arXiv:1602.01196v1 [stat.ME] 3 Feb 2016

# Principal stratification analysis using principal scores

Peng Ding* and Jiannan Lu†

**Abstract**

Practitioners are interested in not only the average causal effect of the treatment on the outcome but also the underlying causal mechanism in the presence of an intermediate variable between the treatment and outcome. However, in many cases we cannot randomize the intermediate variable, resulting in sample selection problems even in randomized experiments. Therefore, we view randomized experiments with intermediate variables as semi-observational studies. In parallel with the analysis of observational studies, we provide a theoretical foundation for conducting objective causal inference with an intermediate variable under the principal stratification framework, with principal strata defined as the joint potential values of the intermediate variable. Our strategy constructs weighted samples based on principal scores, defined as the conditional probabilities of the latent principal strata given covariates, without access to any outcome data. This principal stratification analysis yields robust causal inference without relying on any model assumptions on the outcome distributions. We also propose approaches to conducting sensitivity analysis for violations of the ignorability and monotonicity assumptions, the very crucial but untestable identification assumptions in our theory. When the assumptions required by the classical instrumental variable analysis cannot be justified by background knowledge or cannot be made because of scientific questions of interest, our strategy serves as a useful alternative tool to deal with intermediate variables. We illustrate our methodologies by using two real data examples, and find scientifically meaningful conclusions.

**Keywords:** Causal inference; Exclusion restriction; Intermediate variable; Monotonicity; Nonparametric identification; Principal ignorability; Sensitivity analysis.

*University of California, Berkeley, California, U.S.A. Address for correspondence: Peng Ding, 425 Evans Hall, Berkeley, California 94720, USA. Email: pengdingpku@berkeley.edu

†Microsoft Corporation, Redmond, Washington, U.S.A.

# 1. Causal Inference with Intermediate Variables

When an intermediate variable between the treatment and outcome exists, often researchers are interested in not only the average causal effect of the treatment on the outcome but also the underlying causal mechanism in the presence of the intermediate variable. Naive analysis by conditioning on the observed value of the intermediate variable does not yield valid causal interpretations without imposing strong assumptions. Principal stratification (Frangakis and Rubin 2002), defined as the joint potential values of the intermediate variable under both treatment and control, can be viewed as a pretreatment covariate unaffected by the treatment. Therefore, conditioning on principal stratification yields subgroup causal effects.

The subgroup causal effects classified by principal stratification have clear scientific meanings in various settings. For instance, when the intermediate variable is the actual treatment received, the principal stratification variable indicates the compliance status, and the classical instrumental variable estimator identifies the average causal effect for compliers (Angrist et al. 1996). When the intermediate variable is the indicator for survival status, the only sensible subgroup causal effect on the outcome is the one for survivors who would potentially survive under both treatment and control (Rubin 2006). When the intermediate variable is a surrogate for the outcome, we want to predict the causal effect on the outcome by the causal effect on the surrogate. An ideal surrogate must satisfy the causal necessity that zero effect on the surrogate implies zero effect on the outcome (Frangakis and Rubin 2002) and the causal sufficiency that positive effect on the surrogate implies positive effect on the outcome (Gilbert and Hudgens 2008). Therefore, we can assess these requirements for an ideal surrogate by conducting a principal stratification analysis.

Principal stratification clarifies causal inference with intermediate variables, but it also results in inferential difficulties because of the missingness of the principal stratification variable and the consequential mixture distributions of the observed data. We can sharpen inference about causal effects within principal strata only if we impose some of the following structural or modeling assumptions: (1) *monotonicity* that the treatment has a nonnegative effect on the intermediate variable for each unit (e.g., Angrist et al. 1996; Gilbert and Hudgens 2008); (2) *exclusion restriction* that zero effect on the intermediate variable implies zero effect on the outcome (e.g., Angrist et al. 1996); (3) *Normal outcome distributions* within principal strata (e.g., Zhang et al. 2009; Frumento

et al. 2012); (4) *additional covariates or secondary outcomes* (Ding et al. 2011; Mattei and Mealli 2011; Mattei et al. 2013; Mealli and Pacini 2013; Yang and Small 2016; Jiang et al. 2016). For instance, the classical instrumental variable analysis requires exclusion restriction (Angrist et al. 1996), which may not be justified by background knowledge or cannot be assumed according the scientific questions of interest. Without exclusion restriction, Zhang et al. (2009) and Frumento et al. (2012) assumed Normal outcome models within principal strata, and thus identifiability of the causal effects within principal strata is ensured by identifiability of the Normal Mixture Model. Unfortunately, the results are sensitive to the parametric modeling assumption, and the unbounded likelihood function jeopardizes statistical inference even with correctly specified model (Ding et al. 2011; Frumento et al. 2016). Without these assumptions, however, large sample bounds of causal effects are often too wide to be informative (Zhang and Rubin 2003; Cheng and Small 2006). We will review more applications and further highlight the inferential difficulty of principal stratification without exclusion restriction in Section 2.

Recognizing the salient feature that the intermediate is not randomized even though the treatment is randomized, we take an alternative perspective, viewing the problem as a semi-observational study. For objective causal inference, Rubin (2007, 2008) advocated designs of observational studies without access to the outcome data, which prevents data snooping by selecting favorable outcome models. In parallel with this classical wisdom of propensity scores in observational studies (Rosenbaum and Rubin 1983b), we propose to conduct principal stratification analysis based on principal scores, defined as the conditional probabilities of the latent principal strata given a rich set of covariates that ensure certain ignorability assumptions. Previously, applied researchers (Follman 2000; Hill et al. 2002; Jo and Stuart 2009; Jo et al. 2011; Stuart and Jo 2015) used principal scores to analyze data subject to one-sided noncompliance, and theoretical researchers (Joffe et al. 2007) suggested using principal scores to identify general causal effects within principal strata. We advance the literature by providing the theoretical foundation for using principal scores in the analysis of randomized experiments with intermediate variables. To be more specific, we give the assumptions for identification, extend previous literature to deal with general principal stratification problems beyond one-sided noncompliance, and propose statistically efficient and numerically stable weighting estimators for causal effects. The theoretical results allow for a two-step inferential procedure: we first construct weighted samples without access to the outcome data, and we then obtain simple

weighting estimators for causal effects within principal strata. The whole inferential procedure does not involve any model assumptions of the outcomes, leading to more objective causal inference.

Furthermore, the central role of principal scores relies on certain ignorability and monotonicity assumptions. In parallel with sensitivity analysis in observational studies (Rosenbaum and Rubin 1983a; Rosenbaum 2002), we propose approaches to conducting sensitivity analysis for violations of the ignorability assumptions. Previous literature either dealt with binary outcomes (e.g., Sjölander et al. 2009; Schwartz et al. 2012) or relied on modeling assumptions on the outcomes (e.g., Gilbert et al. 2003), but our strategy deals with general outcomes and relies on less modeling assumptions. Other than a few exceptions (Zhang et al. 2009; Ding et al. 2011; Frumento et al. 2012), most principal stratification analyses assumed monotonicity which might be too restrictive for some applications. Our sensitivity analysis technique further removes the monotonicity assumption, and assesses the impact of its violations. The ignorability and monotonicity assumptions, though crucial for identifying the causal effects of interest, cannot be validated by observed data. Therefore, we argue that principal stratification analyses should always come with sensitivity analysis for violations of these assumptions.

The paper proceeds as follows. Section 2 reviews the basic framework of principal stratification. Section 3 defines principal scores and provides sufficient conditions for identifying causal effects within principal strata. Section 4 highlights the balancing properties of principal scores. Section 5 discusses estimation strategies that are efficient, stable and easy to implement. Section 6 proposes approaches to conducting sensitivity analysis for the ignorability and monotonicity assumptions. We conduct simulation studies in Section 7, apply our methodologies to real data examples in Section 8, and conclude in Section 9. We provide proofs and technical details in the on-line supplementary material.

## 2. Potential Outcomes and Principal Stratification

Consider a randomized controlled experiment with $N$ units. We collect pretreatment covariates $\boldsymbol{X}_i$ for each unit $i$ before the experiment. Let $Z_i$ be the treatment assignment for unit $i$, with $Z_i = 1$ for treatment and $Z_i = 0$ for control. We also collect the outcome of interest $Y_i$ for unit $i$, which can be general (continuous, binary, time-to-event, etc.). In practice, we may also collect

some intermediate variables between the treatment and outcome that are helpful to explain the underlying causal mechanism and treatment effect heterogeneity. We will first focus on the case with a binary intermediate variable $S$, because the binary case has the widest applications as illustrated by examples in a later part of this section. We will also comment on general $S$ later.

We use the potential outcomes framework to define causal effects. Under the Stable Unit Treatment Value Assumption (SUTVA; Rubin 1980), there is only one version of the treatment, and there is no interference between units. The SUTVA allows us to define the potential values of the intermediate variable and outcome for unit $i$ as $S_i(t)$ and $Y_i(t)$ under treatment $t$ for $t = 0$ and 1. Completely randomized experiments satisfy the following treatment assignment mechanism, which we will make use of throughout the paper.

**Assumption 1** (Randomization)**.** $Z \perp\!\!\!\perp \{S(1), S(0), Y(1), Y(0), \boldsymbol{X}\}$.

Frangakis and Rubin (2002) introduced the notion of principal stratification, defined as the joint potential values of the intermediate variable $U_i = \{S_i(1), S_i(0)\} \in \{0, 1\}^2$. For simplicity, we relabel the possible values of $U$, $\{(1, 1), (1, 0), (0, 1), (0, 0)\}$, as $\{ss, s\bar{s}, \bar{s}s, \bar{s}\bar{s}\}$, respectively. Because the principal stratification variable is unaffected by the treatment, inference conditional on $U$ yields a subgroup causal interpretation, captured by the following principal causal effect (PCE):

$$ACE_u = E\{Y(1) - Y(0) \mid U = u\} \quad (u = ss, s\bar{s}, \bar{s}s, \bar{s}\bar{s}).$$

The notion of PCE is not only mathematically sound for causal evaluations, but also of scientific interest in practice. Below, we review some important empirical applications, and discuss the scientific meanings of PCEs in each case.

**Example 1** (Noncompliance)**.** Let $S$ denote the actual treatment received, and noncompliance occurs if the treatment assignment differs from the treatment received. Angrist et al. (1996) called $ss, s\bar{s}, \bar{s}s$, and $\bar{s}\bar{s}$ always-taker, complier, defier, and never-taker, respectively.

**Example 2** (Truncation by death)**.** When some units die before measurements of their outcomes $Y$, the truncation by death problem occurs. Let $S$ be the survival status, with $S = 1$ for survival and $S = 0$ for dead. For dead patients with $S = 0$, the corresponding outcome $Y$ is not well-defined. Rubin (2006) argued that the only scientifically meaningful subgroup causal effect is $ACE_{ss}$, the

survivor average causal effect, defined as the average causal effect among units who will potentially survive under both treatment and control. Other subgroup causal effects are not well defined due to their unmeasured outcome under either treatment or control or both.

There are at least two problems similar to truncation by death. In labor economics where $S$ is the employment status and $Y$ is the income, the only sensible causal effect is $ACE_{ss}$, the average causal effect among the always employed units (Zhang and Rubin 2003, Zhang et al. 2009). In vaccine trials where $S$ is the infection status and $Y$ is a post-infection outcome, we are interested in the causal effect of vaccine on the post-infection outcome among units who would develop infection under both treatment and control (Gilbert et al. 2003; Hudgens and Halloran 2006).

**Example 3** (Surrogate)**.** Surrogate is of great importance in clinical trials, when the measurement of the primary outcome is costly or time-consuming. Let $S$ denote the surrogate for the outcome $Y$. The goal of using the surrogate is to predict the causal effect on the outcome by the causal effect on the surrogate. Frangakis and Rubin (2002) argued that a good surrogate should satisfy the "causal necessity," i.e., whenever the treatment has no effect on the surrogate, it has no effect on the outcome ($ACE_{ss} = 0$ and $ACE_{\bar{s}\bar{s}} = 0$). As a complement, Gilbert and Hudgens (2008) further argued that a good surrogate also should satisfy the "causal sufficiency," i.e., whenever the treatment affects the surrogate, it also affects the outcome ($ACE_{s\bar{s}} \neq 0$ and $ACE_{\bar{s}s} \neq 0$).

In practice, a particular data set may simultaneously have more than one of the problems discussed in Examples 1–3 (Mattei and Mealli 2007, Frumento et al. 2012). In all the examples above, estimation of $ACE_u$ is crucial for the substantive questions of interest. However, the inherent missingness of $U$, due to the ability of measuring only one of $S(1)$ and $S(0)$, jeopardizes the identification of the PCEs without some additional assumptions. In the following, we review some commonly-used assumptions, and discuss their plausibility and limitations.

**Assumption 2** (Strong Monotonicity)**.** $S_i(0) = 0$ for all $i$.

In Example 1 of noncompliance, when the control units have no access to receive the active treatment, Strong Monotonicity holds by the design of experiments. It is sometimes referred to as one-sided noncompliance (Imbens and Rubin 2015), which allows us to rule out the always takers ($U = ss$) and defiers ($U = \bar{s}s$). In the literature on surrogates, Strong Monotonicity is closely related to the "constant biomarker" assumption (Gilbert and Hudgens 2008).

**Assumption 3** (Monotonicity). $S_i(1) \geq S_i(0)$ for all $i$.

Monotonicity rules out only the defiers ($U = \bar{s}s$). In general, we cannot test Monotonicity by the observed data unless $\Pr(S = 1 \mid Z = 1) < \Pr(S = 1 \mid Z = 0)$.

**Assumption 4** (Exclusion Restriction). $Y_i(1) = Y_i(0)$ for $U_i = ss$ and $\bar{s}\bar{s}$.

Exclusion Restriction (ER) implies that $ACE_{ss} = ACE_{\bar{s}\bar{s}} = 0$. In Example 1 of noncompliance, ER is plausible in double-blinded trials because the outcome may be affected only by the treatment received. Angrist et al. (1996) showed that under Monotonicity and ER, the complier average causal effect, $ACE_{s\bar{s}}$, is identified by the ratio of the average causal effects on $Y$ and $S$.

However, in many open-label trials, the treatment assignment may have a "direct effect" on the outcome, and ER may not hold (e.g., Hirano et al. 2000). What is more important, we cannot assume ER in the truncation by death and surrogate problems, because in these settings it is the question of concern to test whether $ACE_{ss}$ or $ACE_{\bar{s}\bar{s}}$ is zero. Imposing ER immediately discards the very scientific question of interest, which is not reasonable. Unfortunately, if we do not impose ER due to either background knowledge or substantive questions of interest, we can no longer nonparametrically identify the PCEs without further assumptions. In this paper, we will discuss alternative sufficient conditions that ensure nonparametric identifiability of the PCEs, and propose estimators that rely on minimal modeling assumptions.

## 3. Nonparametric Identification of Principal Causal Effects

Our identification strategy, in parallel with the notion of propensity score in observational studies, exploits principal scores defined as the probabilities of the latent principal strata given a rich set of pretreatment covariates. Although in the existing literature the principal scores are used in the one-sided noncompliance problem, its rigorous theoretical foundation is lacking, and more importantly it cannot deal with more general cases. We fill in the gap by demonstrating some general identification results based on principal scores.

### 3.1. Principal Scores

Although we cannot uniquely recover the unobserved principal strata indicators, we can create weighted samples based on principal scores:

$$e_u(\boldsymbol{X}) = \Pr(U = u \mid \boldsymbol{X}) \quad (u = ss, s\bar{s}, \bar{s}s, \bar{s}\bar{s}).$$

Define $p_1 = \Pr(S = 1 \mid Z = 1)$ and $p_0 = \Pr(S = 1 \mid Z = 0)$ as the probabilities of $S$ under treatment and control, and the analogous conditional probabilities as $p_1(\boldsymbol{X}) = \Pr(S = 1 \mid Z = 1, \boldsymbol{X})$ and $p_0(\boldsymbol{X}) = \Pr(S = 1 \mid Z = 0, \boldsymbol{X})$ given covariates $\boldsymbol{X}$.

Under Strong Monotonicity, two strata $s\bar{s}$ and $\bar{s}\bar{s}$ exist. The observed data with $(Z = 1, S = 1)$ contain only strata $U = s\bar{s}$, and the observed data with $(Z = 1, S = 0)$ contain only strata $U = \bar{s}\bar{s}$. Therefore, we can use the treatment arm to identify the principal scores by $e_{s\bar{s}}(\boldsymbol{X}) = p_1(\boldsymbol{X})$ and $e_{\bar{s}\bar{s}}(\boldsymbol{X}) = 1 - p_1(\boldsymbol{X})$, and the proportions of the two principal strata by $\pi_{s\bar{s}} = p_1$ and $\pi_{\bar{s}\bar{s}} = 1 - p_1$.

Under Monotonicity, three strata $ss$, $s\bar{s}$, and $\bar{s}\bar{s}$ exist. The observed data with $(Z = 1, S = 0)$ contain only strata $U = \bar{s}\bar{s}$, and the observed data with $(Z = 0, S = 1)$ contain only strata $U = ss$. Therefore, we can identify the principal scores by $e_{ss}(\boldsymbol{X}) = p_0(\boldsymbol{X})$, $e_{\bar{s}\bar{s}}(\boldsymbol{X}) = 1 - p_1(\boldsymbol{X})$, and $e_{s\bar{s}}(\boldsymbol{X}) = p_1(\boldsymbol{X}) - p_0(\boldsymbol{X})$, and the proportions of the three principal strata by $\pi_{ss} = p_0, \pi_{\bar{s}\bar{s}} = 1 - p_1$, and $\pi_{s\bar{s}} = p_1 - p_0$.

The above discussion demonstrates nonparametric identification of principal scores under (Strong) Monotonicity. We postpone the discussion of modeling principal scores to Section 4.2.

### 3.2. Principal Ignorability and Nonparametric Identification

The observed data are mixtures of at most two latent principal strata. Our goal is to disentangle the latent components of the outcome distributions. Although we can view them as the weights for the latent subgroup indicators, principal scores themselves alone are not sufficient to identify the PCEs. The following principal ignorability assumptions (Jo and Stuart 2009; Jo et al. 2011; Stuart and Jo 2015), in parallel with the ignorability assumption in observational studies (Rosenbaum and Rubin 1983b), are sufficient conditions to nonparametrically identify the PCEs.

