Principal stratification with continuous post-treatment variables: nonparametric identification and semiparametric estimation

Sizhu Lu

Joint work Peng Ding and Zhichao Jiang

Department of Statistics, UC Berkeley

CUHK, Shenzhen seminar

May 29, 2024

Overview

Causal inference with post-treatment variables: motivating examples

Principal stratification for post-treatment variables

Principal stratification: difficulties and strategies

Principal ignorability with continuous post-treatment variables

Semiparametric estimation under principal ignorability

Table of Contents

Motivating example 1: partial noncompliance in RCT

Efron and Feldman (1991 JASA)

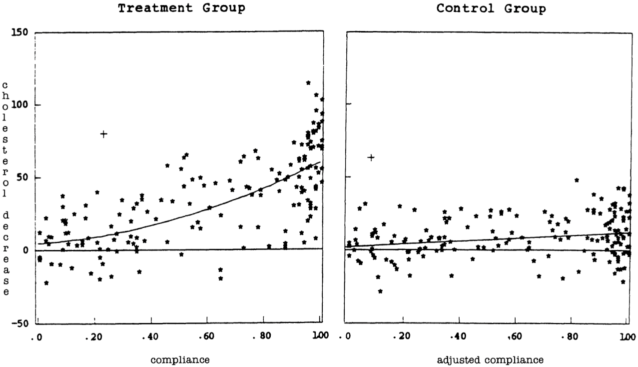

Figure 1. Stanford portion of LRC-CPPT. Left Panel: Treatment group, 164 men (after removal of outlier indicated by +); vertical axis is decrease in total cholesterol; horizontal axis is compliance (the proportion of the nominal choelstyramine dose actually taken). Better compliance leads to larger decreases in total cholesterol, as indicated by the quadratic regression curve. Right Panel: Placebo Control group, 171 men (after removal of outlier +); compliance has been adjusted to match the distribution of compliance in the Treatment group. There is a smaller, but still significant, dose-response relationship between compliance and cholesterol decrease, indicated by the linear regression line.

Motivating example 2: surrogate endpoints

- Gilbert and Hudgens (2008 Biometrics)
  - HIV vaccine trial
  - potential surrogate endpoint: immune response
  - outcome: infection
- Gilbert et al (2015 JCI)
  - herpes zoster vaccine trial
  - potential surrogate endpoint: varicella zoster virus antibody titers
  - outcome: infection

Motivating example 3: heterogeneous effects

- Schwartz et al. (2011 JASA): observational study
  - effect of physical exercise on cardiovascular disease
  - how does the effect vary across levels of BMI
- Alfonsi et al. (2020 Econometrica): RCT in Uganda
  - vocational training on total earnings
  - how does the effect vary across levels of weekly working hours

Table of Contents

Adjusting for post-treatment variables can be tricky

- Notation
  - treatment Z: binary
  - outcome Y
  - post-treatment variable S
- Naive adjustment can be problematic even in RCT

$$
\begin{aligned}
& (Y \mid Z=1, S=s) - (Y \mid Z=0, S=s) \\
= \ & (Y_1 \mid Z=1, S_1=s) - (Y_0 \mid Z=0, S_0=s) \\
= \ & (Y_1 \mid S_1=s) - (Y_0 \mid S_0=s)
\end{aligned}
$$

  - comparing potential outcomes of different units
  - not a causal effect

Principal stratification

proposed by Frangakis and Rubin (2002 Biometrics)

- Conditioning on the observed S is problematic
- Propose to condition on the joint potential values (S_1, S_0):

$$\tau(s_1, s_0) = \mathbb{E}(Y_1 - Y_0 \mid S_1 = s_1, S_0 = s_0)$$

  - (S_1, S_0) acts as a pretreatment covariate, unaffected by treatment
  - $\tau(s_1, s_0)$ has the interpretation of subgroup effect
  - $\tau(s_1, s_0)$ quantifies heterogeneous treatment effect with respect to S

