Tutorial: Principal Stratification for EDM Experiments

Adam C Sales
University of Texas College of Education
1912 Speedway Stop D5700
Austin, Texas, USA
asales@utexas.edu

ABSTRACT

Principal stratification (PS), which measures variation in a causal effect as a function of post-treatment variables, can have wide applicability in educational data mining. Under the PS framework, researchers can model the effect of an intelligent tutor as a function of log data, can account for attrition, and study causal mechanisms. Participants in this tutorial will learn how and when PS works and doesn't work, and will learn three methods of estimating principal effects.

1. PRINCIPAL STRATIFICATION IN EDM RESEARCH

Educational data miners are increasingly interested in causal questions—what interventions work, for whom, and how. Accompanying this interest is the widespread realization that there is no such thing as "the effect": actually, effects can vary widely between individuals. Estimating the differences in effects between types of learners is (in principal) straightforward for types defined prior to the onset of an experiment. But what about learners who use the software in different ways—or, even given the opportunity, don't use it at all? Traditionally, "post-treatment" variables, observed subsequent to treatment assignment, are treated as mediators whose analysis requires the kind of untestable assumptions randomization is supposed to avoid.

Principal stratification (PS) [2] offers a different approach: categorizing learners based on how they *would* (or would not) use the software if given the opportunity. Under the PS approach, an analyst begins by defining types, or "principal strata" of learners based on post-treatment measurements, then estimates the probability each learner is a member of each stratum (conditional on baseline covariates), and finally the average effect of the treatment within each stratum. In a randomized experiment, the final step of the process proceeds from the randomization (and, possibly, testable modeling assumptions). That is, researchers need not assume unconfoundedness, or that all relevant variables have been measured. The result is a principal effect, or separate estimate of an average treatment effect for each usage mode of interest; these may be used to explore causal mechanisms, study the conditions under which software might work better (or worse), learn dosage effects (i.e. does more usage translate to larger effects), and many other applications.

1.1 EDM Questions PS may Help Answer

PS could help address a wide range of research questions in EDM. Some examples are:

- Does the effect of an intervention depend on learners' (measured) emotional state?
- Are some sections of a software more effective than others?
- Do some learner strategies—such as hint usage or mastery learning—correspond to larger effects than others?
- Are there intermediate outcomes, such as mastery speed or error rate, that can serve as good surrogates for a final outcome, such as a post-test?
- Estimating treatment effects after attrition

Each of these questions estimates an average treatment effect for a group of learners which is defined based on variables measured only after the intervention began. This is the type of question principal stratification was designed to answer.

1.2 Estimating Principal Effects

The catch is that principal effects can be difficult to estimate. Estimating effects within principal strata depends on knowing who is in which stratum—for instance, which students in the control condition *would have* been frustrated, had they been assigned to treatment, or which students would have attritted, had they been assigned to the opposite condition—which is unobserved and must be inferred. The most popular and powerful approach begins by assuming a model (typically the normal distribution) for the outcome within each stratum and a model for who is in which stratum (typically logistic regression). Next, it fits a mixture model for those subjects with unobserved stratum membership. For instance, in an experiment comparing students assigned to use an intelligent tutor with students assigned to use traditional curricula, a researcher looking to estimate average effects for high-hint users might model post-test scores for subjects in the control condition as a mixture of two distributions: one for students who would use many hints, and one for students who would not. The success of this approach depends on the fit of the model—misspecified models may yield misleading results—so extensive model checking is necessary. Further, even when the model is correctly specified, its success can depend on factors beyond the researcher's control [1].

Two other approached depend less on modeling assumptions, but may yield less precise estimates. One approach [3] estimates bounds for principal effects, rather than estimating the effects themselves. Another [4], applicable in some PS studies but not others, uses non-parametric techniques to identify plausible candidates for unobserved principal strata, and estimates effects based on those. These approaches are more "automatic" than the model-based approach, in that they do not require careful model fitting and checking, but still require researchers to specify the problem carefully.