**Under Strong Monotonicity** We invoke the following version of Principal Ignorability (PI).

**Assumption 5** (PI). $Y(0) \perp\!\!\!\perp U \mid \boldsymbol{X}$.

A weaker version of the above assumption, shown in Section 6.1, also suffices for our later discussion on identification. Here we use a stronger version for easy interpretation. PI assumes conditional independence of $Y(0)$ and $U$ given $\boldsymbol{X}$, i.e., a random allocation of the principal stratification variable with respect to the control potential outcome given $\boldsymbol{X}$. PI requires an adequate set of covariates $\boldsymbol{X}$, conditional on which there is no difference between the distributions of the control potential outcomes across principal strata $U = s\bar{s}$ and $U = \bar{s}\bar{s}$.

With the identifiability of the principal scores, PI further helps identify $ACE_u$.

**Proposition 1.** Under Strong Monotonicity and PI, we can identify the PCEs by

$$\begin{aligned} ACE_{s\bar{s}} &= E(Y \mid Z = 1, S = 1) - E\{w_{s\bar{s}}(\boldsymbol{X})Y \mid Z = 0\}, \\ ACE_{\bar{s}\bar{s}} &= E(Y \mid Z = 1, S = 0) - E\{w_{\bar{s}\bar{s}}(\boldsymbol{X})Y \mid Z = 0\}, \end{aligned}$$

where $w_{s\bar{s}}(\boldsymbol{X}) = e_{s\bar{s}}(\boldsymbol{X})/\pi_{s\bar{s}}$ and $w_{\bar{s}\bar{s}}(\boldsymbol{X}) = e_{\bar{s}\bar{s}}(\boldsymbol{X})/\pi_{\bar{s}\bar{s}}$.

The treatment group does not involve mixture distributions. The control group is a mixture of two strata $s\bar{s}$ and $\bar{s}\bar{s}$, and Proposition 1 shows that the weight $w_u(\boldsymbol{X})$ is the ratio of the principal score over the marginal proportion of stratum $u$.

**Under Monotonicity** We invoke the following General Principal Ignorability (GPI).

**Assumption 6** (GPI). $Y(z) \perp\!\!\!\perp U \mid \boldsymbol{X}$ for $z = 0$ and 1.

Again, a weaker version of GPI suffices to identify PCEs as discussed in Section 6.1, but the stronger version enjoys easier interpretation. The mathematical form of the above assumption is similar to the ignorability assumption in observational studies (Rosenbaum and Rubin 1983b), with $U$ being the latent principal stratification variable instead of the treatment indicator. Intuitively, the conditional independence of GPI requires enough covariates $\boldsymbol{X}$ remove all "confounding" between $U$ and $Y$. More precisely, conditional on $\boldsymbol{X}$, there is no difference between the distributions of the treatment potential outcomes across strata $U = ss$ and $U = s\bar{s}$, and no difference between the distributions of the control potential outcomes across strata $U = \bar{s}\bar{s}$ and $U = s\bar{s}$. These

interpretations will become more apparent in Section 6.1. See Guo et al. (2014) for a slightly different view on GPI.

**Proposition 2.** Under Monotonicity and GPI, we can identify the PCEs by

$$\begin{aligned} ACE_{s\bar{s}} &= E\{w_{1,s\bar{s}}(\boldsymbol{X})Y \mid Z=1, S=1\} - E\{w_{0,s\bar{s}}(\boldsymbol{X})Y \mid Z=0, S=0\}, \\ ACE_{\bar{s}\bar{s}} &= E(Y \mid Z=1, S=0) - E\{w_{0,\bar{s}\bar{s}}(\boldsymbol{X})Y \mid Z=0, S=0\}, \\ ACE_{ss} &= E\{w_{1,ss}(\boldsymbol{X})Y \mid Z=1, S=1\} - E(Y \mid Z=0, S=1), \end{aligned}$$

where

$$\begin{aligned} w_{1,s\bar{s}}(\boldsymbol{X}) &= \frac{e_{s\bar{s}}(\boldsymbol{X})}{e_{s\bar{s}}(\boldsymbol{X}) + e_{ss}(\boldsymbol{X})} \Big/ \frac{\pi_{s\bar{s}}}{\pi_{s\bar{s}} + \pi_{ss}}, \qquad w_{0,s\bar{s}}(\boldsymbol{X}) = \frac{e_{s\bar{s}}(\boldsymbol{X})}{e_{s\bar{s}}(\boldsymbol{X}) + e_{\bar{s}\bar{s}}(\boldsymbol{X})} \Big/ \frac{\pi_{s\bar{s}}}{\pi_{s\bar{s}} + \pi_{\bar{s}\bar{s}}}, \\ w_{0,\bar{s}\bar{s}}(\boldsymbol{X}) &= \frac{e_{\bar{s}\bar{s}}(\boldsymbol{X})}{e_{s\bar{s}}(\boldsymbol{X}) + e_{\bar{s}\bar{s}}(\boldsymbol{X})} \Big/ \frac{\pi_{\bar{s}\bar{s}}}{\pi_{s\bar{s}} + \pi_{\bar{s}\bar{s}}}, \qquad w_{1,ss}(\boldsymbol{X}) = \frac{e_{ss}(\boldsymbol{X})}{e_{s\bar{s}}(\boldsymbol{X}) + e_{ss}(\boldsymbol{X})} \Big/ \frac{\pi_{ss}}{\pi_{s\bar{s}} + \pi_{ss}}. \end{aligned}$$

The observed data with $(Z=1, S=0)$ and $(Z=0, S=1)$ do not involve mixture distributions. The observed data with $(Z=1, S=1)$ contain a mixture of two strata $s\bar{s}$ and $ss$, and the weight $w_{1,u}(\boldsymbol{X})$ is the probability of stratum $u$ conditional on $(Z=1, S=1, \boldsymbol{X})$ divided by the probability conditional only on $(Z=1, S=1)$. Similar discussion applies to the observed data with $(Z=0, S=0)$ and the weight $w_{0,u}(\boldsymbol{X})$.

# 4. Balancing Properties of Principal Scores

## 4.1. Balancing Properties

Principal scores play a crucial role in the theory developed in the last section. Therefore, it is of practical importance to select a principal score model that is close to the truth. Fortunately, we can use the following balancing conditions for any function of the covariates, $h(\boldsymbol{X})$, to guide our choice of the model for $\Pr(U \mid \boldsymbol{X})$.

**Corollary 1.** Under Strong Monotonicity, we have

$$\begin{aligned} E\{h(\boldsymbol{X}) \mid Z=1, S=1\} &= E\{w_{s\bar{s}}(\boldsymbol{X})h(\boldsymbol{X}) \mid Z=0\}, \\ E\{h(\boldsymbol{X}) \mid Z=1, S=0\} &= E\{w_{\bar{s}\bar{s}}(\boldsymbol{X})h(\boldsymbol{X}) \mid Z=0\}. \end{aligned}$$

**Corollary 2.** Under Monotonicity, we have

$$\begin{aligned} E\{w_{1,s\bar{s}}(\boldsymbol{X})h(\boldsymbol{X}) \mid Z=1, S=1\} &= E\{w_{0,s\bar{s}}(\boldsymbol{X})h(\boldsymbol{X}) \mid Z=0, S=0\}, \\ E\{h(\boldsymbol{X}) \mid Z=1, S=0\} &= E\{w_{0,\bar{s}\bar{s}}(\boldsymbol{X})h(\boldsymbol{X}) \mid Z=0, S=0\}, \\ E\{w_{1,ss}(\boldsymbol{X})h(\boldsymbol{X}) \mid Z=1, S=1\} &= E\{h(\boldsymbol{X}) \mid Z=0, S=1\}. \end{aligned}$$

The above corollaries are direct applications of Propositions 1 and 2. Intuitively, because any functions of the covariates $h(\boldsymbol{X})$ are unaffected by the treatment within principal strata, the "PCEs" on $h(\boldsymbol{X})$ are all zeros. Although simple, the balancing conditions in Corollaries 1 and 2 allow for model checking for principal scores, and are therefore of practical importance. If the balancing conditions above are obviously violated, we need to build a more flexible model to account for the residual dependence of $U$ on $\boldsymbol{X}$. For example, we can add higher order polynomial and interaction terms of the covariates into the Logistic model, until the balancing conditions are well satisfied. This idea is similar to designs of observational studies for achieving objective causal inference (Rubin 2007, 2008). When constructing weighted samples, we do not have access to the outcome data, because we require only $(\boldsymbol{X}, Z, S)$ for creating the principal score estimates. This outcome-free strategy for designs, advocated by Rubin (2007, 2008) and Imbens and Rubin (2015), has the merit of being free of data snooping based on repeated search for favorable outcome models.

### 4.2. Estimating Principal Scores

Although we have nonparametric identification results under the PI assumptions 5 and 6, we can easily deal with only low dimensional and discrete covariates to estimate the principal scores. With high dimensional or continuous covariates, we need to specify models for $\Pr(U \mid \boldsymbol{X})$.

Under Strong Monotonicity, $U$ takes only two values, and we can use a Logistic model for $\Pr(U \mid \boldsymbol{X})$. By Randomization, we can fit a Logistic model of $S$ on $\boldsymbol{X}$ using only the data from

the treatment group, because within arm $Z = 1$, we have $S = 1$ if and only if $U = s\bar{s}$, and $S = 0$ if and only if $U = \bar{s}\bar{s}$.

Under Monotonicity, $U$ takes three values, we can model $\Pr(U \mid \boldsymbol{X})$ as a three-level Multinomial Logistic model, and use the EM algorithm (Dempster et al. 1977) to find the Maximum Likelihood Estimates (MLEs) by treating $U$ as missing data. See the supplementary material for computational details.

In practice, correct specification of the principal score model $\Pr(U \mid \boldsymbol{X})$ is crucial for the validity of the principal causal effect estimation, because misspecification of $\Pr(U \mid \boldsymbol{X})$ may lead to biased estimators for the PCEs. After fitting a principal score model, we can use Corollaries 1 and 2 to check balance of some important covariates and their functions. If the balancing conditions are violated, we can fit a more flexible model (e.g., adding high order polynomials or interaction terms of $\boldsymbol{X}$ into the Logistic models) until the balance conditions are satisfied.

## 5. Modeling the Outcome and Model-Assisted Estimators

Previous identification and sensitivity analysis results assume infinite amounts of data or a known distribution of the observed data. In this section, we discuss finite sample estimators of PCEs. For simplicity, in the main text we will discuss only the estimator for $ACE_{s\bar{s}}$ under Monotonicity. We have similar results for other strata, the cases under Strong Monotonicity, and the cases for sensitivity analysis; we relegate the technical details to the supplementary material.

The identification formulas in Propositions 1–6 immediately give us simple moment estimators by weighting, with $e_u(\boldsymbol{X})$ and $\pi_u$ replaced by their consistent estimators, and the expectations over the population replaced by their sample analogues. In the above discussion about identification and moment estimators for PCEs, we use the covariates to predict latent strata and create weights. In fact, covariates contain useful information about both the principal strata and the outcome distributions. Now we will use covariate adjustment to improve statistical efficiency for estimation. Covariate adjustment is based on the following simple fact that for all $u$ and all fixed vectors $\boldsymbol{\beta}_{z,u}$,

$$ACE_u = E\{Y(1) - \boldsymbol{\beta}_{1,u}^\top \boldsymbol{X} \mid U = u\} - E\{Y(0) - \boldsymbol{\beta}_{0,u}^\top \boldsymbol{X} \mid U = u\} + (\boldsymbol{\beta}_{1,u} - \boldsymbol{\beta}_{0,u})^\top E(\boldsymbol{X} \mid U = u). \quad (1)$$

Treating the "residual" $Y(z) - \boldsymbol{\beta}_{z,u}^\top \boldsymbol{X}$ as a new "potential outcome," we can apply Proposition 2 to identify three expectation terms in formula (1) via the following corollary.

**Corollary 3.** Under Monotonicity and GPI, we have

$$\begin{aligned}
E\{Y(1) - \boldsymbol{\beta}_{1,s\bar{s}}^\top \boldsymbol{X} \mid U = s\bar{s}\} &= E\{w_{1,s\bar{s}}(\boldsymbol{X})(Y - \boldsymbol{\beta}_{1,s\bar{s}}^\top \boldsymbol{X}) \mid Z = 1, S = 1\}, \\
E\{Y(0) - \boldsymbol{\beta}_{0,s\bar{s}}^\top \boldsymbol{X} \mid U = s\bar{s}\} &= E\{w_{0,s\bar{s}}(\boldsymbol{X})(Y - \boldsymbol{\beta}_{0,s\bar{s}}^\top \boldsymbol{X}) \mid Z = 0, S = 0\}, \\
E(\boldsymbol{X} \mid U = s\bar{s}) &= E\{w_{1,s\bar{s}}(\boldsymbol{X})\boldsymbol{X} \mid Z = 1, S = 1\} = E\{w_{0,s\bar{s}}(\boldsymbol{X})\boldsymbol{X} \mid Z = 0, S = 0\}.
\end{aligned}$$

Define $n_{zs} = \#\{i : Z_i = z, S_i = s\}$. The covariate-adjusted estimator for $ACE_{s\bar{s}}$ is

$$\begin{aligned}
\widehat{ACE}_{s\bar{s}}^{\text{adj}} = &\frac{1}{n_{11}} \sum_{\{i:Z_i=1,S_i=1\}} \widehat{w}_{1,s\bar{s}}(\boldsymbol{X}_i)(Y_i - \boldsymbol{\beta}_{1,s\bar{s}}^\top \boldsymbol{X}_i) - \frac{1}{n_{00}} \sum_{\{i:Z_i=0,S_i=0\}} \widehat{w}_{0,s\bar{s}}(\boldsymbol{X}_i)(Y_i - \boldsymbol{\beta}_{0,s\bar{s}}^\top \boldsymbol{X}) \\
&+ \frac{1}{n_{11} + n_{00}} (\boldsymbol{\beta}_{1,s\bar{s}} - \boldsymbol{\beta}_{0,s\bar{s}})^\top \left\{ \sum_{\{i:Z_i=1,S_i=1\}} \widehat{w}_{1,s\bar{s}}(\boldsymbol{X}_i)\boldsymbol{X}_i + \sum_{\{i:Z_i=0,S_i=0\}} \widehat{w}_{0,s\bar{s}}(\boldsymbol{X}_i)\boldsymbol{X}_i \right\}. \quad (2)
\end{aligned}$$

As long as the potential outcomes are correlated with the covariates, the "residual potential outcomes," $Y(z) - \boldsymbol{\beta}_{z,u}^\top \boldsymbol{X}$, will have smaller variances than the original potential outcomes. Therefore, the covariate-adjusted estimator in formula (2) tends to have a smaller asymptotic variance than the unadjusted estimator. Our simulation studies have verified this intuition. Although Corollary 3 and the covariate-adjusted estimator in formula (2) hold for any fixed vectors $\boldsymbol{\beta}_{1,s\bar{s}}$ and $\boldsymbol{\beta}_{0,s\bar{s}}$, we need to choose them in practice. Intuitively, we can choose $\boldsymbol{\beta}_{z,s\bar{s}}$ as the linear regression coefficient of $Y(z)$ onto the space spanned by $\boldsymbol{X}$ for units $U = s\bar{s}$, i.e.,

$$\boldsymbol{\beta}_{z,s\bar{s}} = \{E(\boldsymbol{X}\boldsymbol{X}^\top \mid U = s\bar{s})\}^{-1} E\{\boldsymbol{X}Y(z) \mid U = s\bar{s}\}.$$

Similar to Corollary 3, each component of the above least squares formula is also identifiable.

**Corollary 4.** Under Monotonicity and GPI, we have

$$\begin{aligned}
E\{\boldsymbol{X}Y(1) \mid U = s\bar{s}\} &= E\{w_{1,s\bar{s}}(\boldsymbol{X})\boldsymbol{X}Y \mid Z = 1, S = 1\}, \\
E\{\boldsymbol{X}Y(0) \mid U = s\bar{s}\} &= E\{w_{0,s\bar{s}}(\boldsymbol{X})\boldsymbol{X}Y \mid Z = 0, S = 0\}, \\
E(\boldsymbol{X}\boldsymbol{X}^\top \mid U = s\bar{s}) &= E\{w_{1,s\bar{s}}(\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^\top \mid Z = 1, S = 1\} = E\{w_{0,s\bar{s}}(\boldsymbol{X})\boldsymbol{X}\boldsymbol{X}^\top \mid Z = 0, S = 0\}.
\end{aligned}$$

Therefore, we choose $\boldsymbol{\beta}_{1,s\bar{s}}$ as the weighted least squares regression coefficient of $Y_i$ on $\boldsymbol{X}_i$ using samples with $(Z_i = 1, S_i = 1)$ and weights $w_{1,s\bar{s}}(\boldsymbol{X}_i)$, and $\boldsymbol{\beta}_{0,s\bar{s}}$ as the weighted least squares regression coefficient of $Y_i$ on $\boldsymbol{X}_i$ using samples with $(Z_i = 0, S_i = 0)$ and weights $w_{0,s\bar{s}}(\boldsymbol{X}_i)$.

However, we do not assume that the response surface of $Y(z)$ on $\boldsymbol{X}_i$ is linear, and the consistency of the estimators does not rely on any modeling assumptions about $Y(z)$. Our estimators are essentially moment estimators, and their consistency and asymptotic Normality follow directly from standard arguments of the Law of Large Numbers and Central Limit Theorem. Therefore, they have superior statistical properties compared to principal stratification analysis based on Normal mixture models, which have unbounded likelihood and inaccurate asymptotic Normal approximations as pointed out by Frumento et al. (2016). Furthermore, in our simulation studies shown in the supplementary material, we compare our method and Jo and Stuart (2009)'s method involving outcome modeling, and find that our estimator is not only robust to misspecification of the outcome model but also has smaller standard error.

# 6. Sensitivity Analysis

The theoretical foundation of the identification and estimation relies crucially on Monotonicity and PI, which are fundamentally untestable. In some cases, however, these two assumptions may not be easily justified according to background knowledge. In this section, we propose approaches to conducting sensitivity analysis to assess the impact of violations of Monotonicity or PI.

