# A Bayesian Approach to Principal Stratification with Multiple Binary Post-treatment Covariates

**Bo Liu**
Department of Statistical Science
Duke University
Durham, NC 27705
bl226@duke.edu

**Haoliang Zheng**
Department of Statistical Science
Duke University
Durham, NC 27705
hz228@duke.edu

## Abstract

Abstract

**Key words:** Key words

## 1 Introduction

Randomization, when possible, is desirable in studying the causal effect of one treatment against the other in clinical trials and other experiments. By randomization, the subpopulation receiving either treatment is homogeneous in both observed and unobserved covariates. Hence, the source of bias from the treatment assignment is eliminated, and therefore any difference in outcome between the two treatment groups can be intepreted as the causal effect between the treatments.

In real studies, there may frequently exists post-treatment covariates which are highly correlated with the randomized treatment and have non-ignorable effect on the outcome. The effect of these covariates is not homogeneous within each randomized treatment group, thus introducing bias and increasing difficulty in inference on the causal effect. A common example of such post-treatment covariates is non-compliance, where the actual treatment one received might be opposite from what they are randomized to, as in the studies of add citations here. In these situations, the actual treatment is not randomized, and there is unobserved confounding between the actual treatment and the outcome.

One approach is the standard intention-to-treat (ITT) method, which ignores the actual treatment and compares the outcome between two randomized treatment groups. This preserves the randomization, but the estimand is the effectiveness of the treatment instead of the efficacy, which might be undesirable when the efficacy is of clinical interest. Another direction, pioneered by citation of first IV paper and introduced to the context of non-compliance in a landmark paper add citation Angrist et al, is the instrumental variable (IV) approach. Here, the assigned treatment can be viewed as an instrumental variable, as it is highly correlated with the actual treatment, but might not have direct effect on the outcome. When the no-direct-effect assumption is questionable by domain knowledge, principal stratification add citation is also a feasible approach. The idea is to define principal strata by the potential values of post-treatment covariates under both treatment arms, which is not dependent on the treatment assignment and is determined prior to the treatment. The principal stratification approach identifies the underlying strata and estimates the causal effect within the strata of interest.

In this report, we generalize the idea of principal stratification to multiple binary post-treatment covariates. We also provide a user-friendly software to facilitate Bayesian inference on estimating the causal effect.

Project Report for STA723 Case Study, Duke University.

## 2 Notation

Let data be denoted as $\{(X_i, Y_i, D_i, Z_i)\}_{i=1}^n$, where $X_i \in \mathbb{R}^p$ is the baseline covariates, $Y_i \in \mathbb{F}$ is the outcome variable, $Z_i \in \{0, 1\}$ is the binary treatment variable and $D_i \in \{0, 1\}^d$ is the binary post-treatment covariates. Associated with each unit is a latent variable $S_i$ representing the principal stratum that the unit belongs to. We assume that each unit has a pair of potential post-treatment covariates, namely $D_i(0)$ and $D_i(1)$, and the observed post-treatment covariate is one of these two values, $D_i = Z_i D_i(1) + (1 - Z_i) D_i(0)$. Under this setting, the principal stratum is defined by the potential post-treatment covariates $S_i = (D_i(0), D_i(1))$. We also assume that each unit has a pair of potential outcome defined as $Y_i(0)$ and $Y_i(1)$, and the observed outcome is one realization of these two potential outcomes, $Y_i = Z_i Y_i(1) + (1 - Z_i) Y_i(0)$. We may drop the subscript $i$ when there is no need to emphasize on it.

The number of possible strata grows increasingly with the number of binary post-treatment covariates. When there are $d$ such covariates, both $D(0)$ and $D(1)$ takes values from $\{0, 1\}^d$, and they define $2^{2d} = 4^d$ possible principal strata. Although theoretically reasonable, it is almost impossible to do inference with such a large number of strata, given the normal sample size one would obtain in a clinical study. Therefore, we usually exclude some principal strata, either due to the study design or reasonable justification, and only deal with a subset of all possible principal strata, which is denoted by $\mathcal{S}$. In the simplest setting where $D \in \{0, 1\}$ denotes the actual treatment one receives, one might argue to exclude $S = (1, 0)$ since these are the “strange” people who would always take the opposite treatment to the assigned one. Moreover, when the treatment cannot be acquired otherwise, one also excludes $S = (1, 1)$ since those randomized to the control group can never choose to take the treatment. The reduction in the number of total strata is usually large as $d$ increases with careful assumptions, making it possible to do principal stratification with $d > 1$.

There are some strata where the binary post-treatment covariates stay the same regardless of the assigned treatment. In this case, will the outcome depend on the treatment? It depends on whether all effects of the treatment randomization on the outcome is through the binary post-treatment covariates. However, when multiple post-treatment covariates exist, one direct question is that, if one or some of these covariates remain unchanged under either treatment assignment, will the outcome change? Difficulty emerges since these post-treatment covariates cannot be viewed as a single covariate - some of these covariates may be affecting the outcome directly, but others may be not. To simplify the assumption, we do not dive into these covariates to see which covariate has the direct effect on the outcome; instead, for each stratum, we propose an assumption on whether treatment assignments would affect the outcome. Mathematically, we assume exclusion restriction (ER) on stratum $S = s$ if $p(Y(0) \mid X, S = s) = p(Y(1) \mid X, S = s)$, which is equivalent to $p(Y \mid S = s, X, Z = 0) = p(Y \mid X, S = s, Z = 1)$ when the treatment $Z$ is randomly assigned.

