Faenza BSE study suffers from complications due to
  - ▶ *noncompliance* with the randomly assigned treatment: only 55% of the women assigned to the new treatment complied with their assignment;
  - ▶ *missing outcomes* following treatment noncompliance: only 65% of the women responded to the post-test questionnaire;
  - ▶ “*censoring by death*”: quality of self exams is not only unobserved but also undefined on the sample space for women who do not practice BSE.
- ▶ For illustration, here we will only consider noncompliance and censoring by death.

## Estimands of interest

- ▶ *Causal Estimands on BSE practice outcome*
  - ▶ Intention-To-Treat (ITT) effect;
  - ▶ Complier Average Causal Effect (CACE);
  - ▶ Never-taker Average Causal Effect (NACE).
- ▶ *Causal Estimands on BSE quality outcome*
  - ▶ ITT effect for all women who would practice BSE under both assignments;
  - ▶ average causal effect for compliers who would practice BSE who would practice BSE under both treatments.

## Potential Outcomes

If woman  $i$  in the study ( $i = 1, \dots, N$ ) is to be assigned to treatment  $z$  ( $z = 1$  or  $z = 0$ ), we denote the following:

- ▶ *Indicator of the treatment received:*

$$D_i(z) = \begin{cases} P, & \text{if the woman attends the training program;} \\ p, & \text{if the woman receives only mailed information on BSE.} \end{cases}$$

- ▶ *BSE practice indicator:*  $W_i(z) = \begin{cases} B, & \text{if the woman practices BSE;} \\ b, & \text{otherwise.} \end{cases}$

- ▶ *Potential quality outcome:*

$$Y_i(z) = \begin{cases} H, & \text{if the woman practices BSE with "High" quality;} \\ L, & \text{if the woman practices BSE with "Low" quality;} \\ *, & \text{if the woman does not practice BSE.} \end{cases}$$

# Principal Stratification

- ▶ The variable  $D_i(1)$  defines the compliance behavior of subject  $i$ :
  - ▶ If  $D_i(1) = P$ , then woman  $i$  is a “*complier*”;
  - ▶ if  $D_i(1) = p$ , then woman  $i$  is a “*never-taker*”.
- ▶ The vector  $(W_i(0), W_i(1))$  defines the BSE practice behavior of subject  $i$ .

# Principal Stratification

$PBB = \{i : D_i(1) = P, W_i(0) = B, W_i(1) = B\}$  : compliers who would practice BSE under both treatment arms;

$PbB = \{i : D_i(1) = P, W_i(0) = b, W_i(1) = B\}$  : compliers who would not practice BSE under control but would practice BSE under treatment;

$PBb = \{i : D_i(1) = P, W_i(0) = B, W_i(1) = b\}$  : compliers who would practice BSE under control but would not practice BSE under treatment;

$Pbb = \{i : D_i(1) = P, W_i(0) = b, W_i(1) = b\}$  : compliers who would practice BSE under neither treatment arms;

$pBB = \{i : D_i(1) = p, W_i(0) = B, W_i(1) = B\}$  : never-takers who would practice BSE under both treatment arms;

$pBb = \{i : D_i(1) = p, W_i(0) = b, W_i(1) = B\}$  : never-takers who would not practice BSE under control but would practice BSE under treatment;

$pBb = \{i : D_i(1) = p, W_i(0) = B, W_i(1) = b\}$  : never-takers who would practice BSE under control but would not practice BSE under treatment;