## 6.1. Sensitivity Analysis for Principal Ignorability

The PI assumptions are critical for nonparametric identification of the PCEs as shown in Section 3. They require the observed covariates $\boldsymbol{X}$ capture the key characteristics that affect both the principal stratum and the potential outcomes. They are sufficient conditions to ensure nonparametric identification, which are similar to the ignorability assumption used in causal inference with observational studies (Rosenbaum and Rubin 1983b) and the sequential ignorability assumption used in mediation analysis (cf. VanderWeele 2015). In many cases, the more covariates we observe, the more plausible these assumptions will become. In practice, however, we may not able to collect adequate covariates to remove the "confounding" between the principal stratification and the

outcome variable. Unfortunately, the PI assumptions cannot be validate by the observed data. Although there is a long history of sensitivity analysis in observational studies (e.g., Rosenbaum and Rubin 1983a, Rosenbaum 2002), there are only a few sensitivity analysis techniques for principal stratification analysis with binary outcomes (e.g., Sjölander et al. 2009; Schwartz et al. 2012) and some modeling assumptions (e.g., Gilbert et al. 2003) under Monotonicity. We provide a more general framework to assess the sensitivity of the deviations from the PI assumptions.

**Under Strong Monotonicity** According to its proof in the supplementary material, Proposition 1 holds under a weaker version of PI, $E\{Y(0) \mid U = s\bar{s}, \boldsymbol{X}\} = E\{Y(0) \mid U = \bar{s}\bar{s}, \boldsymbol{X}\}$, which requires the means of the control potential outcomes be the same for strata $U = s\bar{s}$ and $U = \bar{s}\bar{s}$ conditional on covariates $\boldsymbol{X}$. Therefore, our sensitivity analysis is based on the deviation from this weaker assumption, captured by a single sensitivity parameter

$$\varepsilon = \frac{E\{Y(0) \mid U = s\bar{s}, \boldsymbol{X}\}}{E\{Y(0) \mid U = \bar{s}\bar{s}, \boldsymbol{X}\}},$$

where we implicitly assume that $\varepsilon$ does not depend on the covariates $\boldsymbol{X}$. When the outcome is binary, the sensitivity parameter $\varepsilon$ becomes the relative risk of $U$ on the control potential outcome $Y(0)$ given covariates $\boldsymbol{X}$. When $\varepsilon = 1$, the same identification results hold as those under PI. When $\varepsilon \neq 1$, we can identify the PCEs for a fixed value of $\varepsilon$, as shown in the following theorem.

**Proposition 3.** Under Strong Monotonicity, for a fixed value of $\varepsilon$, we can identify the PCEs by

$$\begin{aligned} ACE_{s\bar{s}} &= E(Y \mid Z = 1, S = 1) - E\{w^{\varepsilon}_{s\bar{s}}(\boldsymbol{X})Y \mid Z = 0\}, \\ ACE_{\bar{s}\bar{s}} &= E(Y \mid Z = 1, S = 0) - E\{w^{\varepsilon}_{\bar{s}\bar{s}}(\boldsymbol{X})Y \mid Z = 0\}, \end{aligned}$$

where

$$w^{\varepsilon}_{s\bar{s}}(\boldsymbol{X}) = \frac{\varepsilon e_{s\bar{s}}(\boldsymbol{X})}{\{\varepsilon e_{s\bar{s}}(\boldsymbol{X}) + e_{\bar{s}\bar{s}}(\boldsymbol{X})\}\pi_{s\bar{s}}}, \quad w^{\varepsilon}_{\bar{s}\bar{s}}(\boldsymbol{X}) = \frac{e_{\bar{s}\bar{s}}(\boldsymbol{X})}{\{\varepsilon e_{s\bar{s}}(\boldsymbol{X}) + e_{\bar{s}\bar{s}}(\boldsymbol{X})\}\pi_{\bar{s}\bar{s}}}.$$

Although the principal scores remain the same as in Proposition 1, the new weight $w^{\varepsilon}_{u}(\boldsymbol{X})$ further depends on the deviation from PI, with the principal score $e_{s\bar{s}}(\boldsymbol{X})$ over-weighted by the sensitivity parameter $\varepsilon$.

**Under Monotonicity** According to its proof in the supplementary material, Proposition 2 holds under a weaker version of GPI, i.e., $E\{Y(1) \mid U = s\bar{s}, \boldsymbol{X}\} = E\{Y(1) \mid U = ss, \boldsymbol{X}\}$ and $E\{Y(0) \mid U = s\bar{s}, \boldsymbol{X}\} = E\{Y(0) \mid U = \bar{s}\bar{s}, \boldsymbol{X}\}$, which require the conditional means of $Y(1)$ be the same for strata $U = s\bar{s}$ and $U = ss$, and the conditional means of $Y(0)$ be the same for strata $U = s\bar{s}$ and $U = \bar{s}\bar{s}$ given covariates $\boldsymbol{X}$. Therefore, our sensitivity analysis is based on the deviations from the above weaker assumption, captured by the following two sensitivity parameters:

$$\varepsilon_1 = \frac{E\{Y(1) \mid U = s\bar{s}, \boldsymbol{X}\}}{E\{Y(1) \mid U = ss, \boldsymbol{X}\}}, \quad \varepsilon_0 = \frac{E\{Y(0) \mid U = s\bar{s}, \boldsymbol{X}\}}{E\{Y(0) \mid U = \bar{s}\bar{s}, \boldsymbol{X}\}},$$

where $\varepsilon_1$ and $\varepsilon_0$ are for the potential outcomes under treatment and control, respectively. The sensitivity parameters $\varepsilon_1$ and $\varepsilon_0$ enjoy transparent interpretations, which allows us to select the range of them according to background knowledge. For example, in the flu shot encouragement design with noncompliance discussed in Hirano et al. (2000), it may be reasonable to believe that on average the never-takers ($U = \bar{s}\bar{s}$) are the strongest patients and the always-takers ($U = ss$) are the weakest patients. In this case, the outcome of interest is an indicator of flu related hospital visit, and therefore we can select sensitivity parameters within the range $\varepsilon_1 < 1$ and $\varepsilon_0 > 1$. We will analyze this example in detail in Section 8.1.

For fixed values of the sensitivity parameters $(\varepsilon_1, \varepsilon_0)$, we have the following theorem.

**Proposition 4.** Under Monotonicity, and for fixed values of $(\varepsilon_1, \varepsilon_0)$, we can identify the PCEs by

$$\begin{aligned}
ACE_{s\bar{s}} &= E\{w^{\varepsilon_1}_{1,s\bar{s}}(\boldsymbol{X})Y \mid Z = 1, S = 1\} - E\{w^{\varepsilon_0}_{0,s\bar{s}}(\boldsymbol{X})Y \mid Z = 0, S = 0\}, \\
ACE_{\bar{s}\bar{s}} &= E(Y \mid Z = 1, S = 0) - E\{w^{\varepsilon_0}_{0,\bar{s}\bar{s}}(\boldsymbol{X})Y \mid Z = 0, S = 0\}, \\
ACE_{ss} &= E\{w^{\varepsilon_1}_{1,ss}(\boldsymbol{X})Y \mid Z = 1, S = 1\} - E(Y \mid Z = 0, S = 1),
\end{aligned}$$

where

$$\begin{aligned}
w^{\varepsilon_1}_{1,s\bar{s}}(\boldsymbol{X}) &= \frac{\varepsilon_1 e_{s\bar{s}}(\boldsymbol{X})}{\varepsilon_1 e_{s\bar{s}}(\boldsymbol{X}) + e_{ss}(\boldsymbol{X})} \Big/ \frac{\pi_{s\bar{s}}}{\pi_{s\bar{s}} + \pi_{ss}}, \qquad w^{\varepsilon_0}_{0,s\bar{s}}(\boldsymbol{X}) = \frac{\varepsilon_0 e_{s\bar{s}}(\boldsymbol{X})}{\varepsilon_0 e_{s\bar{s}}(\boldsymbol{X}) + e_{\bar{s}\bar{s}}(\boldsymbol{X})} \Big/ \frac{\pi_{s\bar{s}}}{\pi_{s\bar{s}} + \pi_{\bar{s}\bar{s}}}, \\
w^{\varepsilon_0}_{0,\bar{s}\bar{s}}(\boldsymbol{X}) &= \frac{e_{\bar{s}\bar{s}}(\boldsymbol{X})}{\varepsilon_0 e_{s\bar{s}}(\boldsymbol{X}) + e_{\bar{s}\bar{s}}(\boldsymbol{X})} \Big/ \frac{\pi_{\bar{s}\bar{s}}}{\pi_{s\bar{s}} + \pi_{\bar{s}\bar{s}}}, \qquad w^{\varepsilon_1}_{1,ss}(\boldsymbol{X}) = \frac{e_{ss}(\boldsymbol{X})}{\varepsilon_1 e_{s\bar{s}}(\boldsymbol{X}) + e_{ss}(\boldsymbol{X})} \Big/ \frac{\pi_{ss}}{\pi_{s\bar{s}} + \pi_{ss}}.
\end{aligned}$$

The weights have similar adjustments as in Proposition 3, over-weighting the principal score

$w_{s\bar{s}}(\boldsymbol{X})$ by the sensitivity parameters $\varepsilon_1$ and $\varepsilon_0$ in the treatment group and control group, respectively. $ACE_{s\bar{s}}$ depends on both $\varepsilon_1$ and $\varepsilon_0$, $ACE_{\bar{s}\bar{s}}$ only on $\varepsilon_0$, and $ACE_{ss}$ only on $\varepsilon_1$.

As a side note, Proposition 4 not only allows for sensitivity analysis of possible violations of GPI, but also allows for testing the fundamental assumptions of GPI and ER. To be more specific, if $ACE_{\bar{s}\bar{s}} = 0$ and $\varepsilon_0 = 1$, then

$$E(Y \mid Z = 1, S = 0) = E\{w_{0,\bar{s}\bar{s}}(\boldsymbol{X})Y \mid Z = 0, S = 0\}. \tag{3}$$

The contrapositive states that if we reject (3) by the observed data, then we must reject $ACE_{\bar{s}\bar{s}} = 0$ or $\varepsilon_0 = 1$. Therefore, if we assume $ACE_{\bar{s}\bar{s}} = 0$, then we can test $\varepsilon_0 = 1$; if we assume $\varepsilon_0 = 1$, then we can test $ACE_{\bar{s}\bar{s}} = 0$. Analogous discussion applies to $ACE_{ss} = 0$ and $\varepsilon_1 = 1$. Guo et al. (2014) proposed a parametric likelihood ratio test for GPI under Monotonicity and ER. In fact, Proposition 4 implies tests for compatibility of GPI and ER, which sometimes can be an important initial step in empirical studies when we are unsure about the underlying assumptions. For instance, if we have important covariate information and impose GPI, then Proposition 4 allows us to test ER in the noncompliance setting or the causal necessity in the surrogate problem. If the test is rejected, then we may reject ER and causal necessity, but we may also doubt about the GPI assumption. Although this kind of discordant result does not provide a definite answer, it does warn us of the underlying assumptions and may lead us to conduct more careful analysis or better study design. See Yang et al. (2014) for a concrete example and philosophical discussion about checking compatibility of untestable assumptions in causal inference.

### 6.2. Sensitivity Analysis for Monotonicity

Without Monotonicity, we have all four principal strata, and we cannot even identify their proportions without further assumptions. The inferential difficulties restricts the scope of the current literature to be under Monotonicity. Some exceptions (e.g., Zhang et al. 2009; Ding et al. 2011; Frumento et al. 2012) rely on either strong modeling assumptions or additional information. We take an alternative perspective, and propose an approach to performing sensitivity analysis when Monotonicity is not plausible. We introduce the following sensitivity parameter $\xi$ capturing the

deviation from Monotonicity:

$$\xi = \frac{\Pr(U = \bar{s}s \mid \boldsymbol{X})}{\Pr(U = s\bar{s} \mid \boldsymbol{X})},$$

which is the ratio between the probabilities of strata $U = \bar{s}s$ and $U = s\bar{s}$ conditional on covariates $\boldsymbol{X}$. Furthermore, the conditional ratio is also the marginal ratio of the probabilities, i.e., $\xi = \Pr(U = \bar{s}s)/\Pr(U = s\bar{s})$. The sensitivity parameter can take values from 0 to $\infty$. When $\xi = 0$, we have Monotonicity; when $\xi = 1$, we have equal proportions of $s\bar{s}$ and $\bar{s}s$, and thus zero average causal effect on $S$; when $0 < \xi < 1$, we allow deviation from Monotonicity but still preserve positive average causal effect on $S$; when $\xi > 1$, we have negative average causal effect on $S$. Without loss of generality, we will assume $p_1 - p_0 \geq 0$ and $0 \leq \xi \leq 1$ from now on for sensitivity analysis.

**Proposition 5.** For a fixed sensitivity parameter $\xi$, we can identify the proportions by

$$\pi_{s\bar{s}} = \frac{p_1 - p_0}{1 - \xi}, \quad \pi_{\bar{s}\bar{s}} = 1 - p_0 - \frac{p_1 - p_0}{1 - \xi}, \quad \pi_{ss} = p_1 - \frac{p_1 - p_0}{1 - \xi}, \quad \pi_{\bar{s}s} = \frac{\xi(p_1 - p_0)}{1 - \xi}, \tag{4}$$

which further imply that the sensitivity parameter $\xi$ is bounded by

$$0 \leq \xi \leq 1 - \frac{p_1 - p_0}{\min(p_1, 1 - p_0)} \leq 1. \tag{5}$$

Although $\xi$ is not identifiable, the observed data provide an upper bound for it when the average causal effect on $S$ is non-negative. Therefore, we need to only perform sensitivity analysis within the empirical version of the above bounds of $\xi$.

Analogously, we can show that the principal score $e_u(\boldsymbol{X})$ is identifiable with a known $\xi$, by replacing $p_1$ and $p_0$ in formula (4) by $p_1(\boldsymbol{X})$ and $p_0(\boldsymbol{X})$, respectively. Consequently, GPI is sufficient to identify the PCEs.

**Proposition 6.** Under GPI, and for a fixed $\xi$, we can identify the PCEs by

$$ACE_u \quad = \quad E\{w_{1,u}(\boldsymbol{X})Y \mid Z = 1, S = s(1)\} - E\{w_{0,u}(\boldsymbol{X})Y \mid Z = 0, S = s(0)\},$$

where $s(1)$ and $s(0)$ correspond to the values of $S(1)$ and $S(0)$ of $U = u$, and the weight $w_{z,u}(\boldsymbol{X})$ is defined in the same way as Proposition 2.

Proposition 6 is similar to Proposition 2, except for that all the observed groups defined by

$(Z, S)$ are mixtures of two latent strata.

To end this subsection, we discuss a model strategy for principal scores without Monotonicity. Combining $s\bar{s}$ and $\bar{s}s$ into one category, we define $V_i = U_i$ if $U_i = ss$ or $\bar{s}\bar{s}$, and $V_i = s\&\bar{s}$ if $U_i = s\bar{s}$ or $\bar{s}s$. We can model $\Pr(U \mid \boldsymbol{X})$ by two steps. First, we model $\Pr(V \mid \boldsymbol{X})$ as a three-level Multinomial Logistic regression. Second, we partition the category of $V$, $s\&\bar{s}$, into two sub-categories of $U$, $s\bar{s}$ and $\bar{s}s$, with probabilities $\Pr(U = s\bar{s} \mid V = s\&\bar{s}, \boldsymbol{X}) = 1/(1+\xi)$ and $\Pr(U = \bar{s}s \mid V = s\&\bar{s}, \boldsymbol{X}) = \xi/(1+\xi)$. We show in the supplementary material the EM algorithm for computing the MLE of the above model. After estimating the principal scores, we can apply the weighting and covariate-adjustment method to estimate the PCEs as discussed in Section 5.

## 7. Simulation Studies

To examine the finite sample performance of our estimators, we conduct a series of simulation studies. Let the sample sizes be 500 in all scenarios. For unit $i$, we generate $X_{i1}, \ldots, X_{i4} \overset{i.i.d.}{\sim} N(0,1)$ and $X_{i5} \sim \text{Bern}(1/2)$, and let $\boldsymbol{X}_i = (1, X_{i1}, \ldots, X_{i5})^\top$. We conduct simulations under Strong Monotonicity and Monotonicity, respectively. In each scenario we consider five cases indexed by the parameter $\theta = -1, -0.5, 0, 0.5$, and 1. We postpone the interpretation of $\theta$ until afterwards.

Under Strong Monotonicity, for each $\theta$ we generate principal strata from a Logit model $\text{logit}\Pr(U_i = s\bar{s} \mid \boldsymbol{X}_i) = \boldsymbol{\theta}^\top \boldsymbol{X}_i$, where $\boldsymbol{\theta} = (0, 0.5, 0.5, 1, 1, \theta)^\top$. We generate Normal potential outcomes from $Y_i(1) \mid \boldsymbol{X}_i \sim N\left(\sum_{j=1}^5 X_{ij} + 2 \cdot I_{\{U=s\bar{s}\}} + 1, 1\right)$ and $Y_i(0) \mid \boldsymbol{X}_i \sim N\left(\sum_{j=1}^5 X_{ij} + 2, 1\right)$; Bernoulli potential outcomes from $\text{logit}\Pr\{Y_i(1) = 1 \mid \boldsymbol{X}_i\} = 0.3\sum_{j=1}^5 X_{ij} + I_{\{U=s\bar{s}\}}$ and $\text{logit}\Pr(\{Y_i(0) = 1 \mid \boldsymbol{X}_i\} = 0.3\sum_{j=1}^5 X_{ij} + 0.5$.

Under Monotonicity, for each $\theta$ we generate principal strata from a Multinomial Logit model $\Pr(U_i = u \mid \boldsymbol{X}_i) = \exp(\boldsymbol{\theta}_u^\top \boldsymbol{X}_i)/\sum_{u'} \exp(\boldsymbol{\theta}_{u'}^\top \boldsymbol{X}_i)$ for $u = s\bar{s}, ss, \bar{s}\bar{s}$, where $\boldsymbol{\theta}_{ss} = (0.25, 0.5, 0.5, 1, 1, \theta)$, $\boldsymbol{\theta}_{\bar{s}\bar{s}} = (-0.25, 1, 1, 0.5, 0.5, \theta)$ and $\boldsymbol{\theta}_{s\bar{s}} = \boldsymbol{0}$. We generate Normal potential outcomes from $Y_i(1) \mid \boldsymbol{X}_i \sim N\left(\sum_{j=1}^5 X_{ij} - I_{\{U=\bar{s}\bar{s}\}} + 4, 1\right)$ and $Y_i(0) \mid \boldsymbol{X}_i \sim N\left(\sum_{j=1}^5 X_{ij} + I_{\{U=ss\}} + 1, 1\right)$; Bernoulli potential outcomes from $\text{logit}\Pr\{Y_i(1) = 1 \mid \boldsymbol{X}_i\} = 0.3\sum_{j=1}^5 X_{ij} + 0.25\left(I_{\{U=\bar{s}\bar{s}\}} - 1\right)$ and $\text{logit}\Pr\{Y_i(0) = 1 \mid \boldsymbol{X}_i\} = 0.3\sum_{j=1}^5 X_{ij} + 0.25\left(1 - I_{\{U=ss\}}\right)$. Although the above data generating mechanisms violate GPI, they satisfy its weaker version, i.e., $\varepsilon_1 = \varepsilon_0 = 1$, which also suffices to ensure Proposition 2.