The exclusion of some principal strata reduces the number of strata from $4^d$ to $|\mathcal{S}|$. The model of outcome $Y$ can be then specified by $2|\mathcal{S}|$ conditional distributions

$$p(Y \mid X, S = s, Z = z), \quad s \in \mathcal{S},\ z \in \{0, 1\}.$$

The number of outcome models can be further reduced by the ER assumption, as only one model is needed for each stratum with ER. We index these distinct models by $g = 1, 2, \ldots, G$, and define $G_i = g$ if $p(Y_i \mid X_i, S_i, Z_i)$ corresponds to the outcome model indexed by $g$. $\mathcal{G} = \{1, \ldots, G\}$ is the set of all indices, $|\mathcal{S}| \leq |\mathcal{G}| \leq 2|\mathcal{S}|$.

| $S$ | $Z = 0$ | $Z = 1$ |
|---|---|---|
| $(0, 0)$ | $G = 1$ | $G = 1$ |
| $(0, 1)$ | $G = 2$ | $G = 3$ |
| $(1, 1)$ | $G = 4$ | $G = 4$ |

Table 1: Index of outcome model for flu vaccination example.

*Example - Flu vaccination* Let $Z$ be the encouragement on flu vaccination, and $D$ be the actual vaccination status. $Y$ denotes the hospitalization rate, on which the encouragement of flue vaccination arguably does not have a direct effect. Ruling out people who always do exactly as opposed to encouraged, we have three principal strata $\mathcal{S} = \{(0, 0), (0, 1), (1, 1)\}$. We assume ER on both $(0, 0)$ and $(1, 1)$. These assumptions define four distinct outcome models (Table 1).

# 3 Model

Assume that the data for all units are independent draws from a super population. The joint likelihood of the data $\{X_i, Y_i, Z_i, D_i\}_{i=1}^n$ can be decomposed as the following.

$$\begin{aligned}
p(\{X_i, Y_i, Z_i, D_i\}_{i=1}^n) &= \prod_{i=1}^n p(X_i, Y_i, Z_i, D_i) \\
&= \prod_{i=1}^n \sum_{S_i \in \mathcal{S}} p(X_i, Y_i, Z_i, D_i, S_i) \\
&= \prod_{i=1}^n p(X_i) \sum_{S_i \in \mathcal{S}} p(S_i \mid X_i) p(Z_i \mid X_i, S_i) p(D_i \mid Z_i, X_i, S_i) p(Y_i \mid D_i, Z_i, X_i, S_i) \\
&= \prod_{i=1}^n p(X_i) \sum_{S_i \in \mathcal{S}} p(S_i \mid X_i) p(Z_i \mid X_i, S_i) p(D_i \mid Z_i, S_i) p(Y_i \mid Z_i, X_i, S_i) \\
&= \prod_{i=1}^n p(X_i) \sum_{S_i \in \mathcal{S} : D_i = S_i(Z_i)} p(S_i \mid X_i) p(Z_i \mid X_i, S_i) p(Y_i \mid G_i, X_i).
\end{aligned}$$

In the above decomposition, $p(X_i)$ and $p(Z_i \mid X_i, S_i)$ are not of particular interest. The first one is to model the marginal distribution of observed covariates, which can be done empirically. The second one should be obtainable by the design of the experiment. Hence, we shall specify parametric models for $p(S_i \mid X_i)$ and $p(Y_i \mid G_i, X_i)$.

A common choice of specifying parametric model $p(S_i \mid X_i)$ is a multinomial model

$$p(S_i = s \mid X_i) \propto \exp(X_i^{\mathrm{T}} \boldsymbol{\beta}_s), \quad \forall s \in \mathcal{S}.$$

For technical reasons, we can select some $s_0 \in \mathcal{S}$ and set $\boldsymbol{\beta}_{s_0} = \mathbf{0}$ to avoid non-identifiability issues. Also, for the outcome model, we can specify a parametric model with possible auxillary parameter $\boldsymbol{\theta}_g$ with the form

$$\Pr(y_i \mid X_i, G_i = g) = f(y_i; X_i^{\mathrm{T}} \boldsymbol{\gamma}_g, \boldsymbol{\theta}_g).$$

Common examples of this form including the normal distribution where $Y_i \mid X_i, G_i = g \sim \mathcal{N}(Y_i; X_i^{\mathrm{T}} \boldsymbol{\gamma}_g, \sigma_g^2)$, and Poisson regression model where $Y_i \mid X_i, G_i = g \sim \mathcal{P}ois(X_i^{\mathrm{T}} \boldsymbol{\gamma}_g)$.

The Bayesian model is completed by specifying prior distributions on $\boldsymbol{\beta}_s$, $\boldsymbol{\gamma}_g$ and $\boldsymbol{\theta}_g$.

# 4 Software interface