1.3 My Expertise

For the past three years, I have been working on an NSF-funded project to use the PS framework to study data from the Cognitive Tutor Algebra I effectiveness study. With Dr. John Pane of the RAND Corporation, I have estimated various associations between Cognitive Tutor treatment effects and student usage. This has produced two EDM proceedings papers, [5] and [6]. As part of the project, I have developed a new method for estimating principal effects which expands on [4] and set of new diagnostic and model checking techniques. I have also worked extensively with Neil Heffernan's lab using PS to model data from ASSISTments experiments.

2. TUTORIAL PLAN

2.1 Introduction to Principal Stratification

The beginning of the tutorial will introduce the PS framework. First, we will discuss why principal stratification is necessary: participants will learn to distinguish post-treatment from pre-treatment variables and understand the conceptual and methodological issues with conditioning causal inference on post-treatment variables. Next, we will describe PS framework, so participants understand how it solves the problems with post-treatment conditioning. Finally, we will discuss methods for estimating effects within principal strata: what assumptions they depend on and the source for their identification. We will give a brief overview of the various PS methods that we will explore hands on, in more depth, during the remainder of the tutorial.

2.2 Hands on PS Estimation

The second half of the tutorial will focus on three classes of methods to estimate principal effects: nonparametric bounds, nonparametric randomization inference, and model based PS.

I will provide two real EDM datasets that participants can use for exercises. The first will be a subset of the data from the Cognitive Tutor effectiveness study, comparing subjects assigned to use the Cognitive Tutor to those assigned to traditional curricula. The study produced rich log-data—PS can be used to compare treatment effects between sets of learners who used, or would have used, the tutor differently. The second dataset will come from an experiment run on the ASSISTments platform [7]. I will also give participants the opportunity to bring their own datasets to the tutorial.

The methods will be taught in `R`, a free, open-source language for statistical computing. We will begin with a brief introduction to the software: how to read in data, and how to write and execute simple code.

The bounding portion will be based on [3], which describes a set of bounds on principal effects, depending on available covariates and certain identification assumptions. We will set out a number of real or realistic data scenarios and discuss which bounds may be appropriate when. Next, we will use `R` to calculate the appropriate bounds for principal effects.

The randomization inference portion will be based on [4] and extensions I have developed. They depend on the assumption of monotonicity—that principal stratum membership is directly observable for all members of either the treatment or the control group. I will provide code in `R` to estimate confidence intervals for principal effects with and without covariates the predict stratum membership.

The model based portion will use Bayesian methods, with the JAGS language, via `R` and the `R2Jags` package. We will practice estimating principal effects with pre-written JAGS code (which I will explain) as well as discuss diagnostic tools: model checking, convergence diagnostics, and small simulation studies.

References

[1] A. Feller, E. Greif, L. Miratrix, and N. Pillai. Principal stratification in the twilight zone: Weakly separated components in finite mixture models. *arXiv preprint arXiv:1602.06595*, 2016.

[2] C. E. Frangakis and D. B. Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

[3] L. Miratrix, J. Furey, A. Feller, T. Grindal, and L. C. Page. Bounding, an accessible method for estimating principal causal effects, examined and explained. *arXiv preprint arXiv:1701.03139*, 2017.

[4] T. L. Nolen and M. G. Hudgens. Randomization-based inference within principal strata. *Journal of the American Statistical Association*, 106(494):581–593, 2011.

[5] A. C. Sales and J. F. Pane. Exploring causal mechanisms in a randomized effectiveness trial of the cognitive tutor. In *Proceedings of the 8th International Conference on Educational Data Mining*, 2015.

[6] A. C. Sales, A. Wilks, and J. F. Pane. Student usage predicts treatment effect heterogeneity in the cognitive tutor algebra i program. In *Proceedings of the 9th International Conference on Educational Data Mining*, 2016.

[7] D. Selent, T. Patikorn, and N. Heffernan. Assistments dataset from multiple randomized controlled experiments. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*, pages 181–184. ACM, 2016.