Table of Contents

Principal stratification: conceptually fine, practically hard

- Fundamental problem of causal inference
  - never jointly observe S_1 and S_0
  - $\tau(s_1, s_0)$ is the effect of a latent group
- Never jointly observe Y_1 and Y_0: less problematic

$$\tau(s_1, s_0) = m_1(s_1, s_0) - m_0(s_1, s_0)$$

where

$$m_z(s_1, s_0) = \mathbb{E}(Y_z \mid S_1 = s_1, S_0 = s_0)$$

A famous special case of principal stratification

Instrumental variable (IV) for RCT with noncompliance

- Binary S, treatment received
- Monotonicity $S_1 \geq S_0$: (S_1, S_0) take three values
- Exclusion restriction: $S_1 = S_0 \Rightarrow Y_1 = Y_0$
- Complier average causal effect can be identified

$$\begin{aligned} \tau(1,0) &= \mathbb{E}(Y_1 - Y_0 \mid S_1 = 1, S_0 = 0) \\ &= \frac{\mathbb{E}(Y \mid Z = 1) - \mathbb{E}(Y \mid Z = 0)}{\mathbb{E}(S \mid Z = 1) - \mathbb{E}(S \mid Z = 0)} \end{aligned}$$

Difficulties of principal stratification

- S may not non-binary and even continuous
- Monotonicity $S_1 \geq S_0$ may fail
- Exclusion restriction $S_1 = S_0 \Longrightarrow Y_1 = Y_0$ cannot even be invoked
- We are interested in general $\tau(s_1, s_0)$, not only $\tau(1, 0)$
- Relaxing any of the above assumptions leads to difficulties: $\tau(s_1, s_0)$ is not identifiable without additional assumptions

Some other strategies for principal stratification: Part I

- Model-based approach:

$$(Z, S_1, S_0, Y_1, Y_0 \mid X)$$

  - often with further assumptions on priors of the parameters (Bayesian)
  - identifiability is driven by models
  - JASA: Zhang et al (2009), Jin and Rubin (2008), Schwartz et al (2011)

- Large-sample bounds:
  - $\tau(s_1, s_0)$ partially identified by the observed distribution
  - bounds are often too wide to be informative
  - Zhang and Rubin (2003 JEBS), Lee (2009 REStud)

Some other strategies for principal stratification: Part II

- Auxiliary variables associated with latent (S_1, S_0) but conditionally independent of the outcome
  - secondary outcome, e.g. side effect (Mealli et al 2013 JASA)
  - another vaccine response (Follman 2006 Biometrics)
  - Ding et al (2011 JASA) and Jiang and Ding (2021 StatSci)
  - similar to proximal inference, but has real motivations
- Principal ignorability: $(S_1, S_0) \perp\!\!\!\perp (Y_1, Y_0) \mid X$
  - initially from more applied statistics research
  - theory: Ding and Lu (2017 JRSSB) and Jiang et al (2022 JRSSB)
  - strong and untestable assumption
  - why do we study it? simplicity in implementation, numerically robust

Table of Contents

Assumption – treatment ignorability

$$Z \perp\!\!\!\perp (S_1, S_0, Y_1, Y_0) \mid X$$

- Standard in observational studies, conditional on covariates X
- Ensures identification of $(S_z \mid X)$ and $(Y_z \mid X)$
- Assumes known copula to go from $(S_1 \mid X)$ and $(S_0 \mid X)$ to

$$(s_1, s_0 \mid X) = \mathbb{C}_\rho({}_1(s_1 \mid X),{}_0(s_0 \mid X))$$

  - ensures identifiability of *principal density* $e(s_1, s_0, X)$
  - another strong assumption
  - can vary the copula parameter ρ in sensitivity analysis

Assumption – principal ignorability

$$Y_1 \perp\!\!\!\perp S_0 \mid S_1, X, \quad Y_0 \perp\!\!\!\perp S_1 \mid S_0, X$$

- Slightly stronger than needed in the theory; more elegant
- Treatment ignorability + principal ignorability:

$$\begin{aligned}
\mathbb{E}(Y_1 \mid Z = 1, S_1 = s_1, S_0 = s_0, X) &= \mathbb{E}(Y_1 \mid Z = 1, S_1 = s_1, X) \\
&= \mu_1(s_1, X) \\
\mathbb{E}(Y_0 \mid Z = 0, S_1 = s_1, S_0 = s_0, X) &= \mathbb{E}(Y_0 \mid Z = 0, S_0 = s_, X) \\
&= \mu_0(s_0, X)
\end{aligned}$$