## Principal stratification and associated pattern for potential outcomes

| Principal Stratum | $D_i(1)$ | $W_i(0)$ | $W_i(1)$ | $Y_i(0)$       | $Y_i(1)$       |
|-------------------|----------|----------|----------|----------------|----------------|
| $PBB$             | $P$      | $B$      | $B$      | $\in \{L, H\}$ | $\in \{L, H\}$ |
| $PbB$             | $P$      | $b$      | $B$      | *              | $\in \{L, H\}$ |
| $PBb$             | $P$      | $B$      | $b$      | $\in \{L, H\}$ | *              |
| $Pbb$             | $P$      | $b$      | $b$      | *              | *              |
| $pBB$             | $p$      | $B$      | $B$      | $\in \{L, H\}$ | $\in \{L, H\}$ |
| $pbB$             | $p$      | $b$      | $B$      | *              | $\in \{L, H\}$ |
| $pBb$             | $p$      | $B$      | $b$      | $\in \{L, H\}$ | *              |
| $pbb$             | $p$      | $b$      | $b$      | *              | *              |

## Principal Stratification: More Extensions

- ▶ Continuous intermediate variables: the key is to model the intermediate variable in a parsimonious way, e.g. bivariate Gaussian (Jin and Rubin, 2008, JASA), Gaussian copula (Bartolluci and Grill, 2011, JASA), Dirichlet Process Mixture (Schwartz, Li, Mealli, 2011, JASA)
- ▶ Reduce variance using secondary outcomes and covariates (Mealli and Pacini, 2013, JASA)
- ▶ Sequential treatments (Ricardi, Mattei, Mealli, 2020, JASA)
- ▶ Sensitivity analysis (Schwartz et al. 2012, SIM)

# Additional PS Example: Direct and Indirect Causal Effects

Sjöander et al., 2009; Schwartz et al., 2011

- ▶ Treatment = Physical activity (PA)

$$Z_i = \begin{cases} 0, & \text{if subject } i\text{'s PA level is low;} \\ 1, & \text{if subject } i\text{'s PA level is high.} \end{cases}$$

- ▶ Intermediate Outcome: Body Mass index (BMI) after “assignment” of PA

$$D_i(z) = \begin{cases} H, & \text{if unit } i \text{ is obese (BMI is high) given assignment } z; \\ L, & \text{if unit } i \text{ is not obese (BMI is low) given assignment } z. \end{cases}$$

- ▶ Primary Outcome: CardioVascular Disease (CVD) after “assignment” of PA

$$Y_i(z) = \begin{cases} 1, & \text{if unit } i \text{ reports at least one CVD event before end of follow-up given assignment } z. \\ 0, & \text{if unit } i \text{ remains undiagnosed through follow-up given assignment } z. \end{cases}$$

- ▶ *The total effect of physical activity will be a combination of the direct effect of PA on CVD and the indirect effect mediated by BMI*



- ▶ Subjects who would be obese under both PA levels: BMI is unaffected by PA

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

- ▶ Subjects who would be obese under high PA level and would not be obese under low PA level

$$LH = \{i : D_i(0) = L, D_i(1) = H\}$$

- ▶ Subjects who would not be obese under high PA level and would be obese under low PA level

$$HL = \{i : D_i(0) = H, D_i(1) = L\}$$

- ▶ Subjects would not be obese under both PA levels: BMI is unaffected by PA

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

*A direct causal effect of PA, after controlling for BMI, exists if there is a causal effect of PA on CVD for subjects for whom the treatment does not affect BMI (i.e., basic principal strata HH and LL)*

# Hypothetical Example (1)

| Principal Stratum<br>$S_i$ | Post-trt Var.<br>BMI |          | Pot. Outcome<br>CVD rates (%) |          |
|----------------------------|----------------------|----------|-------------------------------|----------|
|                            | $D_i(1)$             | $D_i(0)$ | $Y_i(1)$                      | $Y_i(0)$ |
| Not Obese                  | $L$                  | $L$      | 10                            | 10       |
| Normal                     | $L$                  | $H$      | 30                            | 50       |
| Obese                      | $H$                  | $H$      | 50                            | 50       |

Assume there are no special units:  $LH = \emptyset$

Equal proportions for each prin stratum

Obs Data from a  
Randomized Study  
Average of  $(D_i^{\text{obs}}, Y_i^{\text{obs}})$   
given assignment

