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

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 [3] and [4]. 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 in a 1928 book by Philip G. Wright and introduced to the context of non-compliance in a landmark paper [1], 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 [2] 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.

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

Project Report for STA640 Causal Inference, Duke University.

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

The software, which generates posterior samples using stan in the backend, has a user-friendly R interface accessible to users with different backgrounds.

The core of the input is `S.model`, `Y.model` and `Y.family`, which specify the multinomial model $p(S_i = s \mid X_i)$ and the outcome model $p(Y_i \mid X_i, G_i = g)$.

(1) `S.model` This argument is specified with the treatment assignment variable Z and all of the binary post-treatment covariates $D_1, \ldots, D_d$. Let $X_1, \ldots, X_p$ be the covariates which the multinomial model is built upon. The `S.model` can be simply specified with syntax

`Z + D1 + ... + Dd ~ X1 + ... + Xp`.

(2) `Y.model` Like `S.model`, this parameter specifies the outcome variable and important covariates to model the outcome. The syntax is as follows.

`Y ~ X1 + ... + Xp`.

(3) `Y.family` This argument takes exactly the same convention as `family` defined in R base `glm()` function. For example, `Y.family = gaussian(link = "identity")` specifies a gaussian model where $Y_i \mid X_i, G_i = g \sim \mathcal{N}(X_i^{\mathrm{T}} \boldsymbol{\gamma}_g, \sigma_g^2)$. The additional parameter σ_g^2 here comes intrinsically with the specified `Y.family`. Similarly, α comes with Gamma family, λ comes with inverse-gaussian family and θ comes with the Cox-survival family.

The set of all strata $\mathcal{S}$ and whether to consider ER for each stratum are also important in specifying the correct mixture model. They are specified by `S` and `ER` respectively.

(4) `S` and `ER` The set of strata in consideration is given by this argument as a vector of stratum indices. The convention to index a stratum is as follows. Recall that $S = (D(0), D(1))$, where both $D(0)$ and $D(1)$ are elements of $\{0,1\}^d$. Hence S can be viewed as an element of $\{0,1\}^{2d}$, or a $2d$-bit zero-one sequence. Then we index S by the integer with the zero-one sequence as binary representation. The argument `ER` is also a vector containing all indices appearing in `S`, within which the ER assumption is admitted.

Finally, we specify the prior distributions of all parameters.

(5) `prior_xxx` There are six additional parameters for prior specification, including `prior_intercept`, `prior_coefficient`, `prior_sigma`, `prior_alpha`, `prior_lambda` and `prior_theta`. When applicable, these parameters can be specified to impose a certain prior on each kind of parameters.

5 Simulation

We provide two simulation studies in this section.

5.1 Simulation 1

This simulation study features a classical non-compliance scenario where defiers ($S = (1,0)$) are excluded from the analysis, and ER is assumed within both always-takers ($S = (1,1)$) and never-takers ($S = (0,0)$).

Specifically, we generate 1000 units. Two covariates X_1, X_2 are independently sampled from $\mathcal{N}(0,1)$. We assign $S \in \{(0,0),(0,1),(1,1)\}$ independently with probability 0.3, 0.5 and 0.2, respectively. The treatment assignment is random to every unit with $p(Z=1) = 0.5$ to mimic a random clinical trial. The outcome Y is binary with $\text{logit}\Pr(Y = 1 \mid S = s, Z = z, X_1, X_2) = \beta_{sz0} + \beta_{sz1}X_1 + \beta_{sz2}X_2$, where

$$\begin{aligned}
\text{logit } \Pr(Y = 1 \mid S = (0,0), Z = z, X_1, X_2) &= X_1 - X_2, \\
\text{logit } \Pr(Y = 1 \mid S = (0,1), Z = z, X_1, X_2) &= 2X_1 - 0.5X_2 + z, \\
\text{logit}\Pr(Y = 1 \mid S = (1,1), Z = z, X_1, X_2) &= X_2.
\end{aligned}$$

We run the sampler with 6 chains and 500 warmup iterations and 500 sampling iterations for each chain. The true values of parameters and the respective posterior distributions are given in Table 2.

	True value	Posterior mean	2.5% quantile	97.5% quantile
$\Pr(S = (0,0))$	0.30	0.28	0.24	0.33
$\Pr(S = (0,1))$	0.50	0.53	0.49	0.56
$\Pr(S = (1,1))$	0.20	0.19	0.17	0.20
$\beta_{(0,1),0,0}$	0.00	-0.02	-0.47	0.41
$\beta_{(0,1),1,0}$	1.00	1.09	0.63	1.61

Table 2: The posterior summary of important coefficients with true values in Simulation 1.