To examine the performance of our estimators with and without (the weaker version of) GPI,

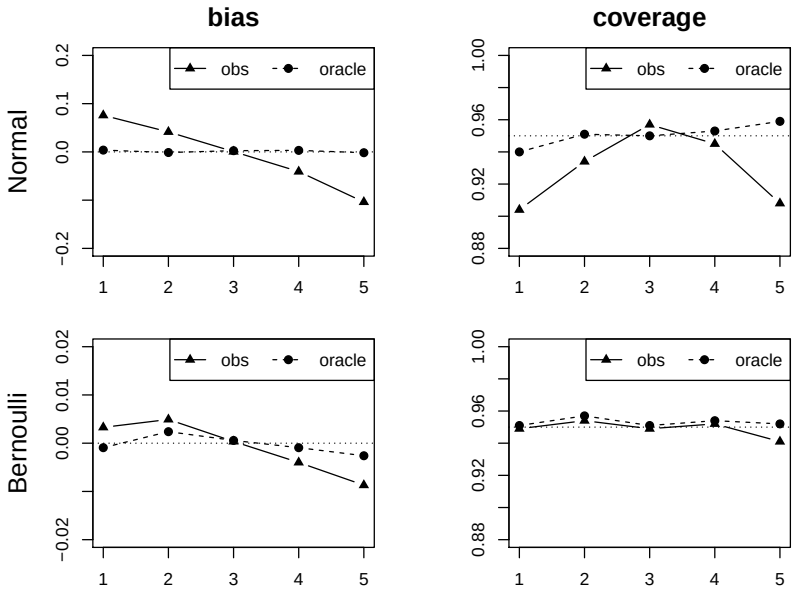
in each simulation scenario we analyze the data with and without the binary covariate $X_{i5}$, and we respectively label the corresponding results as "oracle" and "obs". Without using $X_{i5}$, we can view $\theta$ as a measure of the violation from GPI. In Figure 4, we present only the results for $ACE_{\bar{s}\bar{s}}$ using the model-assisted estimator, because in our simulations the naive weighting estimators are uniformly worse in terms of estimation efficiency. For the ease of presentation, we omit similar results for other principal strata. We use 500 bootstraps to construct 95% confidence intervals, and focus on the average biases and coverage rates over 1000 repeated samplings. With the binary covariate, our estimator has small biases and achieves nominal coverage rates, for both Normal and Bernoulli potential outcomes. Without the binary covariate, our estimators have bias issues for both Normal and Bernoulli potential outcomes when $|\theta|$ approaches one, i.e., GPI is severely violated. The interval estimates under coverage the true parameters for Normal outcomes when $|\theta|$ approaches one, but the coverage properties for Bernoulli outcomes are robust with respect to the violations of GPI. This bias issue, as well as the untestable nature of PI and GPI, warns us that sensitivity analysis with respect to PI and GPI, as proposed in Section 6.1, must be an essential part of any empirical studies using principal scores to analyze principal stratification problems.

Due to the constraint of space, we compare our model-assisted estimator with Jo and Stuart (2009)'s model-based estimator in the supplementary material, showing that our estimator does not lose efficiency compared to full modeling and is robust to model misspecification of the outcome.

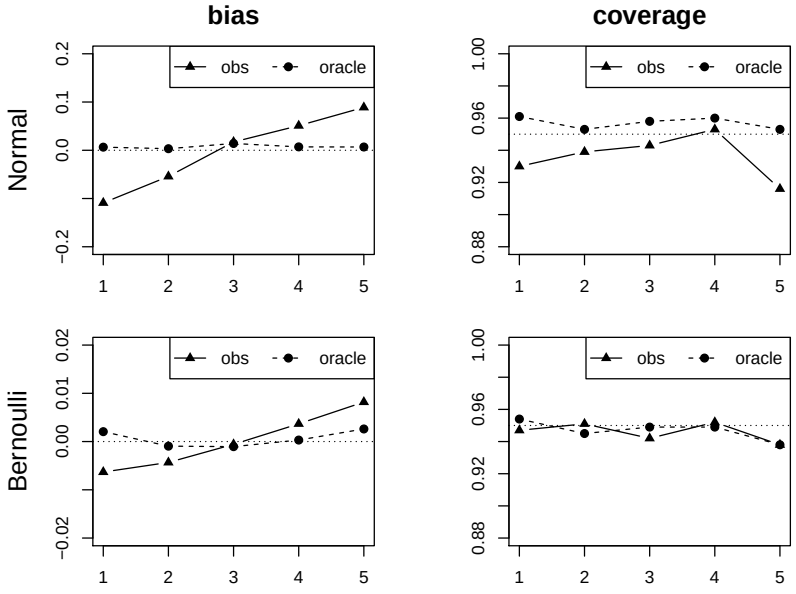
# 8. Applications

## 8.1. An Encouragement Experiment with Noncompliance

In this section, we re-analyze a flu shot encouragement experiment data set previously studied by Hirano et al. (2000). Between 1978 to 1980, a general medicine clinic in Indiana conducted an encouragement experiment, in which participating individuals' physicians were randomly assigned to the treatment arm with computer-generated letters encouraging them to inoculate their patients, or the control arm with no letters. The outcome of interest is the individual's flu-related hospitalization status during the subsequent winter. As in Hirano et al. (2000), we use the data from 1980, with 2893 experimental units. In our analysis, $Z = 1$ if an individual's physician received the letter, and $Z = 0$ otherwise. The intermediate variable $S = 1$ if the individual received the




(a) Under Strong Monotonicity. The horizontal axis shows the case numbers, and "obs" and "oracle" denote the cases with and without the binary covariate, respectively.




(b) Under Monotonicity. The horizontal axis shows the case numbers, and "obs" and "oracle" denote the cases with and without the binary covariate, respectively.

Figure 1: Simulation Results for $ACE_{\bar{s}\bar{s}}$. Each subfigure is a $2 \times 2$ matrix summarizing two repeated sampling properties (average biases and coverage rates of interval estimates).

flu shot, and $S = 0$ otherwise. The outcome of interest $Y = 1$ if the individual was hospitalized for flu-related reasons, and $Y = 0$ otherwise. The Monotonicity assumption is plausible for this data set, because we expect the encouragement letter to have nonnegative effect on taking the flu shot. Because this encouragement experiment is an open-label trial, previous researchers doubted ER due to the possible "direct effect" of the flu shot encouragement on the outcome.

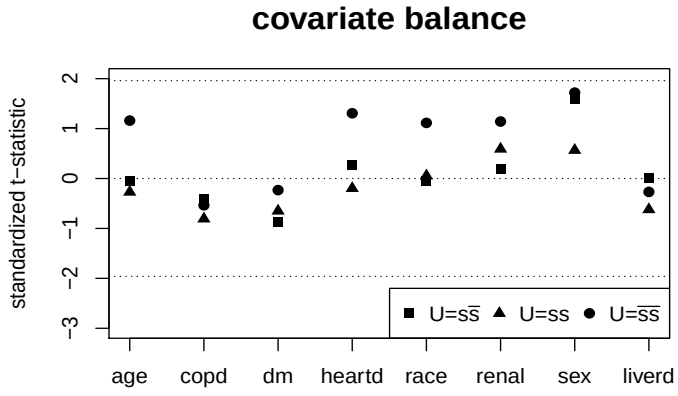
To start our analysis, we use the covariate balancing conditions in Corollary 2 to check the plausibility of the Logistic principal score model. Choosing $h(\boldsymbol{X}) = \boldsymbol{X}$ is reasonable, because all covariates are binary except for "age." The balance checking is equivalent to estimating the PCEs on $h(\boldsymbol{X})$, known to be zero. Therefore, the corresponding "standardized $t$-statistics" should follow standard Normal distributions. Figure 2a shows that the covariates are well balanced. Assuming GPI, we estimate the PCEs with standard errors and 95% confidence intervals in parentheses as:

$$\widehat{ACE}_{s\bar{s}} = -0.018\,[-0.052, 0.016], \quad \widehat{ACE}_{ss} = -0.046\,[-0.091, 0.002], \quad \widehat{ACE}_{\bar{s}\bar{s}} = -0.006\,[-0.030, 0.017].$$

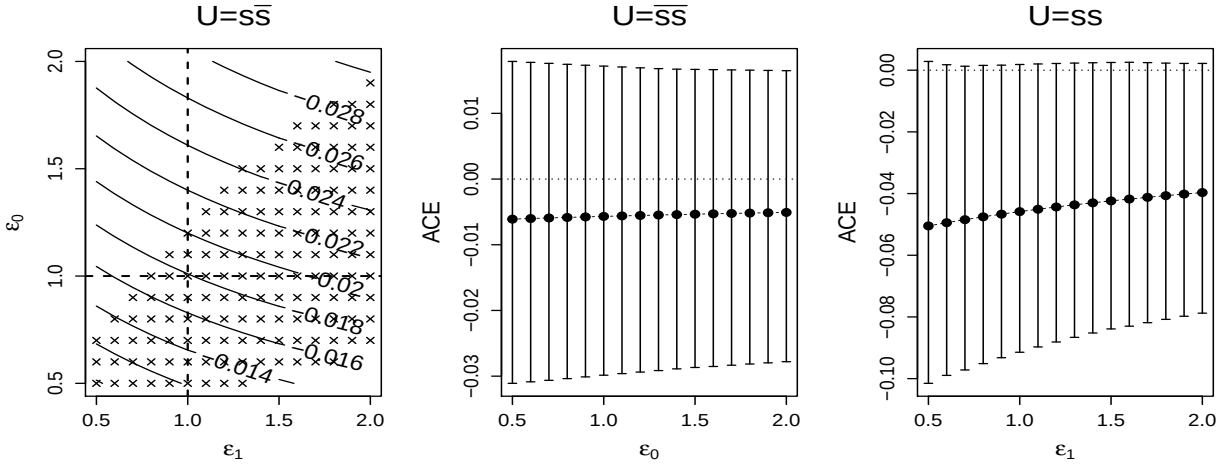
Therefore, for compliers, receiving the encouragement letter will lower the chance of flu related hospital visit by 1.8%, but this effect is not significant. Furthermore, ER seems plausible for never-takers. However, there is some evidence that it does not hold for always-takers, because the upper confidence limit is close to zero. Our findings corroborate Hirano et al. (2000)'s argument that "it is more plausible to impose the exclusion restriction for never-takers than for always-takers." Hirano et al. (2000)'s results required careful analysis, including using data dependent priors with several tuning parameters that account for the background knowledge. Our analysis under GPI yields coherent conclusions as theirs. Therefore, if we believe their prior knowledge and statistical analysis, then GPI seems plausible in this example. At least, there is no obvious contradiction derived from two different analysis, and our results under GPI have meaningful scientific interpretations.

Nevertheless, the data cannot validate GPI, an untestable assumption requiring observed covariates $\boldsymbol{X}$ contain all characteristics related to the latent principal stratum and potential outcomes; Hirano et al. (2000)'s analysis does not contradict GPI but does not prove it either. As advocated in Section 6, we perform sensitivity analysis for GPI, allowing $(\varepsilon_1, \varepsilon_0)$ to vary within $[1/2, 2] \times [1/2, 2]$ with results in Figure 2b. If we are willing to assume that never-takers are the strongest patients and the always-takers are the weakest patients, we can restrict our sensitivity analysis within the

region with $\varepsilon_1 < 1$ and $\varepsilon_0 > 1$. Interestingly, within this range most of the confidence intervals $\widehat{ACE}_{s\bar{s}}$ do not cover zero, suggesting that there is a significant causal effect for compliers. Furthermore, $\widehat{ACE}_{ss}$ and $\widehat{ACE}_{\bar{s}s}$ are relatively robust to $\varepsilon_1$ and $\varepsilon_0$, respectively. The upper confidence limits for $\widehat{ACE}_{ss}$ are always close to zero as $\varepsilon_1$ varies, showing weak evidence for violation of ER for always-takers; the centers of the confidence intervals for $\widehat{ACE}_{\bar{s}\bar{s}}$ are always close to zero as $\varepsilon_0$ varies, suggesting that ER holds for never-takers. Fortunately, although the point and interval estimators vary with the sensitivity parameters, the final conclusions do not change materially.




(a) Covariate Balance Check. The horizontal axis shows the names of the covariates.




(b) Sensitivity Analysis for GPI. The first subfigure shows the contours of the point estimates of $ACE_{s\bar{s}}$ for fixed values of $\varepsilon_1$ and $\varepsilon_0$, where "×" denotes $(\varepsilon_1, \varepsilon_0)$ such that the corresponding interval estimate covers 0. The second and third subfigures show the point and interval estimates of $ACE_{\bar{s}\bar{s}}$ for fixed values of $\varepsilon_0$, and the point and interval estimates of $ACE_{ss}$ for fixed values of $\varepsilon_1$, respectively.

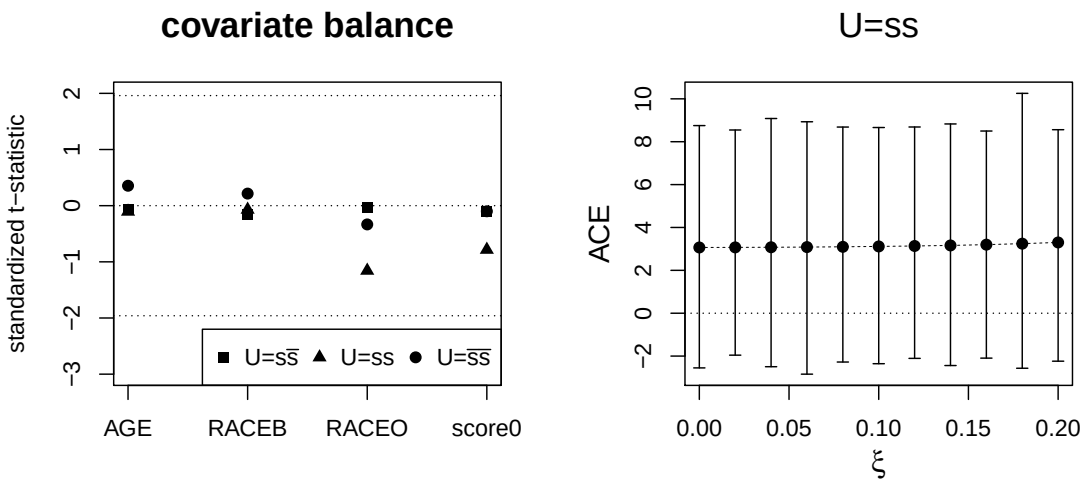
Figure 2: The Flu Shot Encouragement Experiment

## 8.2. A Randomized Trial with Truncation by Death

From October 1999 to January 2003, the Southwest Oncology Group (SWOG) conducted a randomized phase III trial (protocol 99-16) to compare the treatment of docetaxel and estramustine (DE) with mitoxantrone and prednisone (MP) in patients with metastatic, androgen-independent prostate cancer (Petrylak et al. 2004). A total of 674 eligible patients participated in the study between October 1999 and January 2003. Study participants were randomly assigned to the DE arm or the MP arm. The primary outcome is the survival time, and the secondary outcome is the health related quality of life (HRQOL). Petrylak et al. (2004) have reported the overall survival benefit of taking DE over taking MP. In our analysis, we are interested in assessing the causal effect of DE versus MP on the HRQOL one year after receiving the treatment. In our analysis, $Z = 1$ if a patient received DE, and $Z = 0$ if the patient received MP. We use the difference between HRQOL after one year and the baseline HRQOL as the outcome of interest. The survival indicator $S = 1$ if a patient survived after one year.

Because of the truncation by death problem, we are interested in estimating the survivor average causal effect. As in the previous example, we first check the plausibility of the Logistic principal score model. Figure 3a shows that we achieve covariate balance. The point estimate of the SACE is 3.07, but its standard error is 2.976 and the 95% confidence interval $[-2.93, 8.69]$ covers zero. The results show that DE is not significantly more effective than MP to improve the HRQOL of the patients, which is similar to the analysis in Ding et al. (2011). However, applying Zhang et al. (2009)'s Normal mixture model, we obtain point estimate 12.34 with standard error 47.17. The tremendous variability of the estimator is due to the unstable numerical issue and unreliable large sample Normal approximation, as investigated by Frumento et al. (2016).

However, both the treatment and control are active drugs for the prostate cancer, and therefore it is not reasonable to assume that the treatment is more effective than the control for all patients, i.e., Monotonicity may not hold. We perform sensitivity analysis for Monotonicity, and choose the range of the sensitivity parameter $\xi$ based on Proposition 5. We compute from the data that $\widehat{p}_1 = 0.496$ and $\widehat{p}_0 = 0.389$, and therefore $0 \leq \widehat{\xi} \leq 1 - (\widehat{p}_1 - \widehat{p}_0)/\{\min(\widehat{p}_1, 1 - \widehat{p}_0)\} \approx 0.217$. The sensitivity analysis results in Figure 3b show that the point and interval estimates of $ACE_{ss}$ are relatively robust to $\xi$. Furthermore, the interval estimates for $\widehat{ACE}_{ss}$ always cover zero as $\xi$ varies.




(a) Covariate Balance Check. The horizontal axis shows the names of the covariates.

(b) Sensitivity Analysis for Monotonicity. We show the point and interval estimates of $ACE_{ss}$ for fixed values of $\xi$.

Figure 3: The SWOG Randomized Trial

In summary, the sensitivity analysis results confirm our previous conclusions.

## 9. Discussion

In observational studies, causal effects can be estimated by inverse propensity score weighting (Rosenbaum and Rubin 1983b), which may be numerically unstable and have poor finite sample properties. Our estimators, weighted by probabilities themselves, do not suffer from these problems. Researchers (e.g., Bang and Robins 2005) have developed doubly robust methods in observational studies. Similar to our model-assisted estimators, these doubly robust estimators can also be derived from regression estimators in surveys (Cochran 1977). Because of this similarity, it will be interesting to develop doubly robust estimators under the PI assumptions that are consistent when either the principal score or the outcome model is correctly specified.