$Z_i^{\text{obs}} = 1$        $Z_i^{\text{obs}} = 0$

$(L, 20)$

$(L, 10)$

$(H, 50)$

$(H, 50)$

## Hypothetical Example (2)

| Principal Stratum<br>$S_i$ | Post-trt Var.<br>BMI |          | Pot. Outcome<br>CVD rates (%) |          |
|----------------------------|----------------------|----------|-------------------------------|----------|
|                            | $D_i(1)$             | $D_i(0)$ | $Y_i(1)$                      | $Y_i(0)$ |
| Not Obese                  | $L$                  | $L$      | 10                            | 20       |
| Normal                     | $L$                  | $H$      | 30                            | 40       |
| Obese                      | $H$                  | $H$      | 50                            | 60       |

Assume there are no special units:  $LH = \emptyset$

Equal proportions for each prin stratum

Obs Data from a  
Randomized Study  
Average of  $(D_i^{\text{obs}}, Y_i^{\text{obs}})$   
given assignment  
 $Z_i^{\text{obs}} = 1$        $Z_i^{\text{obs}} = 0$

$(L, 20)$

$(L, 20)$

$(H, 50)$

$(H, 50)$

## Additional Examples of Direct and Indirect Effects

- ▶ Evaluating to what extent the causal effects of a new drug treatment having side-effects is be mediated by the effect of taking additional medication to counter its side-effects (Pearl, 2001)
- ▶ The causal effects of a training program on participants' earnings might be mediated by lock-in effects, that is, the loss of labour market experience (Flores & Flores-Lagunes, 2009)
- ▶ Evaluating the extent to which smoking during pregnancy affects the incidence on low birth weight through a shorter gestation time might inform policy makers on promoting drugs that lengthen gestation time (Flores & Flores-Lagunes, 2009)
- ▶ Evaluating to what extent the effect of military service on veterans' earnings is channelled by subsidized higher education (Angrist & Chen, 2008)
- ▶ A special case of mediation (direct and indirect effects) is **Surrogate Endpoints**

## Additional PS Example: Surrogate Endpoints

- ▶ Often in clinical trials the primary outcome may be rare, late-occurring or costly to obtain
- ▶ Instead researchers may rely on ‘surrogate" or “biomarker" endpoints: easier-to-measure variables known to have a strong association with the true endpoint to reliably extract information on the effect of the treatment
- ▶ Example: in a randomized trial to evaluate the efficacy of a HIV vaccine, the primary outcome is HIV infection, which may take place after a considerable period of time. Count of CD4 cells in blood has long been used as a surrogate endpoint for HIV infection
- ▶ There have been different definitions of surrogate endpoints

# Principal Surrogates

- ▶ Frangakis and Rubin (2002) defines “principal surrogates” on based principal stratification; principal surrogates have a causal interpretation
- ▶ For a variable  $D$  measured after treatment but before the true endpoint occurs, define an **associative effect** to be the causal effects for the principal strata with  $d_0 \neq d_1$ , and a **dissociative effect** to be the causal effects for the principal strata with  $d_0 = d_1 = d$
- ▶  **$D$  is a principal surrogate if the dissociative effect is zero for all  $w$**
- ▶ The quality of  $D$  as a surrogate is determined by its associative effects relative to its dissociative effects - “causal effect predictiveness”

# References

- Ding, P. and Lu, J. (2017). Principal stratification analysis using principal scores. *JRSS-B*, 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. *Annals of statistics*, 25, 305-327.
- Jiang, Z., Yang, S. and Ding, P. (2022). Multiply robust estimation of causal effects under principal ignorability. *JRSS-B*, 84, 1423-1445
- 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.
- Liu B, Li F. (2023). PStrata: A R package for principal stratification analysis. [arXiv:2304.02740](https://arxiv.org/abs/2304.02740)

# References

Liu B, Wruck L, Li F. (2024). Principal stratification for noncompliance with time-to-event outcomes. *Biometrics*. 80(1), 1-14.

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. *JRSS-C*. 66(4), 759-776.

Rosenbaum, P. R. (1984). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *JRSS-A*, 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.