5.2 Simulation 2

This simulation study features a more complex scenario where two post-randomization covariates D_1 and D_2 exist. Let the principal strata be defined by $S = (D_1(0), D_2(0), D_1(1), D_2(1))$. In this study, we include five out of 16 possible strata, namely $\mathcal{S} = \{0000, 0001, 0011, 0101, 1111\}$, and assume ER for 0000, 0101 and 1111. We do not include baseline covariates in this study.

We simulate 10,000 data for better identification of the strata. We randomly assign principal stratum S and treatment status Z to each unit, with the stratum-assignment probability being $p = (0.15, 0.2, 0.1, 0.4, 0.15)$ and the treatment assignment probability $P(Z = 1) = 0.5$. The outcome variable Y is sampled from a Gaussian distribution given in Table 3.

	Probability	$Z = 0$	$Z = 1$
$S = 0000$	0.15	$\mathcal{N}(3, 1)$	
$S = 0001$	0.20	$\mathcal{N}(-1, 0.5)$	$\mathcal{N}(-2, 0.5)$
$S = 0011$	0.10	$\mathcal{N}(1, 0.5)$	$\mathcal{N}(4, 0.5)$
$S = 0101$	0.40	$\mathcal{N}(-1, 3)$	
$S = 1111$	0.15	$\mathcal{N}(1, 2)$	

Table 3: Outcome specification for Simulation 2

We run the sampler with 6 chains and 500 warmup iterations and 500 sampling iterations for each chain. The true values of parameters and the respective posterior means are given in Table 4. We omit the credible intervals to assure readability of the table.

	Probability	$Z = 0$	$Z = 1$
$S = 0000$	0.150	$\mathcal{N}(2.98, 1.03)$	
$S = 0001$	0.195	$\mathcal{N}(-0.53, 0.74)$	$\mathcal{N}(-1.98, 0.53)$
$S = 0011$	0.104	$\mathcal{N}(0.28, 0.46)$	$\mathcal{N}(4.00, 0.52)$
$S = 0101$	0.401	$\mathcal{N}(-1.00, 2.98)$	
$S = 1111$	0.150	$\mathcal{N}(1.04, 2.01)$	

Table 4: Estimated probability of strata and outcome models for Simulation 2

Most coefficients given by the posterior mean seem to be very consitent with the true values, with more noticeable discrepancy in the two distributions under $Z = 0$ for $S = 0001$ and $S = 0011$. When $Z = 0$, both observed post-treatment covariates for these two strata are 00, bringing difficulty to identify both strata consistently. In this situation, all information to identify these two strata comes from the subgroup of people with $Z = 1$. Although theoretically identifiable, this could be unstable with a finite sample size. This phenomenon, as described as "weak identifiability" in some literature, suggests that caution should be taken in model specification when the number of principal strata is considerably large relative to the amount of data available.

6 Discussion

The Bayesian approach for inference under principal stratification should be consistent when the model is correctly specified. Furthermore, it can be easily extended to multiple post-treatment covariates and various assumptions as well. However, no extensive use of Bayesian principal stratification approach has been seen in relative literature, despite the flexibility and consistency of the Bayesian approach. One of the hurdles might be the difficulty in programming a sampler to draw samples from the posterior distribution for the above-mentioned model. On the one hand, these models are often mixture of distributions, with the existence of unobservable underlying variables. On the other hand, the posterior distribution varies with the specification of models, the number of strata, and the ER assumptions. Hence, the sampler needs to be designed and programmed on a case-by-case basis. Therefore, researchers in both applied analysis or theoretical study might find it difficult to adopt the Bayesian approach. The design of Stan language faciliates the sampling process by automatically drawing samples from a given posterior distribution. Our software further utilizes the Stan language in the backend to enable automatic posterior sampling in the context of Bayesian principal stratification. With a user-friendly interface, researches are able to do inference with complex principal stratification approaches using R, one of the most commonly used language in statistics.

The simulation studies provided in the previous sections illustrates the use of the software, and shows that the results given by the software are reliable. Currently, the software is still under progress, with random effect models to be available, which appears in nature in many cluster randomized trials. This feature will be released soon afterwards when more simulation tests are conducted.

References

[1] Angrist, J. D. and Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442.

[2] Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

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

[4] 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. *The Annals of Applied Statistics*, 7(4):2336 – 2360.