The theoretical results have demonstrated the two-fold role of the pretreatment covariates. First, the plausibility of the ignorability assumptions rely crucially on adequate covariates. Second, with more covariates that are predictive to the outcome, the covariate-adjusted estimators will be more efficient. Our results suggest that, in the design of randomized experiments, it is important for practitioners to try their best to collect covariates that are predictive to both the latent principal

strata and the potential outcomes, which echoes Jo and Stuart (2009) and Mealli and Pacini (2013).

Although in the main text we focused on the average causal effect within principal strata, our results can be easily extended to general causal measures. For example, we can dichotomize the outcome to identify the distributional causal effects (Ju and Geng 2010). For binary $S$, we have derived clean results and easy-to-implement estimators. For general discrete or continuous $S$, we can likewise derive theoretical results under PI by modifying the weights in Propositions 2 and 5. However, a continuous $S$ results in infinitely many principal strata, which makes it challenging to estimate the principal scores and outcome distributions conditional on continuous variables. We need more structural assumptions on the causal problems (Jin and Rubin 2008; Schwartz et al. 2011) and more sophisticated statistical inferential tools.

Missing data is an important problem that often arises in real data analysis. Our two-step procedure has some advantages if only some outcomes are missing. We can conduct the first step for estimating principal scores without any difficulty, and need only to modify the second weighting step. If the outcome is missing at random, then we can simply weight each observation by the inverse of the conditional probability of being observed given $(Z, S, \boldsymbol{X})$. However, for missing data problem, the key issue is the missing data mechanism. Other missing data mechanisms, e.g., latent ignorability (Frangakis and Rubin 1999), may be more plausible, but the identification becomes challenging. Due to this complication, we leave the missing data problem for future research.

## Acknowledgments

Peng Ding's work is partially supported by grant R305D150040 from Institute of Education Sciences, USA. We are grateful for the comments from Professors Donald Rubin and Tirthankar Dasgupta, and other participants in the "Matched Sampling and Study Designs" seminar at Harvard. We benefit from the suggestions of Avi Feller at Berkeley, Keli Liu at Stanford, and Professors Luke W. Miratrix and Joseph K. Blitzstein at Harvard. The comments from the Joint Editor, the Associate Editor, and two reviewers have helped improve the quality of our paper significantly.

## References


Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables (with discussion). *J. Am. Statist. Ass.*, 91:444–455.

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–973.

Cheng, J. and Small, D. S. (2006). Bounds on causal effects in three-arm trials with non-compliance. *J. R. Statist. Soc. B*, 68:815–836.

Cochran, W. G. (1977). *Sampling Techniques*. New York: John Wiley & Sons, 3rd edition.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Statist. Soc. B*, 39:1–38.

Ding, P., Geng, Z., Yan, W., and Zhou, X. H. (2011). Identifiability and estimation of causal effects by principal stratification with outcomes truncated by death. *J. Am. Statist. Ass.*, 106:1578–1591.

Follman, D. A. (2000). On the effect of treatment among would-be treatment compliers: an analysis of the multiple risk factor intervention trial. *J. Am. Statist. Ass.*, 95:1101–1109.

Frangakis, C. E. and Rubin, D. B. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment-noncompliance and subsequent missing outcomes. *Biometrika*, 86:365–379.

Frangakis, C. E. and Rubin, D. B. (2002). Principal stratification in causal inference. *Biometrics*, 58:21–29.

Frumento, P., Mealli, F., Pacini, B., and Rubin, D. B. (2012). Evaluating the effect of training on wages in the presence of noncompliance, nonemployment, and missing outcome data. *J. Am. Statist. Ass.*, 107:450–466.

Frumento, P., Mealli, F., Pacini, B., and Rubin, D. B. (2016). The fragility of standard inferential approaches in principal stratification models relative to direct likelihood approaches. *Stat. Anal. Data Min.*, in press.

Gilbert, P. B., Bosch, R. J., and Hudgens, M. G. (2003). Sensitivity analysis for the assessment of causal vaccine effects on viral load in HIV vaccine trials. *Biometrics*, 59:531–541.

Gilbert, P. B. and Hudgens, M. G. (2008). Evaluating candidate principal surrogate endpoints. *Biometrics*, 64:1146–1154.

Guo, Z., Cheng, J., Lorch, S. A., and Small, D. (2014). Using an instrumental variable to test for unmeasured confounding. *Stat. Med.*, 33:3528–3546.

Hill, J., Waldfogel, J., and Brooks-Gunn, J. (2002). Differential effects of high-quality child care. *J. Pol. Anal. Manag.*, 21:601–627.

Hirano, K., Imbens, G., Rubin, D. B., and Zhou, X.-H. (2000). Estimating the effect of an influenza vaccine in an encouragement design. *Biostatistics*, 1:69–88.

Hudgens, M. G. and Halloran, M. E. (2006). Causal vaccine effects on binary postinfection outcomes. *J. Am. Statist. Ass.*, 101:51–64.

Imbens, G. W. and Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* New York: Cambridge University Press.

Jiang, Z., Ding, P., and Geng, Z. (2016). Principal causal effect identification and surrogate endpoint evaluation by multiple trials. *J. R. Statist. Soc. B*, in press.

Jin, H. and Rubin, D. B. (2008). Principal stratification for causal inference with extended partial compliance. *J. Am. Statist. Ass.*, 103:101–111.

Jo, B., MacKinnon, D. P., and Vinokur, A. D. (2011). The use of propensity scores in mediation analysis. *Multivar. Behav. Res.*, 46:425–452.

Jo, B. and Stuart, E. A. (2009). On the use of propensity scores in principal causal effect estimation. *Stat. Med.*, 28:2857–2875.

Joffe, M. M., Small, D., and Hsu, C. (2007). Defining and estimating intervention effects for groups that will develop an auxiliary outcome. *Stat. Sci.*, 22:74–97.

Ju, C. and Geng, Z. (2010). Criteria for surrogate end points based on causal distributions. *J. R. Statist. Soc. B*, 72:129–142.

Mattei, A., Li, F., and Mealli, F. (2013). Exploiting multiple outcomes in Bayesian principal stratification analysis with application to the evaluation of a job training program. *Ann. Appl. Stat.*, 7:2336–2360.

Mattei, A. and Mealli, F. (2007). Application of the principal stratification approach to the Faenza randomized experiment on breast self-examination. *Biometrics*, 63:437–446.

Mattei, A. and Mealli, F. (2011). Augmented designs to assess principal strata direct effects. *J. R. Statist. Soc. B*, 73:729–752.

Mealli, F. and Pacini, B. (2013). Using secondary outcomes to sharpen inference in randomized experiments with noncompliance. *J. Am. Statist. Ass.*, 108:1120–1131.

Petrylak, D. P., Tangen, C. M., Hussain, M. H., Lara, P. N. J., Jones, J. A., Taplin, M. E., Burch, P. A., Berry, D., Moinpour, C., Kohli, M., Benson, M. C., Small, E. J., Raghavan, D., and Crawford, E. D. (2004). Docetaxel and estramustine compared with mitoxantrone and prednisone for advanced refractory prostate cancer. *New Engl. J. Med.*, 351:1513–1520.

Rosenbaum, P. R. (2002). *Observational Studies*. New York: Springer, 2nd edition.

Rosenbaum, P. R. and Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. R. Statist. Soc. B*, 45:212–218.

Rosenbaum, P. R. and Rubin, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55.

Rubin, D. B. (1980). Comment on "Randomization analysis of experimental data: the Fisher randomization test" by D. Basu. *J. Am. Statist. Ass.*, 75:591–593.

Rubin, D. B. (2006). Causal inference through potential outcomes and principal stratification: application to studies with 'censoring' due to death (with discussion). *Stat. Sci.*, 21:299–309.

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat. Med.*, 26:20–36.

Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *Ann. Appl. Stat.*, 2:808–840.

Schwartz, S., Li, F., and Reiter, J. P. (2012). Sensitivity analysis for unmeasured confounding in principal stratification settings with binary variables. *Stat. Med.*, 31:949–962.

Schwartz, S. L., Li, F., and Mealli, F. (2011). A Bayesian semiparametric approach to intermediate variables in causal inference. *J. Am. Statist. Ass.*, 106:1331–1344.

Sjölander, A., Humphreys, K., Vansteelandt, S., Bellocco, R., and Palmgren, J. (2009). Sensitivity analysis for principal stratum direct effects, with an application to a study of physical activity and coronary heart disease. *Biometrics*, 65:514–520.

Stuart, E. A. and Jo, B. (2015). Assessing the sensitivity of methods for estimating principal causal effects. *Stat. Methods Med. Res.*, 24:657–674.

VanderWeele, T. J. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford: Oxford University Press.

Yang, F. and Small, D. S. (2016). Using post-quality of life measurement information in censoring by death problems. *J. R. Statist. Soc. B*, 78:299–318.

Yang, F., Zubizarreta, J. R., Small, D. S., Lorch, S., and Rosenbaum, P. R. (2014). Dissonant conclusions when testing the validity of an instrumental variable. *Am. Stat.*, 68:253–263.

Zhang, J. L. and Rubin, D. B. (2003). Estimation of causal effects via principle stratification when some outcomes are truncated by 'death'. *J. Educ. Behav. Stat.*, 28:353–368.

Zhang, J. L., Rubin, D. B., and Mealli, F. (2009). Likelihood-based analysis of causal effects via principal stratification: new approach to evaluating job-training programs. *J. Am. Statist. Ass.*, 104:166–176.

# Supplementary material

A.10 contains the proofs under Strong Monotonicity. A.11 contains the proofs under Monotonicity. A.12 contains the proofs without Monotonicity. A.13 presents the computational details of the EM algorithms for the principal score models. A.14 gives the explicit forms of the weighting and model-assisted estimators. A.15 contains additional simulations comparing our model-assisted estimator with the model-based estimator in Jo and Stuart (2009). We will use $f(\cdot)$ as for a (conditional) probability density function, and the following basic identity of importance sampling to simplify our proofs.

**Lemma A.1.** Assuming existence of moments, $X \sim f_1(x)$ and $Y \sim f_2(y)$, we have

$$E\{g(X)\} = E\left\{\frac{f_1(Y)}{f_2(Y)} g(Y)\right\}. \tag{A.6}$$

## A.10. Proofs of the Propositions Under Strong Monotonicity

To prove Proposition 1, we first need the following lemmas.

**Lemma A.2.** Under Strong Monotonicity, $\boldsymbol{X} \perp\!\!\!\perp U \mid e_u(\boldsymbol{X})$ for $u = s\bar{s}$ and $\bar{s}\bar{s}$.

*Proof of Lemma A.2.* We have $\Pr\{U = s\bar{s} \mid \boldsymbol{X}, e_{s\bar{s}}(\boldsymbol{X})\} = \Pr(U = s\bar{s} \mid \boldsymbol{X}) = e_{s\bar{s}}(\boldsymbol{X})$, and

$$\Pr\{U = s\bar{s} \mid e_{s\bar{s}}(\boldsymbol{X})\} = E[\Pr\{U = s\bar{s} \mid \boldsymbol{X}, e_{s\bar{s}}(\boldsymbol{X})\} \mid e_{s\bar{s}}(\boldsymbol{X})] = E\{e_{s\bar{s}}(\boldsymbol{X}) \mid e_{s\bar{s}}(\boldsymbol{X})\} = e_{s\bar{s}}(\boldsymbol{X}).$$

Therefore, $\Pr\{U = s\bar{s} \mid \boldsymbol{X}, e_{s\bar{s}}(\boldsymbol{X})\} = \Pr\{U = s\bar{s} \mid e_{s\bar{s}}(\boldsymbol{X})\}$, implying $\boldsymbol{X} \perp\!\!\!\perp U \mid e_{s\bar{s}}(\boldsymbol{X})$. Because $e_{\bar{s}\bar{s}}(\boldsymbol{X}) = 1 - e_{s\bar{s}}(\boldsymbol{X})$, other conditional independence also follows. □

**Lemma A.3.** Under Strong Monotonicity and PI, $Y(0) \perp\!\!\!\perp U \mid e_u(\boldsymbol{X})$ for $u = s\bar{s}$ and $\bar{s}\bar{s}$.

*Proof of Lemma A.3.* Applying Law of Iterated Expectation (LIE), we have

$$\begin{aligned}
\Pr\{U = s\bar{s} \mid Y(0), e_{s\bar{s}}(\boldsymbol{X})\} &= E[\Pr\{U = s\bar{s} \mid Y(0), e_{s\bar{s}}(\boldsymbol{X}), \boldsymbol{X}\} \mid Y(0), e_{s\bar{s}}(\boldsymbol{X})] \\
&= E[\Pr\{U = s\bar{s} \mid Y(0), \boldsymbol{X}\} \mid Y(0), e_{s\bar{s}}(\boldsymbol{X})],
\end{aligned}$$

which, by PI, reduces to $E[\Pr\{U = s\bar{s} \mid \boldsymbol{X}\} \mid Y(0), e_{s\bar{s}}(\boldsymbol{X})] = E\{e_{s\bar{s}}(\boldsymbol{X}) \mid Y(0), e_{s\bar{s}}(\boldsymbol{X})\} = e_{s\bar{s}}(\boldsymbol{X})$. According to the proof of Lemma A.2, we also have $\Pr\{U = s\bar{s} \mid e_{s\bar{s}}(\boldsymbol{X})\} = e_{s\bar{s}}(\boldsymbol{X})$, and therefore $\Pr\{U = s\bar{s} \mid Y(0), e_{s\bar{s}}(\boldsymbol{X})\} = \Pr\{U = s\bar{s} \mid e_{s\bar{s}}(\boldsymbol{X})\}$, implying $Y(0) \perp\!\!\!\perp U \mid e_{s\bar{s}}(\boldsymbol{X})$. Other conclusions about conditional independence also hold. □

*Proof of Proposition 1.* In the treatment group, $(Z_i = 1, S_i = 1)$ is equivalent to $(Z_i = 1, U_i = s\bar{s})$, and $(Z_i = 1, S_i = 0)$ is equivalent to $(Z_i = 1, U_i = \bar{s}\bar{s})$. Therefore, it is straightforward to identify

$$E\{Y(1) \mid U = s\bar{s}\} = E(Y \mid Z = 1, S = 1), \quad E\{Y(1) \mid U = \bar{s}\bar{s}\} = E(Y \mid Z = 1, S = 0)$$

by the observed data. The control group is a mixture of $U = s\bar{s}$ and $U = \bar{s}\bar{s}$.

On the one hand, it is relatively easy to show that

$$E\{Y(0) \mid U = u\} = E[E\{Y(0) \mid U = u, e_u(\boldsymbol{X})\} \mid U = u] = E[E\{Y(0) \mid e_u(\boldsymbol{X})\} \mid U = u], \quad \text{(A.7)}$$

according to Lemma A.3. On the other hand, the weighted mean is

$$E\{w_u(\boldsymbol{X})Y \mid Z = 0\} = E\left\{\frac{e_u(\boldsymbol{X})}{\pi_u}Y(0)\right\} = E\left[E\left\{\frac{e_u(\boldsymbol{X})}{\pi_u}Y(0) \mid e_u(\boldsymbol{X})\right\}\right] = E\left[\frac{e_u(\boldsymbol{X})}{\pi_u}E\{Y(0) \mid e_u(\boldsymbol{X})\}\right].$$

Because

$$f\{e_u(\boldsymbol{X}) \mid U = u\} = \frac{f\{e_u(\boldsymbol{X})\}\Pr\{U = u \mid e_u(\boldsymbol{X})\}}{\pi_u} = \frac{f\{e_u(\boldsymbol{X})\}e_u(\boldsymbol{X})}{\pi_u},$$

according to the proof of Lemma A.2, the weighted mean becomes

$$E\{w_u(\boldsymbol{X})Y \mid Z = 0\} = E\left[\frac{f\{e_u(\boldsymbol{X}) \mid U = u\}}{f\{e_u(\boldsymbol{X})\}}E\{Y(0) \mid e_u(\boldsymbol{X})\}\right]. \quad \text{(A.8)}$$

Therefore, formulas (A.7) and (A.8) are tied together by formula (A.6) of importance sampling, yielding $E\{Y(0) \mid U = u\} = E\{w_u(\boldsymbol{X})Y \mid Z = 0\}$. This completes the proof. □

*Proof of Corollary 1.* We can treat $h(\boldsymbol{X})$ as a new "outcome," on which the treatment has zero PCEs. The conclusions follow directly from Proposition 1. □

In order to prove Proposition 3, we need an additional lemma.