Nonparametric identifiability

- Based on principal density and outcome model

$$\mathbb{E}(Y_1 \mid U = s_1 s_0) = \mathbb{E}\left\{ \frac{e(s_1, s_0, X)}{e(s_1, s_0)} \mu_1(s_1, X) \right\}$$

  - notation $U = s_1 s_0$ is the unmeasured latent group

- Based on treatment probability and principal density

$$\mathbb{E}(Y_1 \mid U = s_1 s_0) = \lim_{\epsilon \to 0} \mathbb{E}\left\{ \frac{e(s_1, s_0, X)}{e(s_1, s_0)} \frac{1(s_1 - \epsilon \leq S_1 \leq s_1 + \epsilon)}{2\epsilon \cdot p_1(s_1, X)} \frac{ZY}{\pi(X)}\right.$$

  - notation $\pi(X) = \mathrm{pr}(Z = 1 \mid X)$ is the treatment probability

- Based on treatment probability and outcome model
  - generally difficult: see more details in the paper
  - possible in some special cases, e.g. S is binary

Difficulties of nonparametric estimation

- We can construct nonparametric estimators for $\mathbb{E}(Y_1 \mid U = s_1 s_0)$
- $\mathbb{E}(Y_1 \mid U = s_1 s_0)$ is a local parameter
- Poor finite-sample performance
- Difficult to interpret

Our focus: semiparametric estimation

- Estimation finite-dimensional parameter η_z that minimizes

$$\eta_z = \arg\min_{\eta} \mathbb{E}\left[w_z(S_1, S_0)\{m_z(S_1, S_0) - f_z(S_1, S_0; \eta)\}^2\right]$$

  - notation $m_z(s_1, s_0) = \mathbb{E}(Y_z \mid U = s_1 s_0)$
  - working model $f_z(S_1, S_0; \eta)$ to approximate $m_z(S_1, S_0)$
  - $w_z(S_1, S_0)$ user-specified weight

- An idea appeared in the literature, not so popular (Neugebauer and van der Laan 2007; Kennedy et al. 2019; Ye et al. 2023).

- We will focus on estimating η_z

Table of Contents

First-order condition for η_z

- More explicit formula

$$\eta_z = \arg\min_{\eta} \iint w_z(s_1, s_0)\{m_z(s_1, s_0) - f_z(s_1, s_0; \eta)\}^2 e(s_1, s_0) s_1 s_0$$

- First-order condition for η_z

$$\iint w_z(s_1, s_0)\{m_z(s_1, s_0) - f_z(s_1, s_0; \eta)\}\dot{f}_z(s_1, s_0; \eta) e(s_1, s_0) s_1 s_0 = 0$$

  - some implicit assumptions on the uniqueness of the solution
  - non-degenerate Hessian
  - not directly useful because of unknown $m_z(s_1, s_0)$

Estimation I: principal density and outcome model

- Define

$$D_{z,\mathrm{pd+om}}(Y, S, Z, X; \eta, e, \mu_z)$$
$$= \iint w_z(s_1, s_0)\{\mu_z(s_z, X) - f_z(s_1, s_0; \eta)\}\dot{f}_z(s_1, s_0; \eta)e(s_1, s_0, X)s_1 s_0$$

- Estimating equation

$$\mathbb{E}\{D_{z,\mathrm{pd+om}}(Y, S, Z, X; \eta_z, e, \mu_z)\} = 0$$

- Two step estimation
  - estimate nuisance parameters e, μ_z
  - solve for η_z from the empirical analogue of the estimating equation

Estimation II: treatment probability and principal density

- Define

$$D_{1,\text{tp+pd}}(Y, S, Z, X; \eta, e, \pi)$$
$$= \int w_1(S, s_0)\frac{e(S, s_0, X)}{p_1(S, X)}\frac{Z}{\pi(X)}\{Y - f_1(S, s_0; \eta)\}\dot{f}_1(S, s_0; \eta)s_0$$

- Estimating equation

$$\mathbb{E}\{D_{1,\text{tp+pd}}(Y, S, Z, X; \eta_1, e, \pi)\} = 0$$

- Two step estimation
  - estimate nuisance parameters e, π
  - solve for η_1 from the empirical analogue of the estimating equation