**Lemma A.4.** We can also represent the sensitivity parameter $\varepsilon$ as

$$\varepsilon = \frac{E\{Y(0) \mid U = s\bar{s}, e_u(\boldsymbol{X})\}}{E\{Y(0) \mid U = \bar{s}\bar{s}, e_u(\boldsymbol{X})\}} \quad (u = s\bar{s}, \bar{s}\bar{s}). \tag{A.9}$$

*Proof of Lemma A.4.* We have

$$\begin{aligned} E\{Y(0) \mid U = s\bar{s}, e_u(\boldsymbol{X})\} &= E[E\{Y(0) \mid U = s\bar{s}, e_u(\boldsymbol{X}), \boldsymbol{X}\} \mid U = s\bar{s}, e_u(\boldsymbol{X})] \\ &= E[E\{Y(0) \mid U = s\bar{s}, \boldsymbol{X}\} \mid U = s\bar{s}, e_u(\boldsymbol{X})] \\ &= \varepsilon E[E\{Y(0) \mid U = \bar{s}\bar{s}, \boldsymbol{X}\} \mid U = s\bar{s}, e_u(\boldsymbol{X})] \\ &= \varepsilon E[E\{Y(0) \mid U = \bar{s}\bar{s}, \boldsymbol{X}\} \mid U = \bar{s}\bar{s}, e_u(\boldsymbol{X})], \end{aligned}$$

where the last two lines follow from the definition of $\varepsilon$ and Lemma A.2. We also have

$$\begin{aligned} E\{Y(0) \mid U = \bar{s}\bar{s}, e_u(\boldsymbol{X})\} &= E[E\{Y(0) \mid U = \bar{s}\bar{s}, e_u(\boldsymbol{X}), \boldsymbol{X}\} \mid U = \bar{s}\bar{s}, e_u(\boldsymbol{X})] \\ &= E[E\{Y(0) \mid U = \bar{s}\bar{s}, \boldsymbol{X}\} \mid U = \bar{s}\bar{s}, e_u(\boldsymbol{X})]. \end{aligned}$$

Therefore, we can represent the sensitivity parameter $\varepsilon$ as in formula (A.9). □

*Proof of Proposition 3.* We prove only the conclusion about $ACE_{s\bar{s}}$; conclusions for the others are analogous. We first observe that

$$\begin{aligned} & E\{Y(0) \mid e_{s\bar{s}}(\boldsymbol{X})\} \\ = \; & E\{Y(0) \mid e_{s\bar{s}}(\boldsymbol{X}), U = s\bar{s}\} \Pr\{U = s\bar{s} \mid e_{s\bar{s}}(\boldsymbol{X})\} + E\{Y(0) \mid e_{s\bar{s}}(\boldsymbol{X}), U = \bar{s}\bar{s}\} \Pr\{U = \bar{s}\bar{s} \mid e_{s\bar{s}}(\boldsymbol{X})\} \\ = \; & E\{Y(0) \mid e_{s\bar{s}}(\boldsymbol{X}), U = s\bar{s}\} e_{s\bar{s}}(\boldsymbol{X}) + E\{Y(0) \mid e_{s\bar{s}}(\boldsymbol{X}), U = \bar{s}\bar{s}\} e_{\bar{s}\bar{s}}(\boldsymbol{X}), \end{aligned}$$

according to the proof of Lemma A.2. Further, Lemma A.4 reduces the above result to

$$E\{Y(0) \mid e_{s\bar{s}}(\boldsymbol{X})\} = \left\{ e_{s\bar{s}}(\boldsymbol{X}) + \frac{e_{\bar{s}\bar{s}}(\boldsymbol{X})}{\varepsilon} \right\} E\{Y(0) \mid e_{s\bar{s}}(\boldsymbol{X}), U = s\bar{s}\},$$

which further gives

$$E\{Y(0) \mid e_{s\bar{s}}(\boldsymbol{X}), U = s\bar{s}\} = \left\{e_{s\bar{s}}(\boldsymbol{X}) + \frac{e_{\bar{s}\bar{s}}(\boldsymbol{X})}{\varepsilon}\right\}^{-1} E\{Y(0) \mid e_{s\bar{s}}(\boldsymbol{X})\} = \frac{\varepsilon}{\varepsilon e_{s\bar{s}}(\boldsymbol{X}) + e_{\bar{s}\bar{s}}(\boldsymbol{X})} E\{Y(0) \mid e_{s\bar{s}}(\boldsymbol{X})\}.$$

Therefore, on the one hand we have

$$E\{Y(0) \mid U = s\bar{s}\} = E[E\{Y(0) \mid U = s\bar{s}, e_{s\bar{s}}(\boldsymbol{X})\} \mid U = s\bar{s}] = E\left[\frac{\varepsilon}{\varepsilon e_{s\bar{s}}(\boldsymbol{X}) + e_{\bar{s}\bar{s}}(\boldsymbol{X})} E\{Y(0) \mid e_{s\bar{s}}(\boldsymbol{X})\} \mid U = s\bar{s}\right].$$

On the other hand, the weighted mean can be represented as

$$E\{w^{\varepsilon}_{s\bar{s}}(\boldsymbol{X})Y \mid Z = 0\}/\pi_{s\bar{s}} = E\left[\frac{e_{s\bar{s}}(\boldsymbol{X})}{\pi_{s\bar{s}}} \frac{\varepsilon}{\varepsilon e_{s\bar{s}}(\boldsymbol{X}) + e_{\bar{s}\bar{s}}(\boldsymbol{X})} E\{Y(0) \mid e_{s\bar{s}}(\boldsymbol{X})\}\right].$$

According to the proof of Proposition 1, $e_{s\bar{s}}(\boldsymbol{X})/\pi_{s\bar{s}} = f\{e_u(\boldsymbol{X}) \mid U = u\}/f\{e_u(\boldsymbol{X})\}$ is the importance weight, implying that $E\{Y(0) \mid U = s\bar{s}\} = E\{w^{\varepsilon}_{s\bar{s}}(\boldsymbol{X})Y \mid Z = 0\}/\pi_{s\bar{s}}$. □

## A.11. Proofs of the Propositions Under Monotonicity

To prove Proposition 2, we introduce some lemmas, which rely on the following definitions.

$$e_{1,s\bar{s}}(\boldsymbol{X}) = \frac{e_{s\bar{s}}(\boldsymbol{X})}{e_{s\bar{s}}(\boldsymbol{X}) + e_{ss}(\boldsymbol{X})}, \qquad e_{1,ss}(\boldsymbol{X}) = \frac{e_{ss}(\boldsymbol{X})}{e_{s\bar{s}}(\boldsymbol{X}) + e_{ss}(\boldsymbol{X})},$$
$$e_{0,s\bar{s}}(\boldsymbol{X}) = \frac{e_{s\bar{s}}(\boldsymbol{X})}{e_{s\bar{s}}(\boldsymbol{X}) + e_{\bar{s}\bar{s}}(\boldsymbol{X})}, \qquad e_{0,\bar{s}\bar{s}}(\boldsymbol{X}) = \frac{e_{\bar{s}\bar{s}}(\boldsymbol{X})}{e_{s\bar{s}}(\boldsymbol{X}) + e_{\bar{s}\bar{s}}(\boldsymbol{X})}.$$

**Lemma A.5.** Under Monotonicity, $\Pr(U = u \mid Z = 1, S = 1, \boldsymbol{X}) = e_{1,u}(\boldsymbol{X})$ for $u = s\bar{s}$ and $ss$, and $\Pr(U = u \mid Z = 0, S = 0, \boldsymbol{X}) = e_{0,u}(\boldsymbol{X})$ for $u = s\bar{s}$ and $\bar{s}\bar{s}$.

*Proof of Lemma A.5.* Without essential loss of generality, we show only the case with $u = s\bar{s}$ under the treatment, and other cases are analogous. We have

$$\begin{aligned}\Pr(U = s\bar{s} \mid Z = 1, S = 1, \boldsymbol{X}) &= \frac{\Pr(Z = 1, S = 1, U = s\bar{s} \mid \boldsymbol{X})}{\Pr(Z = 1, S = 1 \mid \boldsymbol{X})} \\ &= \frac{\Pr(Z = 1, U = s\bar{s} \mid \boldsymbol{X})}{\Pr(Z = 1, U = s\bar{s} \mid \boldsymbol{X}) + \Pr(Z = 1, U = ss \mid \boldsymbol{X})}.\end{aligned}$$

Randomization implies $Z \perp\!\!\!\perp (U, \boldsymbol{X})$, and therefore

$$\Pr(U = s\bar{s} \mid Z = 1, S = 1, \boldsymbol{X}) = \frac{\Pr(U = s\bar{s} \mid \boldsymbol{X})}{\Pr(U = s\bar{s} \mid \boldsymbol{X}) + \Pr(U = ss \mid \boldsymbol{X})} = \frac{e_{s\bar{s}}(\boldsymbol{X})}{e_{s\bar{s}}(\boldsymbol{X}) + e_{ss}(\boldsymbol{X})} = e_{1,s\bar{s}}(\boldsymbol{X}).$$

□

**Lemma A.6.** Under Monotonicity, $\boldsymbol{X} \perp\!\!\!\perp U \mid \{Z = 1, S = 1, e_{1,u}(\boldsymbol{X})\}$ for $u = s\bar{s}$ and $ss$, and $\boldsymbol{X} \perp\!\!\!\perp U \mid \{Z = 0, S = 0, e_{0,u}(\boldsymbol{X})\}$ for $u = s\bar{s}$ and $\bar{s}\bar{s}$.

*Proof of Lemma A.6.* We prove only the case with $u = s\bar{s}$ under the treatment. Conditional on $(Z = 1, S = 1)$, the principal strata can take only two values $s\bar{s}$ and $ss$, and the proof here follows similar arguments as the proof of Lemma A.2.

First, $\Pr\{U = s\bar{s} \mid \boldsymbol{X}, Z = 1, S = 1, e_{1,s\bar{s}}(\boldsymbol{X})\} = \Pr(U = s\bar{s} \mid \boldsymbol{X}, Z = 1, S = 1) = e_{1,s\bar{s}}(\boldsymbol{X})$. Second, by LIE we have

$$\begin{aligned}
\Pr\{U = s\bar{s} \mid Z = 1, S = 1, e_{1,s\bar{s}}(\boldsymbol{X})\} &= E[\Pr\{U = s\bar{s} \mid Z = 1, S = 1, e_{1,s\bar{s}}(\boldsymbol{X}), \boldsymbol{X}\} \mid Z = 1, S = 1, e_{1,s\bar{s}}(\boldsymbol{X})] \\
&= E\{e_{1,s\bar{s}}(\boldsymbol{X}) \mid Z = 1, S = 1, e_{1,s\bar{s}}(\boldsymbol{X})\} = e_{1,s\bar{s}}(\boldsymbol{X}).
\end{aligned}$$

Therefore, $\Pr\{U = s\bar{s} \mid \boldsymbol{X}, Z = 1, S = 1, e_{1,s\bar{s}}(\boldsymbol{X})\} = \Pr\{U = s\bar{s} \mid Z = 1, S = 1, e_{1,s\bar{s}}(\boldsymbol{X})\}$, and the conditional independence $\boldsymbol{X} \perp\!\!\!\perp U \mid \{Z = 1, S = 1, e_{1,u}(\boldsymbol{X})\}$ follows. □

**Lemma A.7.** Under Monotonicity, GPI implies that

$$Y(1) \perp\!\!\!\perp U \mid (Z = 1, S = 1, \boldsymbol{X}), \quad Y(0) \perp\!\!\!\perp U \mid (Z = 0, S = 0, \boldsymbol{X}).$$

*Proof of Lemma A.7.* We prove only the first part; the second part is analogous.

GPI implies $Y(1) \perp\!\!\!\perp U \mid \boldsymbol{X}$, and therefore $Y(1) \perp\!\!\!\perp \{U, \mathbf{1}_{(U = s\bar{s} \text{ or } ss)}\} \mid \boldsymbol{X}$. Furthermore, we have $Y(1) \perp\!\!\!\perp U \mid \{\mathbf{1}_{(U = s\bar{s} \text{ or } ss)}, \boldsymbol{X}\}$. Because Randomization ensures that $Z$ is independent of all the potential outcomes and covariates, we have $Y(1) \perp\!\!\!\perp U \mid \{Z, \mathbf{1}_{(U = s\bar{s} \text{ or } ss)}, \boldsymbol{X}\}$, which further implies $Y(1) \perp\!\!\!\perp U \mid \{Z = 1, \mathbf{1}_{(U = s\bar{s} \text{ or } ss)} = 1, \boldsymbol{X}\}$ or equivalently $Y(1) \perp\!\!\!\perp U \mid (Z = 1, S = 1, \boldsymbol{X})$. □

**Lemma A.8.** Under Monotonicity and GPI, we have $Y(1) \perp\!\!\!\perp U \mid \{Z = 1, S = 1, e_{1,u}(\boldsymbol{X})\}$ for $u = s\bar{s}$ and $ss$, and $Y(0) \perp\!\!\!\perp U \mid \{Z = 0, S = 0, e_{0,u}(\boldsymbol{X})\}$ for $u = s\bar{s}$ and $\bar{s}\bar{s}$.

*Proof of Lemma A.8.* With Lemmas A.5–A.7, the proof follows from the same logic as that of Lemma A.3. □

*Proof of Proposition 2.* It is straightforward to obtain

$$E\{Y(1) \mid U = \bar{s}\bar{s}\} = E(Y \mid Z = 1, S = 0), \quad E\{Y(0) \mid U = ss\} = E(Y \mid Z = 0, S = 1).$$

Without loss of generality, we show only $E\{Y(1) \mid U = s\bar{s}\} = E\{w_{1,s\bar{s}}(\boldsymbol{X})Y \mid Z = 1, S = 1\}$.

First, by Randomization we have

$$E\{Y(1) \mid U = s\bar{s}\} = E\{Y(1) \mid Z = 1, U = s\bar{s}\} = E\{Y(1) \mid Z = 1, S = 1, U = s\bar{s}\}.$$

By LIE and Lemma A.8, we have

$$\begin{aligned} E\{Y(1) \mid U = s\bar{s}\} &= E[E\{Y(1) \mid Z = 1, S = 1, U = s\bar{s}, e_{1,s\bar{s}}(\boldsymbol{X})\} \mid Z = 1, S = 1, U = s\bar{s}] \\ &= E[E\{Y(1) \mid Z = 1, S = 1, e_{1,s\bar{s}}(\boldsymbol{X})\} \mid Z = 1, S = 1, U = s\bar{s}]. \end{aligned}$$

Second, we have

$$E\{w_{1,s\bar{s}}(\boldsymbol{X})Y \mid Z = 1, S = 1\} = E[w_{1,s\bar{s}}(\boldsymbol{X})E\{Y(1) \mid Z = 1, S = 1, e_{1,s\bar{s}}(\boldsymbol{X})\} \mid Z = 1, S = 1].$$

Also we have

$$f\{e_{1,s\bar{s}}(\boldsymbol{X}) \mid U = s\bar{s}, Z = 1, S = 1\} = \frac{f\{e_{1,s\bar{s}}(\boldsymbol{X}) \mid Z = 1, S = 1\} \Pr\{U = s\bar{s} \mid Z = 1, S = 1, e_{1,s\bar{s}}(\boldsymbol{X})\}}{\Pr(U = s\bar{s} \mid Z = 1, S = 1)}.$$

From the proof of Lemma A.6, we have $\Pr\{U = s\bar{s} \mid Z = 1, S = 1, e_{1,s\bar{s}}(\boldsymbol{X})\} = e_{1,s\bar{s}}(\boldsymbol{X})$. Simple algebra gives $\Pr(U = s\bar{s} \mid Z = 1, S = 1) = \pi_{s\bar{s}}/(\pi_{s\bar{s}} + \pi_{ss})$. Consequently, the weight satisfies

$$w_{1,s\bar{s}}(\boldsymbol{X}) = \frac{e_{1,s\bar{s}}(\boldsymbol{X})}{\pi_{s\bar{s}}/(\pi_{s\bar{s}} + \pi_{ss})} = \frac{\Pr\{U = s\bar{s} \mid Z = 1, S = 1, e_{1,s\bar{s}}(\boldsymbol{X})\}}{\Pr(U = s\bar{s} \mid Z = 1, S = 1)} = \frac{f\{e_{1,s\bar{s}}(\boldsymbol{X}) \mid U = s\bar{s}, Z = 1, S = 1\}}{f\{e_{1,s\bar{s}}(\boldsymbol{X}) \mid Z = 1, S = 1\}},$$

which implies $E\{Y(1) \mid U = s\bar{s}\} = E\{w_{1,s\bar{s}}(\boldsymbol{X})Y \mid Z = 1, S = 1\}$ according to Lemma A.1. □

*Proof of Corollary 2.* The theorem follows from Proposition 2 and zero PCEs on $h(\boldsymbol{X})$. □

In order to prove Proposition 4, we need an additional lemma.

**Lemma A.9.** We can also represent the sensitivity parameters as

$$\varepsilon_1 = \frac{E\{Y(1) \mid U = s\bar{s}, e_{1,u}(\boldsymbol{X})\}}{E\{Y(1) \mid U = \bar{s}\bar{s}, e_{1,u}(\boldsymbol{X})\}} \quad (u = s\bar{s}, ss); \qquad \varepsilon_0 = \frac{E\{Y(0) \mid U = s\bar{s}, e_{0,u}(\boldsymbol{X})\}}{E\{Y(0) \mid U = \bar{s}\bar{s}, e_{0,u}(\boldsymbol{X})\}} \quad (u = s\bar{s}, \bar{s}\bar{s}).$$

*Proof of Lemma A.9.* It follows from Lemma A.6 and the method in the proof of Lemma A.4. □

*Proof of Proposition 4.* We prove only that $E\{Y(1) \mid U = s\bar{s}\} = E\{w^{\varepsilon_1}_{1,s\bar{s}}(\boldsymbol{X})Y \mid Z = 1, S = 1\}$; other conditional expectations of the potential outcomes are analogous.

Randomization and LIE allow us to obtain that

$$\begin{aligned} E\{Y(1) \mid U = s\bar{s}\} &= E\{Y(1) \mid Z = 1, U = s\bar{s}, S = 1\} \\ &= E[E\{Y(1) \mid Z = 1, U = s\bar{s}, S = 1, e_{1,s\bar{s}}(\boldsymbol{X})\} \mid Z = 1, U = s\bar{s}, S = 1]. \end{aligned}$$

From the proofs of Lemmas A.6–A.9, we also have

$$\begin{aligned} & E\{Y(1) \mid Z = 1, S = 1, e_{1,s\bar{s}}(\boldsymbol{X})\} \\ = \; & E\{Y(1) \mid Z = 1, U = s\bar{s}, S = 1, e_{1,s\bar{s}}(\boldsymbol{X})\} \Pr\{U = s\bar{s} \mid Z = 1, S = 1, e_{1,s\bar{s}}(\boldsymbol{X})\} \\ & + E\{Y(1) \mid Z = 1, U = \bar{s}\bar{s}, S = 1, e_{1,s\bar{s}}(\boldsymbol{X})\} \Pr\{U = \bar{s}\bar{s} \mid Z = 1, S = 1, e_{1,s\bar{s}}(\boldsymbol{X})\} \\ = \; & E\{Y(1) \mid Z = 1, U = s\bar{s}, S = 1, e_{1,s\bar{s}}(\boldsymbol{X})\} e_{1,s\bar{s}}(\boldsymbol{X}) + E\{Y(1) \mid Z = 1, U = \bar{s}\bar{s}, S = 1, e_{1,s\bar{s}}(\boldsymbol{X})\} e_{1,\bar{s}\bar{s}}(\boldsymbol{X}) \\ = \; & \{e_{1,s\bar{s}}(\boldsymbol{X}) + e_{1,\bar{s}\bar{s}}(\boldsymbol{X})/\varepsilon_1\} E\{Y(1) \mid Z = 1, U = s\bar{s}, S = 1, e_{1,s\bar{s}}(\boldsymbol{X})\}. \end{aligned}$$

Consequently, we obtain that

$$E\{Y(1) \mid Z = 1, U = s\bar{s}, S = 1, e_{1,s\bar{s}}(\boldsymbol{X})\} = \{e_{1,s\bar{s}}(\boldsymbol{X}) + e_{1,\bar{s}\bar{s}}(\boldsymbol{X})/\varepsilon_1\}^{-1} E\{Y(1) \mid Z = 1, S = 1, e_{1,s\bar{s}}(\boldsymbol{X})\},$$

implying that

$$E\{Y(1) \mid U = s\bar{s}\} = E\left[\frac{\varepsilon_1}{\varepsilon_1 e_{1,s\bar{s}}(\boldsymbol{X}) + e_{1,\bar{s}\bar{s}}(\boldsymbol{X})} E\{Y(1) \mid Z = 1, S = 1, e_{1,s\bar{s}}(\boldsymbol{X})\} \mid Z = 1, U = s\bar{s}, S = 1\right].$$

On the other hand, the weighted mean can be represented as

$$
\begin{aligned}
& E\{w_{1,s\bar{s}}^{\varepsilon_1}(\boldsymbol{X})Y \mid Z=1, S=1\} \\
= \; & E[w_{1,s\bar{s}}^{\varepsilon_1}(\boldsymbol{X}) E\{Y(1) \mid Z=1, S=1, e_{1,s\bar{s}}(\boldsymbol{X})\} \mid Z=1, S=1] \\
= \; & E\left[\frac{e_{1,s\bar{s}}(\boldsymbol{X})}{\pi_{s\bar{s}}/(\pi_{s\bar{s}}+\pi_{ss})} \frac{\varepsilon_1}{\varepsilon_1 e_{1,s\bar{s}}(\boldsymbol{X}) + e_{1,\bar{s}\bar{s}}(\boldsymbol{X})} E\{Y(1) \mid Z=1, S=1, e_{1,s\bar{s}}(\boldsymbol{X})\} \mid Z=1, S=1\right].
\end{aligned}
$$

The proof of Proposition 2 shows that $e_{1,s\bar{s}}(\boldsymbol{X})/\{\pi_{s\bar{s}}/(\pi_{s\bar{s}}+\pi_{ss})\}$ is exactly the importance weight $f\{e_{1,s\bar{s}}(\boldsymbol{X}) \mid U = s\bar{s}, Z=1, S=1\}/f\{e_{1,s\bar{s}}(\boldsymbol{X}) \mid Z=1, S=1\}$, and the conclusion follows from Lemma A.1 and the above expressions of $E\{Y(1) \mid U = s\bar{s}\}$ and $E\{w_{1,s\bar{s}}^{\varepsilon_1}(\boldsymbol{X})Y \mid Z=1, S=1\}$. □

## A.12. Proofs of the Results Without Monotonicity

*Proof of Proposition 5.* The observed data impose the following restrictions:

$$
\begin{cases}
\pi_{ss} + \pi_{s\bar{s}} &= p_1, \\
\pi_{ss} + \pi_{\bar{s}s} &= p_0, \\
\pi_{ss} + \pi_{\bar{s}s} + \pi_{s\bar{s}} + \pi_{\bar{s}\bar{s}} &= 1, \\
\pi_{\bar{s}s} - \xi\pi_{s\bar{s}} &= 0.
\end{cases}
\Longrightarrow
\begin{cases}
0 \le \pi_{s\bar{s}} = (p_1 - p_0)/(1-\xi) \le 1, \\
0 \le \pi_{ss} = p_1 - (p_1 - p_0)/(1-\xi) \le 1, \\
0 \le \pi_{\bar{s}s} = \xi(p_1 - p_0)/(1-\xi) \le 1, \\
0 \le \pi_{\bar{s}\bar{s}} = (1-p_0) - (p_1 - p_0)/(1-\xi) \le 1,
\end{cases}
$$

which further imply that $0 \le \xi \le 1 - (p_1 - p_0)/\min(p_1, 1-p_0)$. □

*Proof of Proposition 6.* The same reasoning of the proof of Proposition 2 applies here. □

## A.13. EM Algorithms for Estimating Principal Scores

### A.13.1. With Monotonicity

Because $U$ takes three values, we can model $\Pr(U \mid \boldsymbol{X})$ as a three-level Multinomial Logistic model:

$$
\Pr(U = u \mid \boldsymbol{X}) = \frac{\exp(\boldsymbol{\theta}_u^\top \boldsymbol{X})}{\sum_{u'} \exp(\boldsymbol{\theta}_{u'}^\top \boldsymbol{X})}, \quad (u = s\bar{s}, ss, \bar{s}\bar{s})
$$

where $\boldsymbol{\theta}_{s\bar{s}} = \mathbf{0}$ for identification. Although we cannot fully observe $U$, we can and use the EM algorithm to find the MLEs by treating $U$ as missing data.

In the E-step, we first calculate the conditional probabilities of latent strata given the data $(Z_i = 1, S_i = 1, \boldsymbol{X}_i)$ and the parameters $(\boldsymbol{\theta}^k_{s\bar{s}}, \boldsymbol{\theta}^k_{\bar{s}\bar{s}})$: if $(Z_i = 1, S_i = 1)$, then

$$\Pr(U_i = s\bar{s} \mid -) = \frac{1}{1 + \exp(\boldsymbol{\theta}^{k\top}_{ss} \boldsymbol{X}_i)}, \quad \Pr(U_i = ss \mid -) = \frac{1}{1 + \exp(-\boldsymbol{\theta}^{k\top}_{ss} \boldsymbol{X}_i)};$$

if $(Z_i = 1, S_i = 1)$, then $U_i = \bar{s}\bar{s}$; if $(Z_i = 0, S_i = 1)$, then $U_i = ss$; if $(Z_i = 0, S_i = 0)$, then

$$\Pr(U_i = s\bar{s} \mid -) = \frac{1}{1 + \exp(\boldsymbol{\theta}^{k\top}_{\bar{s}\bar{s}} \boldsymbol{X}_i)}, \quad \Pr(U_i = \bar{s}\bar{s} \mid -) = \frac{1}{1 + \exp(-\boldsymbol{\theta}^{k\top}_{\bar{s}\bar{s}} \boldsymbol{X}_i)}.$$

We then create weighted samples: for each $i$ with $(Z_i = 1, S_i = 1)$, we create two observations $(U_i = s\bar{s}, \boldsymbol{X}_i)$ and $(U_i = ss, \boldsymbol{X}_i)$ with weights $w^k_i = \Pr(U_i = s\bar{s} \mid -)$ and $w^k_i = \Pr(U_i = ss \mid -)$, respectively; for each $i$ with $(Z_i = 1, S_i = 0)$, we create one observation $(U_i = \bar{s}\bar{s}, \boldsymbol{X}_i)$ with weight $w^k_i = 1$; for each $i$ with $(Z_i = 0, S_i = 1)$, we create one observation $(U_i = ss, \boldsymbol{X}_i)$ with weight $w^k_i = 1$; for each $i$ with $(Z_i = 0, S_i = 0)$, we create two observations $(U_i = s\bar{s}, \boldsymbol{X}_i)$ and $(U_i = \bar{s}\bar{s}, \boldsymbol{X}_i)$ with weights $w^k_i = \Pr(U_i = s\bar{s} \mid -)$ and $w^k_i = \Pr(U_i = ss \mid -)$, respectively.

In the M-step, we fit a Multinomial Logistic model $\Pr(U \mid \boldsymbol{X})$ based on the weight samples created above, and update the parameters to be $(\boldsymbol{\theta}^{k+1}_{s\bar{s}}, \boldsymbol{\theta}^{k+1}_{\bar{s}\bar{s}})$.

### A.13.2. Without Monotonicity

We define $V_i = U_i$ if $U_i = ss$ or $\bar{s}\bar{s}$, and $V_i = s\&\bar{s}$ if $U_i = s\bar{s}$ or $\bar{s}s$. First, we model $\Pr(V \mid \boldsymbol{X})$ as a three-level Multinomial Logistic regression:

$$\Pr(V = v \mid \boldsymbol{X}) = \frac{\exp(\boldsymbol{\theta}^\top_v \boldsymbol{X})}{\sum_{v'} \exp(\boldsymbol{\theta}^\top_{v'} \boldsymbol{X})}, \quad (v = s\&\bar{s}, ss, \bar{s}\bar{s})$$

where $\boldsymbol{\theta}_{s\&\bar{s}} = \mathbf{0}$. Second, we partition the category of $V$, $s\&\bar{s}$, into two sub-categories of $U$, $s\bar{s}$ and $\bar{s}s$, with probabilities $\Pr(U = s\bar{s} \mid V = s\&\bar{s}, \boldsymbol{X}) = 1/(1 + \xi)$ and $\Pr(U = \bar{s}s \mid V = s\&\bar{s}, \boldsymbol{X}) = \xi/(1 + \xi)$. We use the EM algorithm to find the MLEs by treating $U$ as missing data.

In the E-step, we first calculate the conditional probabilities of latent strata given the data $(Z_i, S_i, \boldsymbol{X}_i)$ and the parameters $(\boldsymbol{\theta}^k_{ss}, \boldsymbol{\theta}^k_{\bar{s}\bar{s}})$: if $(Z_i = 1, S_i = 1)$, then

$$\Pr(U_i = s\bar{s} \mid -) = \frac{1}{1 + (1 + \xi)\exp(\boldsymbol{\theta}^{k\top}_{ss} \boldsymbol{X}_i)}, \quad \Pr(U_i = ss \mid -) = \frac{1}{1 + \exp(-\boldsymbol{\theta}^{k\top}_{ss} \boldsymbol{X}_i)/(1 + \xi)};$$

if $(Z_i = 1, S_i = 0)$, then

$$\Pr(U_i = \bar{s}\bar{s} \mid -) = \frac{1}{1 + \xi \exp(-\boldsymbol{\theta}_{\bar{s}\bar{s}}^{k\top} \boldsymbol{X}_i)/(1+\xi)}, \quad \Pr(U_i = \bar{s}s \mid -) = \frac{1}{1 + (1+\xi)\exp(\boldsymbol{\theta}_{\bar{s}\bar{s}}^{k\top} \boldsymbol{X}_i)/\xi};$$

if $(Z_i = 0, S_i = 1)$, then

$$\Pr(U_i = ss \mid -) = \frac{1}{1 + \xi \exp(-\boldsymbol{\theta}_{ss}^{k\top} \boldsymbol{X}_i)/(1+\xi)}, \quad \Pr(U_i = \bar{s}s \mid -) = \frac{1}{1 + (1+\xi)\exp(\boldsymbol{\theta}_{ss}^{k\top} \boldsymbol{X}_i)/\xi};$$

if $(Z_i = 0, S_i = 0)$, then

$$\Pr(U_i = s\bar{s} \mid -) = \frac{1}{1 + (1+\xi)\exp(\boldsymbol{\theta}_{\bar{s}\bar{s}}^{k\top} \boldsymbol{X}_i)}, \quad \Pr(U_i = \bar{s}\bar{s} \mid -) = \frac{1}{1 + \exp(-\boldsymbol{\theta}_{\bar{s}\bar{s}}^{k\top} \boldsymbol{X}_i)/(1+\xi)}.$$

We then create a set of weighted samples: for each $i$ with $(Z_i = 1, S_i = 1)$, we create two observations $(V_i = s\&\bar{s}, \boldsymbol{X}_i)$ and $(V_i = ss, \boldsymbol{X}_i)$ with weights $w_i^k = \Pr(U_i = s\bar{s} \mid -)$ and $w_i^k = \Pr(U_i = ss \mid -)$, respectively; for each $i$ with $(Z_i = 1, S_i = 0)$, we create two observations $(V_i = \bar{s}\bar{s}, \boldsymbol{X}_i)$ and $(V_i = s\&\bar{s}, \boldsymbol{X}_i)$ with weights $w_i^k = \Pr(U_i = \bar{s}\bar{s} \mid -)$ and $w_i^k = \Pr(U_i = \bar{s}s \mid -)$; for each $i$ with $(Z_i = 0, S_i = 1)$, we create two observations $(V_i = ss, \boldsymbol{X}_i)$ and $(V_i = s\&\bar{s}, \boldsymbol{X}_i)$ with weight $w_i^k = \Pr(U_i = ss \mid -)$ and $\Pr(U_i = \bar{s}s \mid -)$; for each $i$ with $(Z_i = 0, S_i = 0)$, we create two observations $(V_i = s\&\bar{s}, \boldsymbol{X}_i)$ and $(V_i = \bar{s}\bar{s}, \boldsymbol{X}_i)$ with weights $w_i^k = \Pr(U_i = s\bar{s} \mid -)$ and $w_i^k = \Pr(U_i = ss \mid -)$, respectively.

In the M-step, we fit a Multinomial Logistic model $\Pr(V \mid \boldsymbol{X})$ based on the weight samples created above, and update the parameters to be $(\boldsymbol{\theta}_{s\bar{s}}^{k+1}, \boldsymbol{\theta}_{\bar{s}\bar{s}}^{k+1})$.

## A.14. Explicit Forms of the Estimators

We present explicit forms of moment estimators by weighting and model-assisted estimators via covariate adjustment in Section 5. Let $N_1$ and $N_0$ be the treatment and control sample sizes.

### A.14.1. With Strong Monotonicity

With PI, by Proposition 1 the moment estimators for PCEs are

$$\begin{aligned}
\widehat{ACE}_{s\bar{s}} &= \frac{1}{n_{11}} \sum_{\{i:Z_i=1,S_i=1\}} Y_i - \frac{1}{N_0} \sum_{\{i:Z_i=0\}} \widehat{w}_{s\bar{s}}(\boldsymbol{X}_i) Y_i, \\
\widehat{ACE}_{\bar{s}\bar{s}} &= \frac{1}{n_{10}} \sum_{\{i:Z_i=1,S_i=0\}} Y_i - \frac{1}{N_0} \sum_{\{i:Z_i=0\}} \widehat{w}_{\bar{s}\bar{s}}(\boldsymbol{X}_i) Y_i,
\end{aligned}$$

and the model-assisted estimators for PCEs are

$$\begin{aligned}
\widehat{ACE}^{\text{adj}}_{s\bar{s}} &= \frac{1}{n_{11}} \sum_{\{i:Z_i=1,S_i=1\}} (Y_i - \boldsymbol{\beta}_{1,s\bar{s}}^\top \boldsymbol{X}_i) - \frac{1}{N_0} \sum_{\{i:Z_i=0\}} \widehat{w}_{s\bar{s}}(\boldsymbol{X}_i)(Y_i - \boldsymbol{\beta}_{0,s\bar{s}}^\top \boldsymbol{X}) \\
&\quad + \frac{1}{n_{11}+N_0} (\boldsymbol{\beta}_{1,s\bar{s}} - \boldsymbol{\beta}_{0,s\bar{s}})^\top \left( \sum_{\{i:Z_i=1,S_i=1\}} \boldsymbol{X}_i + \sum_{\{i:Z_i=0\}} \widehat{w}_{s\bar{s}}(\boldsymbol{X}_i) \boldsymbol{X}_i \right), \\
\widehat{ACE}^{\text{adj}}_{\bar{s}\bar{s}} &= \frac{1}{n_{10}} \sum_{\{i:Z_i=1,S_i=0\}} (Y_i - \boldsymbol{\beta}_{1,\bar{s}\bar{s}}^\top \boldsymbol{X}_i) - \frac{1}{N_0} \sum_{\{i:Z_i=0\}} \widehat{w}_{\bar{s}\bar{s}}(\boldsymbol{X}_i)(Y_i - \boldsymbol{\beta}_{0,\bar{s}\bar{s}}^\top \boldsymbol{X}) \\
&\quad + \frac{1}{n_{10}+N_0} (\boldsymbol{\beta}_{1,\bar{s}\bar{s}} - \boldsymbol{\beta}_{0,\bar{s}\bar{s}})^\top \left( \sum_{\{i:Z_i=1,S_i=0\}} \boldsymbol{X}_i + \sum_{\{i:Z_i=0\}} \widehat{w}_{\bar{s}\bar{s}}(\boldsymbol{X}_i) \boldsymbol{X}_i \right).
\end{aligned}$$

We choose $\boldsymbol{\beta}_{1,s\bar{s}}$ as the linear regression coefficients of $Y_i$ on $\boldsymbol{X}_i$ using samples with $(Z_i = 1, S_i = 1)$, $\boldsymbol{\beta}_{0,s\bar{s}}$ as the weighted linear regression coefficients of $Y_i$ on $\boldsymbol{X}_i$ using samples with $Z_i = 0$ and weights $w_{s\bar{s}}(\boldsymbol{X}_i)$, $\boldsymbol{\beta}_{1,\bar{s}\bar{s}}$ as the linear regression coefficients of $Y_i$ on $\boldsymbol{X}_i$ using samples with $(Z_i = 1, S_i = 0)$, and $\boldsymbol{\beta}_{0,\bar{s}\bar{s}}$ as the weighted linear regression coefficients of $Y_i$ on $\boldsymbol{X}_i$ using samples with $Z_i = 0$ and weights $w_{\bar{s}\bar{s}}(\boldsymbol{X}_i)$.

Without PI, we need only to change the estimates of the weights for a fixed sensitivity parameter $\varepsilon$ as in Proposition 3, and obtain estimators of the same forms as above.