- Analogous results for η_0

Estimation III: Doubly robust estimation

- Estimating equation

$$\mathbb{E}\{\ell_1 + \ell_2 + \ell_3\} = 0$$

- $\ell_1 = D_{1,\mathrm{pd+om}}(Y, S, Z, X; \eta_1, e, \mu_1)$
- ℓ_2: correction term with details on the next page
- ℓ_3 is similar to $D_{1,\mathrm{tp+pd}}(Y, S, Z, X; \eta_1, e, \pi)$:

$$\ell_3 = \int w_1(S, s_0) \frac{e(S, s_0, X)}{p_1(S, X)} \frac{Z}{\pi(X)} \{Y - \mu_1(S, X)\} \dot{f}_1(S, s_0; \eta) s_0$$

  - $Y - \mu_1(S, X)$, not $Y - f_1(S, s_0; \eta)$

Estimation III: Doubly robust estimation, ℓ_2

- Define

$$
\begin{aligned}
\nu_1(s_1, s_0, X) &= w_1(s_1, s_0)\{\mu_1(s_1, X) - f_1(s_1, s_0; \eta)\}\dot{f}_1(s_1, s_0; \eta)e(s_1, s_0, \\
r_u(s_1, s_0, S, X) &= 1 - \frac{\dot{c}_u(s_1, s_0, X)}{c(s_1, s_0, X)}\{1(S \leq s_1) - F_1(s_1, X)\} \\
r_v(s_1, s_0, S, X) &= 1 - \frac{\dot{c}_v(s_1, s_0, X)}{c(s_1, s_0, X)}\{1(S \leq s_0) - F_0(s_0, X)\}
\end{aligned}
$$

- ℓ_2 equals

$$
\begin{aligned}
&\frac{Z}{\pi(X)}\left\{\frac{\int \nu_1(S, s_0, X)s_0}{p_1(S, X)} - \iint \nu_1(s_1, s_0, X)r_u(s_1, s_0, S, X)s_1 s_0\right\} \\
&+\frac{1 - Z}{1 - \pi(X)}\left\{\frac{\int \nu_1(s_1, S, X)s_1}{p_0(S, X)} - \iint \nu_1(s_1, s_0, X)r_v(s_1, s_0, S, X)s_1 s_0\right\}
\end{aligned}
$$

Estimation III: Doubly robust estimation

- Where does it come from?
  - efficient influence function (EIF)
  - semiparametric theory (Bickel et al 1993)
- Theoretical properties
  - consistent if either treatment probability or the outcome model is correct, given that the principal density model is correct
  - semiparametrically efficient if three models are correct
- Complicated in general; explicit formulas under linear working model

$$f_z(s_1, s_0; \eta_z) = \eta_z' g(s_1, s_0)$$

Application (Alfonsi et al., 2020 Econometrica)

- RCT among disadvantaged youth entering the labor market in Uganda
- 6-month training program on labor market outcomes
- Y: workers' total earnings 48 months later
- S: total number of hours worked in a specific week 36 months later
- X: pretreatment covariates
- $\tau(s_1, s_0)$: how does the effect of the training program on total earnings vary across hours worked?

A simple working model

- Linear working model

$$f_z(s_1, s_0; \eta_z) = \beta_z(s_1 - s_0) + \alpha_z$$

- Average causal effect model

$$\tau(s_1, s_0; \eta) = \beta_\tau(s_1 - s_0) + \alpha_\tau$$

where $\beta_\tau = \beta_1 - \beta_0$ and $\alpha_\tau = \alpha_1 - \alpha_0$
  - α_τ: effect of Z on Y even if Z does not affect S
  - β_τ: how does effect of Z on Y related to the effect of Z on S?

Estimates under the linear working model

- ρ: sensitivity parameter in the copula
- tp+pd seems an outlier: unstable weighting estimators

		eif	tp+pd	pd+om
$\rho = 0$	$\hat{\eta}_\tau$	0.048	0.050	0.049
		(0.008)	(0.008)	(0.008)
	$\hat{\alpha}_\tau$	0.524	0.206	0.490
		(0.500)	(0.590)	(0.515)
$\rho = 0.5$	$\hat{\eta}_\tau$	0.042	0.173	0.050
		(0.022)	(0.121)	(0.013)
	$\hat{\alpha}_\tau$	0.652	0.255	0.684
		(0.658)	(1.794)	(0.642)

Discussion

- Sensitivity analysis with respect to principal ignorability
- Multiple post-treatment variables