### A.14.2. With Monotonicity

With GPI, by Proposition 2 the moment estimators for PCEs are

$$\begin{aligned}
\widehat{ACE}_{s\bar{s}} &= \frac{1}{n_{11}} \sum_{\{i:Z_i=1,S_i=1\}} \widehat{w}_{1,s\bar{s}}(\boldsymbol{X}_i) Y_i - \frac{1}{n_{00}} \sum_{\{i:Z_i=0,S_i=0\}} \widehat{w}_{0,s\bar{s}}(\boldsymbol{X}_i) Y_i, \\
\widehat{ACE}_{\bar{s}\bar{s}} &= \frac{1}{n_{10}} \sum_{\{i:Z_i=1,S_i=0\}} Y_i - \frac{1}{n_{00}} \sum_{\{i:Z_i=0,S_i=0\}} \widehat{w}_{0,\bar{s}\bar{s}}(\boldsymbol{X}_i) Y_i, \\
\widehat{ACE}_{ss} &= \frac{1}{n_{11}} \sum_{\{i:Z_i=1,S_i=1\}} \widehat{w}_{1,ss}(\boldsymbol{X}_i) Y_i - \frac{1}{n_{01}} \sum_{\{i:Z_i=0,S_i=1\}} Y_i,
\end{aligned}$$

and the model-assisted estimators for PCEs are

$$\begin{aligned}
\widehat{ACE}^{\text{adj}}_{s\bar{s}} &= \frac{1}{n_{11}} \sum_{\{i:Z_i=1,S_i=1\}} \widehat{w}_{1,s\bar{s}}(\boldsymbol{X}_i)(Y_i - \boldsymbol{\beta}_{1,s\bar{s}}^\top \boldsymbol{X}_i) - \frac{1}{n_{00}} \sum_{\{i:Z_i=0,S_i=0\}} \widehat{w}_{0,s\bar{s}}(\boldsymbol{X}_i)(Y_i - \boldsymbol{\beta}_{0,s\bar{s}}^\top \boldsymbol{X}) \\
&\quad + \frac{1}{n_{11}+n_{00}} (\boldsymbol{\beta}_{1,s\bar{s}} - \boldsymbol{\beta}_{0,s\bar{s}})^\top \left( \sum_{\{i:Z_i=1,S_i=1\}} \widehat{w}_{1,s\bar{s}}(\boldsymbol{X}_i)\boldsymbol{X}_i + \sum_{\{i:Z_i=0,S_i=0\}} \widehat{w}_{0,s\bar{s}}(\boldsymbol{X}_i)\boldsymbol{X}_i \right), \\
\widehat{ACE}^{\text{adj}}_{\bar{s}\bar{s}} &= \frac{1}{n_{10}} \sum_{\{i:Z_i=1,S_i=0\}} (Y_i - \boldsymbol{\beta}_{1,\bar{s}\bar{s}}^\top \boldsymbol{X}_i) - \frac{1}{n_{00}} \sum_{\{i:Z_i=0,S_i=0\}} \widehat{w}_{0,\bar{s}\bar{s}}(\boldsymbol{X}_i)(Y_i - \boldsymbol{\beta}_{0,\bar{s}\bar{s}}^\top \boldsymbol{X}) \\
&\quad + \frac{1}{n_{10}+n_{00}} (\boldsymbol{\beta}_{1,\bar{s}\bar{s}} - \boldsymbol{\beta}_{0,\bar{s}\bar{s}})^\top \left( \sum_{\{i:Z_i=1,S_i=0\}} \boldsymbol{X}_i + \sum_{\{i:Z_i=0,S_i=0\}} \widehat{w}_{0,\bar{s}\bar{s}}(\boldsymbol{X}_i)\boldsymbol{X}_i \right), \\
\widehat{ACE}^{\text{adj}}_{ss} &= \frac{1}{n_{11}} \sum_{\{i:Z_i=1,S_i=1\}} \widehat{w}_{1,ss}(\boldsymbol{X}_i)(Y_i - \boldsymbol{\beta}_{1,ss}^\top \boldsymbol{X}_i) - \frac{1}{n_{01}} \sum_{\{i:Z_i=0,S_i=1\}} (Y_i - \boldsymbol{\beta}_{0,ss}^\top \boldsymbol{X}) \\
&\quad + \frac{1}{n_{10}+n_{01}} (\boldsymbol{\beta}_{1,ss} - \boldsymbol{\beta}_{0,ss})^\top \left( \sum_{\{i:Z_i=1,S_i=1\}} \widehat{w}_{1,ss}(\boldsymbol{X}_i)\boldsymbol{X}_i + \sum_{\{i:Z_i=0,S_i=1\}} \boldsymbol{X}_i \right).
\end{aligned}$$

We choose $\boldsymbol{\beta}_{1,u}$ for $u = s\bar{s}$ and $ss$ as the weighted linear regression coefficients of $Y_i$ on $\boldsymbol{X}_i$ using samples with $(Z_i = 1, S_i = 1)$ and weights $w_{1,u}(\boldsymbol{X}_i)$, $\boldsymbol{\beta}_{1,\bar{s}\bar{s}}$ as the linear regression coefficients of $Y_i$ on $\boldsymbol{X}_i$ using samples with $(Z_i = 1, S_i = 0)$, $\boldsymbol{\beta}_{0,ss}$ as the linear regression coefficients of $Y_i$ on $\boldsymbol{X}_i$ using samples with $(Z_i = 0, S_i = 1)$, and $\boldsymbol{\beta}_{0,u}$ for $u = s\bar{s}$ and $\bar{s}\bar{s}$ as the weighted linear regression coefficients of $Y_i$ on $\boldsymbol{X}_i$ using samples with $(Z_i = 0, S_i = 0)$ and weights $w_{0,u}(\boldsymbol{X}_i)$.

Without GPI, we need only to change the estimates of the weights for fixed sensitivity parameters $(\varepsilon_1, \varepsilon_0)$.

### A.14.3. Without Monotonicity

By Proposition 6 the estimators for PCEs are

$$\begin{aligned}
\widehat{ACE}_{s\bar{s}} &= \frac{1}{n_{11}} \sum_{\{i:Z_i=1,S_i=1\}} \widehat{w}_{1,s\bar{s}}(\boldsymbol{X}_i)Y_i - \frac{1}{n_{00}} \sum_{\{i:Z_i=0,S_i=0\}} \widehat{w}_{0,s\bar{s}}(\boldsymbol{X}_i)Y_i, \\
\widehat{ACE}_{\bar{s}\bar{s}} &= \frac{1}{n_{10}} \sum_{\{i:Z_i=1,S_i=0\}} \widehat{w}_{1,\bar{s}\bar{s}}(\boldsymbol{X}_i)Y_i - \frac{1}{n_{00}} \sum_{\{i:Z_i=0,S_i=0\}} \widehat{w}_{0,\bar{s}\bar{s}}(\boldsymbol{X}_i)Y_i, \\
\widehat{ACE}_{ss} &= \frac{1}{n_{11}} \sum_{\{i:Z_i=1,S_i=1\}} \widehat{w}_{1,ss}(\boldsymbol{X}_i)Y_i - \frac{1}{n_{01}} \sum_{\{i:Z_i=0,S_i=1\}} \widehat{w}_{0,ss}(\boldsymbol{X}_i)Y_i, \\
\widehat{ACE}_{\bar{s}s} &= \frac{1}{n_{10}} \sum_{\{i:Z_i=1,S_i=0\}} \widehat{w}_{1,\bar{s}s}(\boldsymbol{X}_i)Y_i - \frac{1}{n_{01}} \sum_{\{i:Z_i=0,S_i=1\}} \widehat{w}_{0,\bar{s}s}(\boldsymbol{X}_i)Y_i,
\end{aligned}$$

and the model-assisted estimators for PCEs are

$$\begin{aligned}
\widehat{ACE}^{\text{adj}}_{s\bar{s}} &= \frac{1}{n_{11}} \sum_{\{i:Z_i=1,S_i=1\}} \widehat{w}_{1,s\bar{s}}(\boldsymbol{X}_i)(Y_i - \boldsymbol{\beta}_{1,s\bar{s}}^\top \boldsymbol{X}_i) - \frac{1}{n_{00}} \sum_{\{i:Z_i=0,S_i=0\}} \widehat{w}_{0,s\bar{s}}(\boldsymbol{X}_i)(Y_i - \boldsymbol{\beta}_{0,s\bar{s}}^\top \boldsymbol{X}) \\
&\quad + \frac{1}{n_{11}+n_{00}} (\boldsymbol{\beta}_{1,s\bar{s}} - \boldsymbol{\beta}_{0,s\bar{s}})^\top \left( \sum_{\{i:Z_i=1,S_i=1\}} \widehat{w}_{1,s\bar{s}}(\boldsymbol{X}_i)\boldsymbol{X}_i + \sum_{\{i:Z_i=0,S_i=0\}} \widehat{w}_{0,s\bar{s}}(\boldsymbol{X}_i)\boldsymbol{X}_i \right), \\
\widehat{ACE}^{\text{adj}}_{\bar{s}\bar{s}} &= \frac{1}{n_{10}} \sum_{\{i:Z_i=1,S_i=0\}} \widehat{w}_{1,\bar{s}\bar{s}}(\boldsymbol{X}_i)(Y_i - \boldsymbol{\beta}_{1,s\bar{s}}^\top \boldsymbol{X}_i) - \frac{1}{n_{00}} \sum_{\{i:Z_i=0,S_i=0\}} \widehat{w}_{0,\bar{s}\bar{s}}(\boldsymbol{X}_i)(Y_i - \boldsymbol{\beta}_{0,s\bar{s}}^\top \boldsymbol{X}) \\
&\quad + \frac{1}{n_{10}+n_{00}} (\boldsymbol{\beta}_{1,\bar{s}\bar{s}} - \boldsymbol{\beta}_{0,\bar{s}\bar{s}})^\top \left( \sum_{\{i:Z_i=1,S_i=0\}} \widehat{w}_{1,\bar{s}\bar{s}}(\boldsymbol{X}_i)\boldsymbol{X}_i + \sum_{\{i:Z_i=0,S_i=0\}} \widehat{w}_{0,\bar{s}\bar{s}}(\boldsymbol{X}_i)\boldsymbol{X}_i \right), \\
\widehat{ACE}^{\text{adj}}_{ss} &= \frac{1}{n_{11}} \sum_{\{i:Z_i=1,S_i=1\}} \widehat{w}_{1,ss}(\boldsymbol{X}_i)(Y_i - \boldsymbol{\beta}_{1,ss}^\top \boldsymbol{X}_i) - \frac{1}{n_{01}} \sum_{\{i:Z_i=0,S_i=1\}} \widehat{w}_{0,ss}(\boldsymbol{X}_i)(Y_i - \boldsymbol{\beta}_{0,ss}^\top \boldsymbol{X}) \\
&\quad + \frac{1}{n_{11}+n_{01}} (\boldsymbol{\beta}_{1,ss} - \boldsymbol{\beta}_{0,ss})^\top \left( \sum_{\{i:Z_i=1,S_i=1\}} \widehat{w}_{1,ss}(\boldsymbol{X}_i)\boldsymbol{X}_i + \sum_{\{i:Z_i=0,S_i=1\}} \widehat{w}_{0,ss}(\boldsymbol{X}_i)\boldsymbol{X}_i \right), \\
\widehat{ACE}^{\text{adj}}_{\bar{s}s} &= \frac{1}{n_{10}} \sum_{\{i:Z_i=1,S_i=0\}} \widehat{w}_{1,\bar{s}s}(\boldsymbol{X}_i)(Y_i - \boldsymbol{\beta}_{1,\bar{s}s}^\top \boldsymbol{X}_i) - \frac{1}{n_{01}} \sum_{\{i:Z_i=0,S_i=1\}} \widehat{w}_{0,\bar{s}s}(\boldsymbol{X}_i)(Y_i - \boldsymbol{\beta}_{0,\bar{s}s}^\top \boldsymbol{X}) \\
&\quad + \frac{1}{n_{10}+n_{01}} (\boldsymbol{\beta}_{1,\bar{s}s} - \boldsymbol{\beta}_{0,\bar{s}s})^\top \left( \sum_{\{i:Z_i=1,S_i=0\}} \widehat{w}_{1,\bar{s}s}(\boldsymbol{X}_i)\boldsymbol{X}_i + \sum_{\{i:Z_i=0,S_i=1\}} \widehat{w}_{0,\bar{s}s}(\boldsymbol{X}_i)\boldsymbol{X}_i \right).
\end{aligned}$$

We choose $\boldsymbol{\beta}_{1,u}$ for $u = s\bar{s}$ and $ss$ as the weighted linear regression coefficients of $Y_i$ on $\boldsymbol{X}_i$ using samples with $(Z_i = 1, S_i = 1)$ and weights $w_{1,u}(\boldsymbol{X}_i)$, $\boldsymbol{\beta}_{1,u}$ for $u = s\bar{s}$ and $\bar{s}\bar{s}$ as the weighted linear regression coefficients of $Y_i$ on $\boldsymbol{X}_i$ using samples with $(Z_i = 1, S_i = 0)$ and weights $w_{1,u}(\boldsymbol{X}_i)$, $\boldsymbol{\beta}_{0,u}$

for $u = \bar{s}s$ and $ss$ as the weighted linear regression coefficients of $Y_i$ on $\boldsymbol{X}_i$ using samples with $(Z_i = 0, S_i = 1)$ and weights $w_{0,u}(\boldsymbol{X}_i)$, and $\boldsymbol{\beta}_{0,u}$ for $u = s\bar{s}$ and $\bar{s}\bar{s}$ as the weighted linear regression coefficients of $Y_i$ on $\boldsymbol{X}_i$ using samples with $(Z_i = 0, S_i = 0)$ and weights $w_{0,u}(\boldsymbol{X}_i)$.

## A.15. Additional Simulations

We compare the finite sample performances of our model-assisted estimator with Jo and Stuart (2009)'s linear model-based estimator. The later assumes Strong Monotonicity, and we conduct simulations accordingly.

Let the sample size be 500. We generate $X_{i1}, \ldots, X_{i4} \overset{i.i.d.}{\sim} N(0,1)$ and $X_{i5} \sim \text{Bern}(1/2)$, and let $\boldsymbol{X}_i = (1, X_{i1}, \ldots, X_{i5})^\top$. We consider five cases indexed by the parameter $\theta = -1, -0.5, 0, 0.5, 1$. For each $\theta$, we generate principal strata by logit $\Pr(U_i = s\bar{s} \mid \boldsymbol{X}_i) = \boldsymbol{\theta}^\top \boldsymbol{X}_i$, where $\boldsymbol{\theta} = (0, 0.5, 0.5, 1, 1, \theta)^\top$. We generate potential outcomes using a linear model as in Jo and Stuart (2009) by $Y_i(1) \mid \boldsymbol{X}_i \sim N\left(\sum_{j=1}^5 X_{ij} + 2 \cdot I_{\{U=s\bar{s}\}} + 1, 1\right)$ and $Y_i(0) \mid \boldsymbol{X}_i \sim N\left(\sum_{j=1}^5 X_{ij} + 2, 1\right)$; using a quadratic model by $Y_i(1) \mid \boldsymbol{X}_i \sim N\left(\sum_{j=1}^5 X_{ij} + \sum_{j=1}^4 X_{ij}^2 + 2 \cdot I_{\{U=s\bar{s}\}} + 1, 1\right)$ and $Y_i(0) \mid \boldsymbol{X}_i \sim N\left(\sum_{j=1}^5 X_{ij} + \sum_{j=1}^4 X_{ij}^2 + 2, 1\right)$.

To examine the performance of the estimators with and without PI, in each scenario we analyze the data with and without the binary covariate $X_{i5}$, and respectively label the results as "oracle" and "obs." Without using $X_{i5}$, $\theta$ is a measure of the violation from PI. Figure 4 presents only the results for $ACE_{s\bar{s}}$, and omits similar results for other principal strata. We report the biases and standard errors of the estimators over 1000 repeated samplings.

For linear case which favors the estimator in Jo and Stuart (2009), with the binary covariate both estimators has small biases, and without the binary covariate both estimators have similar bias issues. However, in both scenarios our estimator has smaller standard error. For the quadratic case in which the estimator in Jo and Stuart (2009) is mis-specified, with or without the binary covariate the estimator in Jo and Stuart (2009) suffers from severe bias issues, but our estimator has small biases. Again, in both scenarios our estimator has smaller standard error.

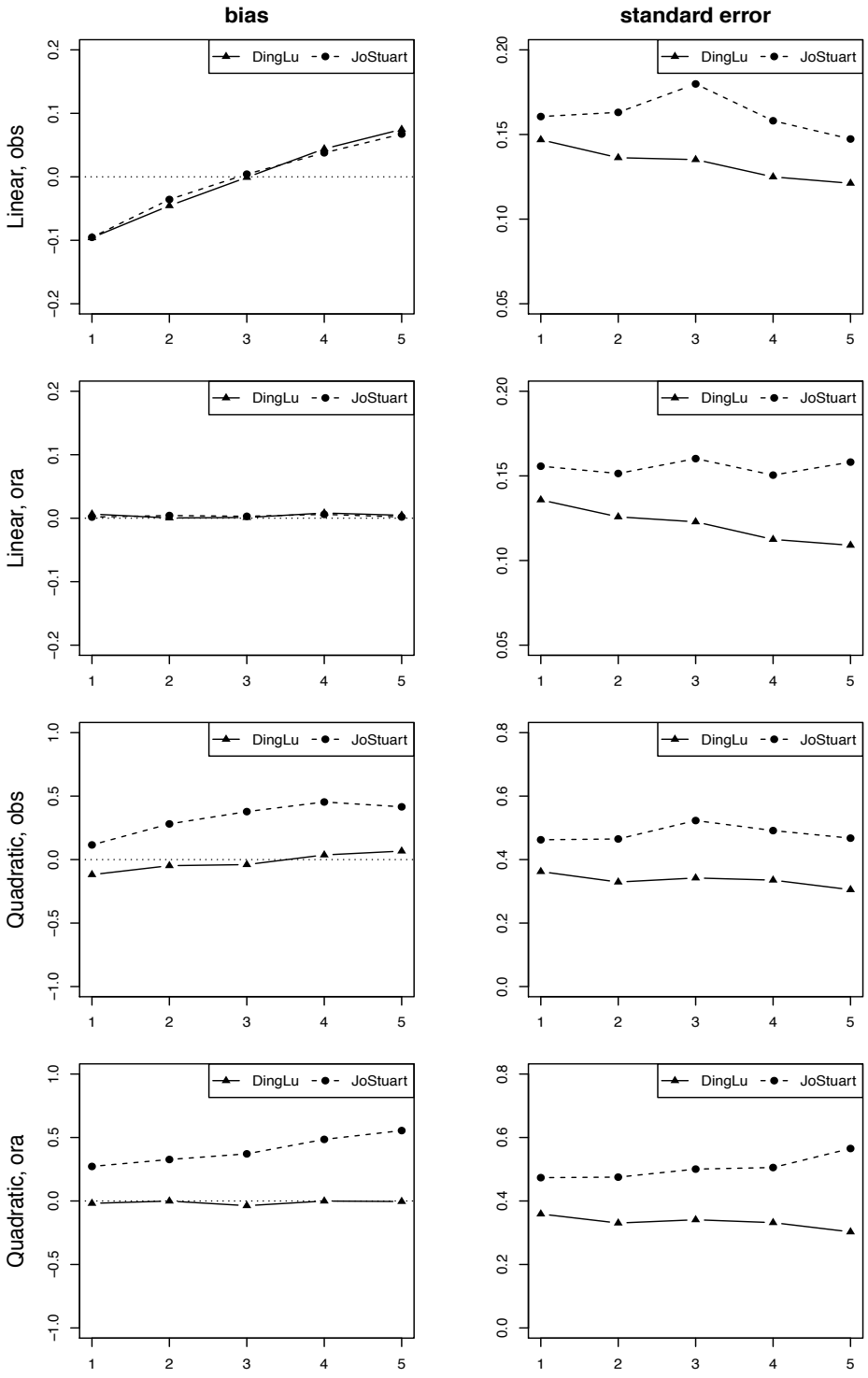



Figure 4: Comparison between our estimator ("DingLu") and Jo and Stuart (2009)'s estimator ("JoStuart") in terms of bias and standard error. The first two rows have linear outcome models, the last two rows have quadratic outcome models. The labels "oracle" and "obs" correspond to analysis with and without the binary covariate that ensures PI with other covariates. All of them are under Strong Monotonicity. The horizontal axis shows the case numbers.